Applied Statistical Modelling 1: introduction

Prof.ssa Chiara Seghieri
Laboratorio di Management e Sanità, Istituto di Management, L'EMbeDS
Scuola Superiore Sant'Anna, Pisa
c.seghieri@santannapisa.it

Instructor: Dr Claudio Mazzi
Laboratorio di Management e Sanità, Istituto di Management, L'EMbeDS
Scuola Superiore Sant'Anna, Pisa
c.mazzi@santannapisa.it

About this course…

Beyond the formulas and theoretical concepts that are learned in a statistics course (and texts), we will analyze and discuss about the concepts and methods of inferential statistics that most frequently find application in social sciences.

Final aim of the course:
help students gain an understanding of the rationale behind many statistical methods, as well as an appreciation of the use and misuse of statistics across examples from the social sciences.

Encouraging critical thinking!

> *Even more important than learning about statistical techniques is the development of what might be called a capability for* ***statistical thinking****.*

(***Preface*** di G. E. P. **Box,** W. G. **Hunter** e J. S. **Hunter** del 1978 ***Statistics for Experimenters. An Introduction to Design, Data Analysis, and Model Building***, John Wiley and Sons, Inc., New York).

Final Assignment

Students will be asked to think about a statistical question and apply statistical methods learned in class to analyze data in order to answer to the identified question.

Findings and conclusions from the analysis will be discussed in class.

In practice:

- Divide in groups (max 2 students per group)
- Think about a topic you would like to explore and the research question
- Look for data on the selected topic (either existing or new collected data)
- Perform statistical analysis to answer to your research question
- Present your findings in about 10 slides in a scientific manner

1 - What is statistics

Statistics is…

the science concerned with developing and studying methods for **collecting**, **analyzing**, **presenting** and **drawing conclusions** from data.

Statistics is a highly interdisciplinary field; research in statistics finds applicability in virtually all scientific fields and research questions in the various scientific fields motivate the development of new statistical methods and theory.

Statistics in the context of a general process of investigation:

1. Identify a research question or problem.
2. Collect relevant data on the topic.
3. Analyze the data.
4. Interpret results and form a conclusion.

That is, statistics has three primary components:

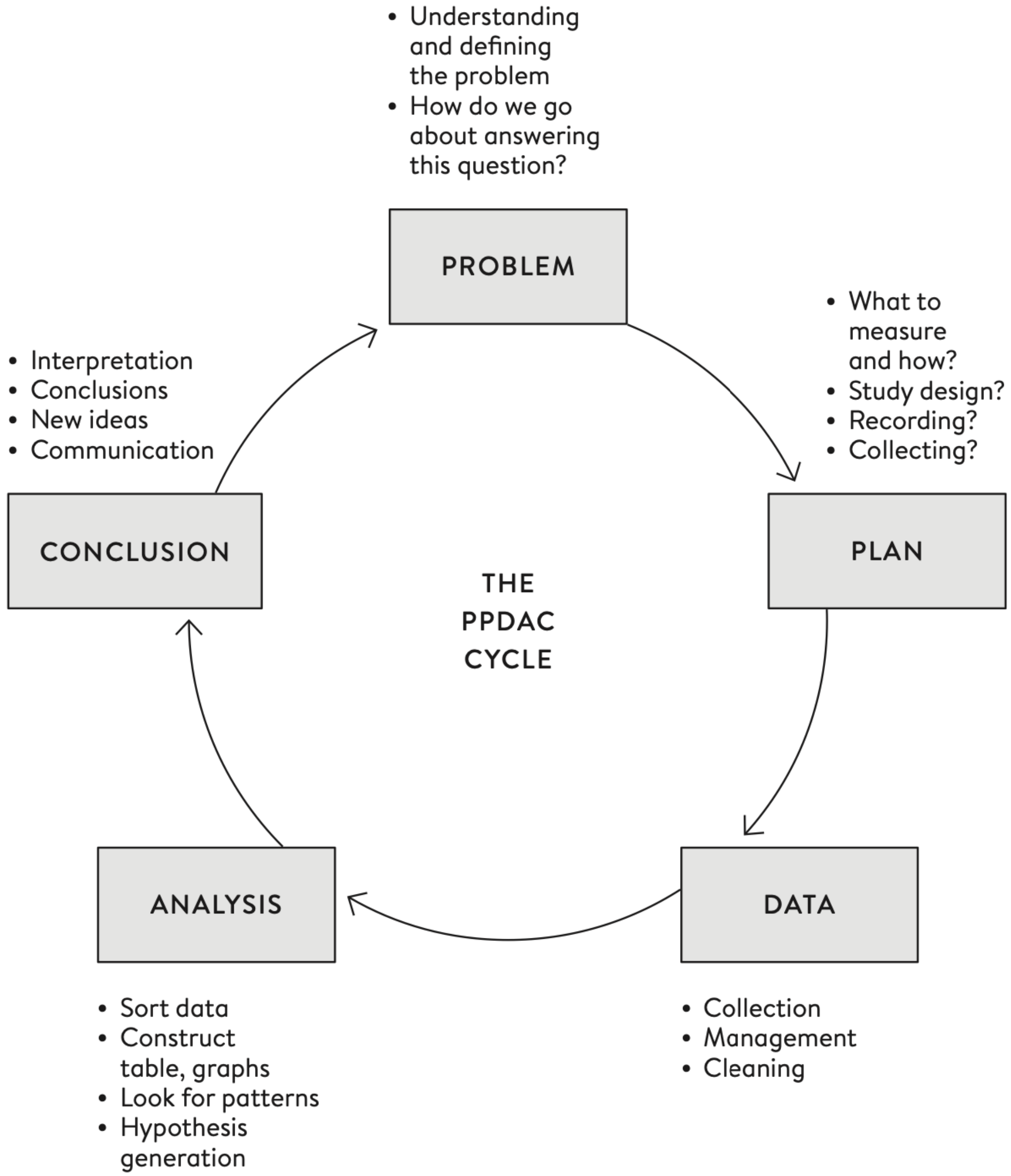
How best can we collect data?
How should it be analyzed?
And what can we infer from the analysis?

A well-written statistical question refers to:

a **population** of interest (collective phenomenon),
a **measurement** of interest,
and anticipates answers that **vary** (the phenomenon varies among the subjects of the population - anticipates **variability** in the response).

the **statistics aims** at describing the phenomenon and/or looking for regular pattern (try to explain most of the variation).

The data Cycle. Statistics helps answer real questions and support decision making



THE PPDAC CYCLE

PROBLEM

- Understanding and defining the problem
- How do we go about answering this question?

PLAN

- What to measure and how?
- Study design?
- Recording?
- Collecting?

DATA

- Collection
- Management
- Cleaning

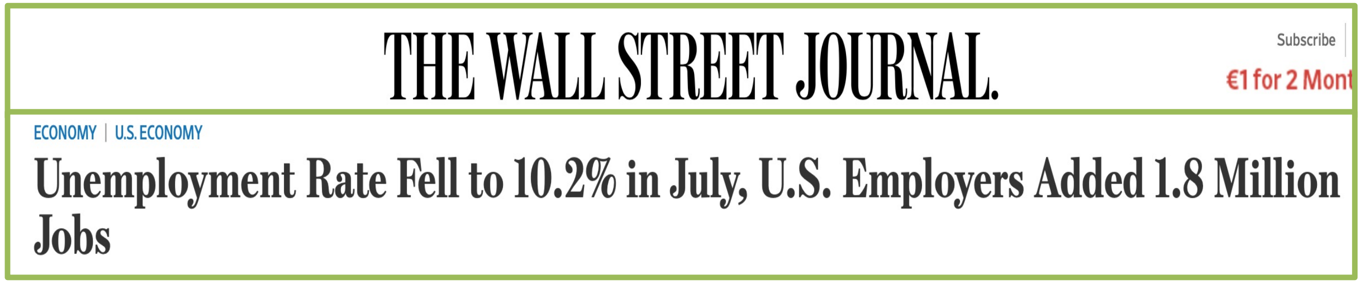
ANALYSIS

- Sort data
- Construct table, graphs
- Look for patterns
- Hypothesis generation

CONCLUSION

- Interpretation
- Conclusions
- New ideas
- Communication

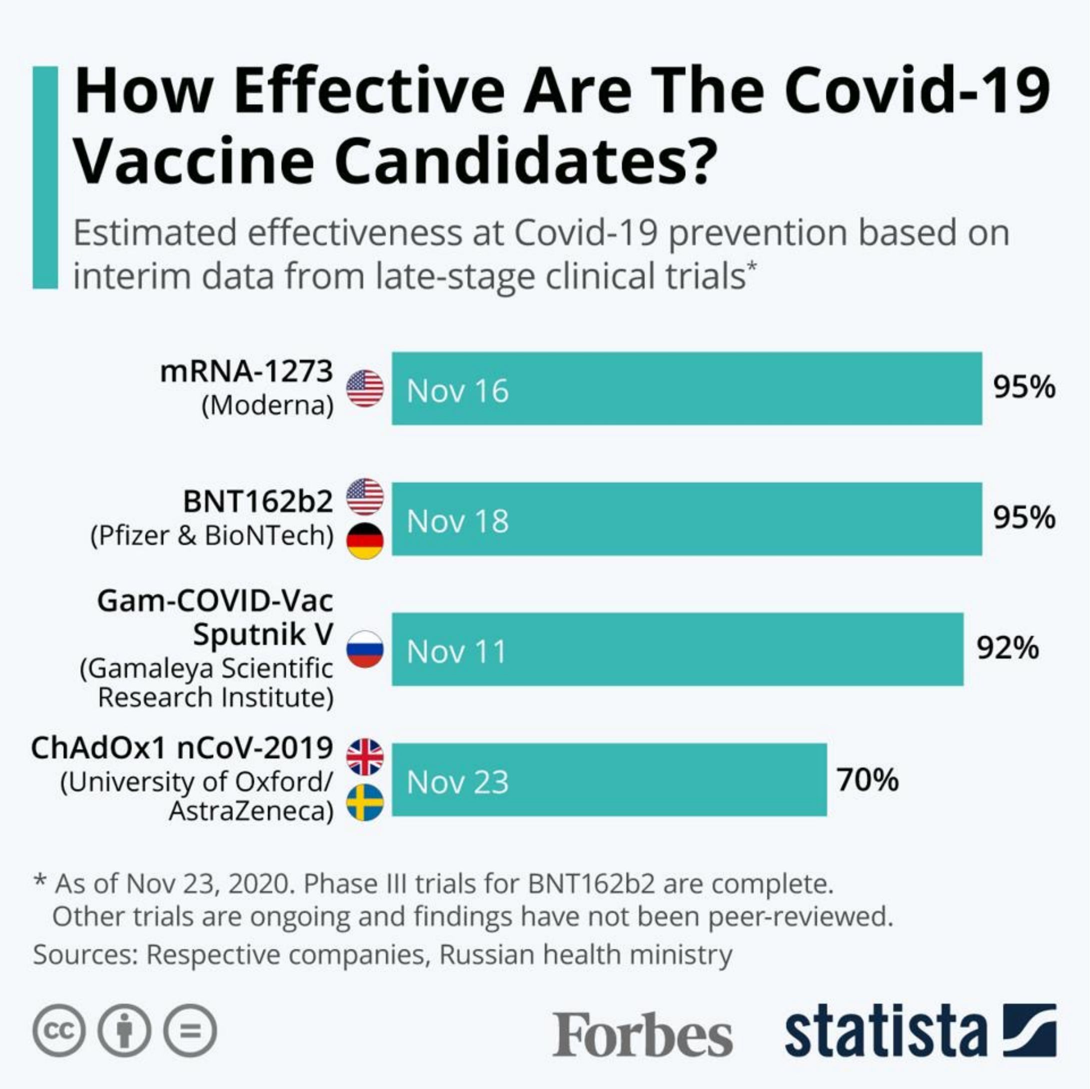
2 – Types of studies



THE WALL STREET JOURNAL.

ECONOMY | U.S. ECONOMY

Unemployment Rate Fell to 10.2% in July, U.S. Employers Added 1.8 Million Jobs



How Effective Are The Covid-19 Vaccine Candidates?

Estimated effectiveness at Covid-19 prevention based on interim data from late-stage clinical trials*

* As of Nov 23, 2020. Phase III trials for BNT162b2 are complete. Other trials are ongoing and findings have not been peer-reviewed.

Sources: Respective companies, Russian health ministry

Forbes statista

Type of Studies

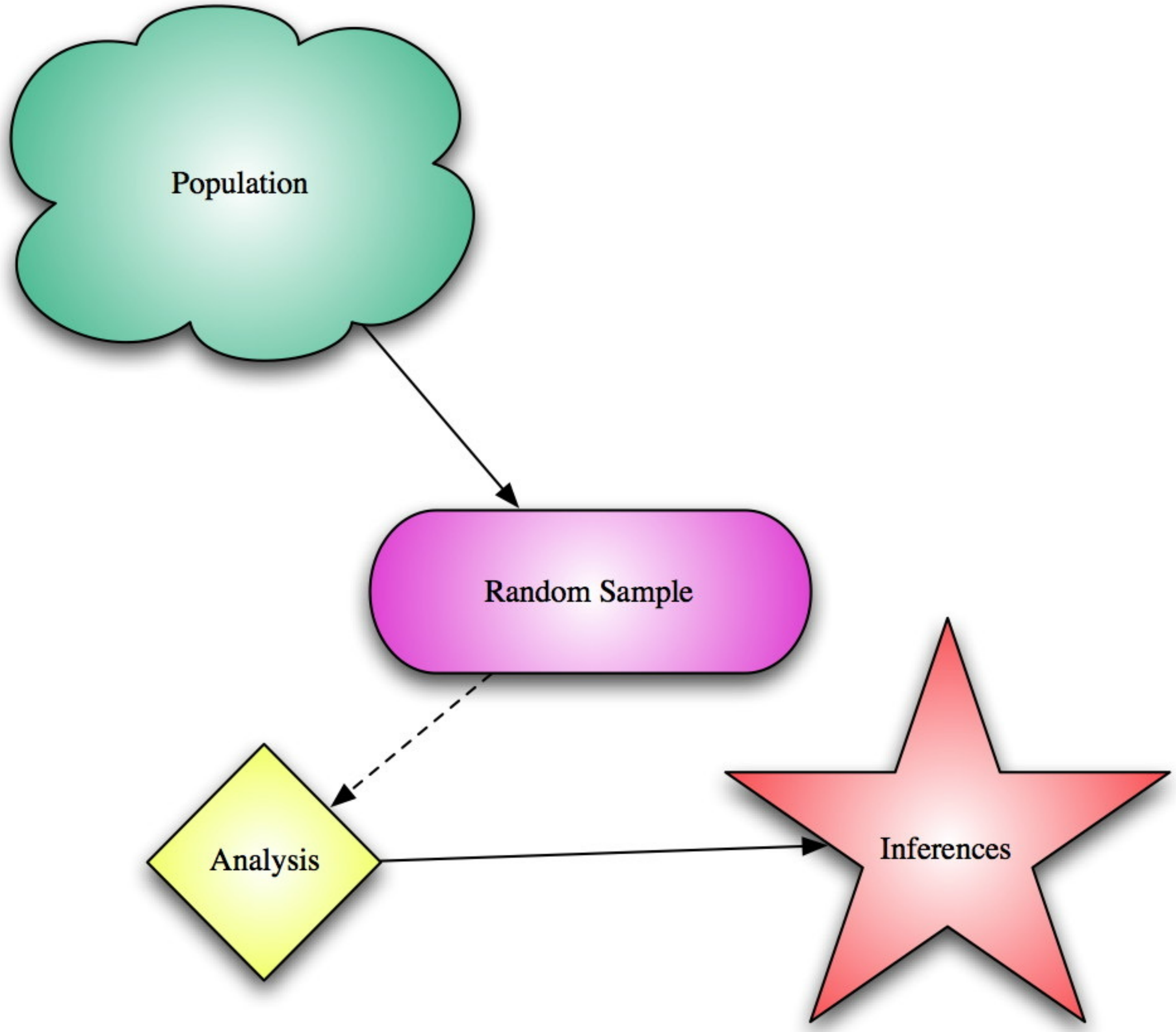
There are two primary types of data collection: **observational studies** and **experiments**.

Researchers perform an observational study when they collect data in a way that does not directly interfere with how the data arise. For instance, researchers may collect information using surveys, reviewing medical or company records, or follow a cohort of many similar individuals to consider why certain diseases might develop.
In each of these cases, the researchers try not to interfere with the natural order of how the data arise.

In general, observational studies can provide evidence of a naturally occurring association between variables, but they cannot show a causal connection.

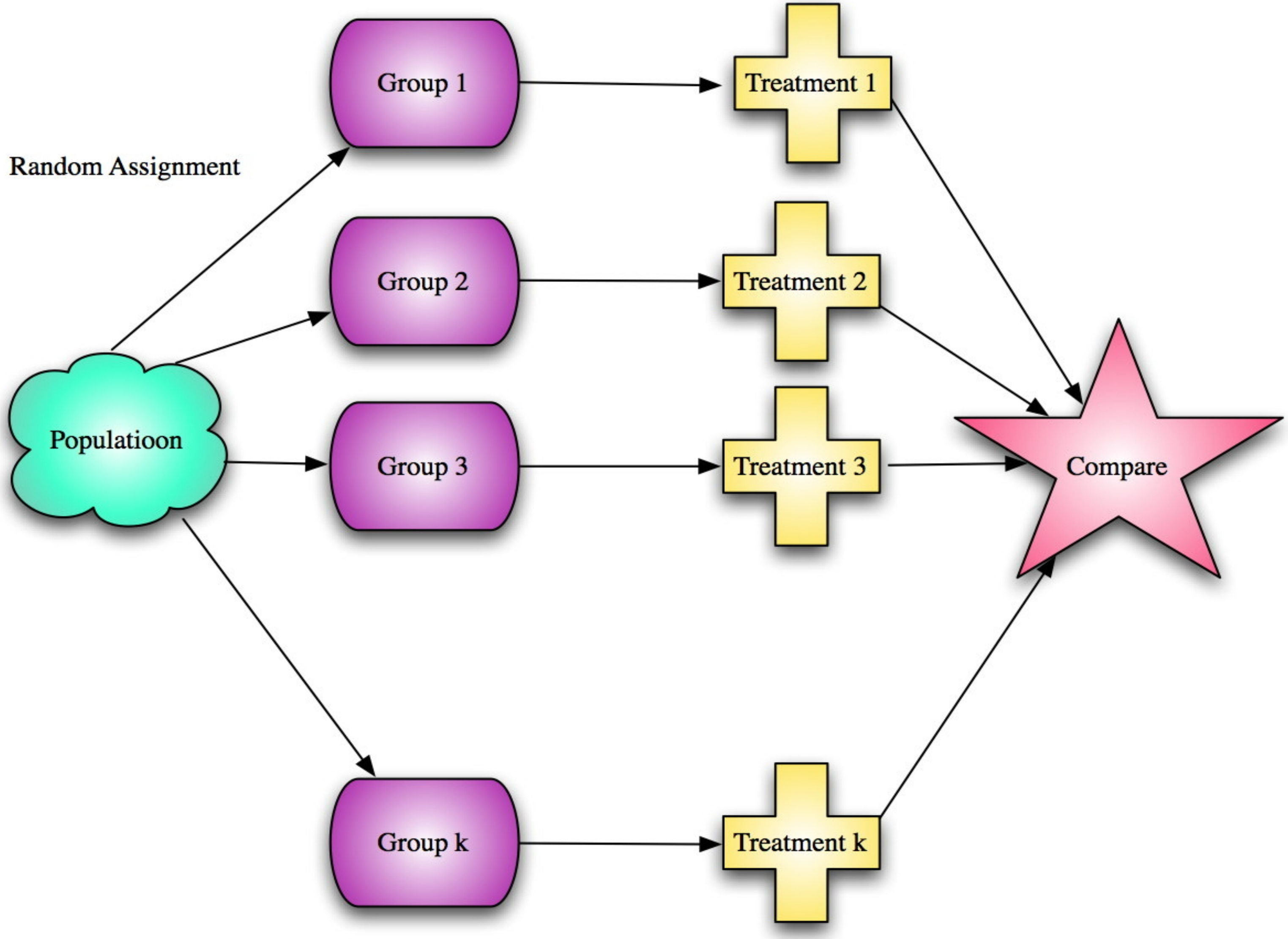
When researchers want to establish a causal connection, they conduct an experiment.

Observational studies



In an **experiment**, one variable is manipulated to create treatment conditions. A second variable is observed and measured to obtain scores for a group of individuals in each of the treatment conditions. The measurements are then compared to see if there are differences between treatment conditions. All other variables are controlled to prevent them from influencing the results.

The goal of an **experiment** is to demonstrate a cause-and-effect relationship between two variables; that is, to show that changing the value of one variable causes changes to occur in a second variable.


Random Assignment
Populatioon
Group 1
Group 2
Group 3
Group k
Treatment 1
Treatment 2
Treatment 3
Treatment k
Compare

Experiments

In the **experiment**, the investigator controls or modifies the environment and observes the effect on the variable under study.

In a randomized experiment (Randomized Control Trails – RCTs) investigators randomly assign the treatments to the experimental units (people, animals, plots of land, etc.) to study whether the treatment causes change in the response.

It is more likely to yield unbiased estimates of causal effects than typical observational studies.

Natural Experiments

In particular research domains, the randomized control trial (RCT) is considered to be the only means for obtaining reliable estimates of the true impact of an intervention. However, an RCT design would often not be considered ethical, politically feasible, or appropriate for evaluating the impact of many policy, programme,...

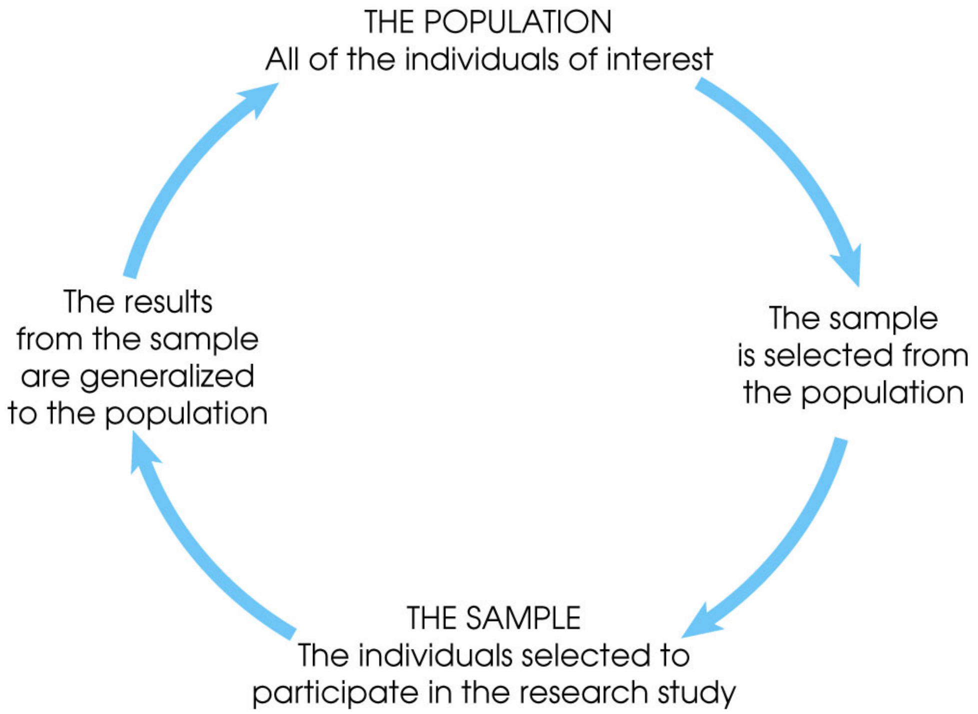
As such, researchers must use alternative yet robust research methods for determining the impact of such interventions. The evaluation of natural experiments (i.e. an intervention not controlled or manipulated by researchers), using various experimental and non-experimental design options can provide an alternative to the RCT.

3 – Population and sample

STATISTICS

Descriptive Statistics:
methods of organizing, summarizing, and presenting data in an informative way

Inferential Statistics:
methods for using sample data to make general conclusions (inferences) about populations using probability theory and **summarise uncertain**ty



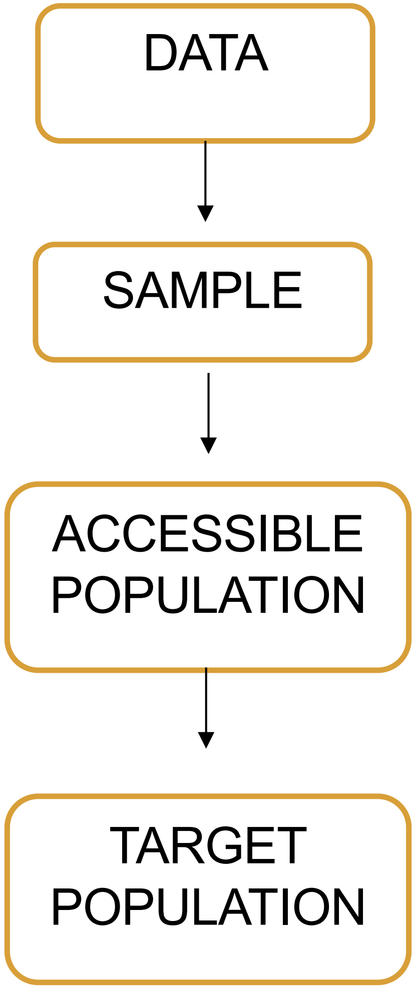
THE POPULATION
All of the individuals of interest

The sample
is selected from
the population

THE SAMPLE
The individuals selected to
participate in the research study

The results
from the sample
are generalized
to the population

Inductive Inference Process



DATA

SAMPLE

ACCESSIBLE POPULATION

the portion of the population to which the researcher has reasonable access; may be a subset of the target population

TARGET POPULATION

The entire group of people or objects to which the researcher wishes to generalize the study findings

Obtaining good samples

✓For valid statistical inference the sample must be representative of the population.

✓Typically it is hard to tell whether a sample is representative of the population.

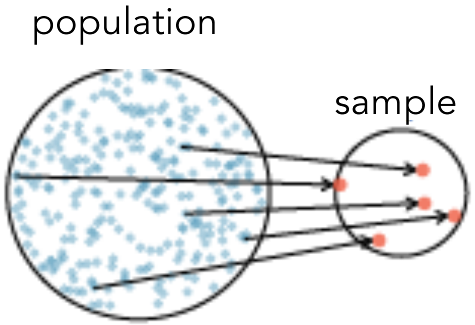
✓The only guarantee for that comes from the method used to select the sample (sampling method)→**probability sampling**

✓There are several sampling methods that guarantee representativeness.

Obtaining good samples

- If observational data are not collected in a random framework from a population, these statistical methods – the estimates and errors associated with the estimates – are not reliable.

- Most commonly used random sampling techniques are *simple*, *stratified*, and *cluster* sampling.



Simple random sampling

The most basic random sample is called a **simple random sample**: each case in the population has an equal chance of being included and there is no implied connection between the cases in the sample.

Begin with a population of size N and randomly draws n units from the population in a way that ensures that the probability of any one unit being drawn for the sample is 1/N.

Procedure:

Assign a number to each member of the population.
Random numbers can be generated by a random number table, software program or a calculator.
Members of the population that correspond to these numbers become members of the sample.

Simple random sampling

We pick samples randomly to reduce the chance we introduce biases. If someone is permitted to pick and choose exactly which individuals were included in the sample, it is entirely possible that the sample could be skewed to that “person's interests”, which may be entirely unintentional. This introduces bias into a sample. Sampling randomly helps resolve this problem.

Even when people are picked at random, e.g. for surveys, caution must be exercised if the non-response rate is high. For instance, if only 30% of the people randomly sampled for a survey actually respond, then it is unclear whether the results are representative of the entire population.

NON PROBABILITY SAMPLES:

Convenience sampling is the selection of study subjects because they are accessible for one reason or another to the researcher. This common form of sampling is often used for rapid market analysis or projection of political elections.

The findings of this form of research cannot be generalized to a cohort or population other than the sample group. A further issue with convenience sampling is that there is an inability to determine potential sampling error.

A special form of convenience sampling is **quota sampling** in which sample subjects are selected from a baseline convenience sample that meet demographic characteristics of the target population.

Purposeful sampling is often used in phenomenon-based qualitative research. In this type of sampling, the researcher uses their own expertise or judgement to select a sample that may represent a target population. Unless the participants in a purposeful sample are selected using random sampling, purposeful sampling is a form of non-probability research.

Often, purposeful or convenience sampling groups are expanded by **snowball sampling**. Those who have been selected to participate in the sample are asked to recruit other potential sample members that they may be affiliated with.

Often, we have **access to data for the entire population**, we did not do any sampling, we have all the data, and there is no more we could collect.

Think for instance to administrative data, such as of the number of births that occur each year, the examination results for a particular class, or data on all the countries of the world – none of these can be considered as a sample from an actual population.

We might then think about a **METAPHORICAL POPULATION**:

«…The idea of a metaphorical population is challenging, and it may be best to think of what we have observed as having been drawn from some imaginary space of possibilities. For example, the history of the world is what it is, but we can imagine history having played out differently, and we happen to have ended up in just one of these possible states of the world. This set of all the alternative histories can be considered a metaphorical population.»

you could view the measurements from these data as one concrete manifestation of an imaginary process that generated the results.

4 – Types of data

How will the data be sourced?Secondary and Primary Data Collection

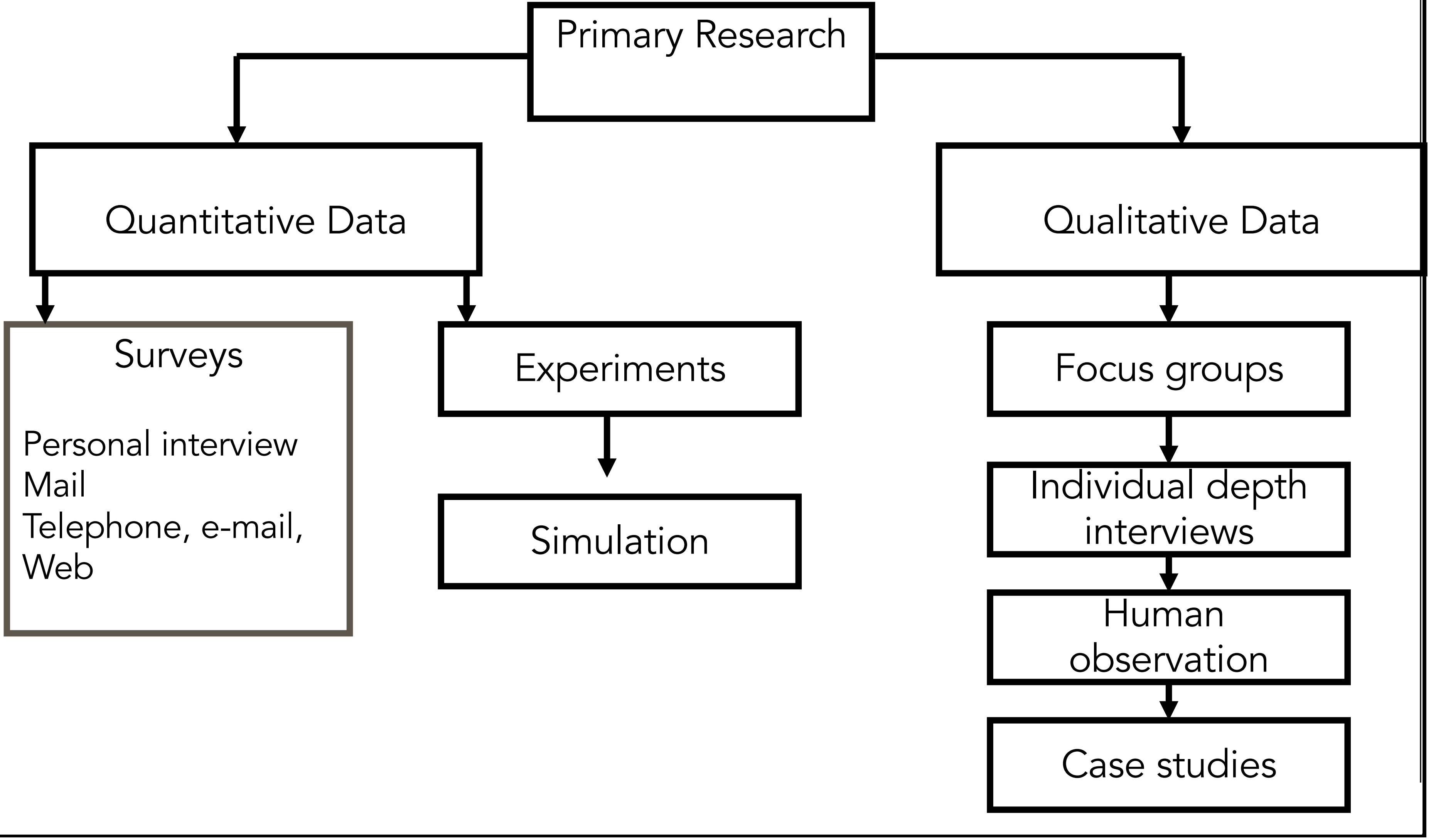
- **Primary:**

  Data collected for the first time ("new" data), to answer specific questions. Primary data comes from the researcher for the purpose of the specific purpose it hand.

- **Secondary:**

  Published information available from other sources that has already been gathered. Collected by others and re-used. Often (but not always) collected for a different use

Primary Research Methods & Techniques



Primary Research

Quantitative Data

Qualitative Data

Surveys

Personal interview
Mail
Telephone, e-mail,
Web

Experiments

Simulation

Focus groups

Individual depth interviews

Human observation

Case studies

Secondary data: basic characteristics

- Secondary data tend to emerge from three principal kinds of collection processes:
  - Survey data: collection for research purposes, coherent research design, well-defined sampling process, intent to generalize
  - Administrative data: collection for program administration or routine record-keeping. Routinely collected.
  - Census
  - qualitative sources (qualitative official documents, twitter,…)

- **Secondary data may be available either as:**
  - Microdata: individual level records for a unit of analysis
  - Aggregate data: summary counts or statistics across multiple units (cities, households, regions,…)

- **Secondary data may be available either as:**
  - Cross-sectional: data collected at a single point in time
  - Longitudinal data: data collected for the same unit of observation at multiple points in time

Different Types of data

- Cross-Sectional Data
- Time Series Data
- Panel Data

Cross-sectional data

- Cross-section data are data on one or more variables collected *at the same point in time.*

  - *Examples:*

✓ *Survey data- questionnaire (microdata).*

✓ *Macro data relating to different economic entities: countries, banks at a particular point in time.*

– Only source of variation is across individuals (or whatever the unit of observation).

Time series data

A time series is a set of observations on the values that a variable takes *at different times.*

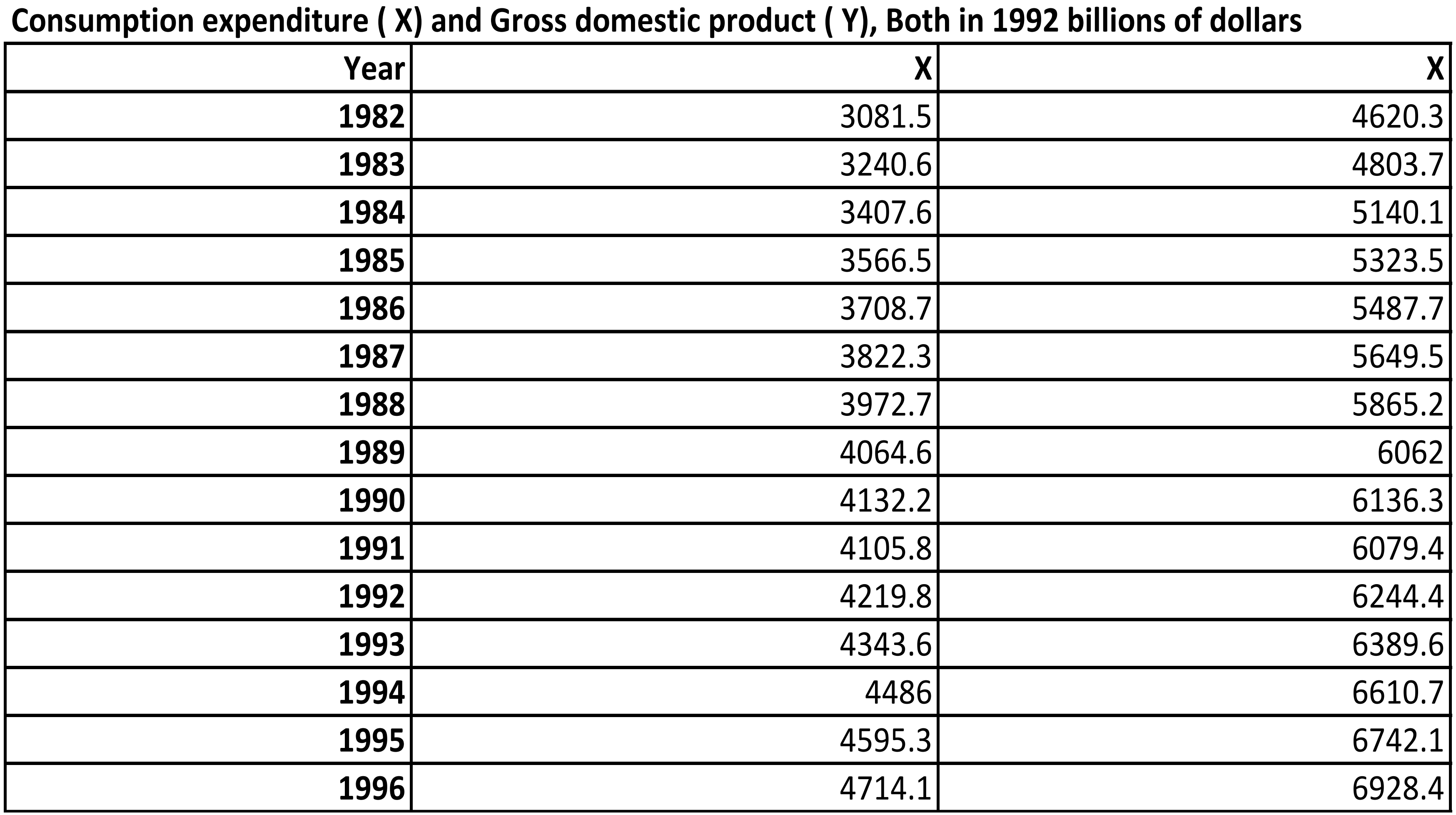
*D*ata may be collected at regular time intervals:

- ◦Minutely and Hourly- collected literally continuously
- ◦Daily- e.g., Financial time series-Stock prices, exchange rates; weather reports- rainfall, temperature
- ◦……
- ◦Monthly- e.g., consumer price index
- ◦Quarterly- e.g., GDP
- ◦…..
- ◦Annually- e.g., Fiscal data

- *Data matrix*

time	variable 1	variable 2	variable 4	etc
t0	x	x	x	x
t1	x	x	x	x
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
T	X	X	X	X

Example: Consumption and Income in US (annual)



Consumption expenditure (X) and Gross domestic product (Y), Both in 1992 billions of dollars

Year	X	X
1982	3081.5	4620.3
1983	3240.6	4803.7
1984	3407.6	5140.1
1985	3566.5	5323.5
1986	3708.7	5487.7
1987	3822.3	5649.5
1988	3972.7	5865.2
1989	4064.6	6062
1990	4132.2	6136.3
1991	4105.8	6079.4
1992	4219.8	6244.4
1993	4343.6	6389.6
1994	4486	6610.7
1995	4595.3	6742.1
1996	4714.1	6928.4

Panel data

- Combination of both time and cross-section data

- *micropanel* data: where a cross-sectional unit (say, individual, family, firm) is surveyed over time.
- Surveying same individual over time is able to provide useful information on the dynamics of individual/household/firm behavior

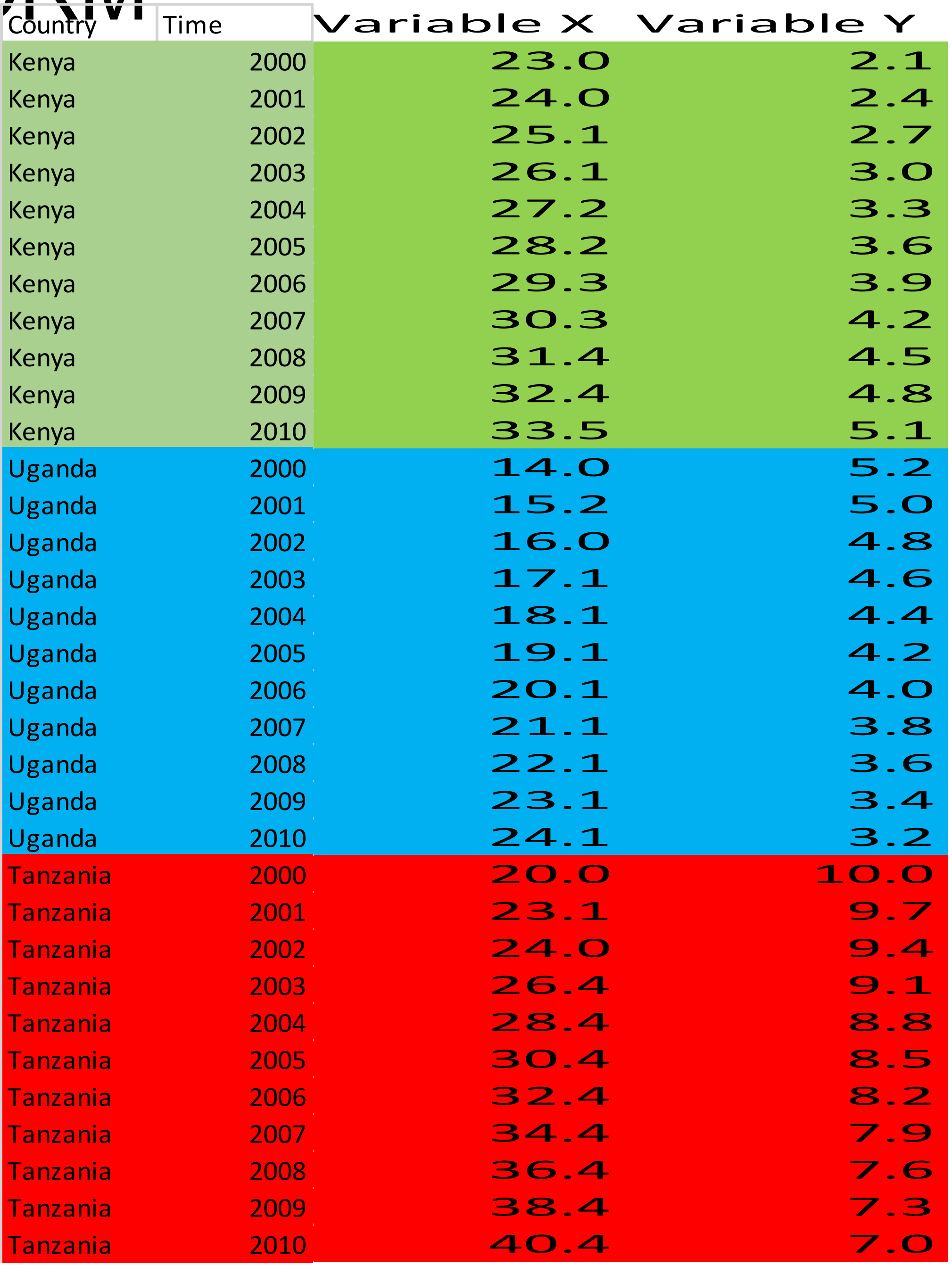
- Common example: Labor Force Surveys
  - Take information about individuals
  - Usually contains time invarying for any individual (race, sex, education level)
  - Usually contains time varying for any given individual (employed last week)

  - Can use both "within" (for an individual over time) and "between variation (across individuals in a given time)

Example 1

	Variable X			Variable Y		
	Kenya	Uganda	Tanzania	Kenya	Uganda	Tanzania
2000	23.0	14.0	20.0	2.1	5.2	10.0
2001	24.0	15.2	23.1	2.4	5.0	9.7
2002	25.1	16.0	24.0	2.7	4.8	9.4
2003	26.1	17.1	26.4	3.0	4.6	9.1
2004	27.2	18.1	28.4	3.3	4.4	8.8
2005	28.2	19.1	30.4	3.6	4.2	8.5
2006	29.3	20.1	32.4	3.9	4.0	8.2
2007	30.3	21.1	34.4	4.2	3.8	7.9
2008	31.4	22.1	36.4	4.5	3.6	7.6
2009	32.4	23.1	38.4	4.8	3.4	7.3
2010	33.5	24.1	40.4	5.1	3.2	7.0

- LONG FORM



Country	Time	Variable X	Variable Y
Kenya	2000	23.0	2.1
Kenya	2001	24.0	2.4
Kenya	2002	25.1	2.7
Kenya	2003	26.1	3.0
Kenya	2004	27.2	3.3
Kenya	2005	28.2	3.6
Kenya	2006	29.3	3.9
Kenya	2007	30.3	4.2
Kenya	2008	31.4	4.5
Kenya	2009	32.4	4.8
Kenya	2010	33.5	5.1
Uganda	2000	14.0	5.2
Uganda	2001	15.2	5.0
Uganda	2002	16.0	4.8
Uganda	2003	17.1	4.6
Uganda	2004	18.1	4.4
Uganda	2005	19.1	4.2
Uganda	2006	20.1	4.0
Uganda	2007	21.1	3.8
Uganda	2008	22.1	3.6
Uganda	2009	23.1	3.4
Uganda	2010	24.1	3.2
Tanzania	2000	20.0	10.0
Tanzania	2001	23.1	9.7
Tanzania	2002	24.0	9.4
Tanzania	2003	26.4	9.1
Tanzania	2004	28.4	8.8
Tanzania	2005	30.4	8.5
Tanzania	2006	32.4	8.2
Tanzania	2007	34.4	7.9
Tanzania	2008	36.4	7.6
Tanzania	2009	38.4	7.3
Tanzania	2010	40.4	7.0

AGGREGATED DATA

Macro level quantitative studies analyse relationships between aggregate level characteristics indexes.

The unit of analysis is the state, the community or some other aggregations of units.

Most of this studies rely heavily on published statistics (World bank, OECD, WHO,…)

Examples of Research questions:

What are the political, social and economic causes of inequality?
Why inequality at national level is increasing and it is increasing more in some places than in others?
What are the social and economic factors that influence economic development at national or regional level?

What are the impact of national/regional policies? in health, education, environment….

AGGREGATED DATA: some issues

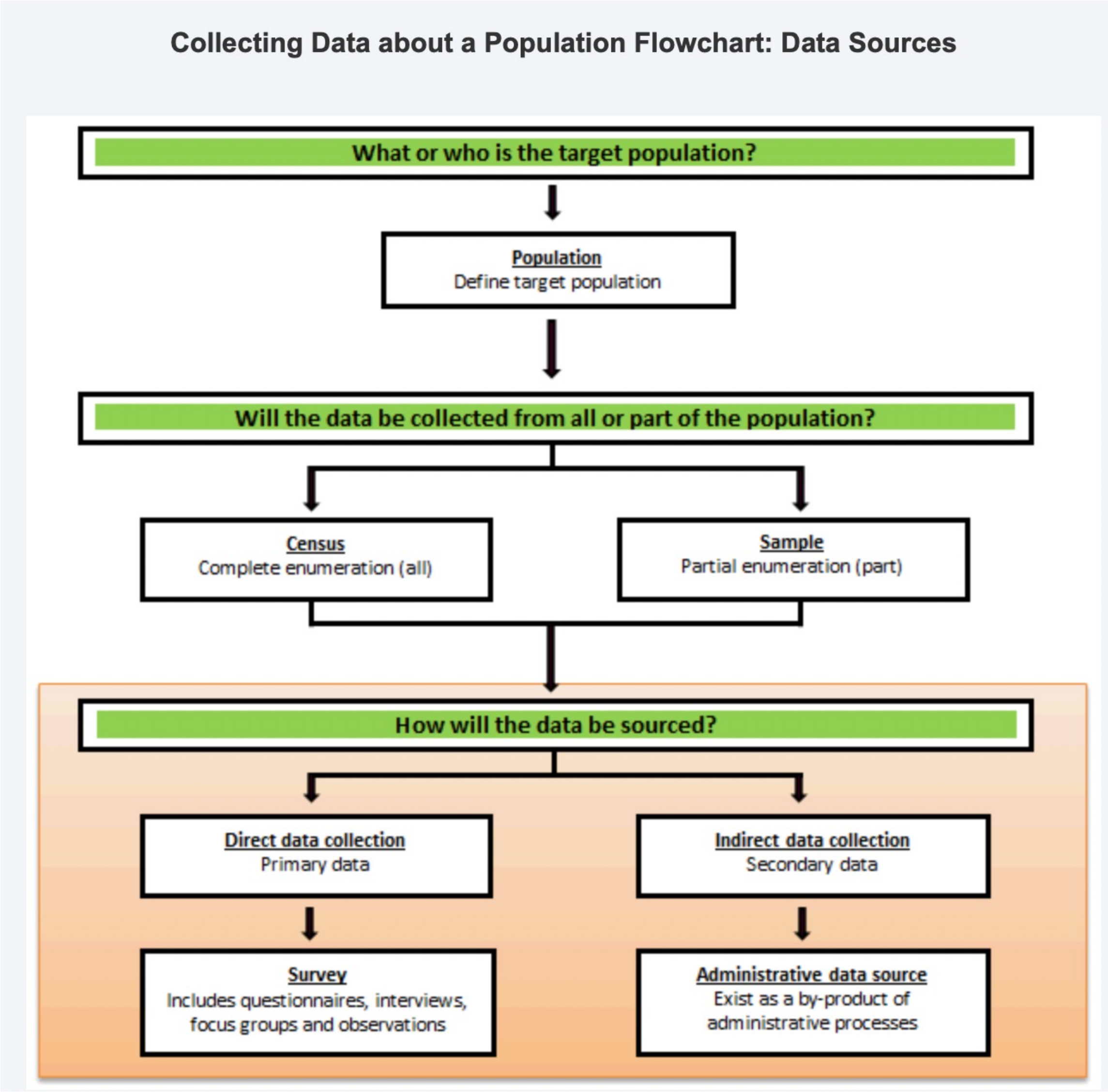
- Trustable

- Comparable

- **Ecological fallacy**: cannot infer about relationships at disaggregated level (i.e. relationship between income and health at individual level might be different).

- **Causality**: difficult to detect

Let’s concentrate on observational studies



Collecting Data about a Population Flowchart: Data Sources

What or who is the target population?

Population
Define target population

Will the data be collected from all or part of the population?

Census
Complete enumeration (all)

Sample
Partial enumeration (part)

How will the data be sourced?

Direct data collection
Primary data

Indirect data collection
Secondary data

Survey
Includes questionnaires, interviews, focus groups and observations

Administrative data source
Exist as a by-product of administrative processes

Australian Bureau of Statistics

✓**Parameter:** fixed (often unknown) number that summarize a characteristics of the the population (average, proportion,…). It is based on all the elements within that population.

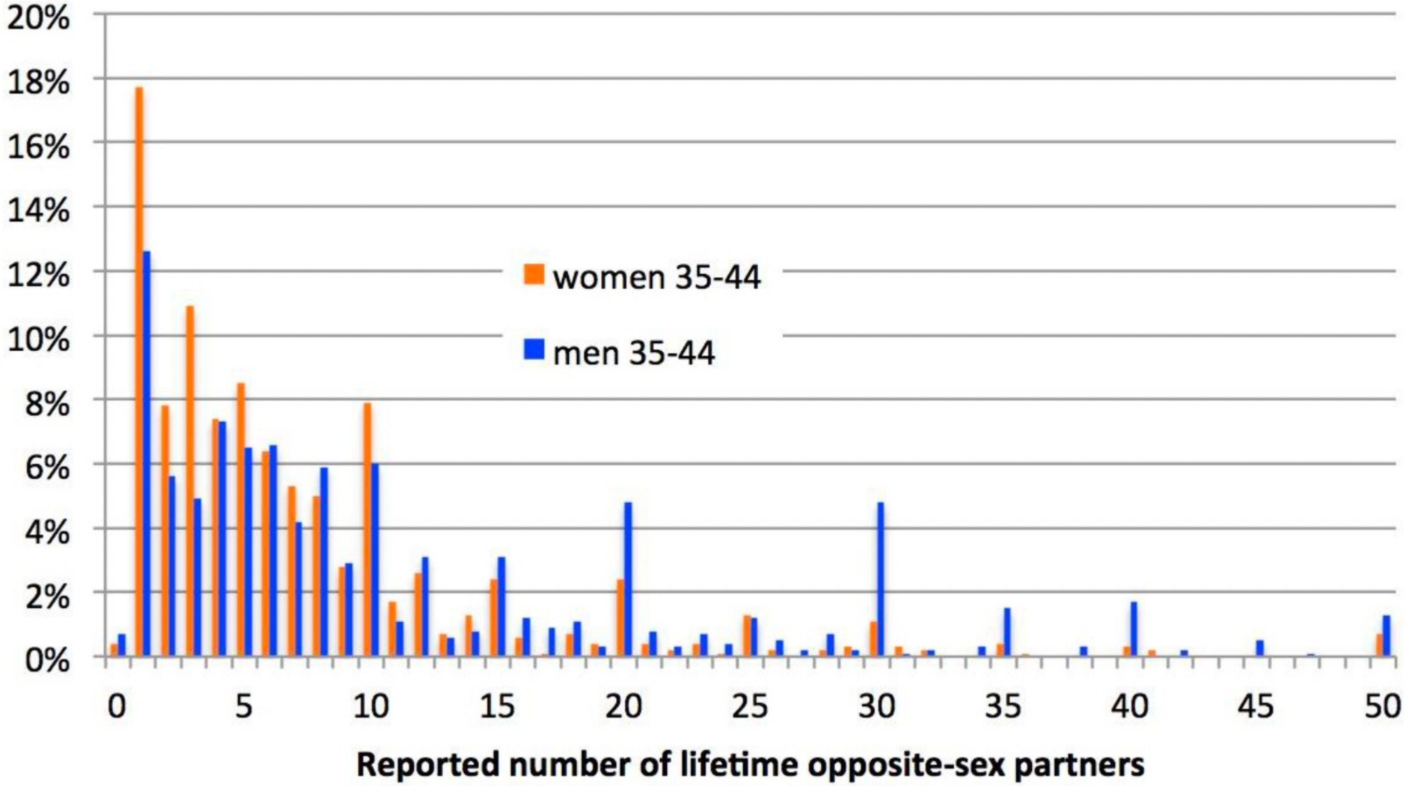
✓**Statistics:** known number that summarize a characteristics of the sample. A statistic is often used to point estimate the parameter in the population.

It is important to note that a sample statistic can differ from sample to sample whereas a population parameter is constant for a population!

How many sexual partners have people in Britain had in their lifetime?

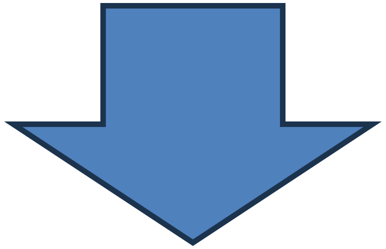
- **Problem**: cannot know this as a fact
- **Plan**: survey in which people are carefully asked about the sexual activity (Natsal)
- **Data**: reports of numbers of partners
- **Analysis**: plotting and summary statistics

How many sexual partners do people report?



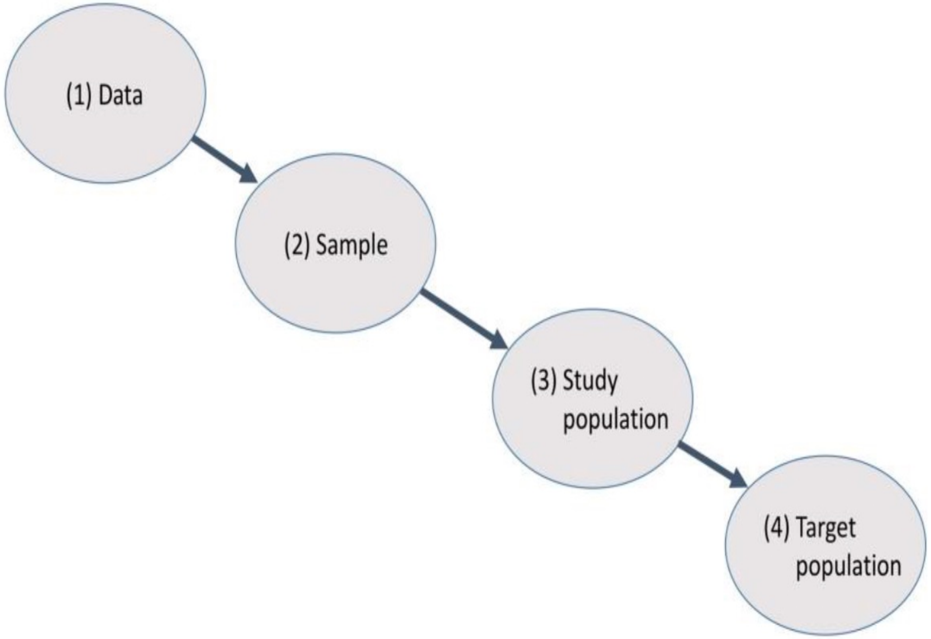
Reported number of sexual partners in lifetime	Men aged 35–44	Women aged 35–44
Mean	14.3	8.5
Median	8	5
Mode	1	1
Range	0 to 500	0 to 550
Inter-quartile range	4 to 18	3 to 10
Standard deviation	24.2	19.7

The answers to the survey are used to make conclusions about the sexual activity of the **general population** in GB



INFERENCE

INDUCTION PROCESS



From 1 to 2: measurement issues.
We want raw data to be **reliable and valid.**

From 2 to 3: **internal validity**, is the sample really reflection the study phenomenon (representative sample?).

From 3 to 4: **external validity**

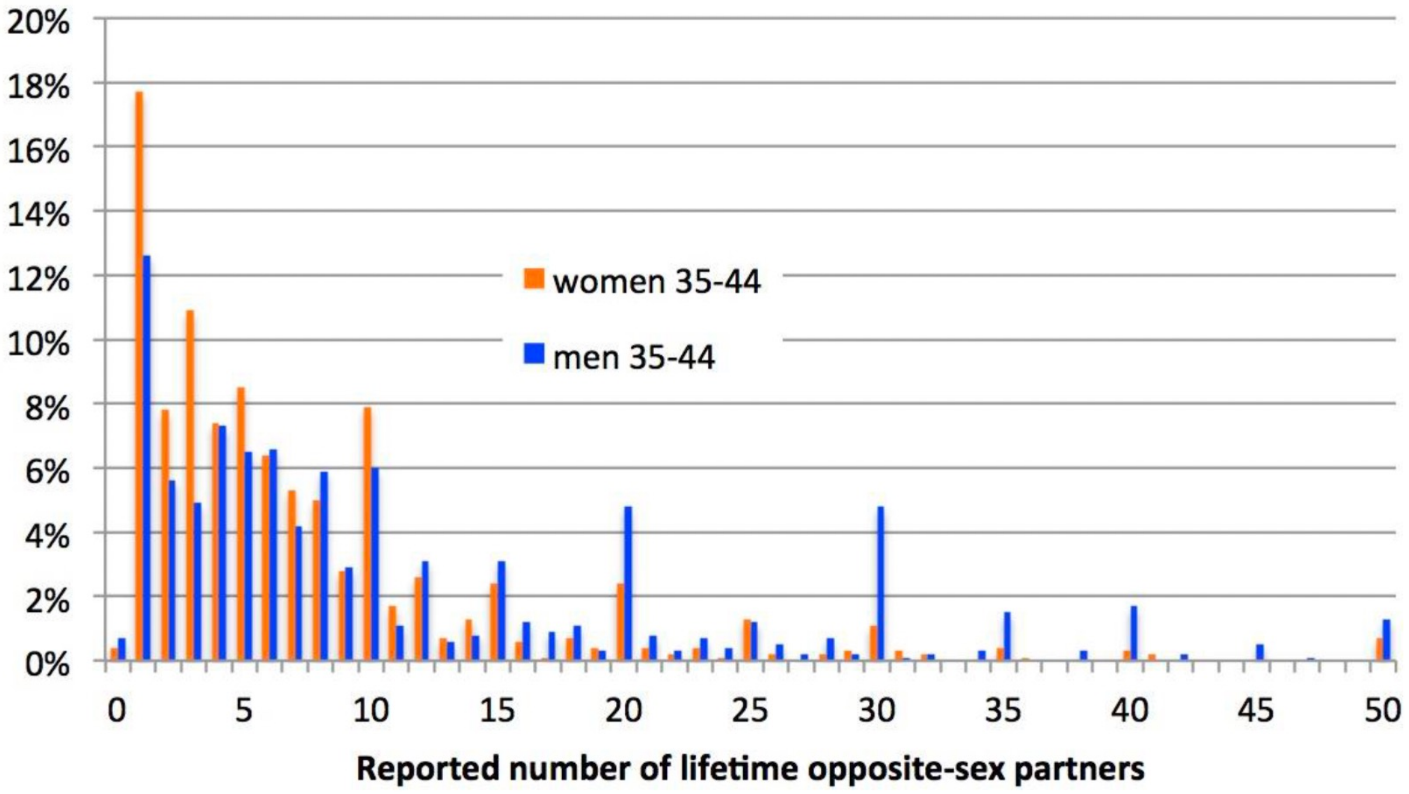
Inference and bias

How many sexual partners have people in Britain ***really*** *had in their lifetime?*

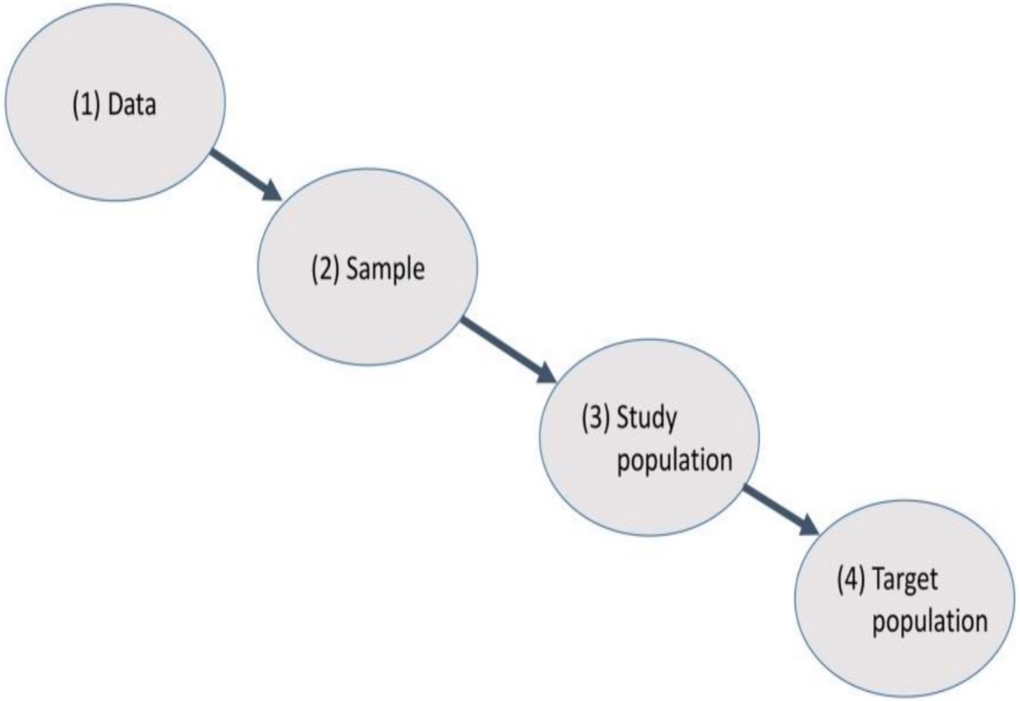
Reported number of sexual partners in lifetime	Men aged 35–44	Women aged 35–44
Mean	14.3	8.5
Median	8	5
Mode	1	1
Range	0 to 500	0 to 550
Inter-quartile range	4 to 18	3 to 10
Standard deviation	24.2	19.7

- **Conclusions**: can we generalise this to the whole population?????

How many sexual partners do people report?



Induction: the stages in generalising from data



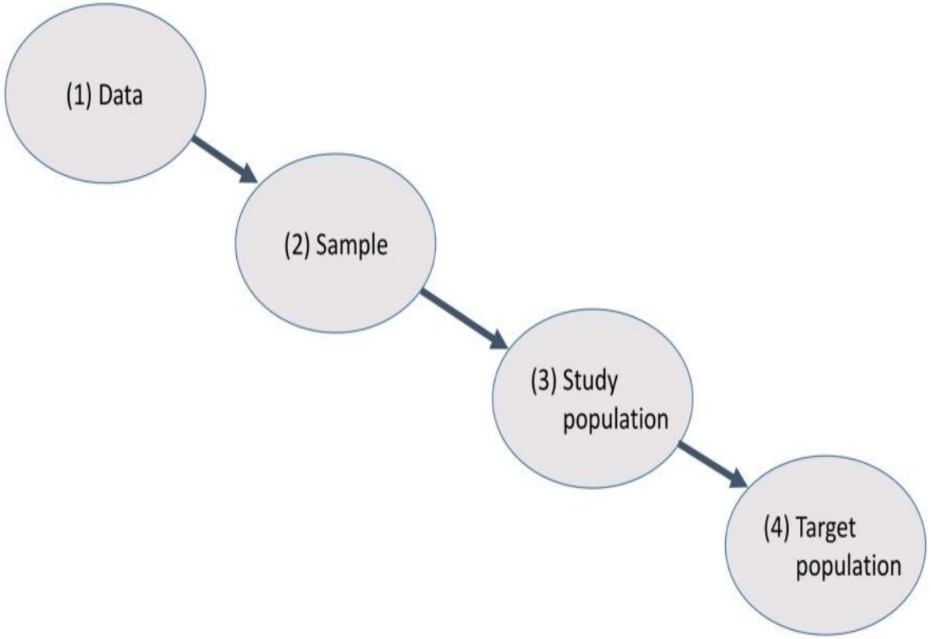
- ***1 to 2.*** How reliable are the reports?
- *Poor memory, social acceptability bias etc*

- ***2 to 3.*** How representative is the sample of those eligible for the study?
- *Random sampling of families (soup), 66% response*

- ***3 to 4.*** How close does the study population match the target population?
- *No people in institutions, etc*

5 - Types and sources of error in statistical data

INDUCTION PROCESS



From 1 to 2: measurement issues. We want raw data to be reliable and valid

From 2 to 3: internal validity, is the sample really reflection the study phenomenon (representative sample)

From 3 to 4: external validity

What is sampling error? (1/2)

Sampling error occurs as a result of using a sample from a population, rather than conducting a census (complete enumeration) of the population.

It refers to the difference between an estimate for a population based on data from a sample and the 'true' value for that population which would result if a census were taken. Sampling errors do not occur in a census, as the census values are based on the entire population.

What is sampling error? (1/2)

Because a sample is typically only a part of the whole population, sample data provide only limited information about the population. As a result, sample statistics are generally imperfect representatives of the corresponding population parameters.

The discrepancy (natural difference that exist by chance) between a sample statistic and its population parameter is called **sampling error**.

Defining and measuring sampling error is a large part of inferential statistics.

Because a sample is typically only a part of the whole population, sample data provide only limited information about the population. As a result, sample statistics are generally imperfect representatives of the corresponding population parameters.

The discrepancy (natural difference that exist by chance) between a sample statistic and its population parameter is called **sampling error**.

Defining and measuring sampling error is a large part of inferential statistics.

Point Estimate

We use the statistic from a sample as a ***point estimate*** for a population parameter.

Point estimates will not match population parameters exactly, but they are our best guess, given the data.

Key questions

- Sample statistics ***vary*** from sample to sample. (they will not match the parameter exactly)

- *KEY QUESTION*: For a given sample statistic, what are plausible values for the population parameter? How much uncertainty surrounds the sample statistic?

- *KEY ANSWER*: It depends on how much the statistic varies from sample to sample!

Sampling error can be measured
and controlled in random samples

What is non-sampling error? (1/3)

Non-sampling error is caused by factors other than those related to sample selection. They arise during data collection activities.

Non-sampling error can occur at any stage of a census or sample study and are not easily identified or quantified.

What is non-sampling error? (2/3)

Non-sampling error can include (but is not limited to):

Coverage error: this occurs when a unit in the sample is incorrectly excluded or included, or is duplicated in the sample (e.g. a field interviewer fails to interview a selected household or some people in a household).

Non-response error: this refers to the failure to obtain a response from some unit because of absence, non-contact, refusal, or some other reason. Non-response can be complete non-response (i.e. no data has been obtained at all from a selected unit) or partial non-response (i.e. the answers to some questions have not been provided by a selected unit).

What is non-sampling error? (2/3)

Response error: this refers to a type of error caused by respondents intentionally or accidentally providing inaccurate responses. This occurs when concepts, questions or instructions are not clearly understood by the respondent; when there are high levels of respondent burden and memory recall required; and because some questions can result in a tendency to answer in a socially desirable way (giving a response which they feel is more acceptable rather than being an accurate response).

•**Interviewer error:** this occurs when interviewers incorrectly record information; are not neutral or objective; influence the respondent to answer in a particular way; or assume responses based on appearance or other characteristics.

•**Processing error:** this refers to errors that occur in the process of data collection, data entry, coding, editing and output.

Examples of question wording which may contribute to non-sampling error.

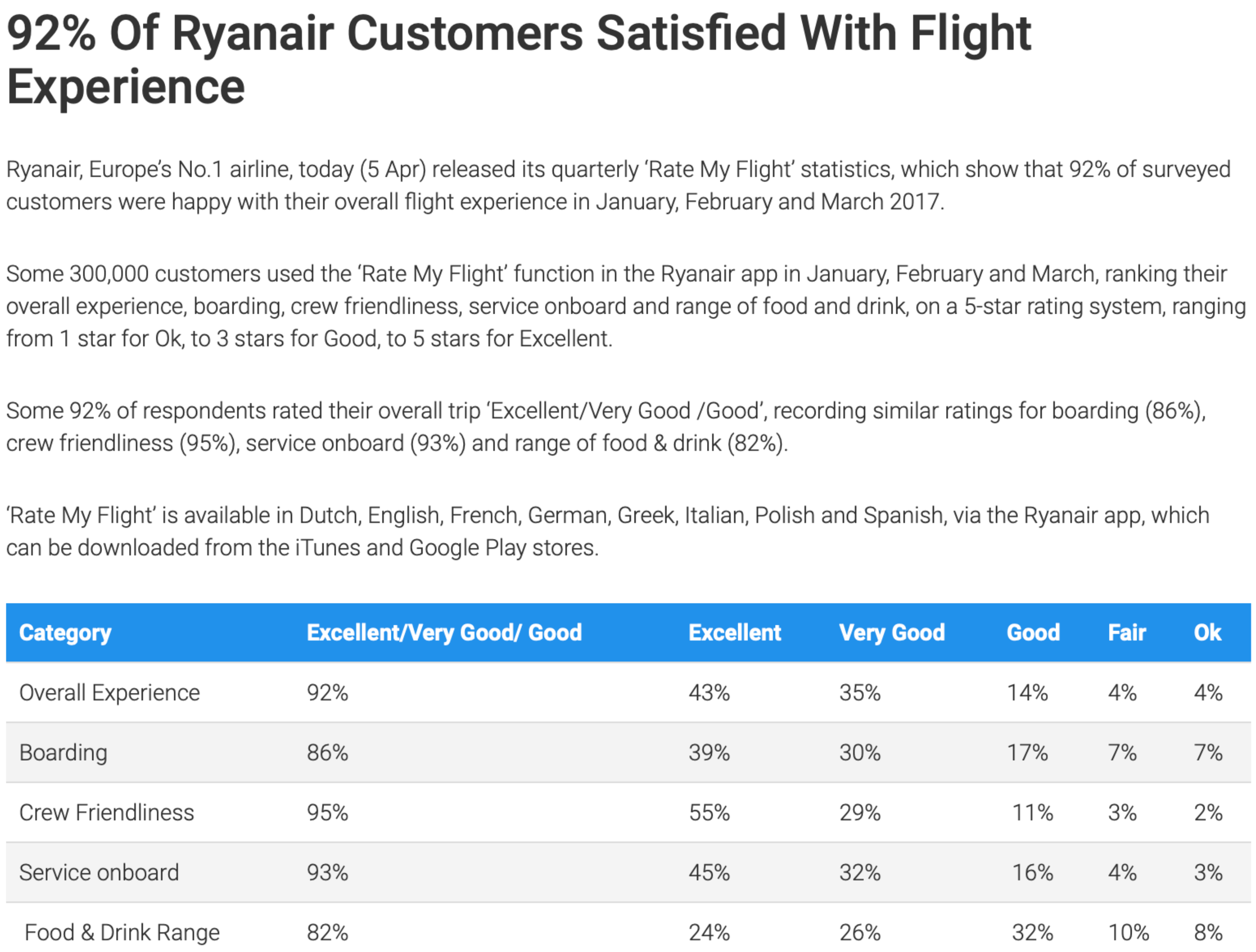
Memory recall:
"How many kilometres did you travel in July last year?"

Socially desirable questions:
"Do you regularly recycle your waste paper and plastics?"

Under reporting:
"How many glasses of alcohol do you drink per week?"

Double-barrelled question:
"Are you happy with the price of, and services offered by, your gym membership?"

Biased survey questions: positive (negative) framing



92% Of Ryanair Customers Satisfied With Flight Experience

Ryanair, Europe's No.1 airline, today (5 Apr) released its quarterly 'Rate My Flight' statistics, which show that 92% of surveyed customers were happy with their overall flight experience in January, February and March 2017.

Some 300,000 customers used the 'Rate My Flight' function in the Ryanair app in January, February and March, ranking their overall experience, boarding, crew friendliness, service onboard and range of food and drink, on a 5-star rating system, ranging from 1 star for Ok, to 3 stars for Good, to 5 stars for Excellent.

Some 92% of respondents rated their overall trip 'Excellent/Very Good /Good', recording similar ratings for boarding (86%), crew friendliness (95%), service onboard (93%) and range of food & drink (82%).

'Rate My Flight' is available in Dutch, English, French, German, Greek, Italian, Polish and Spanish, via the Ryanair app, which can be downloaded from the iTunes and Google Play stores.

Category	Excellent/Very Good/ Good	Excellent	Very Good	Good	Fair	Ok
Overall Experience	92%	43%	35%	14%	4%	4%
Boarding	86%	39%	30%	17%	7%	7%
Crew Friendliness	95%	55%	29%	11%	3%	2%
Service onboard	93%	45%	32%	16%	4%	3%
Food & Drink Range	82%	24%	26%	32%	10%	8%

https://corporate.ryanair.com/news/170405-92-of-ryanair-customers-satisfied-with-flight-experience/

The greater the error the less reliable are the results of the study.

A credible data source will have measures in place throughout the data collection process to minimise the amount of error and will also be transparent about the size of the expected error so that users can decide whether the data are 'fit for purpose'.