Applied Statistics

Feb 02 2020

Prof.ssa Chiara Seghieri,

Laboratorio di Management e Sanità, Istituto di Management,
Scuola Superiore Sant'Anna, Pisa
c.seghieri@santannapisa.it

DI COSA CI OCCUPEREMO IN QUESTO CORSO?

Delle tecniche statistiche applicate alle scienze sociali.

Al di là delle formule e dei concetti teorici che si imparano in un corso di Statistica di base, prendendo come riferimento esempi pratici, analizzeremo concetti e metodi dell'analisi statistica univariata e multivariata che più frequentemente trovano applicazione nell'analisi delle popolazioni.

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

(Dal ***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).

Statistics is…

THE WALL STREET JOURNAL.

Subscribe | €1 for 2 Mont

ECONOMY | U.S. ECONOMY

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

World's Best Cities To Live In 2019

Global Finance **selects the world's 10 best cities to live in based on four reputable rankings.**

FINANCIAL TIMES

The gap in giving

How charities are changing strategy to cope. Page 7

Welcome to the new Middle Ages

Comment, Page 9

House price falls forecast to continue

Jump start Britain thaws out while ski competition hots up

BBC | Sign in | News | Sport | Reel | Worklife | Travel

NEWS

Home | US Election | Coronavirus | Video | World | UK | Business | Tech

Business | Market Data | Global Trade | Companies | Entrepreneurship | Tech

Number of Americans in poverty hits record high

MENU | CERCA

la Repubblica

ABBONATI | QUOTIDIANO R | ACCEDI

La classifica degli ospedali: i top al Nord. In Toscana cure migliori. Napoli maglia nera

Le strutture lombarde più forti degli scandali: in sei nella top ten. Nella nostra elaborazione sui dati Agenas la maglia nera è del Federico II

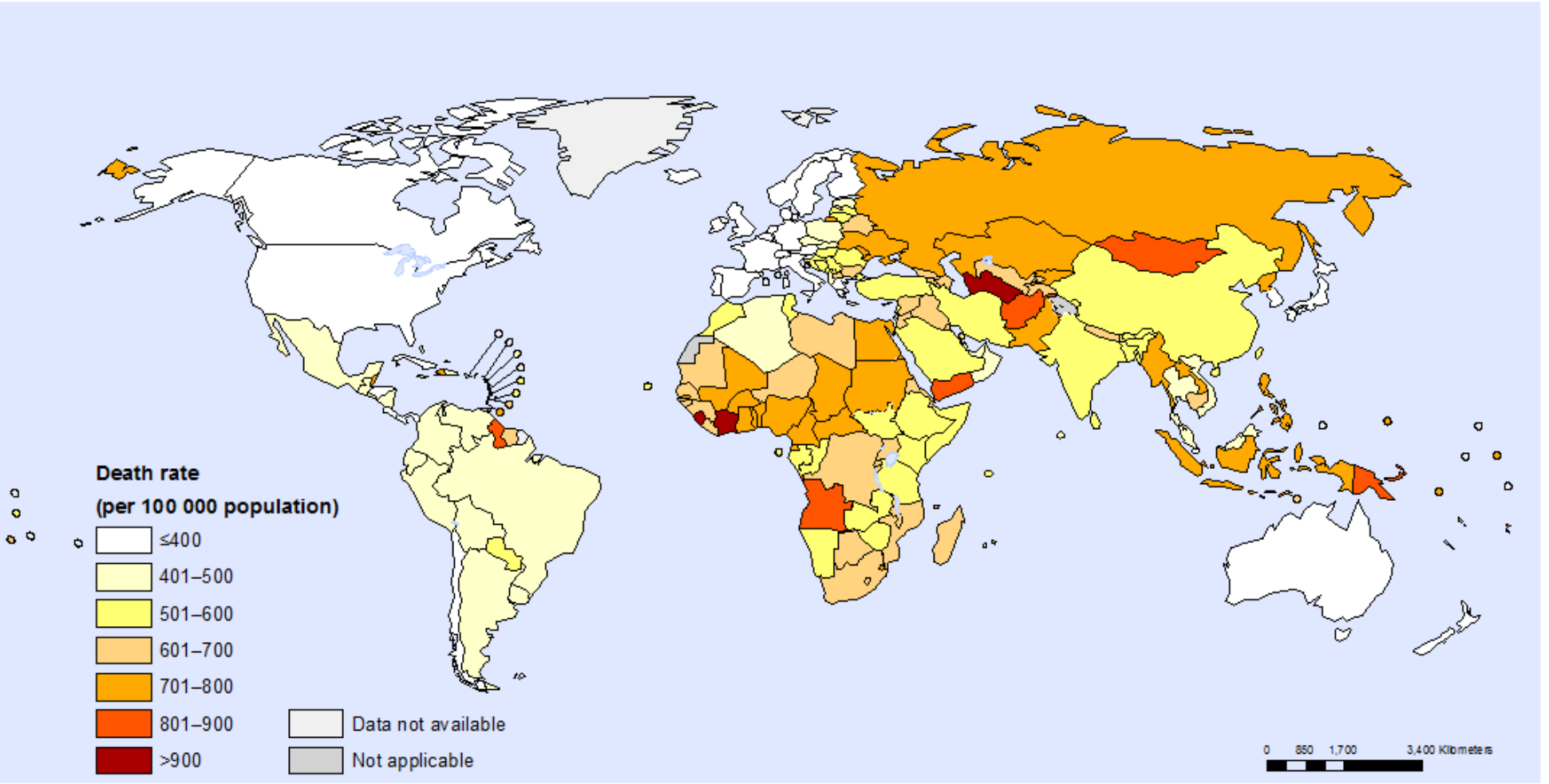
di MICHELE BOCCI e FABIO TONACCI

03 OTTOBRE 2013 | PUBBLICATO PIÙ DI UN ANNO FA | 3 MINUTI DI LETTURA

Antibiotic use before cancer treatment cuts survival time - study

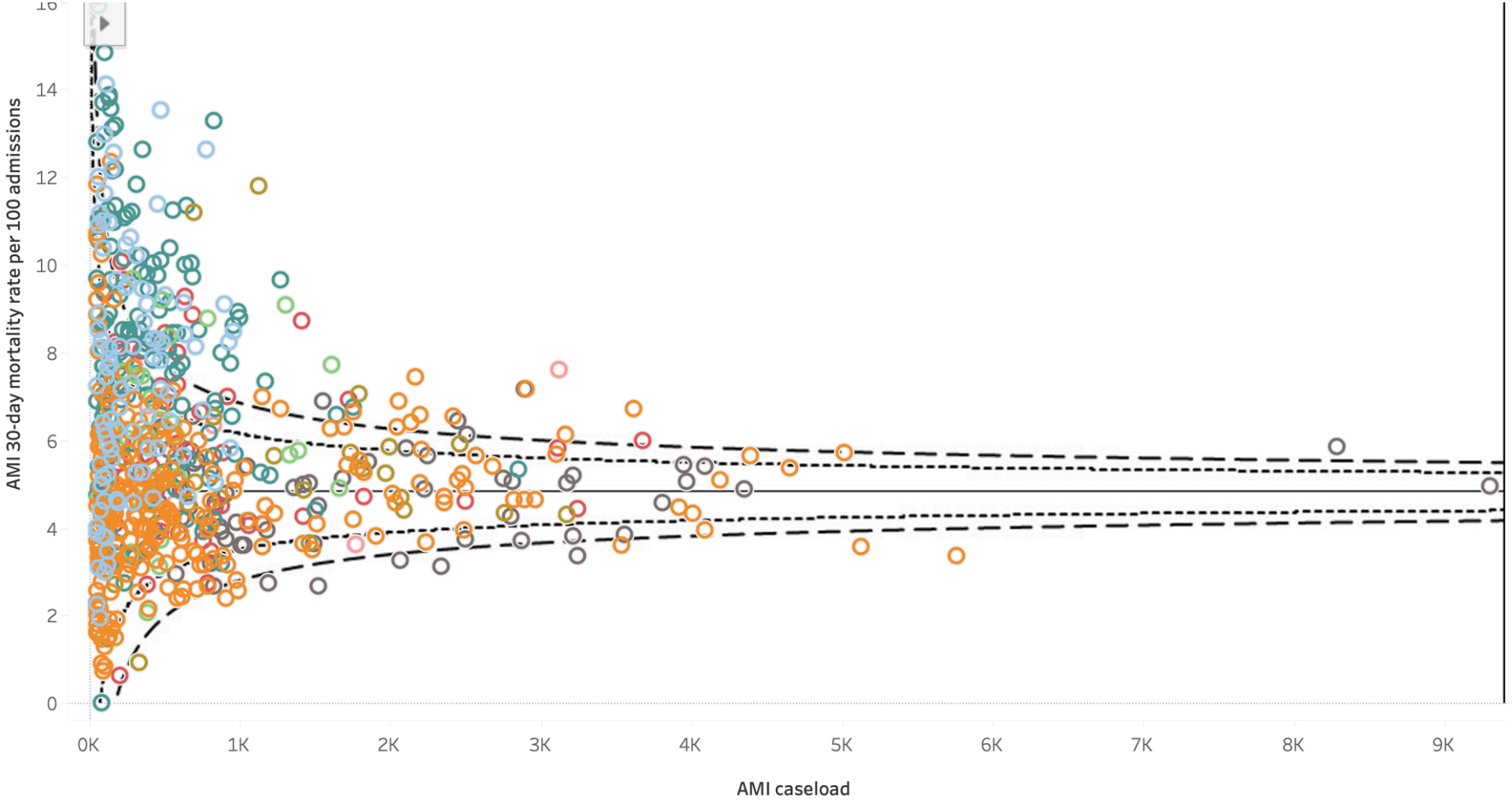
Patients live longer if they do not take antibiotics in month before immunotherapy

Deaths due to noncommunicable diseases: age-standardized death rate (per 100 000 population)
Both sexes, 2015



Data Source: World Health Organization
Map Production: Information Evidence and Research (IER)
World Health Organization

World Health Organization



	Belgium	Canada	Ireland	Israel	Korea	Norway	Slovenia	Sweden
Total hospitals	92	260	30	26	163	63	11	67
Hospitals above 99.7% control limit	29	15	6	3	65	11	2	5
Hospitals above 95% control limit	40	28	10	6	91	22	5	7
Hospitals below 95% control limit		26	1	2	2	2	1	13
Hospitals below 99.7% control limit		5		1				5

OECD data 2019

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.

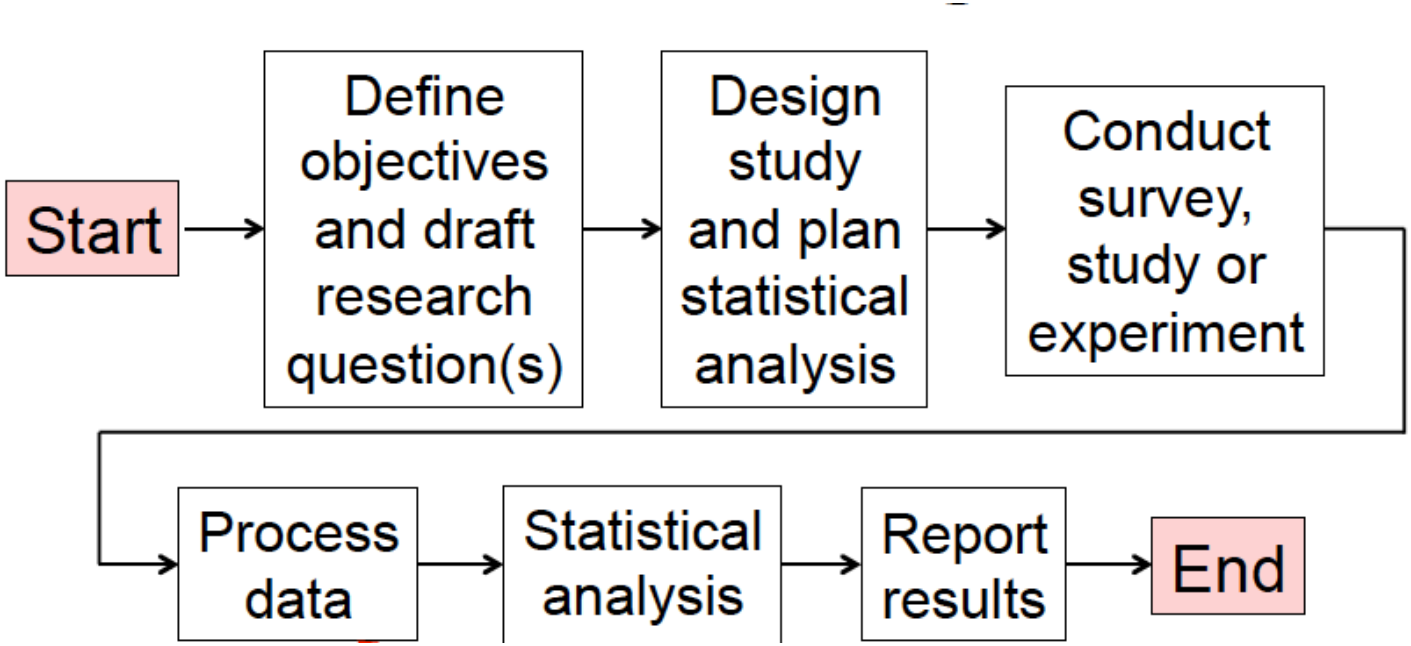
Statistics in the context of a general process of investigation:

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:

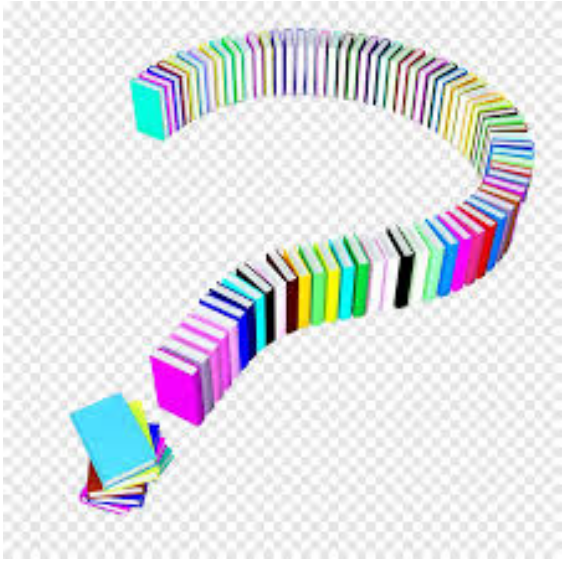
How best can we collect data?
How should it be analyzed?
And what can we infer from the analysis?

The research study process



What is a statistical question?

question that require statistical analysis for the answers



A well-written statistical question refers to:

a **population** of interest (collective phenomenon),
a **variable** 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).

Examples of statistical questions in social sciences

Do neighbourhoods with high rates of rented accommodation contain high rates of unemployed persons?

Are there differences in sports participation by ethnic group?

Are wealthy people happier?

Do women get paid less than men?

Do higher values of GDP correspond to higher life expectancy?

The population

The choice of the statistical population is dictated by the objective of the study.

The population is made of statistical units/subjects (i.e. animals, objects, individuals,…)

It can be composed of the entire **population** (universe) or of a subset of it (sample).

A **variable** is a characteristic or condition of a study subject that can change or take on different values. Height, age, amount of income, grades obtained at school are all examples of variables. Variables may be classified into various categories.

Variables can be categorical or numerical. Most research begins with a general question about the relationship between two variables for a specific group of individuals.

A parameter is?

A Statistic is?

✓**Parameter:** fixed (often unknown) number that summarize a characteristics of the the population (i.e. mean age, median income,…). It is based on all the elements within that population.

✓**Statistic:** 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!

Notation

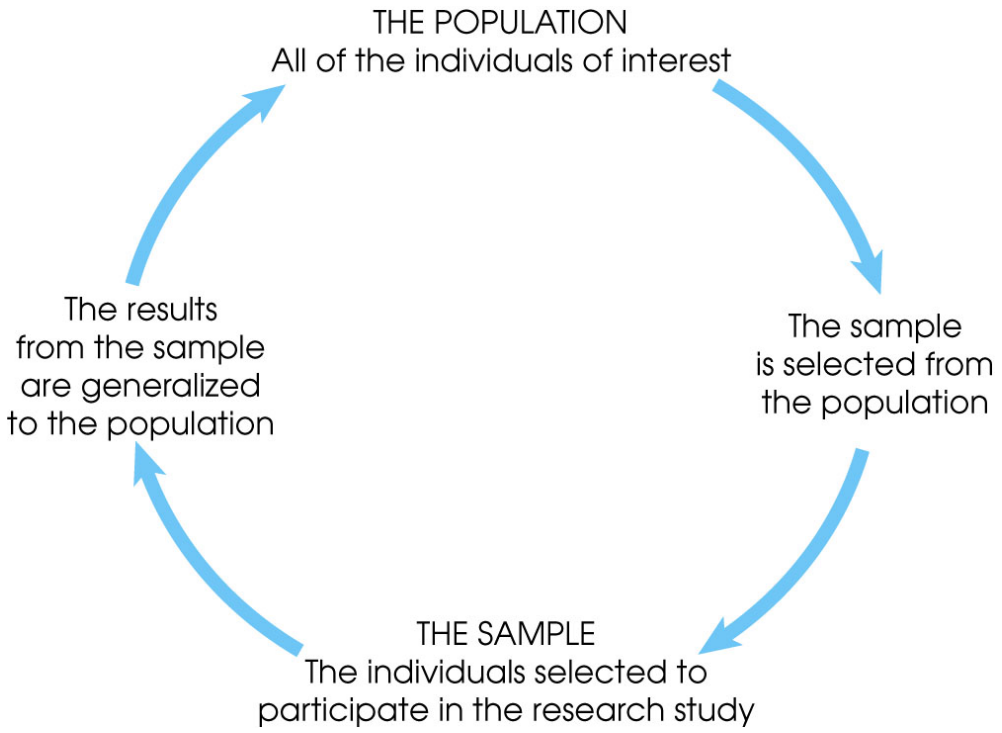
	Parameter		Statistic	
Mean	μ	mu	$\bar{x}$	x-bar
Proportion	p		$\hat{p}$	p-hat
Std. Dev.	σ	sigma	s	
Correlation	ρ	rho	r	

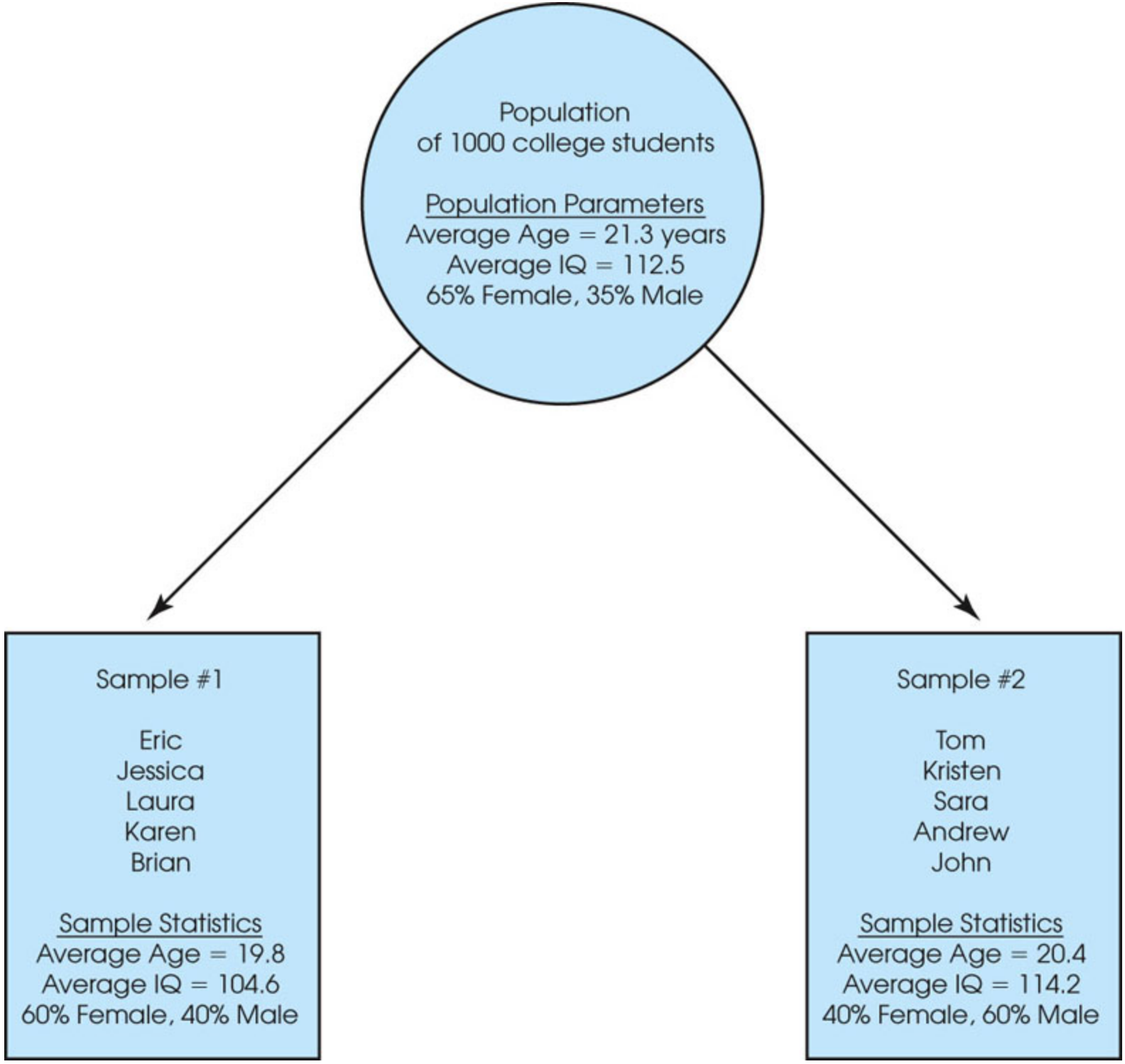
N represents population size
n represents sample size

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 uncertainty**

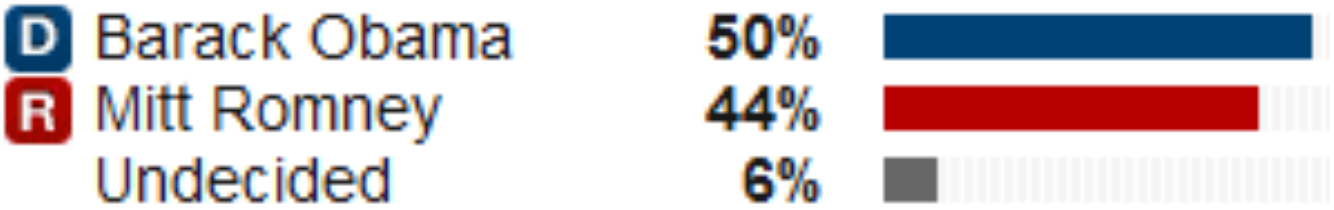



Population
of 1000 college students
Population Parameters
Average Age = 21.3 years
Average IQ = 112.5
65% Female, 35% Male
Sample #1
Eric
Jessica
Laura
Karen
Brian
Sample Statistics
Average Age = 19.8
Average IQ = 104.6
60% Female, 40% Male
Sample #2
Tom
Kristen
Sara
Andrew
John
Sample Statistics
Average Age = 20.4
Average IQ = 114.2
40% Female, 60% Male

Example: Election Polls

- Over the weekend (9/7/12 – 9/9/12), 1000 registered voters were asked who they plan to vote for in the 2012 presidential election

- What proportion of voters plan to vote for Obama?



$\hat{p} = 0.50$ $p = ???$

http://www.politico.com/p/2012-election/polls/president

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.

Example: Election Polls

- Actually, several polls were conducted over the weekend (9/7/12 – 9/9/12):

http://www.politico.com/p/2012-election/polls/president

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.

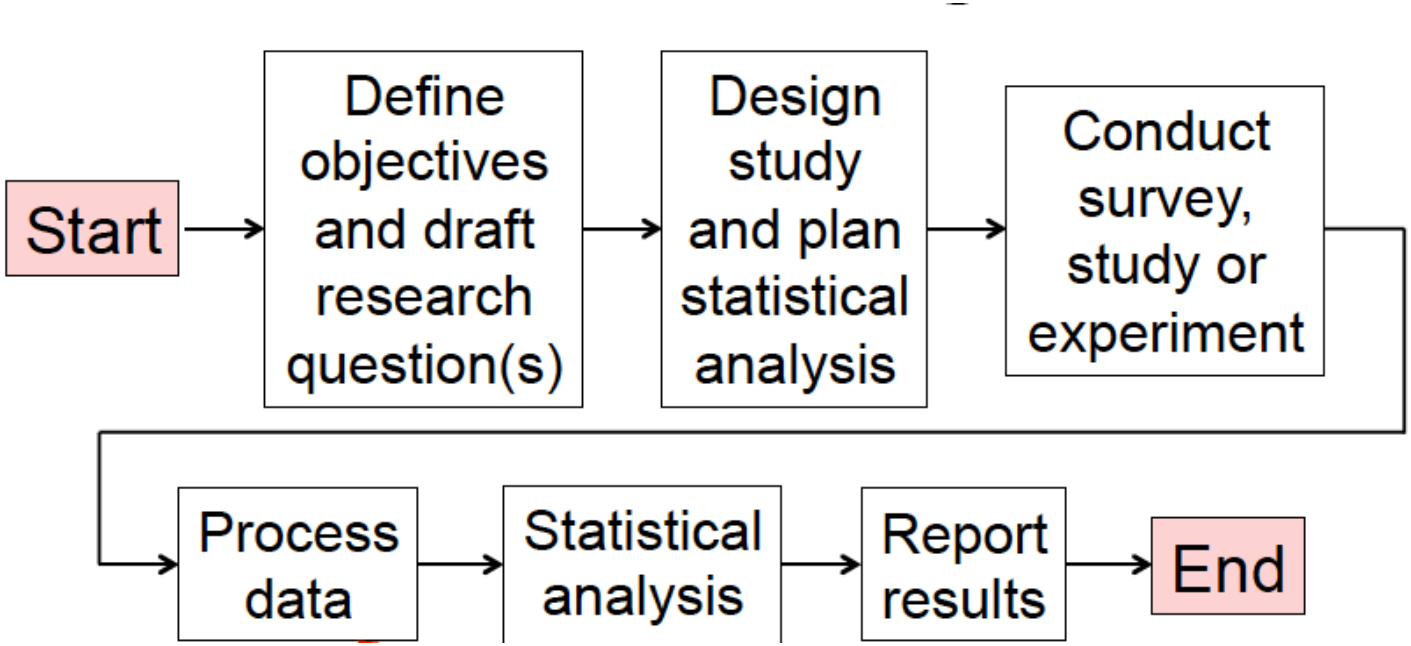
Key questions

- 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!

The research study process



Examples of statistical questions in social sciences

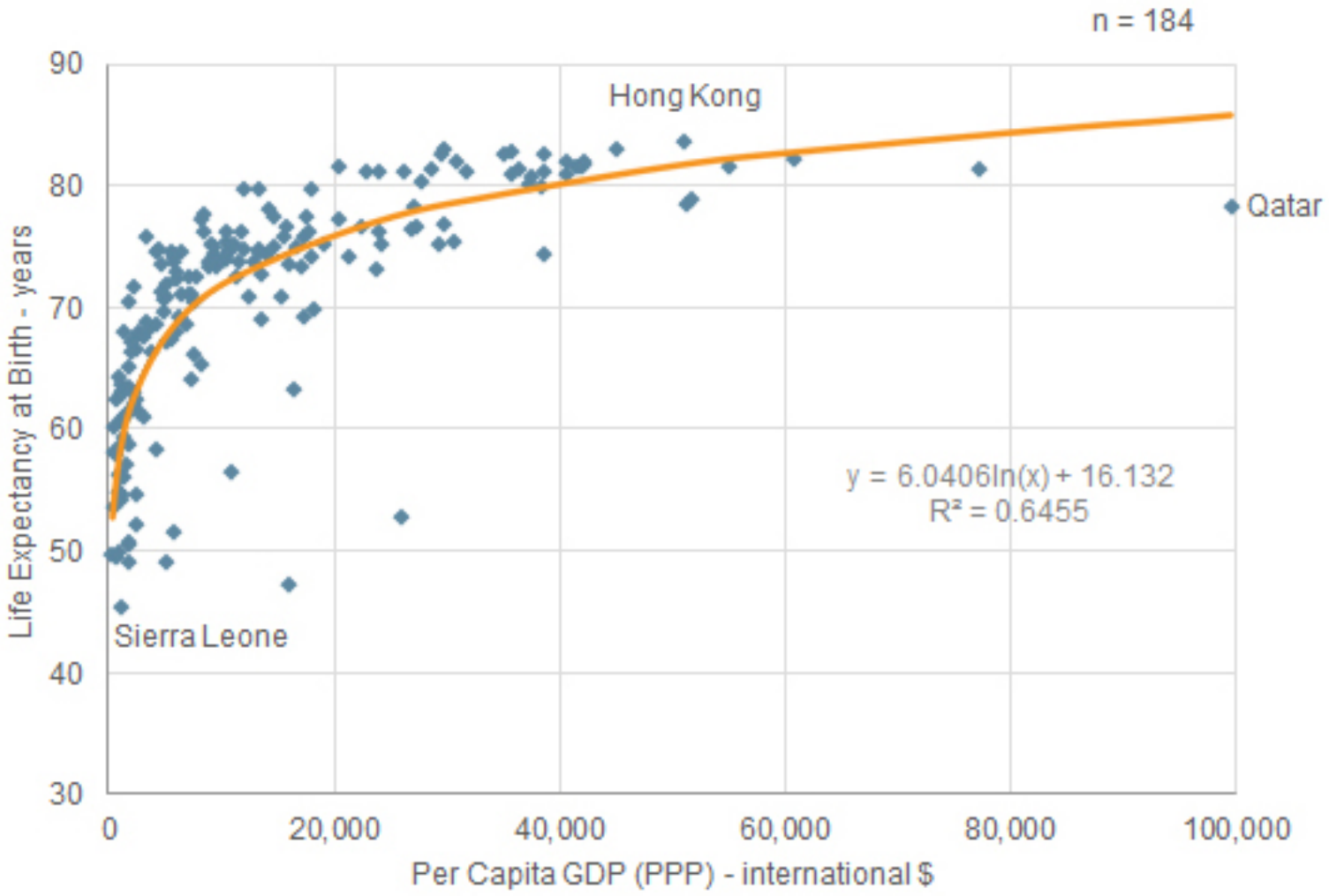
Do neighbourhoods with high rates of rented accommodation contain high rates of unemployed persons?

Are there differences in sports participation by ethnic group?

Are wealthy people happier?
Do women get paid less than men?

Do higher values of GDP correspond to higher life expectancy?

Do higher values of GDP correspond to higher life expectancy?



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 among countries?
Why inequality at national level is increasing and it is increasing more in some places than in others?
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

Data Matrix: aggregated data

City	State	Region	divorces/1000	Educaton	Hhinquality	change	poor	population	n_homicides
Sterling Heig	MI	Midwest	7.461	12.6	0.28	77.6	3.1	109000	1
Sunnyvale	CA	West	10.096	13.2	0.35	11.1	3.7	106600	3
Concord	CA	West	9.287	12.9	0.33	21.2	4.6	103300	3
Fullerton	CA	West	9.976	13.2	0.41	18.7	4.7	102000	2
Independenc	MO	Midwest	10.077	12.5	0.35	0.2	4.9	111800	4
Tempe	AZ	West	12.724	14	0.38	68	5.5	106700	4
Milwaukee	WI	Midwest	6.662	12.6	0.39	-11.3	6.8	636200	50
Tulsa	OK	South	13.603	12.8	0.42	9.3	7.4	360900	31
Honolulu	HI	West	8.109	12.7	0.44	12.4	7.4	365000	33
Virginia Bea	VA	South	7.705	12.8	0.36	52.3	7.7	262200	10
Allentown	PA	N.East	5.604	12.3	0.39	-5.6	8.4	103800	4
Portland	OR	West	10.605	12.8	0.43	-3.6	8.5	366400	32
Albuquerque	NM	West	13.965	12.9	0.4	35.7	9.3	331800	21
Peoria	IL	Midwest	9.931	12.6	0.43	-2.2	9.4	124200	5
Erie	PA	N.East	6.614	12.3	0.39	-7.8	10.2	119100	6
Salt Lake	UT	West	10.268	12.9	0.45	-7.3	10.5	163000	10
Dallas	TX	South	11.96	12.7	0.45	7.1	10.8	904100	271
Berkeley	CA	West	9.287	16.1	0.5	-9.4	11.7	103300	9
Columbus	GA	South	12.613	12.3	0.43	9.3	14.5	169400	16
Rochester	NY	N.East	6.387	12.3	0.42	-18.1	14.5	241700	26

Each row of a data matrix corresponds to a unit, each column corresponds to a variable. Data matrices (nxp) are convenient for recording data as well as analyzing data using a computer. Convention: p denotes the number of variables in a dataset, n denotes the number of study subjects

Data Matrix: individual level data

wave	country	hid	pid	pd001	age	sex	maritalstatu	pe001	personalincome	healthstatus
w2 surve	spain	6068101	60681101	1948	47	male	married	paid emp	2400695	good
w6 surve	denmark	5445702	54457103	1974	25	female	married	paid emp	129000	very goo
w3 surve	spain	5882101	58821101	1934	62	male	married	paid emp	7350000	na
w3 surve	spain	3612101	36121101	1924	72	male	married	retired	1820000	bad
w1 surve	italy	97301	973101	1949	45	male	married	paid emp	40100	good
w6 surve	italy	614001	6140102	1945	54	female	married	housewor	0	very goo
w5 surve	italy	779601	7796103	1971	27	female	never ma	paid emp	12900	good
w4 surve	italy	545301	5453102	1965	32	female	married	self-emp	0	good
w1 surve	spain	5153101	51531103	1946	48	female	widowed	housewor	447996	good
w1 surve	spain	13813101	1.38E+08	1961	33	male	married	paid emp	1458000	fair
w6 surve	ireland	921001	9210101	1942	57	male	married	self-emp	7968	good
w5 surve	italy	352201	3522102	1930	68	female	married	retired	26640	fair
w1 surve	spain	3587101	35871101	1930	64	male	married	retired	1850426	good
w4 surve	ireland	1732601	17326102	1955	42	female	married	paid emp	8976	very goo
w6 surve	spain	2391101	23911101	1951	48	male	married	paid emp	1546726	good
w5 surve	denmark	264601	2646101	1919	79	female	widowed	retired	120612	very goo

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 (*the so-called real time quote*)
- 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 (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

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

Data Sources: Secondary and Primary Data Collection

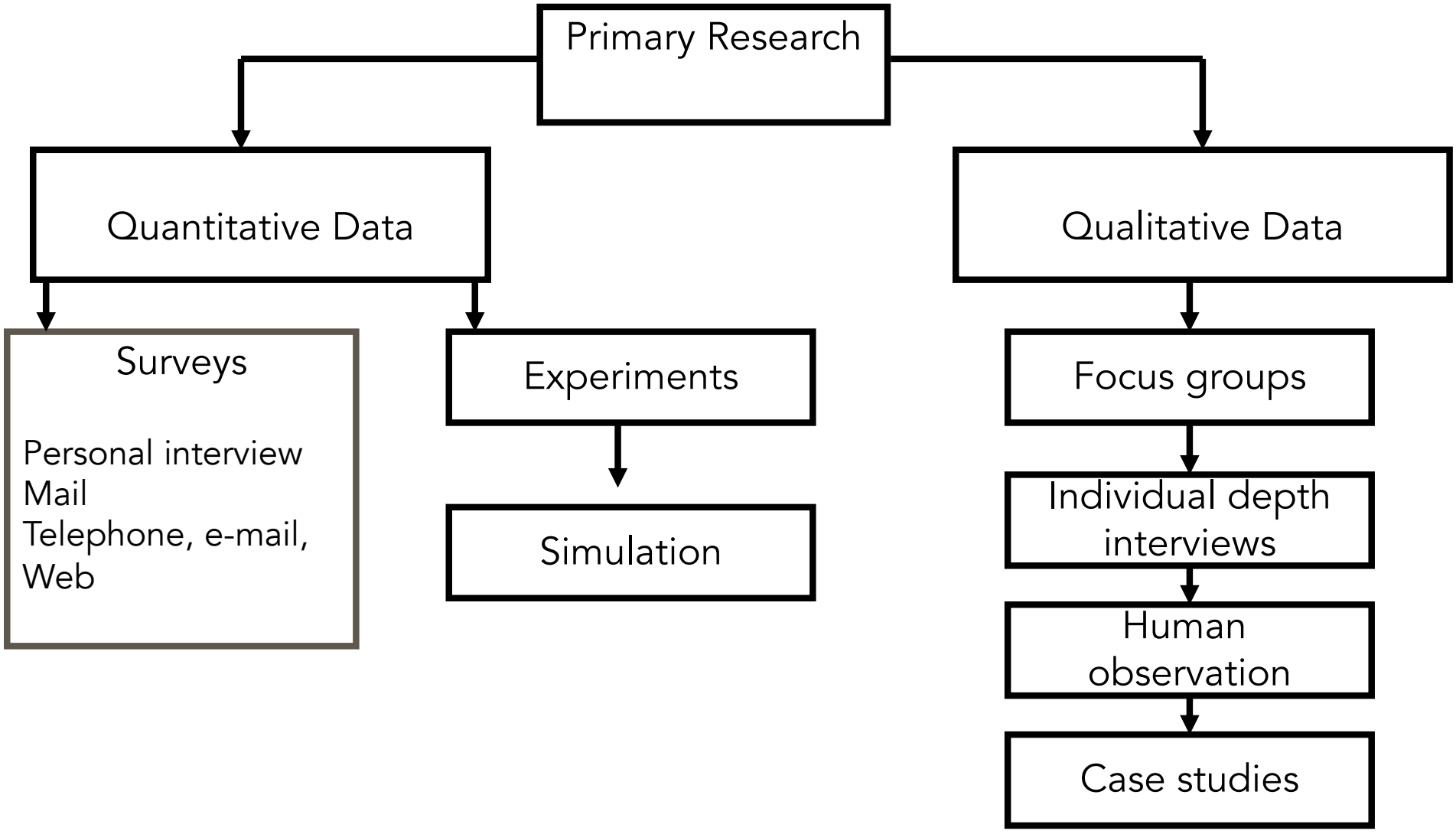
- **Primary:**

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



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

Data Characteristics

Survey Data Characteristics:

- Well defined sampling process
- Individual opinions often gathered

Administrative data characteristics:

- Restricted universe, but can have large amounts of data (millions of observations)
- Data collected only for program administration
- Often linkable to other data
- Rarely includes participant opinion

Datasources (observational studies)

Economics & Socio-demographic
International economic, social, agricultural and health data from the OECD (www.oecd.org) or Eurostat (epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home).
Demographic information from government statistics bureaus in Australia (www.abs.gov.au) or Canada (www.statcan.ca) or Italy (www.istat.it) or US (www.census.gov).
The world bank (http://data.worldbank.org/italian)

Education
U.S. education data is available from the National Center for Education Statistics (http://nces.ed.gov/).

Energy
The U.S. Energy Information Administration provides worldwide usage data and demand forecasts for most any energy source (http://www.eia.gov/).

Health
US health statistics at the Centers for Disease Control (http://www.cdc.gov/nchs/datawh.htm).
International health statistics from the WHO (http://www.who.int/whosis/en/).

Environement:
UNEP - UN Environment Programme

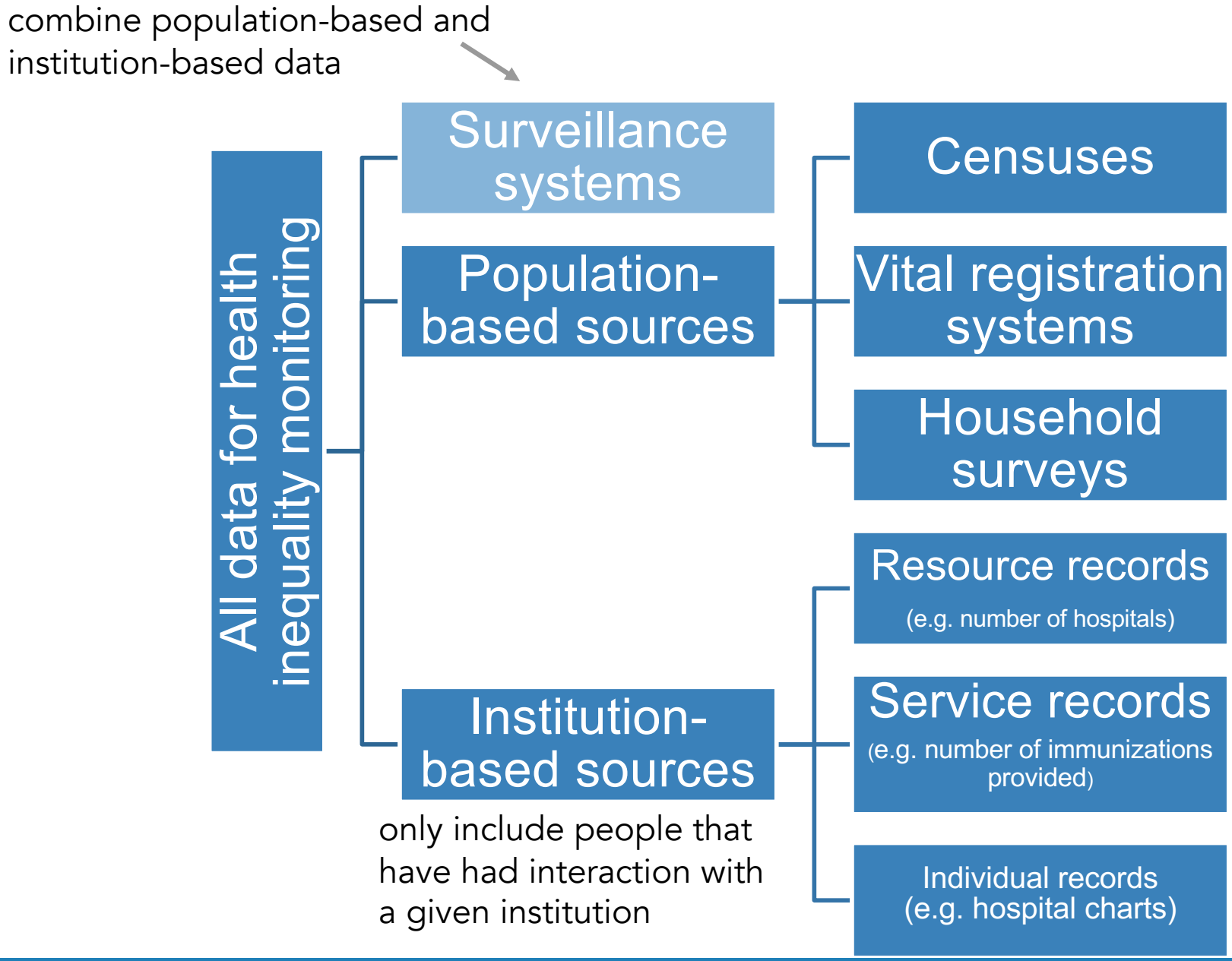
data from different sources could be combined!

Health inequality monitoring: with a special focus on low- and middle-income countries

Data sources

World Health Organization

Data source types



Observational studies and sampling strategies

Type of Studies

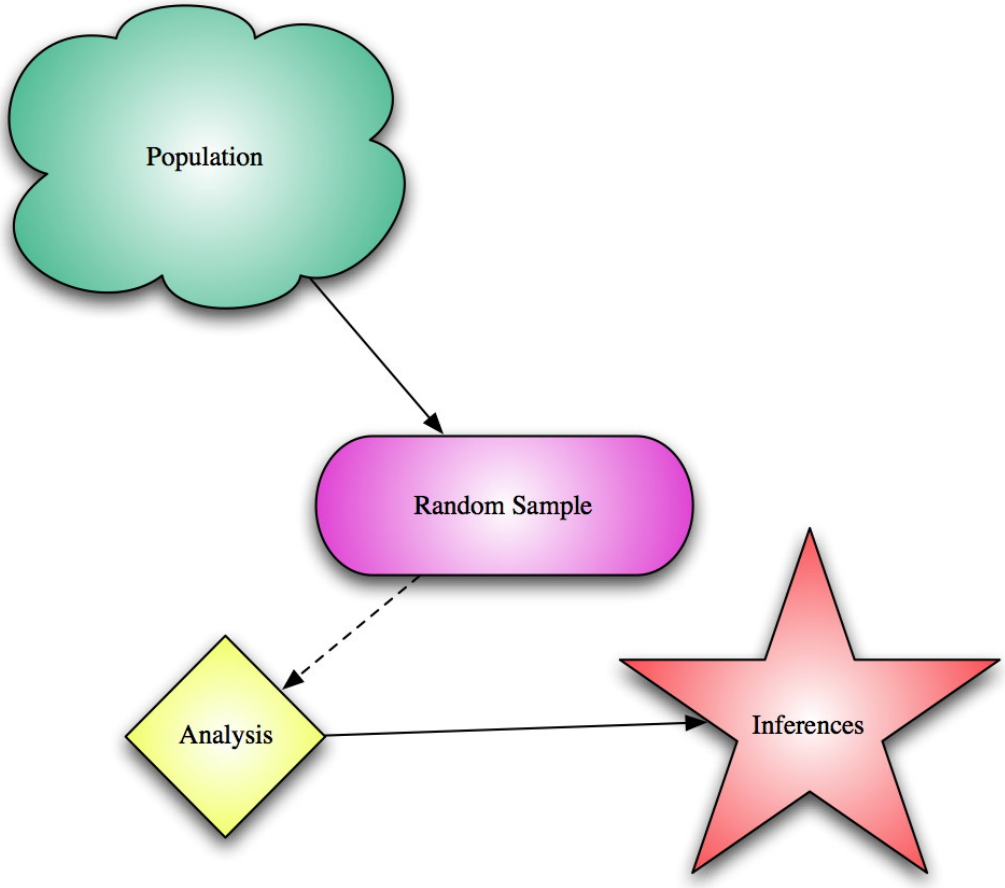
There are two primary types of data collection: **observational studies** and **experiments**.

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



Observational studies

- Researchers collect data in a way that does not directly interfere with how the data arise.

- Results of an observational study can generally be used to establish an association between variables.

There are many important questions that observational studies cannot help us answer:
Does smoking cause lung cancer? Is a new medication for treating migraine headaches more effective than the current treatment that doctors most often prescribe?

Observational studies: Prospective vs. Retrospective Studies

A prospective study identifies individuals and collects information as events unfold.

- Example: The Nurses Health Study has been recruiting registered nurses and then collecting data from them using questionnaires since 1976.

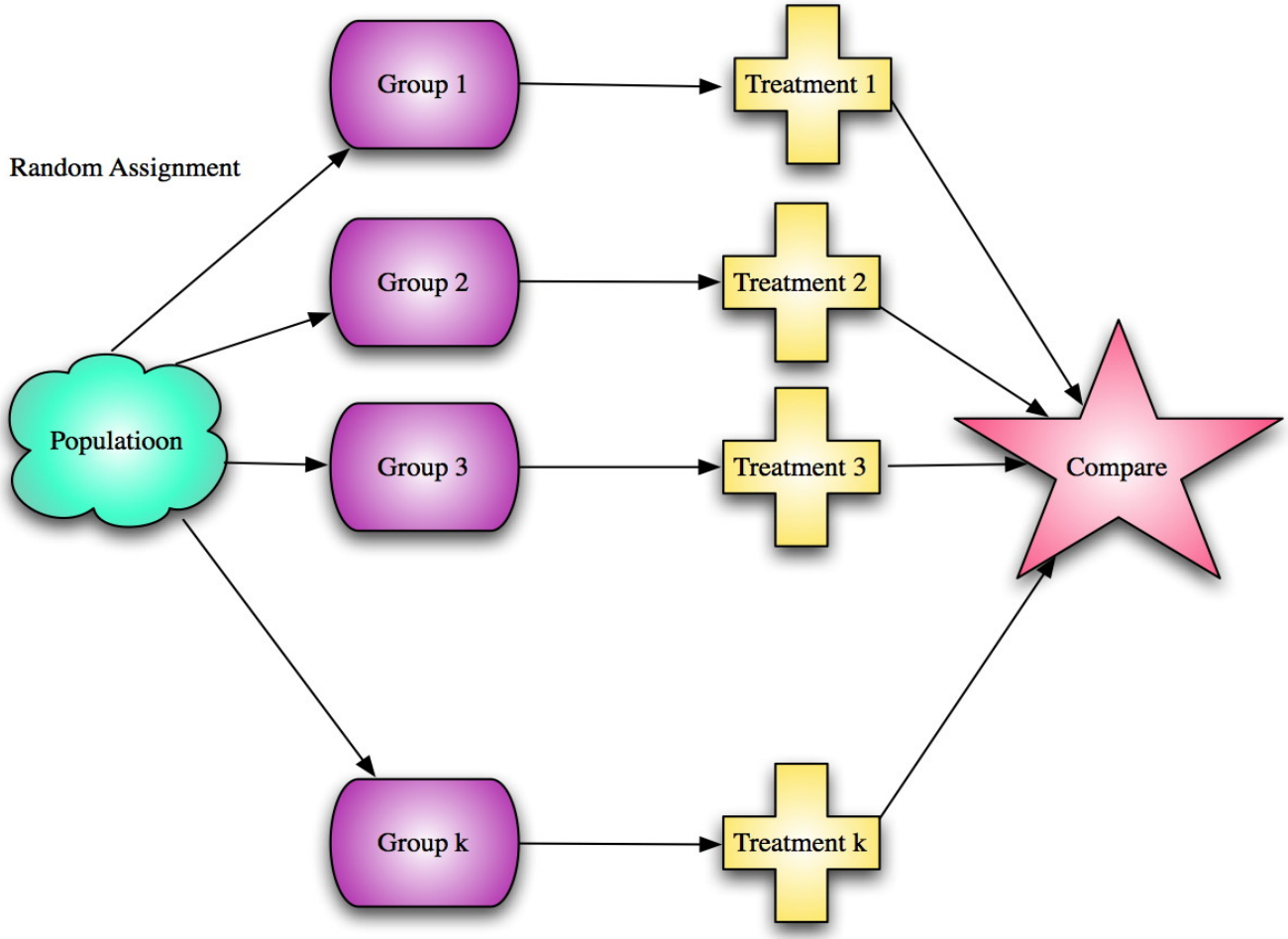
Retrospective studies collect data after events have taken place.

- Example: Researchers reviewing past events in medical records.

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.


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

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

Impact evaluation analysis

Impact evaluation is an assessment of how the intervention being evaluated affects outcomes, whether these effects are intended or unintended. The proper analysis of impact requires a counterfactual of what those outcomes would have been in the absence of the intervention.
The counterfactual represents how programme participants would have performed in the absence of the program

- Problem: Counterfactual cannot be observed
- Solution: We need to "mimic" or construct the counterfactual

Different impact evaluation methodologies differ in how they construct the counterfactual.

The principal methods are:

1. Randomized (Social) Experiments,
2. Differences-in-Differences,
3. Propensity Score Matching
4. Instrumental Variable Methods
5. Regression discontinuity….

Impact evaluation

In 2004/05 in Malawi a severe drought led to a very poor corn harvest. Almost 5 million people (38% of the population) needed emergency food aid.

Policy: introducing fertilizer subsidies. → fertilizer is more affordable → use of fertilizer increases → soil able to support bigger harvest → famine ends.

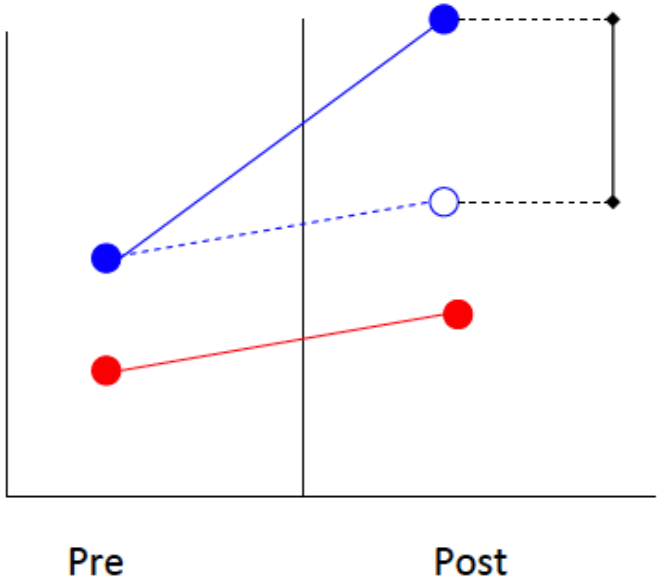
2006 and 2007: record-breaking maize harvests in Malawi. Was this dramatic turnaround the result of the subsidy? How can we tell what the true impact of the subsidies was?

Who can be our control group (counterfactual)?

- Pre-post: Compare the 2007 Malawi harvest to the 2005 Malawi harvest.
- Difference in difference: – Compare the change in Zambian harvests between 2005 - 2007 to the change in Malawian harvests over the same period

Difference in Difference

Whatever happened to the control group over time is what would have happened to the treatment group in the absence of the program.



Effect of program difference-in-difference (taking into account pre-existing differences between T & C and general time trend)

https://opportunityinsights.org/wp-content/uploads/2020/06/tracker-summary.pdf

Census

- Wouldn't it be better to just include everyone study the phenomenon in the entire population instead of on a sample of it?

    - This is called a *census*.

Census

There are problems with taking a census:

- It can be difficult to complete a census: there always seem to be some individuals who are hard to locate or hard to measure. Populations rarely stand still. Even if you could take a census, the population changes constantly, so it's never possible to get a perfect measure.
- Taking a census may be more complex than sampling.
- It is expensive and time consuming

Illegal Immigrants Reluctant To Fill Out Census Form

by PETER O'DOWD

March 31, 2010 4:00 AM

from KJZZ

Listen to the Story

Morning Edition

3 min 48 sec

+ Playlist

↓ Download

There is an effort underway to make sure Hispanics are accurately counted in the 2010 Census. Phoenix has some of the country's "hardest-to-count" districts. Some Latinos, especially illegal residents, fear that participating in the count will expose them to immigration raids or government harassment.

http://www.npr.org/templates/story/story.php?storyId=125380052

Designing a Statistical Study:

1. Identify the variable(s) of interest and the population of the study.
2. Develop a detailed plan for **collecting data**. If you use a sample, make sure the sample **is representative of the population**.
3. Collect the data.
4. Describe the data, using descriptive statistics techniques.
5. Interpret the data and make decisions about the population using inferential statistics.
6. Identify any possible errors.

Sampling methods

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.

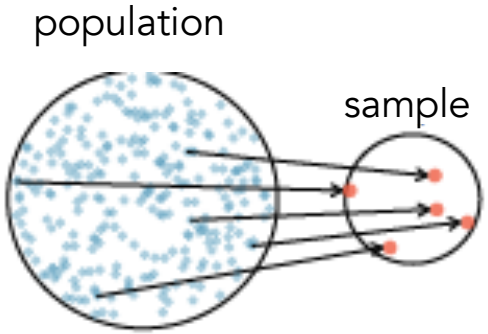
✓The only guarantee for that comes from the method used to select the sample (sampling method)→**probability sampling**

✓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.

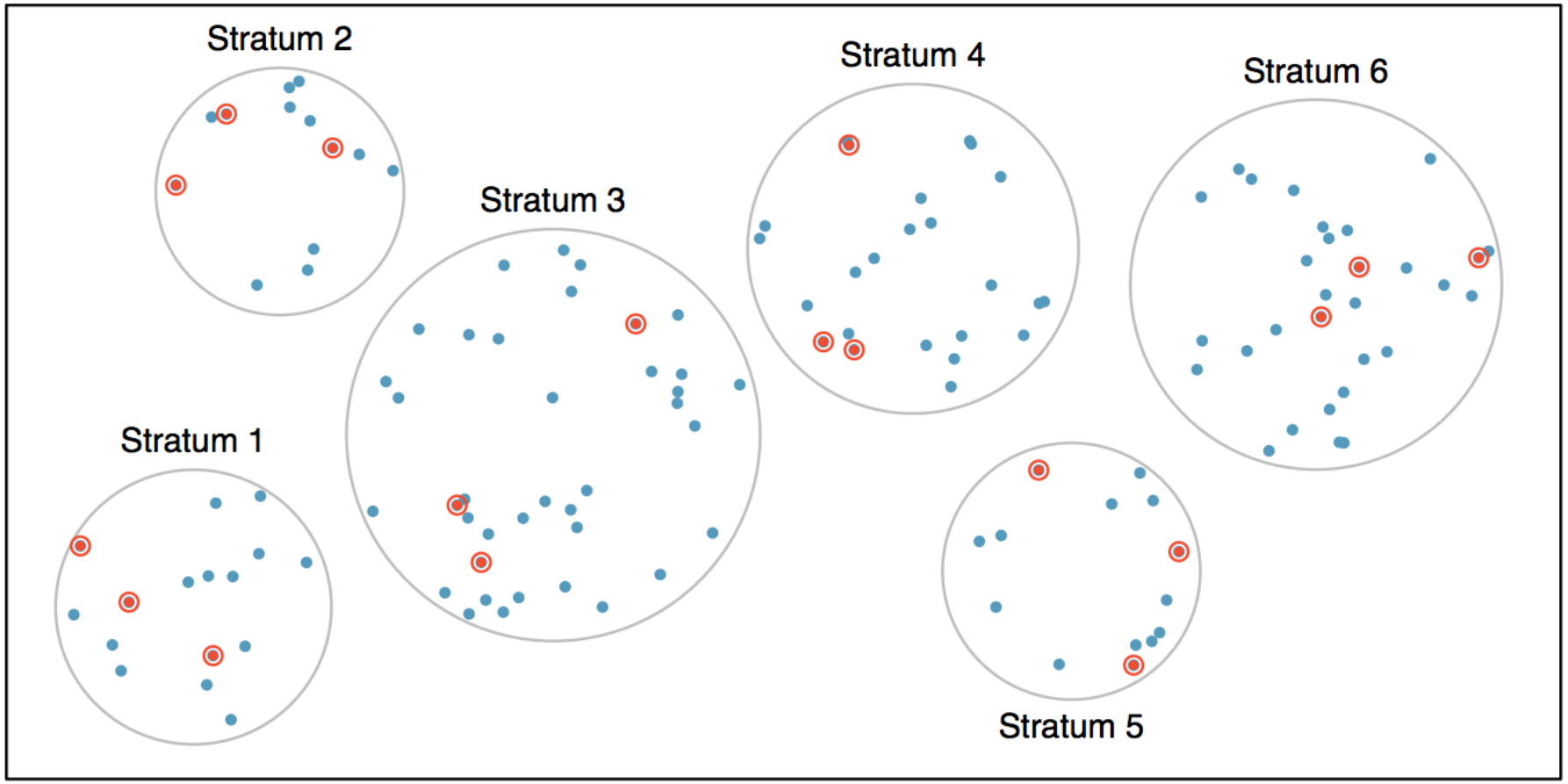
Stratified random sampling

The population is divided into groups called strata.
The strata are chosen so that similar cases are grouped together (e.g. age classes, gender,…), then a second sampling method, usually simple random sampling, is employed within each stratum.

Stratified sampling is especially useful when the cases in each stratum are very similar with respect to the outcome of interest. It ensures that various segments of the population are represented in the sample.

The downside is that analyzing data from a stratified sample is a more complex task than analyzing data from a simple random sample.

Strata are made up of similar observations. We take a simple random sample from each stratum.



Cluster and multistage random sampling

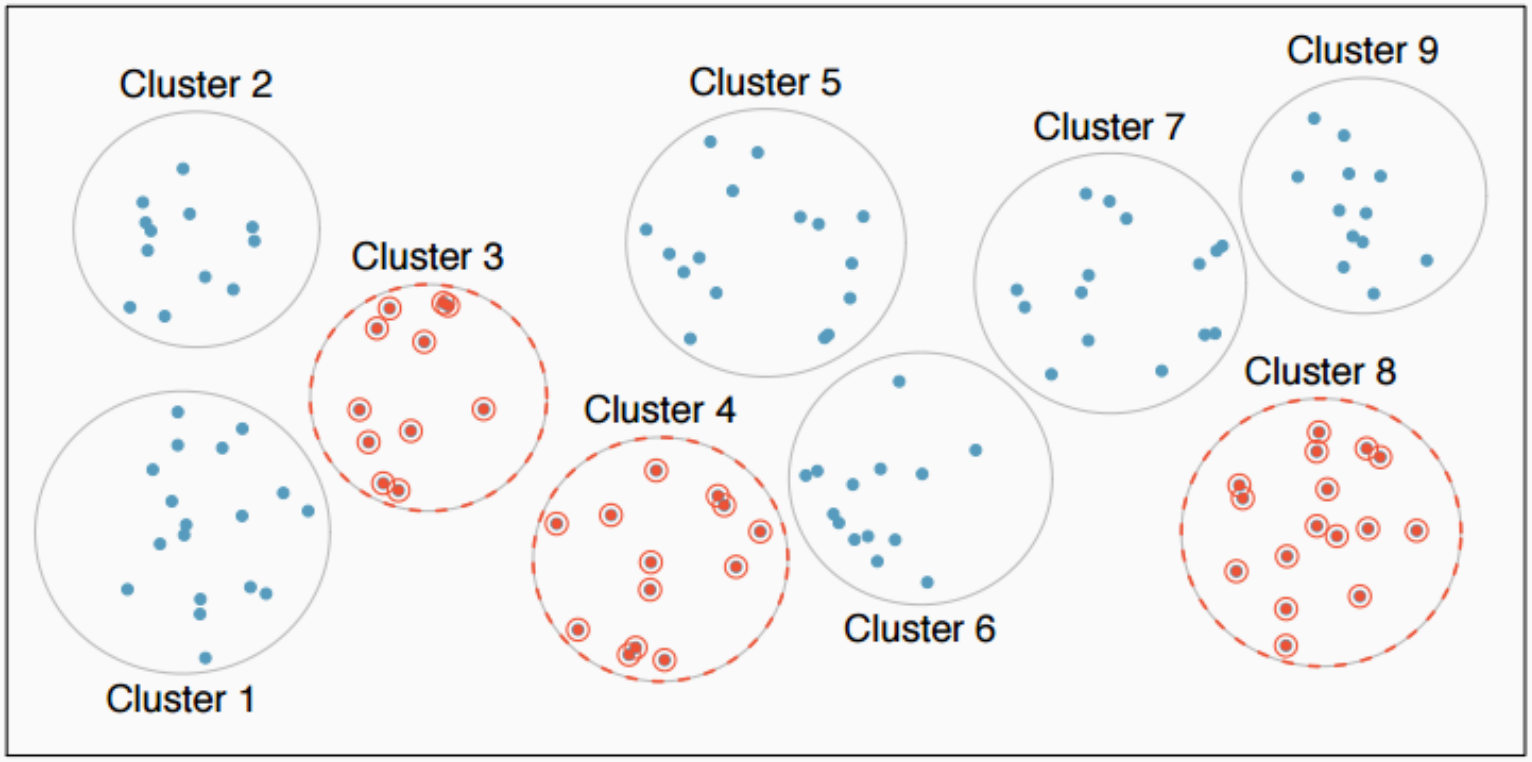
we break up the population into many groups (usually naturally occurring groups like municipalities, classes, hospitals,…), called clusters. Then we sample a fixed number of clusters and include all observations from each of those clusters in the sample.
Usually all clusters have similar characteristic.

A multistage sample is like a cluster sample, but rather than keeping all observations in each cluster, we collect a random sample within each selected cluster.

Sometimes cluster or multistage sampling can be more economical than the alternative sampling techniques.

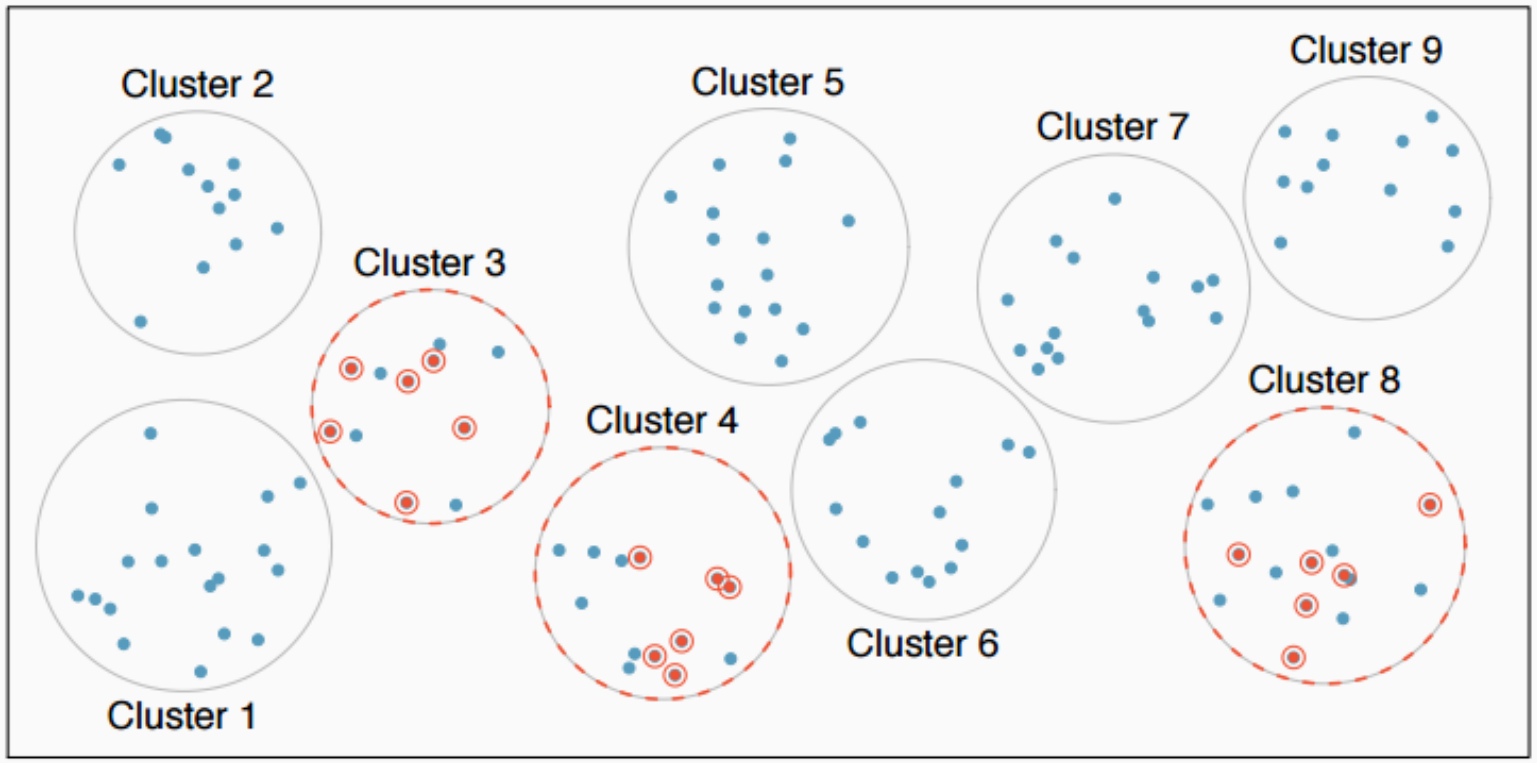
Cluster Sample

Clusters are usually not made up of homogeneous observations. We take a simple random sample of clusters, and then sample all observations in that cluster. Usually preferred for economical reasons.



Multistage Sample

Clusters are usually not made up of homogeneous observations. We take a simple random sample of clusters, and then take a simple random sample of observations from the sampled clusters



Cluster and multistage random sampling

Also, unlike stratified sampling, these approaches are most helpful when there is a lot of case-to-case variability within a cluster but the clusters themselves don't look very different from one another.

A downside of these methods is that more advanced techniques are typically required to analyze the data.

Example:

we are interested in estimating the malaria rate in a rural area of India. There are 60 villages in that area each more or less similar to the next. Our goal is to test 300 individuals for malaria. What sampling method should be employed?

A simple random sample would likely draw individuals from all villages, which could make data collection extremely expensive. Stratified sampling would be a challenge since it is unclear how we would build strata of similar individuals. Cluster sampling or multistage sampling seem like very good ideas.

If we decided to use multistage sampling, we might randomly select half of the villages, then randomly select 10 people from each. This would probably reduce our data collection costs substantially in comparison to a simple random sample, and the cluster sample would still give us reliable information, even if we would need to analyze the data with slightly more advanced methods.

Nonprobability Samples

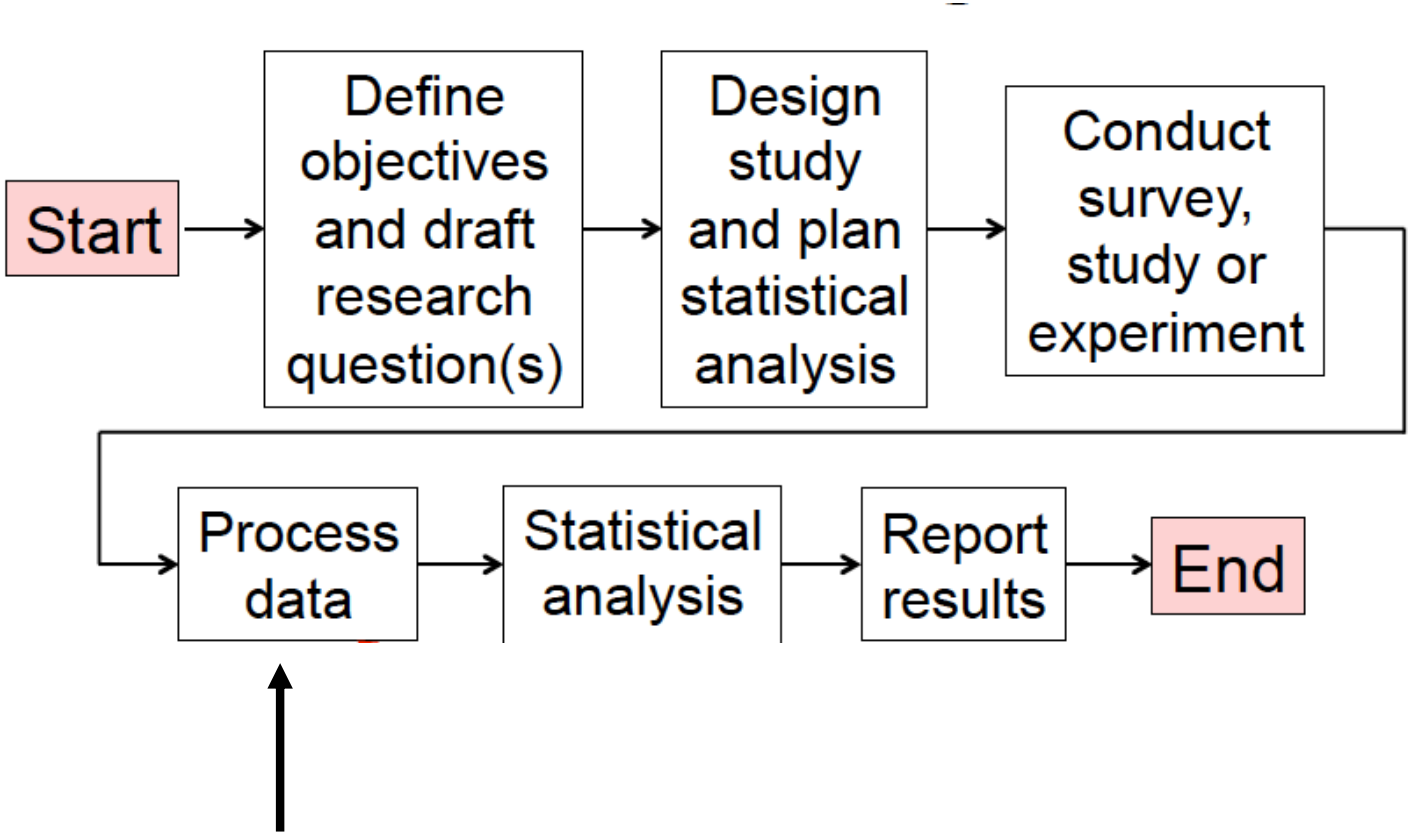
samples where the probability that every unit is in the sample cannot be known. The researcher cannot correct for any unrepresentativeness in the sampling mechanism.
Some examples:

Convenience Sampling: the researcher samples whatever units come most readily to hand.

Snowball Sampling: The researcher selects a unit, and then other units with some relationship to that first unit are sampled, and so forth. In such a design, even if the initial sample is random / probabilistic, the probabilities of selection for subsequent units cannot be known by the researcher (even if, for example, they are known to the units themselves).

DATA & DESCRIPTIVES

The research study process



This normally involves creating, as first step, a spreadsheet of raw data in any software (i.e. excel, stata, sas,…) as data matrix form.

Data Matrix (nxp): individual (person) level

wave	country	hid	pid	pd001	age	sex	maritalstatu	pe001	personalincome	healthstatus
w2 surve	spain	6068101	60681101	1948	47	male	married	paid emp	2400695	good
w6 surve	denmark	5445702	54457103	1974	25	female	married	paid emp	129000	very goo
w3 surve	spain	5882101	58821101	1934	62	male	married	paid emp	7350000	na
w3 surve	spain	3612101	36121101	1924	72	male	married	retired	1820000	bad
w1 surve	italy	97301	973101	1949	45	male	married	paid emp	40100	good
w6 surve	italy	614001	6140102	1945	54	female	married	housewor	0	very goo
w5 surve	italy	779601	7796103	1971	27	female	never ma	paid emp	12900	good
w4 surve	italy	545301	5453102	1965	32	female	married	self-emp	0	good
w1 surve	spain	5153101	51531103	1946	48	female	widowed	housewor	447996	good
w1 surve	spain	13813101	1.38E+08	1961	33	male	married	paid emp	1458000	fair
w6 surve	ireland	921001	9210101	1942	57	male	married	self-emp	7968	good
w5 surve	italy	352201	3522102	1930	68	female	married	retired	26640	fair
w1 surve	spain	3587101	35871101	1930	64	male	married	retired	1850426	good
w4 surve	ireland	1732601	17326102	1955	42	female	married	paid emp	8976	very goo
w6 surve	spain	2391101	23911101	1951	48	male	married	paid emp	1546726	good
w5 surve	denmark	264601	2646101	1919	79	female	widowed	retired	120612	very goo

Each row of a data matrix corresponds to a unit, each column corresponds to a variable. Data matrices are convenient for recording data as well as analyzing data using a computer. Convention: p denotes the number of variables in a dataset, n denotes the number of study subjects

Data Matrix (nxp): aggregate level (cities)

City	State	Region	divorces/1000	Educaton	Hhinquality	change	poor	population	n_homicides
Sterling Heig	MI	Midwest	7.461	12.6	0.28	77.6	3.1	109000	1
Sunnyvale	CA	West	10.096	13.2	0.35	11.1	3.7	106600	3
Concord	CA	West	9.287	12.9	0.33	21.2	4.6	103300	3
Fullerton	CA	West	9.976	13.2	0.41	18.7	4.7	102000	2
Independenc	MO	Midwest	10.077	12.5	0.35	0.2	4.9	111800	4
Tempe	AZ	West	12.724	14	0.38	68	5.5	106700	4
Milwaukee	WI	Midwest	6.662	12.6	0.39	-11.3	6.8	636200	50
Tulsa	OK	South	13.603	12.8	0.42	9.3	7.4	360900	31
Honolulu	HI	West	8.109	12.7	0.44	12.4	7.4	365000	33
Virginia Bea	VA	South	7.705	12.8	0.36	52.3	7.7	262200	10
Allentown	PA	N.East	5.604	12.3	0.39	-5.6	8.4	103800	4
Portland	OR	West	10.605	12.8	0.43	-3.6	8.5	366400	32
Albuquerque	NM	West	13.965	12.9	0.4	35.7	9.3	331800	21
Peoria	IL	Midwest	9.931	12.6	0.43	-2.2	9.4	124200	5
Erie	PA	N.East	6.614	12.3	0.39	-7.8	10.2	119100	6
Salt Lake	UT	West	10.268	12.9	0.45	-7.3	10.5	163000	10
Dallas	TX	South	11.96	12.7	0.45	7.1	10.8	904100	271
Berkeley	CA	West	9.287	16.1	0.5	-9.4	11.7	103300	9
Columbus	GA	South	12.613	12.3	0.43	9.3	14.5	169400	16
Rochester	NY	N.East	6.387	12.3	0.42	-18.1	14.5	241700	26

Williams-Flewelling, 1988, American Sociological Review

Data Richness

- You should always use the richest (most detailed) data available because it will give more accurate results
- Here, the Age data is richer than the Age Category data
- However, there might be ethical issues in obtaining detailed data
- Here, the respondents might feel embarrassed to give their exact age

Age	Age Category
29	25-29
50	40+
27	25-29
27	25-29
31	30-30
24	18-24
31	30-30
32	30-30
34	30-30
17	18-24

Types of variables

Qualitative (or categorical) data - the characteristic being studied is nonnumeric.
Examples: gender, religious affiliation, state of birth, eye color.

Quantitative (or numerical) data - information is reported numerically. A number is assigned as a quantitative value representing count or measurement. Mathematical operations are possible!
Examples: income, height, number of children in a family.

Quantitative Variables:

can be classified as either **discrete** or **continuous**.

Discrete variables: can only assume a finite number of values – no decimals.

EXAMPLE: the number of bedrooms in a house (1,2,3,…,etc).

Continuous variable: can assume any value within a specified range.

EXAMPLE: the weight or the height of students

Price is a quantitative continuous variable since it can take a wide range of numerical values, and it is sensible to add, subtract, or take averages with those values.
On the other hand, a variable reporting telephone area codes cannot be classified as quantitative since their average, sum, and difference have no clear meaning
The number passengers in a train variable is also quantitative, although it seems to be a little different than price. The variable passengers can only take whole positive numbers (1, 2, ...) since it is not possible to have 4.5 passengers. The variable passengers is said to be discrete since it only can take numerical values with jumps (e.g. 3 or 4, but not any number in between).

Categorical (qualitative) Variables:

have values that describe labels or attributes. Even if the categories can be placed in a natural order, they have no magnitude or units. There are two major scales for categorical variables:

1.Nominal variables have categories with no distinct or defined order. For example:

1. gender
2. favorite color
3. nationality

2.Ordinal variables have an inherent order. For example:

1. Likert scales (strongly disagree, disagree, neutral, agree, strongly agree)
2. t-shirt size (small, medium, large)

Note: Ordinal categorical variables are often aggregated to create scales in humanities research and can be treated as numeric if they have a sufficient amount of variation in values.

Proportions, Rates & Ratios

Ratio

A ratio can be written as one number divided by another (a fraction) of the form a/b – Both a and b refer to the frequency of some event or occurrence

For example: clinicians to patients or beds to clients

In district X, there are 600 nurses and 200 clinics. What is the ratio of nurses to clinics?
600 / 200 = 3 nurses per clinic, a ratio of 3:1

Consider a class that has 20 male students and 80 female students. We can think about this in several ways. We could express this simply as the ratio of men to women and write the relationship as 20/80 or simplify this to a 1:4 ratio (or 1/4 ratio). This indicates that for every man, there are four women.

Proportion

A proportion is a ratio in which the numerator is a subset (or part) of the denominator and can be written as a/(a+b) (it is a relative frequency!)

It is used to compare part of the whole, such as proportion of all clients of a bank who are less than 35 years old.

In the class with with 20 men and 80 women, the total class size is 100, and the proportion of men is 20/100 or 20%. The proportion of women is 80/100 or 80%. In both of these proportions the size of part of the class is being related to the size of the entire class. The class above conveniently had a total size of 100, but this usually isn't the case.

Allow to compare different groups, facilities, countries that may have different denominators!

Rate

A rate is a ratio of the form a*/ (a+b)
where:

a* = the frequency of events during a certain time period
a+b = the number at risk of the event during that time period

Infant mortality rate (IMR) = number of infant deaths per 1,000 live births during a calendar year

Fertility rate = number of live births per 1,000 women aged 15–44 during a calendar year

Death rate (all causes), crude (per 1,000 people)

In 2015 municipality A counts 150 number of deaths over a population of 6000

$$\frac{150}{6000} = .002 \times 1000 = 2$$

Crude mortality rate per 1,000 residents

CONFOUNDING

When comparing units frequently we encounter a problem that involves comparing the results of populations that have different structures with respect to background characteristics.
An example of this is comparing mortality figures for populations with a different age distribution. For ex countries with a young population will usually have lower mortality rates than countries with a much older population. In this case, a country's gross (crude) mortality rate is therefore not a good indicator of the health of its citizens.
Only when the data are examined for age effects, by only comparing individuals in the same age class, is it possible to make a fair comparison.

Mortality *per 1000* by State, 1991

		Alaska		**Florida**	
i	Age	Deaths	Pop. **(×1000)**	Deaths	Pop. **(×1000)**
1	0–4	122	57	2,177	915
2	5–24	144	179	2,113	3,285
3	25–44	382	222	8,400	4,036
4	45–64	564	88	21,108	2,609
5	65–74	406	16	30,977	1,395
6	75+	582	7	71,483	1,038
	TOTAL	2,200	569	136,258	13,278

Crude rate, Alaska

$$cR_{Alask.} = \frac{2200}{569} = 3.9$$

Crude rate, Florida

$$cR_{Florida} = \frac{136{,}258}{13{,}278} = 10.3$$

Age Distributions

Age	AK	%	FL	%
0-4	57	10%	915	7%
5-24	179	31%	3285	25%
25-44	222	39%	4036	30%
45-64	88	15%	2609	20%
65-74	16	3%	1395	11%
>75	7	1%	1038	8%
TOTAL	569	100%	13278	100%

What can we do about confounding?

Like-to-like (strata-specific) comparisons
(e.g., 80-year old to 80-year old)

Mathematical adjustments:

1. Direct (use as weights the demographic composition of a standard population) and indirect standardization (use as weights the specific rates of a standard population)
2. Regression models

Direct Adjustment

$$aR_{direct} = \sum w_i r_i$$

where

$$w_i = \frac{N_i}{N}$$

$N_i \equiv$ reference population size, strata i

$N \equiv$ reference population total size

$r_i \equiv$ rate, study population, strata i

aR_{direct} is a weighted average of strata-specific rates

"Standard Million" 1991 Reference

Weight, strata i (w_i) = proportion in reference pop = N_i / N

i	Age	N_i	w_i
1	0–4	76,158	0.076158
2	5–24	286,501	0.286501
3	25–44	325,971	0.325971
4	45–64	185,402	0.185402
5	65–74	72,494	0.072494
6	75+	53,474	0.053474
	$\Sigma\rightarrow$	N = 1,000,000	1.000000

Alaska, Direct Adjustment

(Rates are per 1000)

i	Age	Rate r_i	Weights w_i	Product $w_i \cdot r_i$
1	0–4	2.14	0.076158	0.16297814
2	5–24	0.80	0.286501	0.22920080
3	24–44	1.72	0.325971	0.56067012
4	45–64	6.40	0.185402	1.18657280
5	65–74	25.38	0.072494	1.83989772
6	75+	83.14	0.053474	4.44582836
			$\Sigma w_i \cdot r =$	8.42514792

$$aR_{Alask.} = \sum w_i r_i = 0.163 + 0.223 + \cdots + 4.45 \approx 8.43$$

Florida, Direct Adjustment

(Rates are per 1000)

i	Rate r_i	Weights w_i	Product $w_i \cdot r_i$
1	2.38	0.076158	0.18126
2	0.64	0.286501	0.18336
3	2.08	0.325971	0.67802
4	8.09	0.185402	1.49990
5	22.21	0.072494	1.61009
6	68.87	0.053474	3.68275
		$\Sigma w_i \cdot r =$	7.83538

$$aR_{Florida} = 0.181 + 0.183 + \cdots + 3.683 = 7.84$$

Conclusions

cR_{FL} (10.3) > cR_{AK} (3.9)
aR_{FL} (7.8) < aR_{AK} (8.4)
Age confounded the crude comparison
State (E) associated with age (C)
Age (C) is independent risk factor for death (D)

MDGs: 8 goals, 18 targets, 48 indicators

Goal 1. Eradicate extreme poverty and hunger
Goal 2. Achieve universal primary education
Goal 3. Promote gender equality and empower women
Goal 4. Reduce child mortality
Goal 5. Improve maternal health
Goal 6. Combat HIV/AIDS, malaria and other diseases
Goal 7. Ensure environmental sustainability
Goal 8. Develop a Global Partnership for Development

For each goal: one or several targets; one or several indicators
However, several key areas identified have not been captured adequately or at all

Education is vital to meet all of the development goals

Example:

Millennium Development Goals

Goal 2: Achieve universal primary education in selected countries.

Goal 3: Promote gender equality and empower women.

Education Indicators and Data Analysis
UNESCO Institute for Statistics
www.uis.unesco.org http://uis.unesco.org/en/home

MDGs and education

Goal 2: Achieve universal primary education

Target 3. Ensure that, by 2015, children everywhere, boys and girls alike, will be able to complete a full course of primary schooling

6. Net enrolment ratio in primary education
7. Proportion of pupils starting grade 1 who reach grade 5
8. Literacy rate of 15-24-year-olds

Goal 3: Promote gender equality and empower women

(including)

Target 4. Eliminate gender disparity in primary and secondary education, preferably by 2005, and to all levels of education no later than 2015

(including)

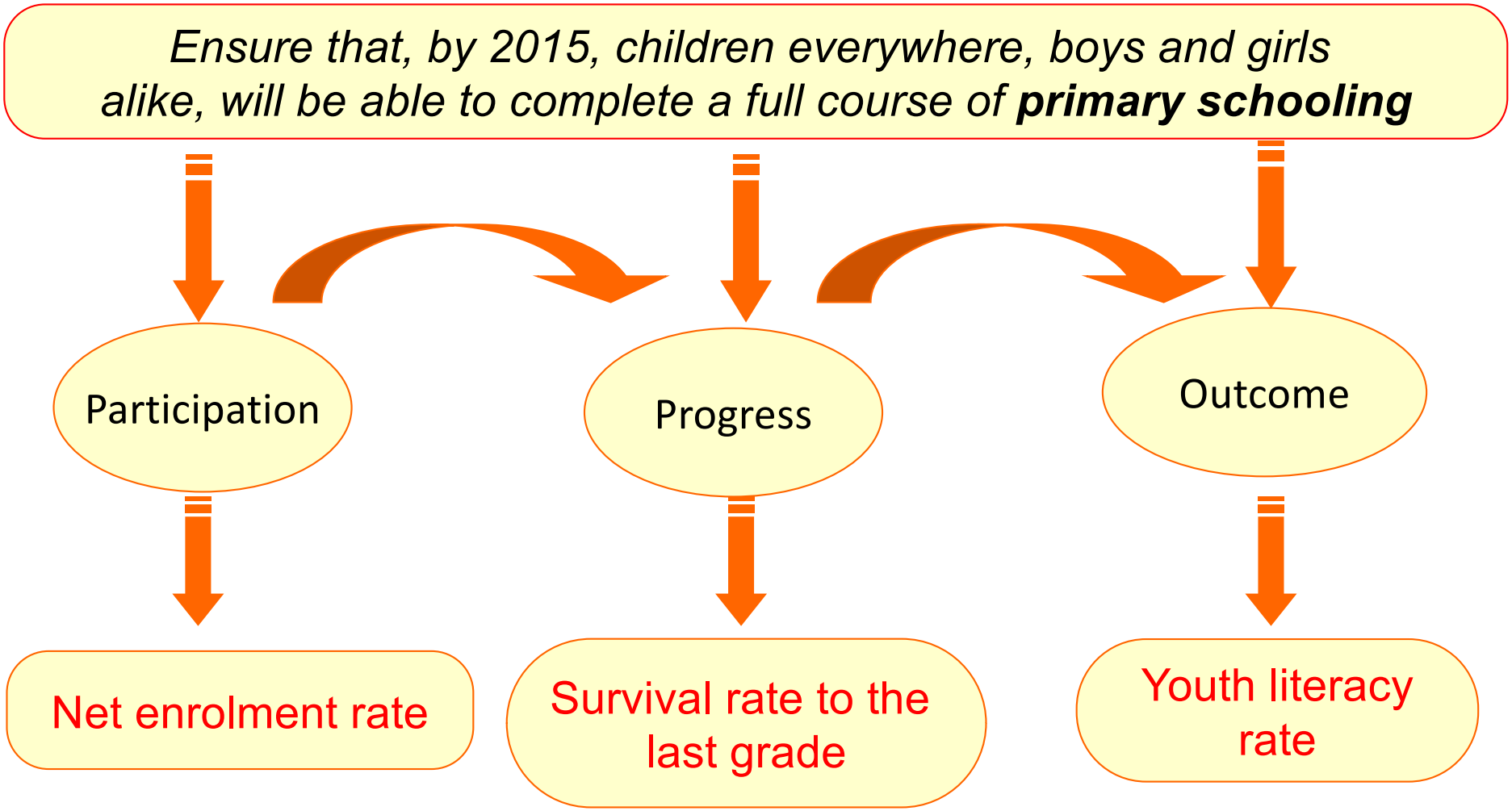
9. Ratio of girls to boys in primary, secondary and tertiary education
10. Ratio of literate females to males of 15-to-24-year-olds

Goal 2

Achieve universal primary education

Target: Ensure that, by 2015, children everywhere, boys and girls alike, will be able to complete a full course of primary schooling.

Monitoring indicators:



Participation: Net enrolment rate (NER)

Definition: Percentage of children of the official primary age group who are enrolled in primary education.

Calculation: Divide the number of pupils of the official primary age group who are enrolled in primary education by the population of the same age group and multiply the result by 100.

Net enrolment rate (NER)

$$NER_h^t = \frac{E_{h,a}^t}{P_{h,a}^t} * 100$$

Where:

NER_h^t Net Enrolment Rate at level of education **h** in school year **t**

$E_{h,a}^t$ Enrolment of the population of age group **a** at level of education h in school year **t**

$P_{h,a}^t$ Population in age group **a** which officially corresponds to level of education **h** in school year **t**

Example: If the entrance age for primary education is 7 years with a duration of 6 years then **a** is (7-12) years.

Republic of Moldova (2011)

Entry age: 7 year old
Duration: 4 years

Official age group:
7-10

$$SAP_{7-10} = 147{,}897$$

Enrolment in official age group =129,870

$$NER = \frac{129{,}870}{147{,}897} * 100 = 87.8\%$$

Age	Population	Enrolment in primary education
5	37,472	19
6	37,484	5,088
7	36,206	32,111
8	36,373	33,983
9	37,196	33,084
10	38,122	30,692
11	39,200	3,027
12	40,777	296
13	43,147	68
14	46,737	47
15	49,511	21
Total		138,436

Progress

Indicator 2.2: Proportion of pupils starting grade 1 who reach the last grade of primary

Survival rate to the last grade of primary education

Definition: Percentage of a cohort of pupils enrolled in the first grade who are expected to reach the last grade of primary education, regardless the repetition.

Rationale: This indicator measures an education system's success in retaining students from one grade to the next as well as its internal efficiency. Various factors account for poor performance on this indicator, including low quality of schooling, discouragement over poor achievement and the direct and indirect costs of schooling. Students' progress to higher grades may also be limited by the availability of teachers, classrooms and educational materials.

Survival rate to the last grade of primary education

Calculation: The survival rate is calculated on the basis of the reconstructed cohort method, which uses data on enrolment and repeaters for two consecutive years.

This method makes three assumptions:
dropouts never return to school;
promotion, repetition and dropout rates remain constant over the entire period in which the cohort is enrolled in school;
the same rates apply to all pupils enrolled in a given grade, regardless of whether they previously repeated a grade.

Survival rate to the last grade of primary education

Interpretation: Indicator values range from 0% (none of the pupils starting grade 1 reach the last grade) to 100% (all of the pupils reach the last grade). Survival rate approaching 100 per cent indicate a high level of retention and a low incidence of dropout. It is important to note that it does not imply that all children of school age complete primary education.

Outcome: Youth literacy rate (15-24 years)

Definition: Percentage of people aged 15 to 24 years who can both read and write with understanding a short, simple statement on their everyday life.

Calculation: Divide the number of people aged 15 to 24 years who are literate by the total population in the same age group and multiply the result by 100.

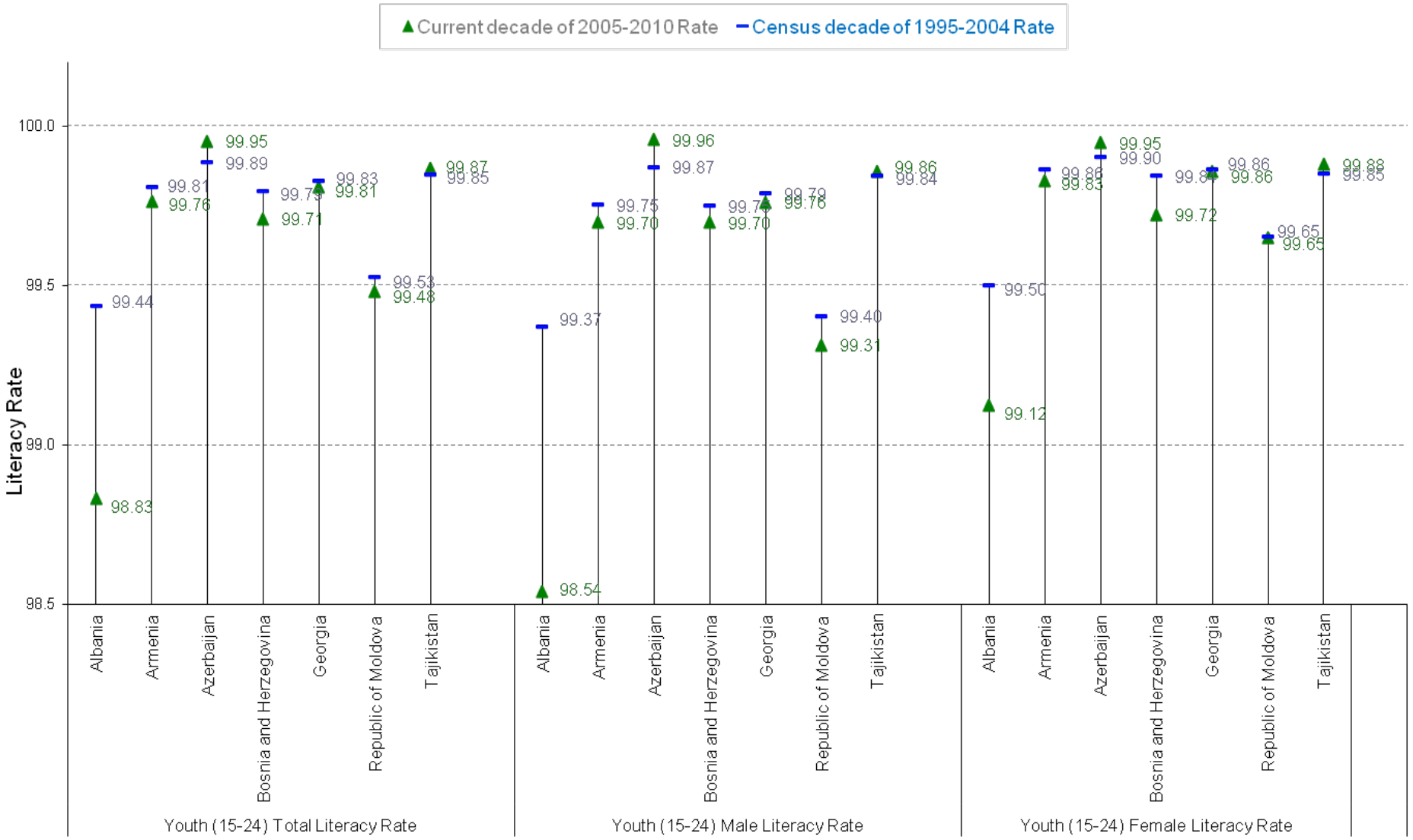
The youth literacy rate reflects the outcomes of the primary education system over the previous 10 years, and is often seen as a proxy measure of social progress and economic achievement

Youth literacy rate (15-24 years)

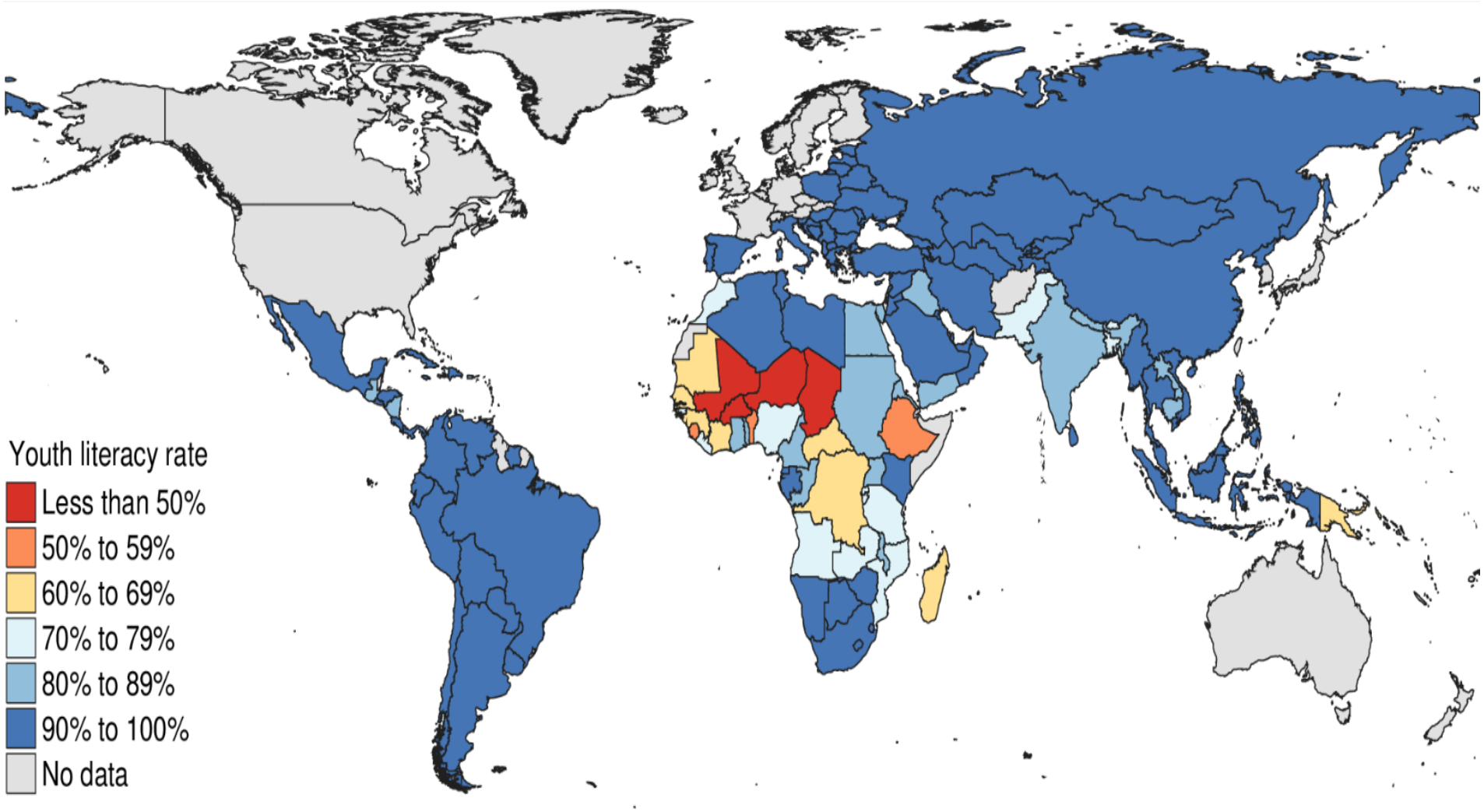
Interpretation: The indicator ranges from 0% (all youth are illiterate) to 100% (all youth are literate). Literacy rates below 100 per cent indicate the need to increase school participation and education quality.

Rationale: The youth literacy rate reflects the outcomes of the primary education system over the previous 10 years and is often seen as a proxy measure of social progress and economic achievement. The literacy rate is the complement of the illiteracy rate. It is not a measure of the adequacy of the literacy levels needed for individuals to function and participate in a society (functional literacy).

Youth literacy rate (15-24 years)



Youth literacy rate (15-24 years)



Youth literacy rate (15-24 years)

Limitations: Some countries apply definitions and criteria for literacy which are different from the international standard defined above, or change definitions between censuses.

Practices for identifying literates and illiterates during actual census enumeration may also vary. Errors in literacy self-declaration can affect the reliability of the statistics.

Goal 3

Promote gender equality and empower women

Target: Eliminating gender disparity in primary and secondary education, preferably by 2005, and in all levels of education no later that 2015

Monitoring indicator:

Eliminating gender disparities by 2005 in primary and secondary education, and at all levels no later than 2015

Equal participation of boys and girls

Gender parity index of Gross enrolment ratio

Gender parity index (GPI)

Definition: Ratio of female to male values of a given indicator.
Purpose: The GPI measures progress towards gender parity in education participation and/or learning opportunities available for girls in relation to those available to boys.
Calculation: Divide the female value of an indicator by the male value of the same indicator.

Gender parity index (GPI)

$$GPI_{GER} = \frac{GER_{Female}}{GER_{Male}}$$

Where:

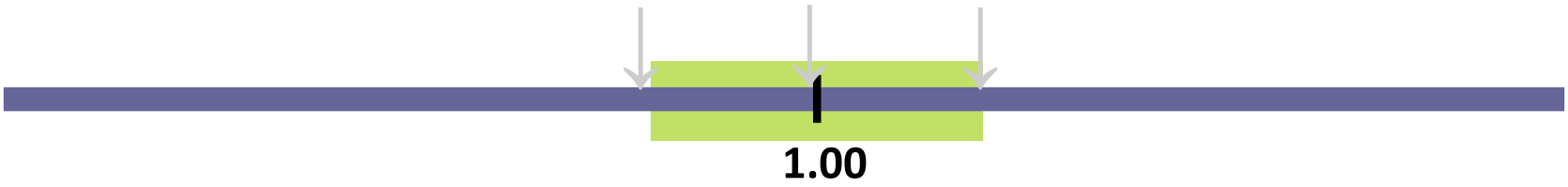
GPI_{GER} Gender parity index for Gross enrolment ratio

GER_{Female} Gross enrolment ratio for female

GER_{Male} Gross enrolment ratio for male

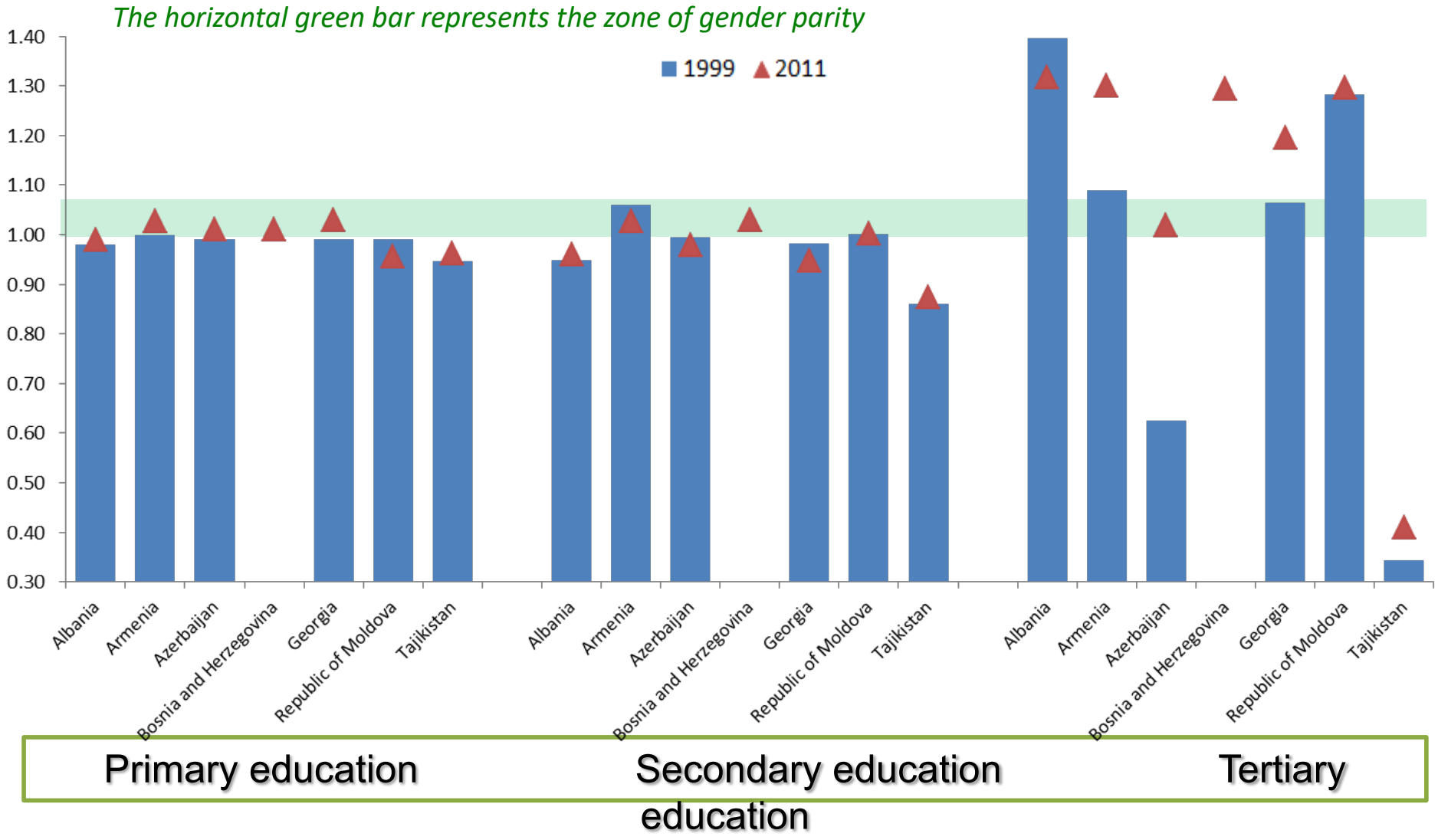
Measuring gender parity
GER for Primary education, 2011

Gender parity index – an index of 1.00 is perfect parity, and 0.97 – 1.03 is considered a zone of gender parity



	Tajikistan	Albania	Georgia
Girls GER	98.4	85.4	108.1
Boys GER	102.4	86.4	105.0
GPI	0.96	0.99	1.03

Gender parity index by level of education, 1999 and 2011



Indicators help to

"to describe as much detail as possible about a system to help understand, compare, predict, improve, and innovate."
(The Good Indicators Guide)

- **describe** situations
- **measure** trends over time
- provide a yardstick whereby facilities / teams/providers of a service can **compare** themselves to others
- **monitor performance:** progress towards defined targets

Composite Indicators

A composite indicator is the *mathematical combination of individual indicators that represent different dimensions of a concept* whose description is the objective of the analysis.

The construction of composite indicators involves stages where subjective judgement has to be made i.e. the selection of indicators, the aggregation method, the weights of the indicators, etc.

These subjective choices can be used to manipulate the results. It is, thus, important to identify the sources of subjective or imprecise assessment.

Saisana and Tarantola, 2002

Creating Composite Indicators: steps

1. Identify individual components
2. Weight the components
3. Measure and Combine
4. Quality assessment

Different type of weighting…

1. Equal weights
2. Weights based on principal component analysis and factor analysis
3. Based on Regression analysis
4. Weights based on public/expert opinion
5. …..

Different methods for combining the items…

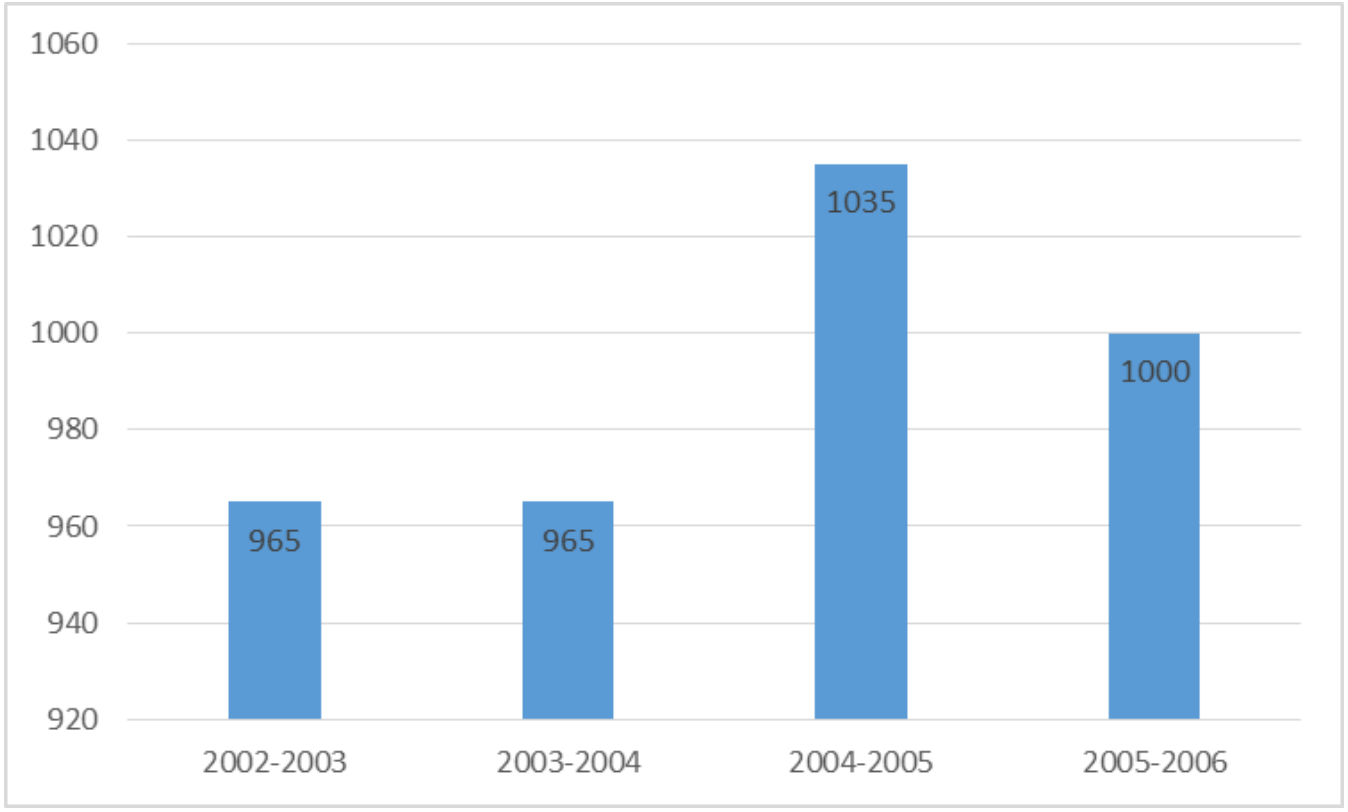
1. Weighted sum
2. Principal component/factor analysis
3. Regression
4. Structural equation modeling
5. ….

Client Satisfaction Composite Score

Factor	Weighting
Survey Score	**+10**
Open Comments –Positive	**+5**
Open Comments – Negative	**-10**
New Contracts	**+10**
Contract Renewals	**+25**
Contract Cancellations	**-100**
Consults	**+5**
Complaints	**-10**

Year	Survey	Positive Opinions	Negative Opinions	New Contracts	Contract Renewals	Contract Cancellations	Consults	Complaints
2002-2003	90	24	6	0	0	0	5	2
2003-2004	85	22	10	4	0	0	5	0
2004-2005	85	22	6	6	0	0	3	0
2005-2006	85	20	2	2	1	0	4	1

Client Satisfaction Composite Score



Euro Health Consumer Index

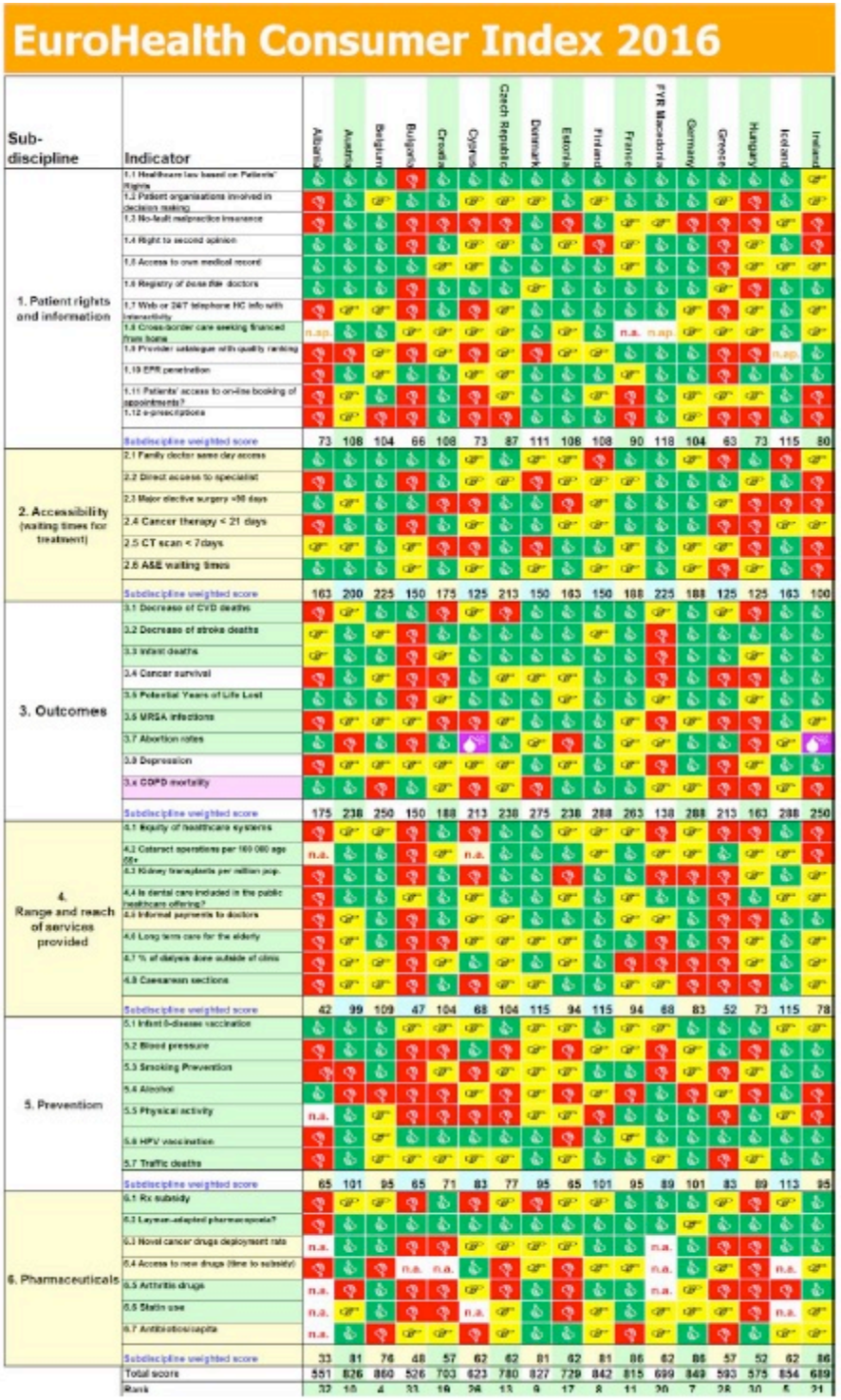
The aim has been to select a limited number of indicators, within a definite number of evaluation areas, which in combination can present a telling tale of how the healthcare consumer is being served by the respective systems.

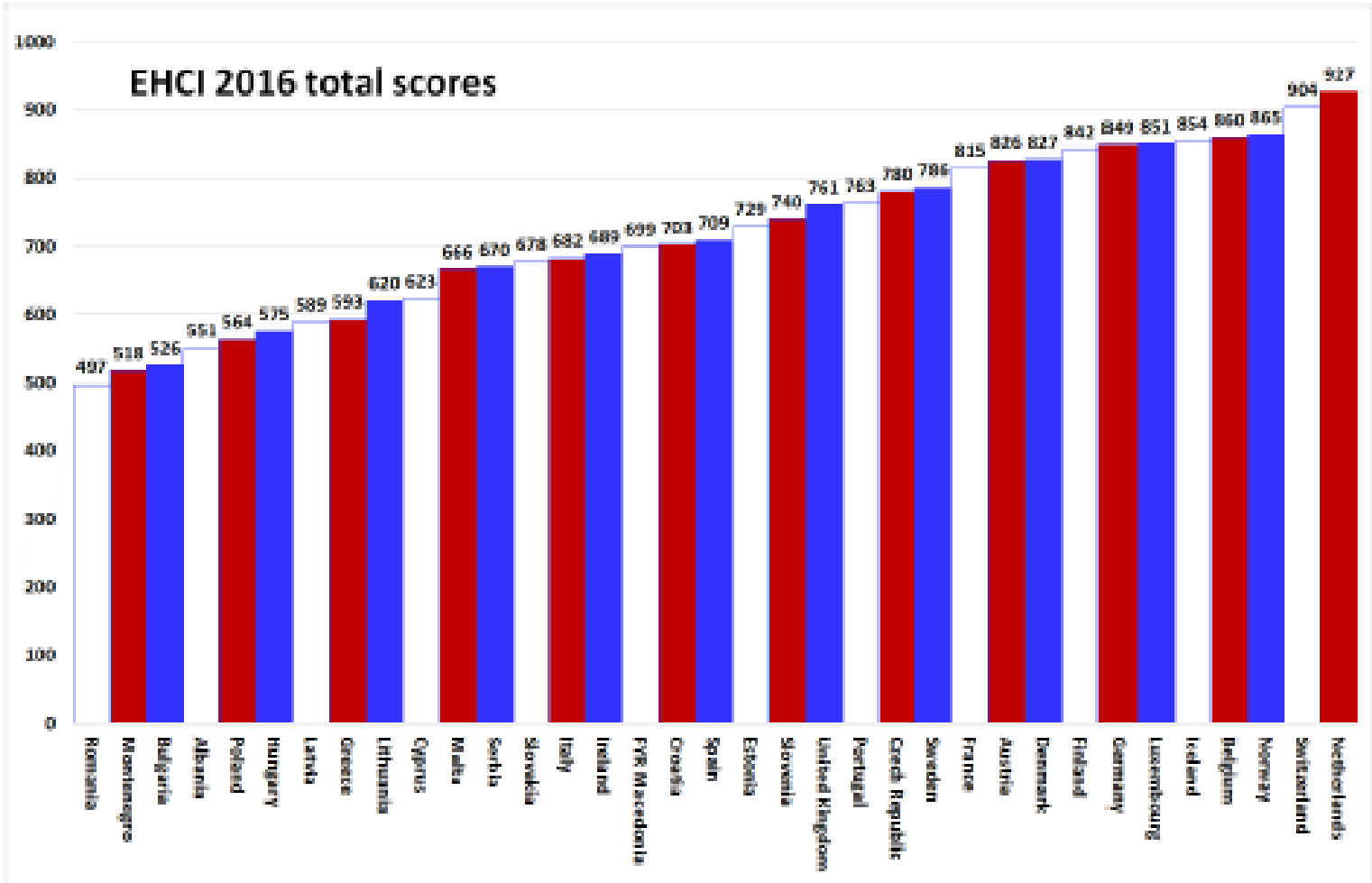
Comparing healthcare systems performance in 35 countries.

http://www.healthpowerhouse.com/

Sub-discipline	Relative weight ("All Green" score contribution to total maximum score of 1000)
1.Patient rights, information and e-Health	125
2.Accessibility (Waiting time for treatment)	225
3.Outcomes	300
4.Range and reach of services ("Generosity")	125
5.Prevention	125
6.Pharmaceuticals	100
Total sum of weights	**1000**

The accessibility and outcomes sub disciplines were decided as the main candidates for higher weight coefficients based mainly on discussions with expert panels and experience from a number of patient survey studies.




EHCI 2016 total scores
1000
900
800
700
600
500
400
300
200
100
0
497
518
526
551
564
575
589
593
620
623
666
670
678
682
689
699
703
709
729
740
761
763
780
786
815
826
827
843
849
851
854
860
865
904
927
Romania
Montenegro
Bulgaria
Albania
Poland
Hungary
Latvia
Greece
Lithuania
Cyprus
Malta
Serbia
Slovakia
Italy
Ireland
FYR Macedonia
Croatia
Spain
Estonia
Slovenia
United Kingdom
Portugal
Czech Republic
Sweden
France
Austria
Denmark
Finland
Germany
Luxembourg
Iceland
Belgium
Norway
Switzerland
Netherlands

http://www.oecdbetterlifeindex.org