Introduction to Statistics

Prof.ssa Chiara Seghieri,

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

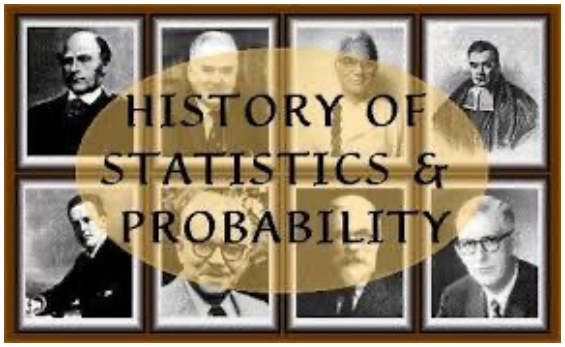
Statistics is…

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.

The history in brief



Stat[e]istics - originally conceived as the science of the state - the collection and analysis of facts about a country.

The collection of forms of census data goes back into ages, rulers "were interested in keeping track of their people, money and key events (such as wars and the flooding of the Nile). In Rome the first census dates back in 578 B.C.

Another thread in the development of modern statistics was the foundations of probability, with its origins in games of chance - Pascal (1623–1662) and later Bernoulli (1654–1705). And afterwards, Bayes in 1764 and Laplace (1749-1827).
But **formal methodological techniques** for gathering and analysing data (sample surveys and census) appeared in the **nineteenth century**.

The Royal Statistical Society began in 1834, the American Statistical Association was formed in 1839. In 1926 the National Institute of Statistics was founded in Italy.

Starting from 70's thanks to the use of computers there has been a significant and rapid development of statistics.

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 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

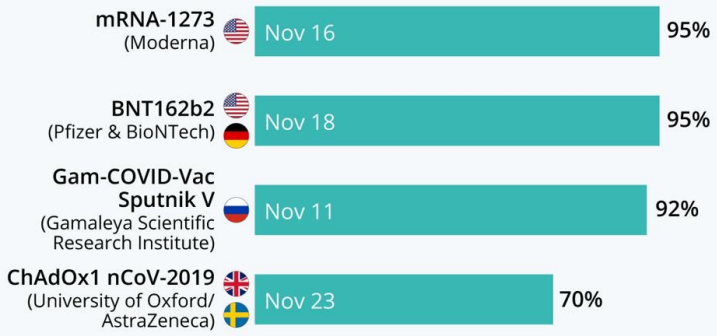
Global Finance **selects the world's 10 best cities to live in based on four reputable rankings.**

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

Number of Americans in poverty hits record high

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

FINANCIAL TIMES

The gap in giving

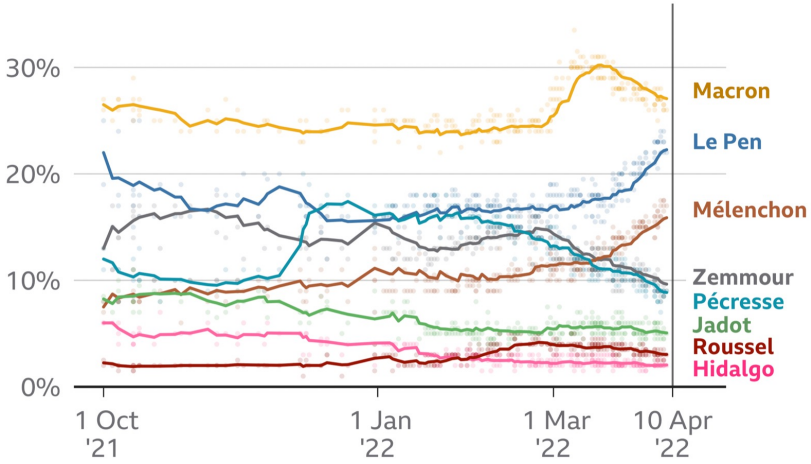
Welcome to the new Middle Ages

House price falls forecast to continue

Jump start Britain blasts out while ski competition bids

French presidential election polling

Voting intention for first round, dots show individual polls, lines show weighted average

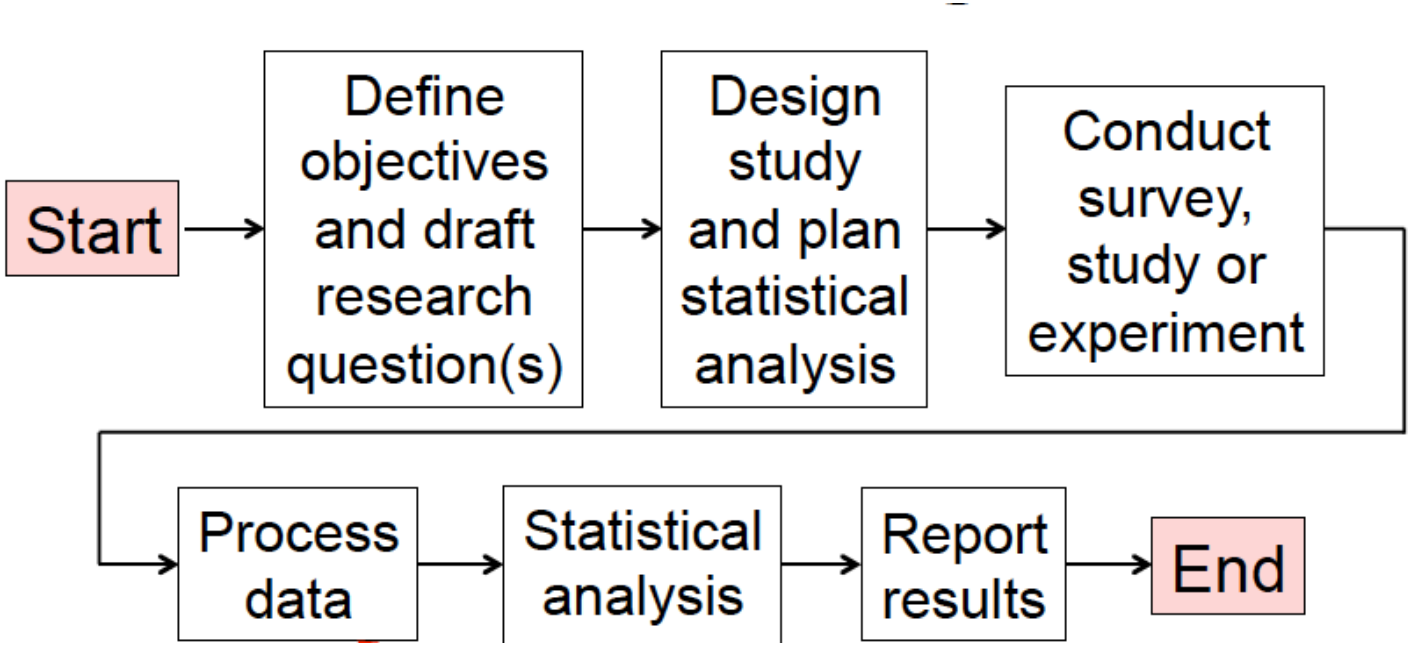


Lines show an average of polls, weighted to give the most recent polls more influence. Companies that poll more often are given less influence. Each individual poll has a margin of error, usually around three percentage points.

Source: NSPPolls, latest poll 8 Apr

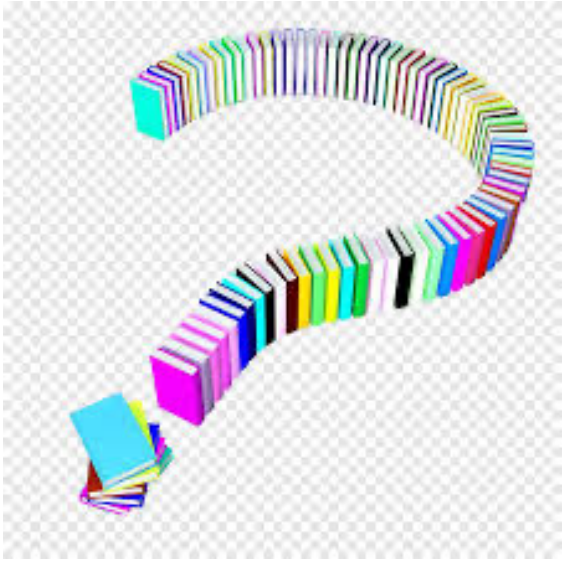
BBC

The research study process



What is a statistical question?

question that require statistical analysis for the answers

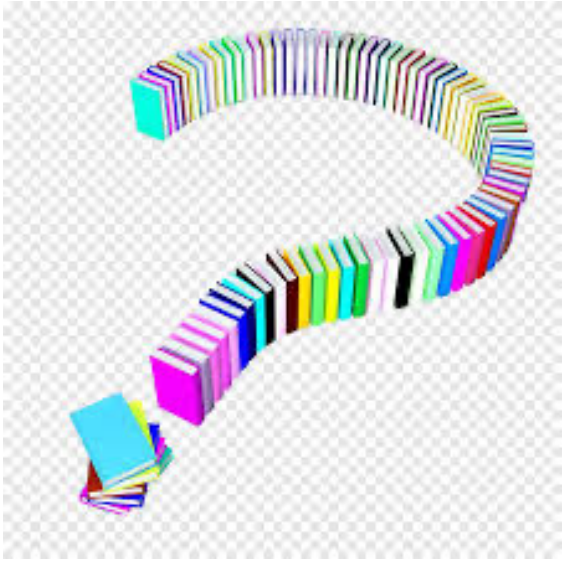


A well-written statistical question refers to:

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

“How old is my professor of statistics?” is a statistical question?



“How old is my professor of statistics?” is a statistical question?

it is not a statistical question because there is only a single subject, and hence no variability, we don’t need any statistical methods to answer to it.

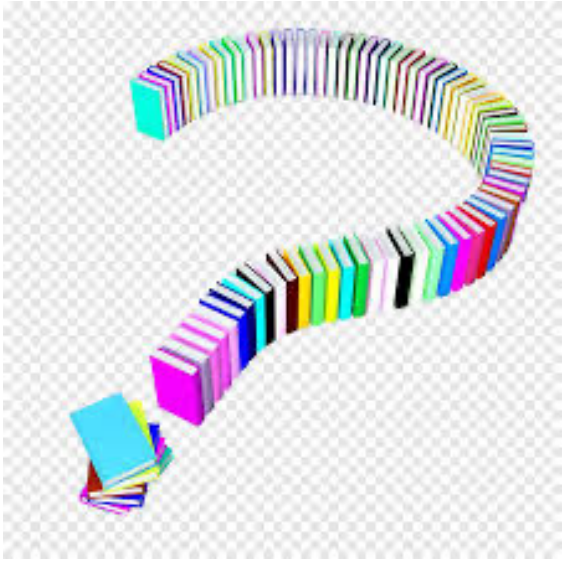
“How old are the associate professors in our university?”

is a statistical question:

“associate professors in our university” is the population,
“age” is the measurement variable,
and we expect several ages.

We might want to find a type of central tendency here (mean age, on average the professors age is…)

Is the question “Do people like pizza?” a well formulated statistical question?



Is the question “Do people like pizza?” a well formulated statistical question?

The question might be of interest but as it is formulated is too broad. It is unclear exactly what the population is.

A better version would be: “Of all teenagers in Pisa, who likes pizza with ham?”

The population is…?
The measurement is…?

A better version would be: “Of all teenagers in Pisa, who likes pizza with ham?”

The population is “teenagers in Pisa,”

the measurement is “like or don’t like pizza with ham,” and

we would expect some people to like this type of pizza and some not to like it

OTHER EXAMPLES…

How do boys and girls of first year classes of primary schools in Italy compare regarding the ability to read?
Population is…
measurement is…
we would expect ….

What is the longest-lasting brand of AA batteries?
Population is …
measurement is …
we would expect ….

In social sciences

Are wealthy people happier?
Is society becoming more tolerant of diversity?
How do people cope with financial hardship?
Do people with higher qualifications earn more?
Does volunteering increase your sense of wellbeing?
Do women get paid less than men?

The Educational Correlates of Voting: A Cross-sectional Study of Finnish Undergraduates' Turnout in the 2014 European Parliament Election

(1995) finds that political sophistication is a central mediator between educational attainment and turnout. Taken together, these studies suggest that higher education makes electoral participation easier and more meaningful because it translates into greater internal *political efficacy*, or a stronger belief in one's own competence to influence the political system (Campbell et al. 1960, 479). This discussion leads to our first hypothesis:

H1: HE students with higher degrees of internal political efficacy are more likely to vote than HE students with lower degrees of efficacy.

Scandinavian Political Studies, Vol. 42 – No. 1, 2019

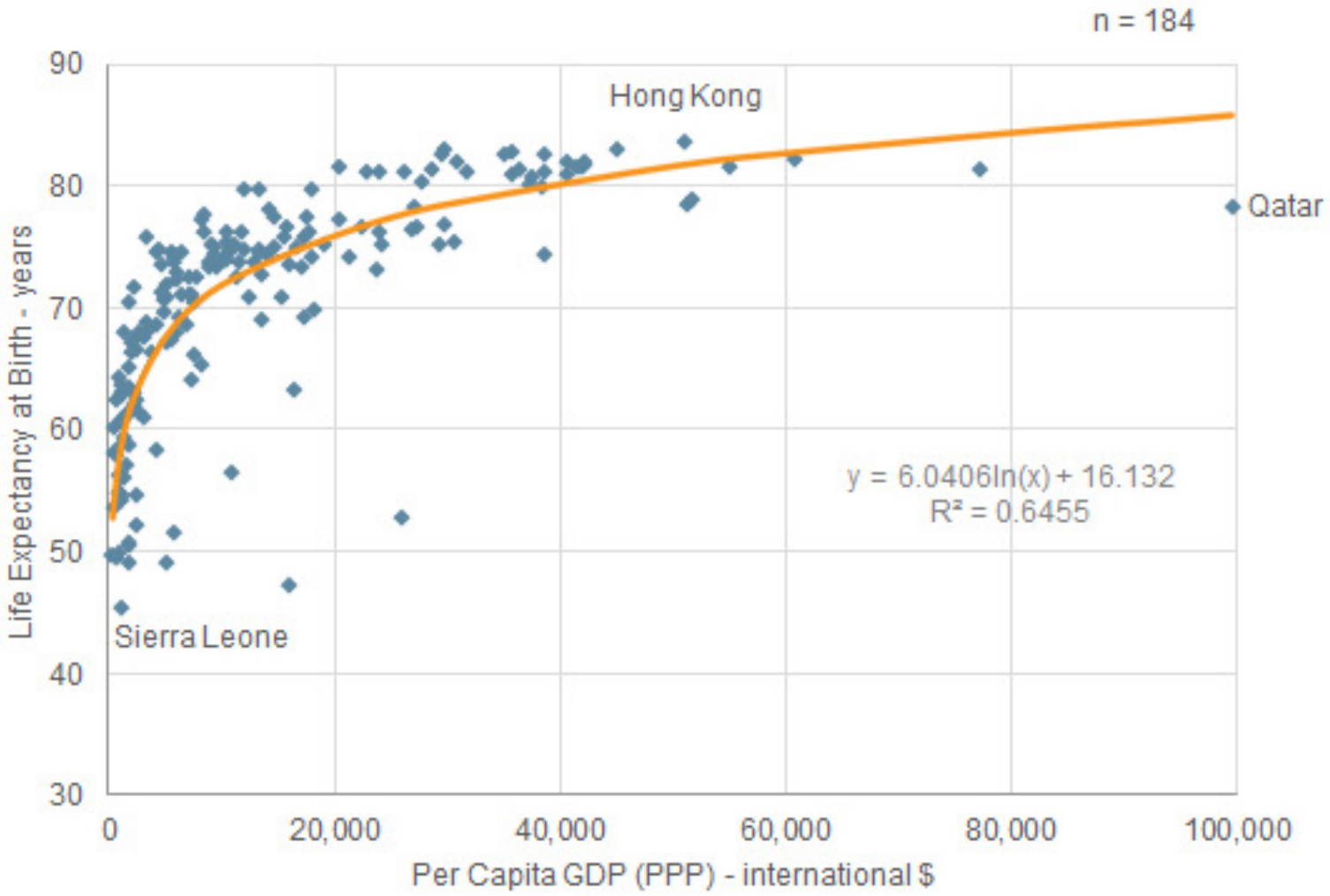
Results from a 2020 field experiment encouraging voting by mail

Experimental Design

We seek to answer three research questions. We are primarily interested in 1) whether postcards from local officials increase applications to vote by mail and secondarily 2) whether the postcards' wording matters or 3) their effect is stronger among those who had received four postcards as part of a prior 2019 experiment. The University of Pennsylvania Institutional Review Board approved this study (832927).

PNAS January 26, 2021 118 (4) e2021022118

Do higher values of GDP correspond to higher life expectancy?



Population? Measure?

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

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 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

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!

The Glossary of today

The population and sample

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,…)

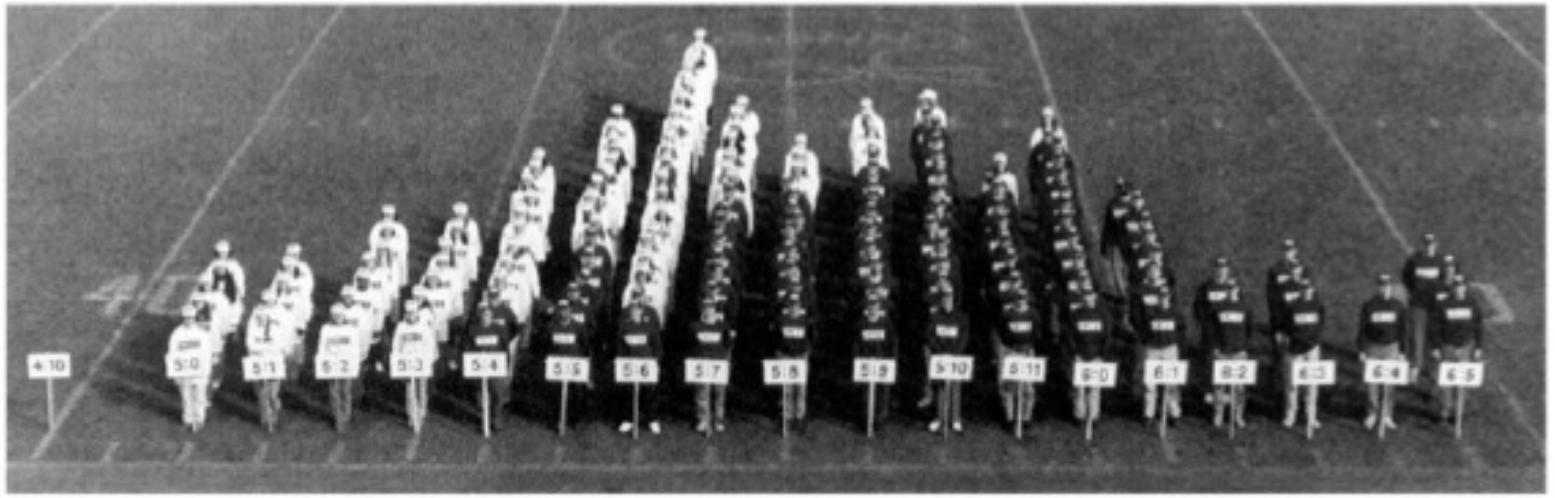
It can be composed of the entire **population** (universe) or of a subset of it (**sample**).

When is a population identified?

The population needs to be clearly identified at the beginning of a study.
The study should be based on a clear understanding of who or what is of interest, as well as the type of information required from that population.

Analysis of the height of one individual (single statistical unit)

Analysis of the height of a population



Example:

1. In a web survey, 2500 adults in the Italy were asked if they thought there was evidence of any financial crisis. 80% of the adults said yes.

2. The Ministry of Economics conducts a weekly surveys of approximately 900 gasoline stations to determine the average price per gallon of gasoline. In 2017, the average price was $2.75 per gallon.

Which is the population?
Which is the sample?

A **variable** is a characteristic or condition of a study subject that can be measured or counted. It changes or take on different values. Values may vary between data units in a population, and may change in value over time.

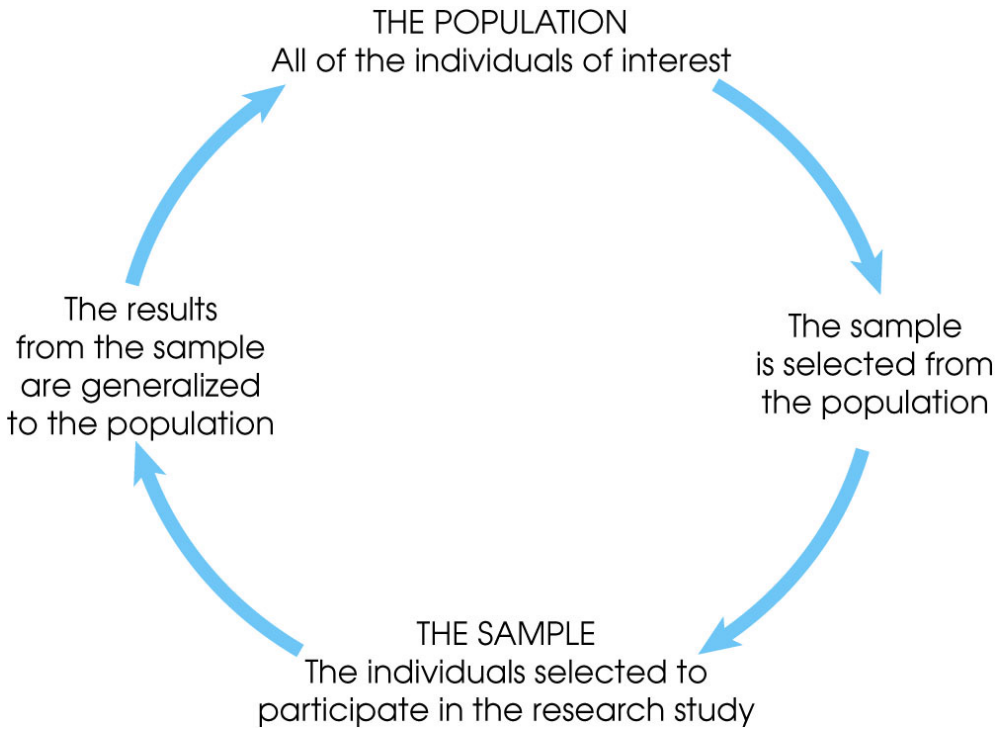
Height, age, income, country of birth, grades obtained at school are all examples of variables. Variables may be classified into various categories.

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.

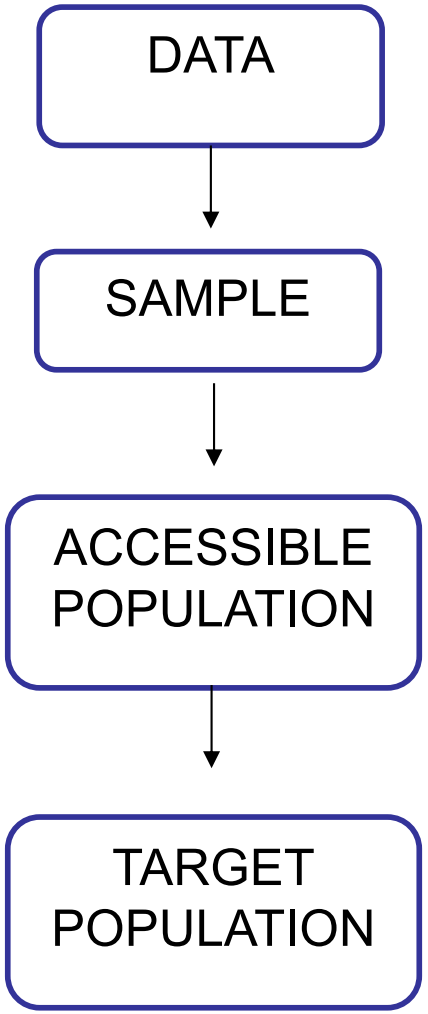
STATISTICS

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**



Inductive Inference Process



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

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

✓**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.

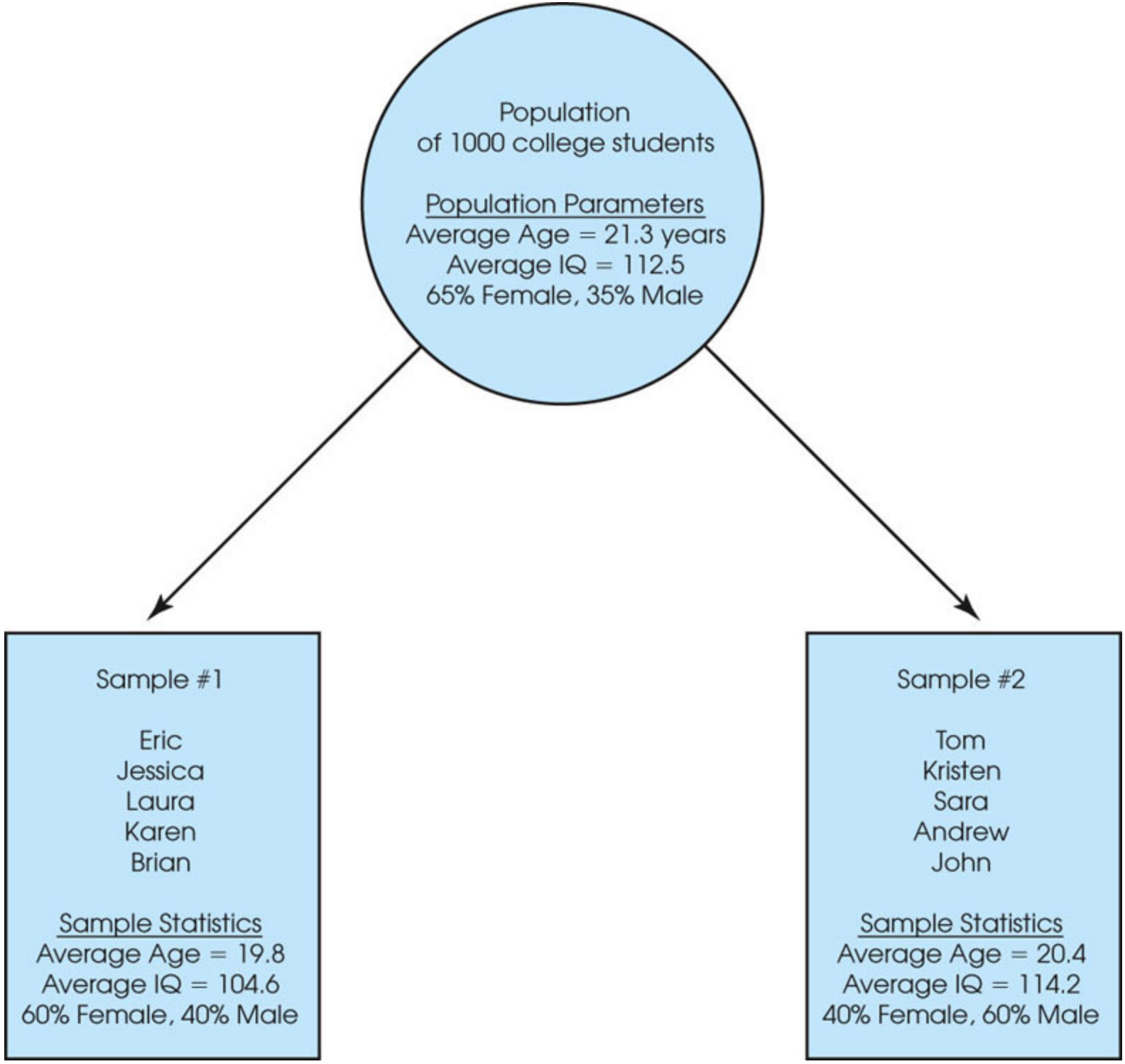
✓**Statistics:** known number that summarize a characteristics of the sample. A statistic is often used to point estimate the parameter in the population.

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

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


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

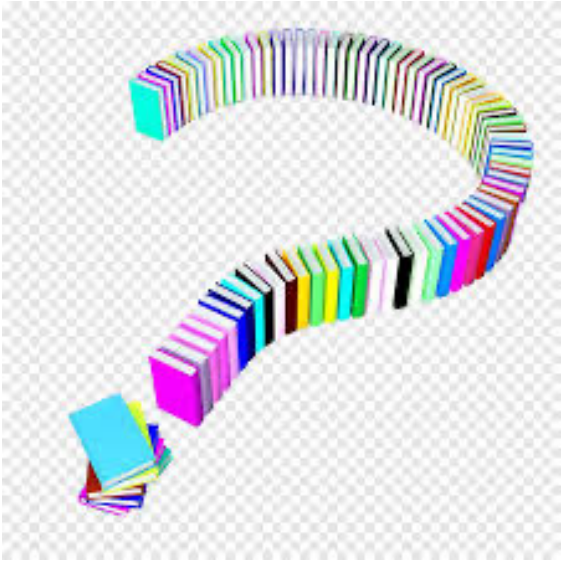
The Educational Correlates of Voting: A Cross-sectional Study of Finnish Undergraduates' Turnout in the 2014 European Parliament Election

(1995) finds that political sophistication is a central mediator between educational attainment and turnout. Taken together, these studies suggest that higher education makes electoral participation easier and more meaningful because it translates into greater internal *political efficacy*, or a stronger belief in one's own competence to influence the political system (Campbell et al. 1960, 479). This discussion leads to our first hypothesis:

H1: HE students with higher degrees of internal political efficacy are more likely to vote than HE students with lower degrees of efficacy.

Scandinavian Political Studies, Vol. 42 – No. 1, 2019

Macro or micro?
Population?
Sample?
Variables?



Practice:

- ✓ divide in groups
- ✓ choose one statistical question
- ✓ Discuss about of the elements of the statistical process (population, sample, variables), how you might collect data and describe the variability in the data you might expect.

Observational studies and sampling strategies

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:

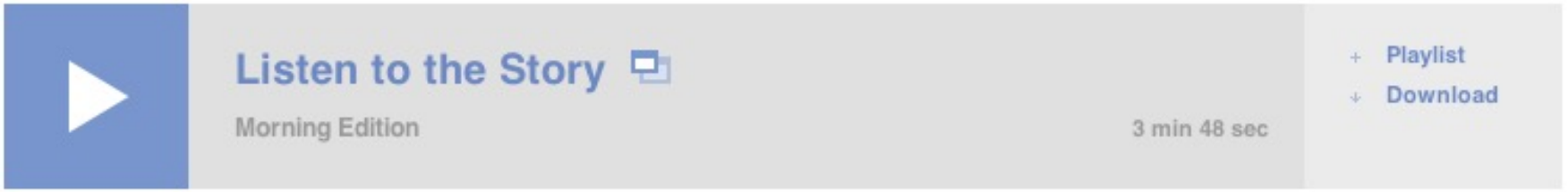
- It can be difficult to complete a census: there always seem to be some individuals who are hard to locate or hard to measure. Populations rarely stand still. Even if you could take a census, the population changes constantly, so it's never possible to get a perfect measure.
- Taking a census may be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



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

Observational studies: Obtaining good samples

- Almost all statistical methods are based on the notion of implied randomness.

- If observational data are not collected in a random framework from a population, 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.

When possible, random selection should be used to choose samples in research studies! Random selection helps ensure a sample is representative of the population from which it was chosen. More practically, random selection allows researchers to generalize sample results to some larger population of interest.

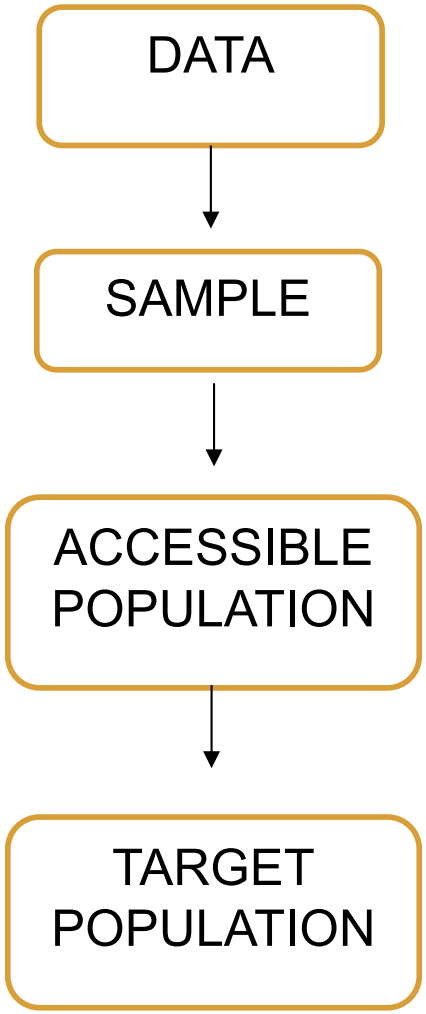
The most basic random sample is called a **simple random sample**: each case in the population has an equal chance of being included and there is no implied connection between the cases in the sample.

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.

Types and sources of error in statistical data

Inductive Inference Process



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

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

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.

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.

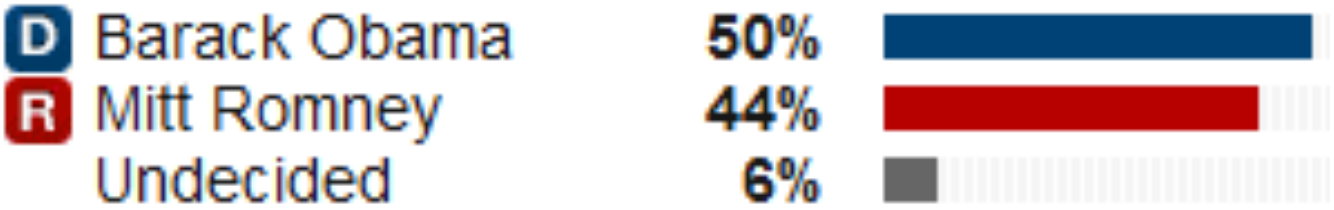
The discrepancy (natural difference that exist by chance) between a sample statistic and its population parameter is called **sampling error**.

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

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

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? (3/3)

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.

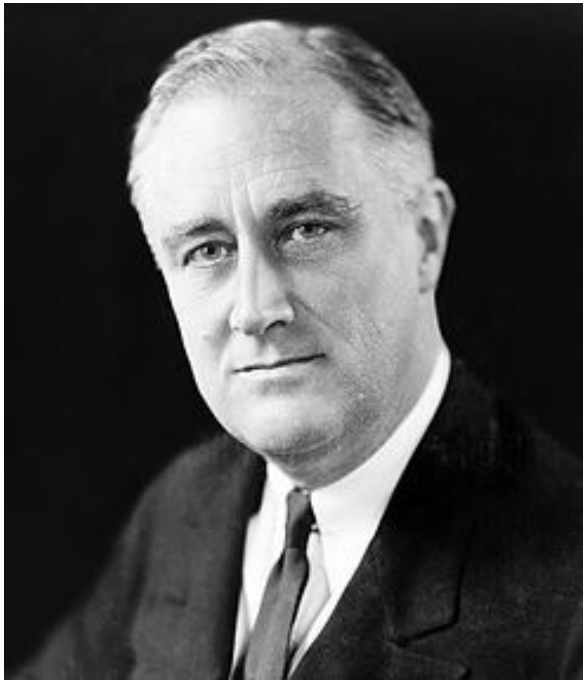
The greater the error the less reliable are the results of the study.

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'.

Sampling bias example:
Landon vs. FDR

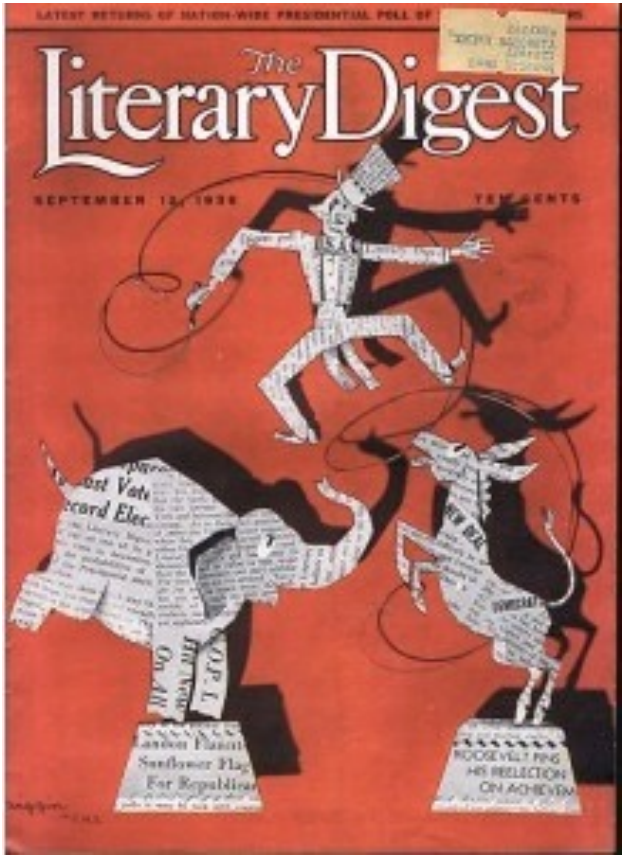
A historical example of a biased sample yielding misleading results

In 1936, Landon sought the Republican presidential nomination opposing the re-election of Franklin D. Roosevelt.



The Literary Digest Poll

- The Literary Digest polled about 10 million Americans, and got responses from about 2.4 million.
- The poll showed that Landon would likely be the overwhelming winner and FDR would get only 43% of the votes.
- Election result: Roosvelt won, with 62% of the votes.
- The magazine was completely discredited because of the poll



The Literary Digest Poll - what went wrong?

- The magazine had surveyed
  - its own readers,
  - registered telephone users.

- These groups had incomes well above the national average of the day (remember, this is Great Depression era) which resulted in lists of voters far more likely to support Republicans than a truly *typical* voter of the time, i.e. the sample was not representative of the American population at the time.

Large samples are preferable, but...

- The Literary Digest election poll was based on a sample size of 2.4 million, which is huge, but since the sample was *biased*, the sample did not yield an accurate prediction.

Non sampling error example: biased survey questions

"Poll Finds 1 out of 3 Americans Open to Doubt There was a Holocaust," Los Angeles Times, April 20, 1993, the results from a Roper poll created a big stir.

But a closer look at the question asked reveals a potential problem with question bias: "Does it seem possible or does it seem impossible to you that the Nazi extermination of the Jews never happened?" The compound structure of the sentence and the use of the double negative makes the question confusing. 22% reported that it was possible that the Holocaust didn't happen, and 12% didn't know.

A year later, Roper repeated the survey, keeping all other questions the same and rewording this one: "Does it seem possible to you that the Nazi extermination of the Jews never happened, or do you feel certain that it happened?" This time only 1% reported that it was possible that it didn't happen, and 8% didn't know.

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/

"1 in 4 youths abused, survey finds" 10/4/94 San Francisco Examiner

1 in 4 youths abused, survey finds

CHICAGO One in four adolescents said they were physically or sexually abused within the past year, according to a new survey.

The telephone survey of 2,000 children ages 10 to 16, suggests, "We're not doing a very good job of counting and tracking the problem," said David Finkelhor, a sociologist at the University of New Hampshire and co-author of the study in the October issue of Pediatrics.

The assault rate reported by the participants was 15.6 percent, three times higher than the 5.2 percent reported in the National Crime Survey in 1991, the study said. The survey's rate for rape was 0.5 percent, five times higher than the federal estimate of 0.1 percent.

from "Children as victims of violence: a national survey," D. Finkelhor and J. Dziuba-Leatherman, Pediatrics 94, 413–420 (1994)

Some limitations

- ✓ Nonresponse.
- ✓ Difficulties in generalizing to the real world: only those between 10-16 years were studied.
- ✓ Uncertainty of effectiveness of a telephone interview (time constraints, no knowledge of body language)
- ✓ Exclusion of some high risk children such as those without telephones, those in juvenile correctional and mental health facilities, those with disabilities, and those who were angry/alienated to participate.
- ✓ Possible lack of disclosure of intimate victimizations to a stranger interviewer.
- ✓ Definitions are broad. The NYS estimate may be higher because it includes assault from siblings and other nonparents family members which the Finkelhor and Dziuba-Leatherman nonfamily assault rate does not cover

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

- 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 the explanatory and response variables.

There are many other 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?

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.

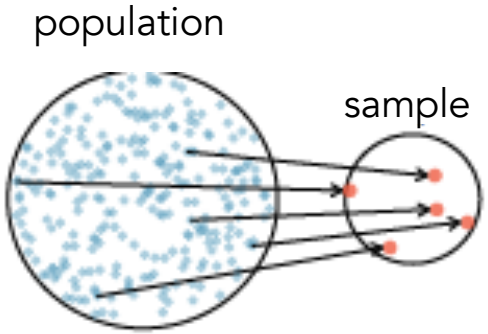
✓The only guarantee for that comes from the method used to select the sample (sampling method)→**probability sampling**

✓There are several sampling methods that guarantee representativeness.

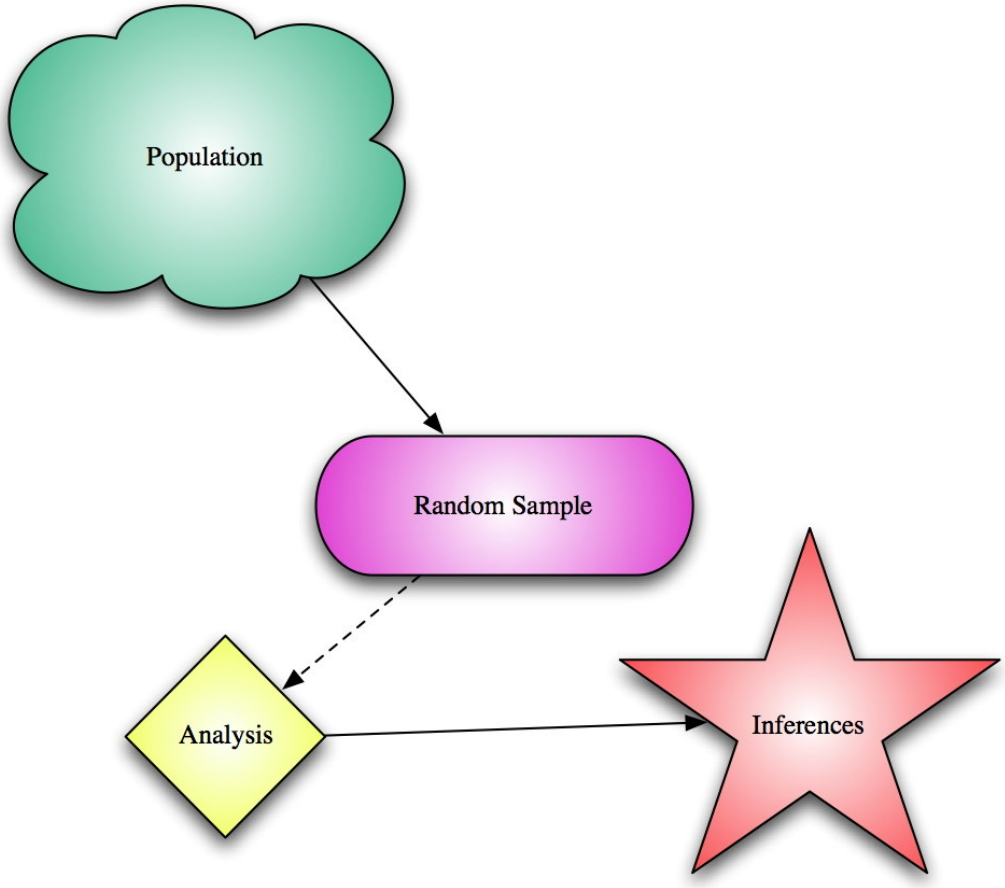
Obtaining good samples

- 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.



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.

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.

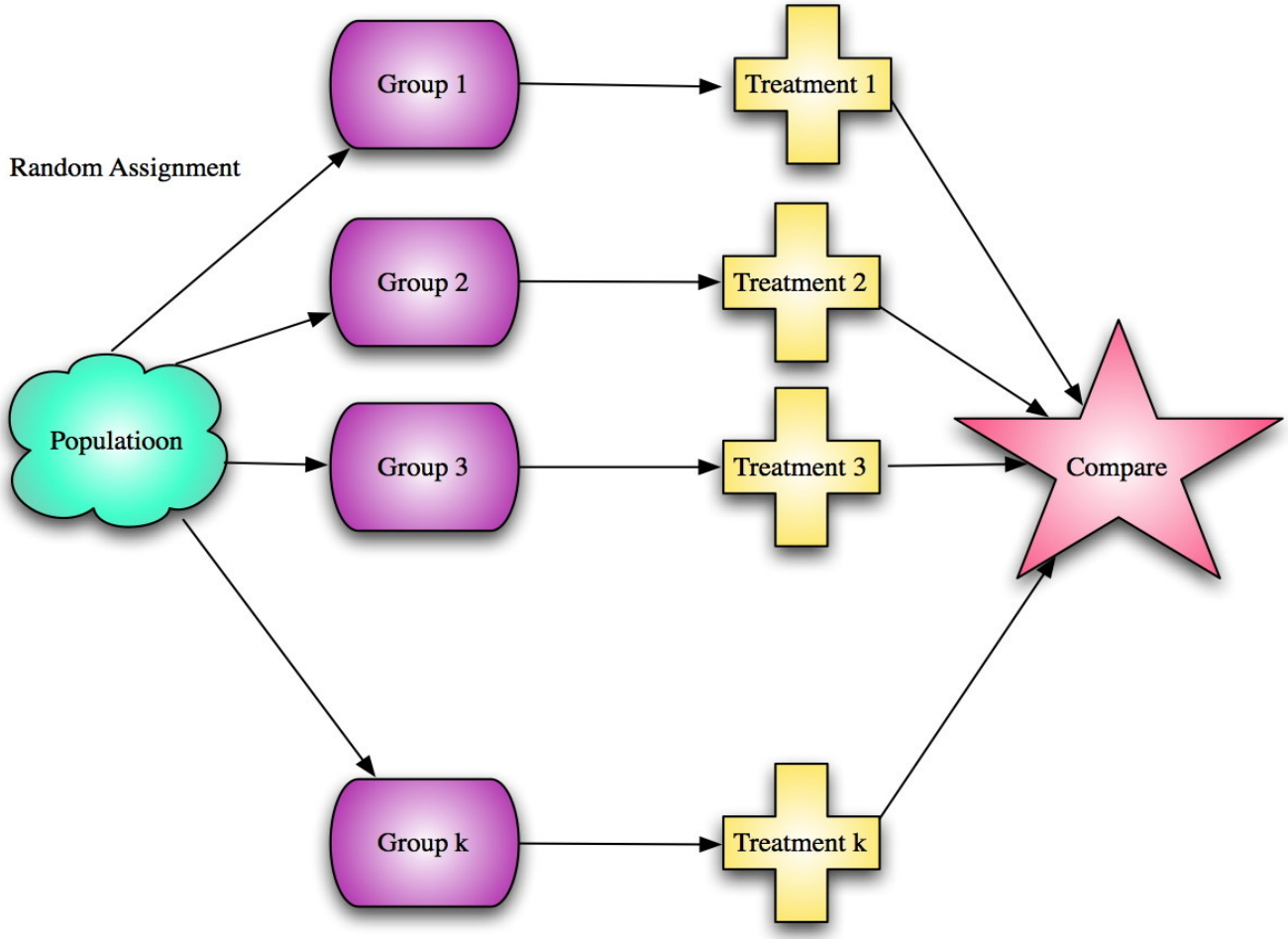
Retrospective studies collect data after events have taken place.

- 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.

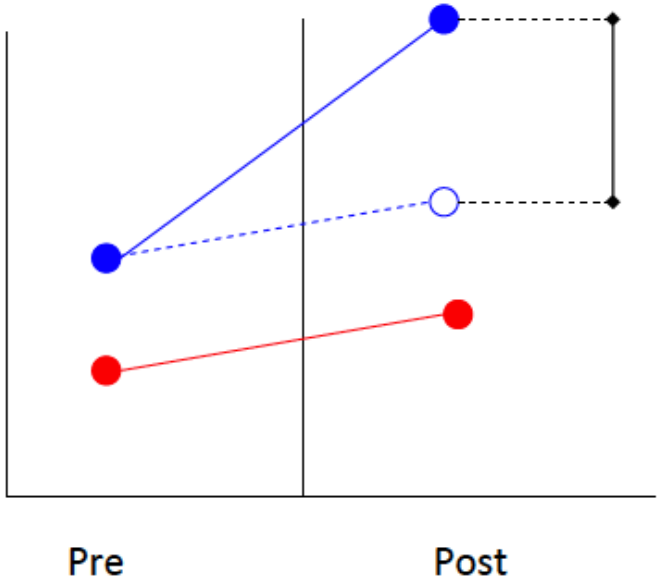
2006 and 2007: record-breaking maize harvests in Malawi. Was this dramatic turnaround the result of the subsidy? How can we tell what the true impact of the subsidies was?

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

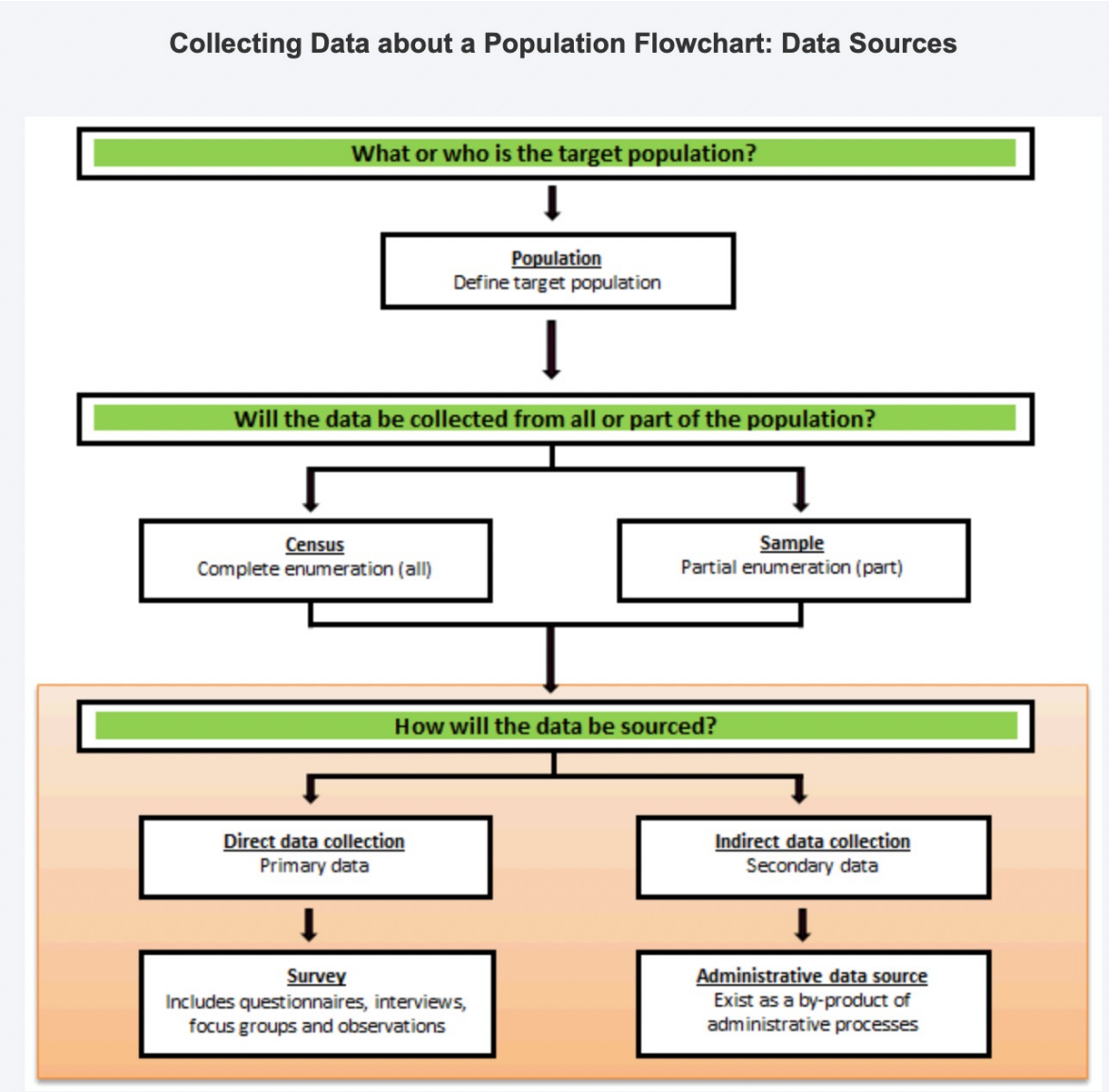
Whatever happened to the control group over time is what would have happened to the treatment group in the absence of the program.



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

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

Let's concentrate on observational studies



Australian Bureau of Statistics

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

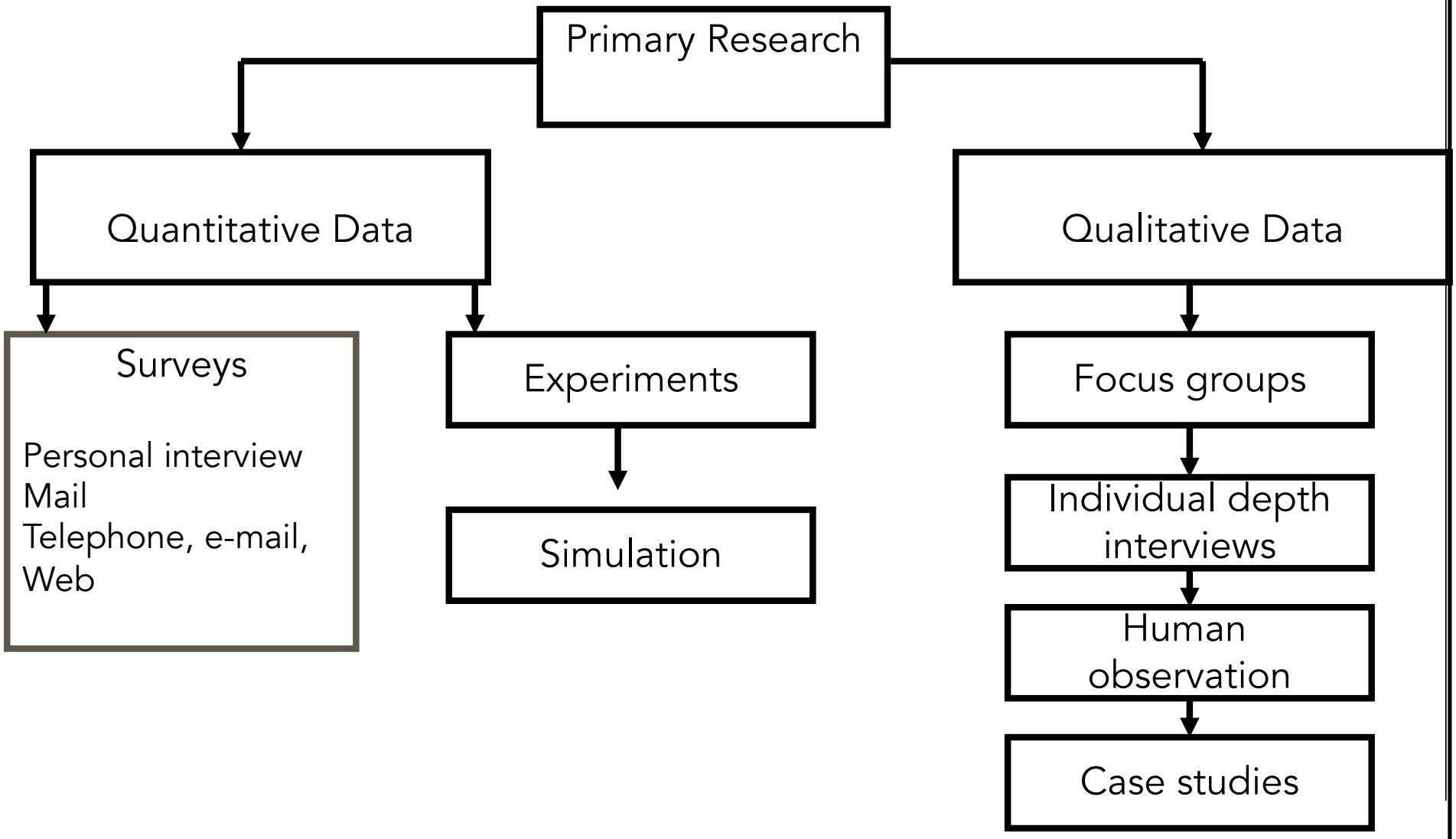
- Primary:

  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



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
- Usually fewer observations
- Individual opinions and characteristics 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

Practice

Research question: Do most people wash their hands after using the bathroom?

Not according to a December 2005 newspaper article titled "Many Adults Report Not Washing Their Hands When They Should, and More People Claim to Wash Their Hands than Who Actually Do".

The article was based on two studies that were done in August 2005. In the first study, 1,013 U.S. adults were asked questions about their hand-washing habits by telephone. In the second study, observers watched and recorded the actual hand-washing behaviours of 6,336 adults in public restrooms in four major U.S. cities. Both studies were carried out by Harris Interactive, a company that specializes in these kinds of statistical research.

Results:

While 91% of surveyed adults *claimed* to always wash their hands after using the bathroom, only 83% of the adults in the second study did so.

In the survey, 94% of women claimed to always wash their hands after using the bathroom, compared with 88% of men. In the second study, 90% of the women actually washed their hands, compared with 75% of men.

A study similar to the second one was done in 2003 revealed that 78% of the adults observed actually washed their hands after using the bathroom. In that study, 83% of the women and 74% of the men were observed washing their hands.

Based on these studies, what can we conclude?

Can we conclude that 83% of *all* U.S. adults always wash their hands after using the bathroom?

In the Harris Interactive survey, people were contacted by telephone. One of the questions the interviewers asked was, “How often do you wash your hands after using a public restroom?”

Which U.S. adults were not included in this study?

The survey estimated that 91% of all U.S. adults would claim that they always wash their hands after using the bathroom. Do you think this estimate is too high, too low, or about right?

Several people refused to participate in the survey. Give a reason that this might happen.

In any survey, it is possible that some people will not answer a question accurately or honestly. Thinking about the hand-washing survey, do you think this is likely to happen? Explain your answer

The second study of hand washing was conducted at a baseball field in Atlanta, a museum and an aquarium in Chicago, a bus and train terminal in New York, and a farmers' market in San Francisco.

a) Observers in the public bathrooms combed their hair or put on make-up at one of the available sinks while they were watching individuals' hand-washing behaviors. If the observation had been done by hidden camera instead (with no observer present), do you think the percent who washed their hands would have been greater than, less than, or about the same as 83%?

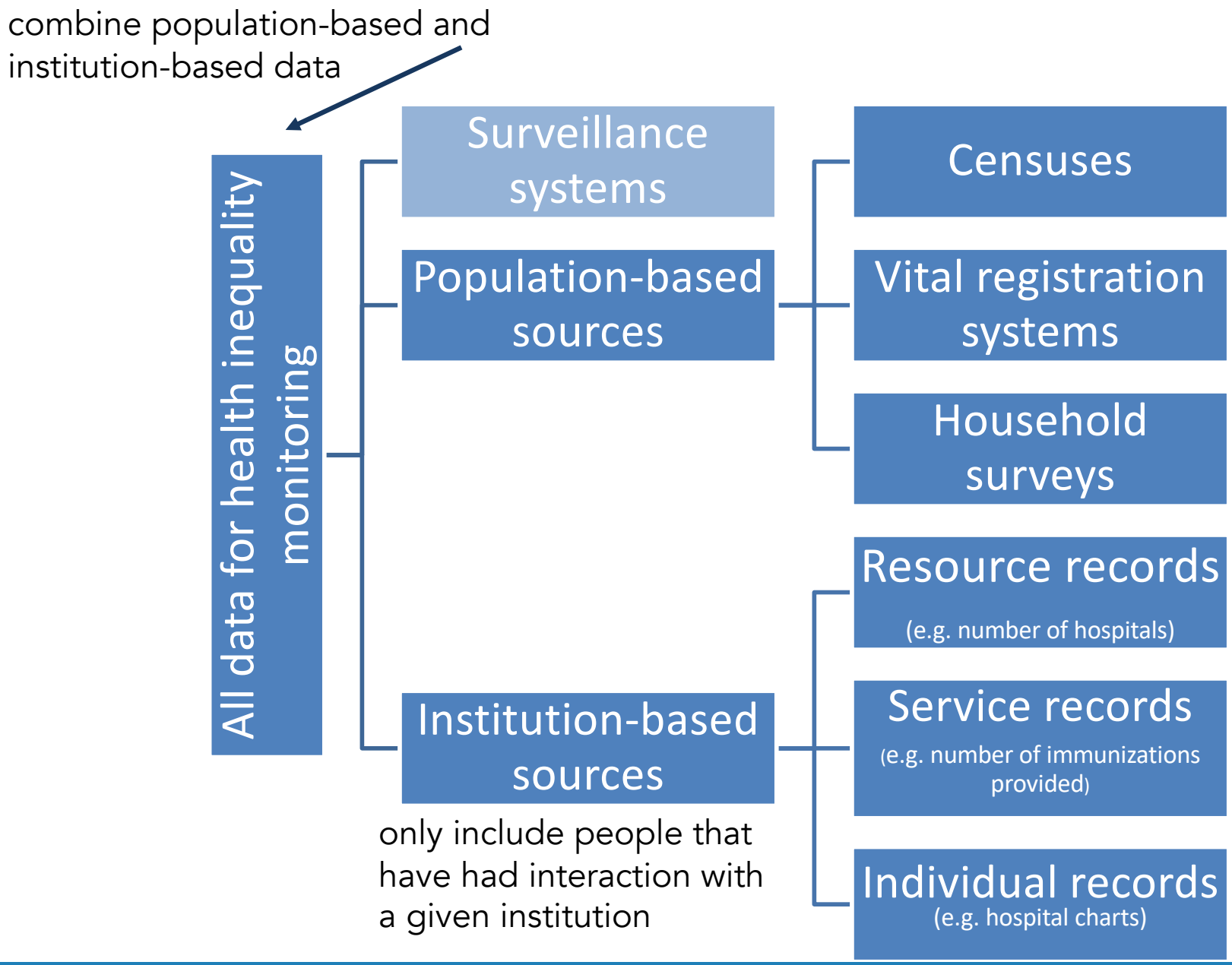
b) Suppose the observational study had been conducted using hidden cameras in the homes of the same 6,336 adults. Do you think the percent of these individuals who washed their hands would have been greater than, less than, or about the same as 83%?

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

Data sources

World Health Organization

Data source types



Population-based sources: censuses

- National population and household censuses are implemented every 10 years in most countries
- Data cover the entire population (or nearly so), providing accurate denominator counts for population subgroups
- Censuses may be a good source of information about equity stratifiers, but usually contain only limited information on health
- Possible improvements:
  - include mortality and cause of death questions

Population-based sources: vital registration systems

- For example, national birth or death registries
- These systems are the best and most-reliable source for fertility, mortality, life expectancy and cause-of-death statistics
- Vital registration systems can often be linked to information on sex, geographical region, occupation, education
- Incomplete in many low- and middle-income countries
- Possible improvements:
  - Expand coverage
  - Include cause of death, birth weight and gestational age
  - Include at least one socioeconomic indicator

EXTRA INFORMATION

Global status of vital registration systems

- In 2009…
  - only 25% of the world population lived in countries where at least 90% of births and deaths are registered
  - only 34 countries (representing 15% of the global population) had high-quality cause-of-death data
  - 74 countries lacked data altogether about births and deaths
  - In the WHO African Region 42 out of 46 countries had no death registration data

Source: World Health Organization. *World Health Statistics 2012*. Geneva, World Health Organization, 2012.

Population-based sources: household surveys

- Currently the most common and overall most reliable data source for health inequality monitoring in low- and middle-income countries
- Data are representative for a specific population (often national)
- Have rich data on a specific health topic as well as living standards and other complementary variables
- Often repeated over time, allowing for measurement of time trends
- Conducted in multiple countries, allowing for benchmarking
- Sampling and non-sampling errors can be important
- Survey may not be representative of small subpopulations of interest
- Possible improvements:
  - Repeat surveys on a regular basis
  - Enhance comparability over time and between countries by harmonizing survey questions
  - Increase sample sizes

EXTRA INFORMATION

Examples of multinational household survey programmes

Survey name	Organization	Website
AIDS Indicator Survey (AIS)	United States Agency for International Development	http://www.measuredhs.com/What-We-Do/Survey-Types/AIS.cfm
Demographic and Health Survey (DHS)	United States Agency for International Development	http://www.measuredhs.com/
Living Standards Measurement Study (LSMS)	World Bank	http://go.worldbank.org/IPLXWMCNJ0
Malaria Indicator Survey (MIS)	United States Agency for International Development	http://www.malariasurveys.org/
Multiple Indicator Cluster Survey (MICS)	United Nations Children's Fund	http://www.unicef.org/statistics/index_24302.html
Study on Global Ageing and Adult Health (SAGE)	World Health Organization	http://www.who.int/healthinfo/systems/sage/en/
World Health Survey (WHS)	World Health Organization	http://www.who.int/healthinfo/survey/en/index.html

EXTRA INFORMATION

Data availability in low- and middle-income countries

- Household surveys are the main data source in many low- and middle-income countries
- Health inequality monitoring in low- and middle-income countries is limited to the health indicators for which data are available
  - Often outcome or impact indicators
- Certain health topics may be challenging to monitor, particularly those related to inputs and processes and outputs, which are usually collected from institution-based sources

Institution-based data sources

- Data are readily and quickly available (medical charts, police records, employment records and school records,…)
- Can be used at lower administrative levels (e.g. district level)
- Data may be fragmented or of poor quality
- Often data cannot be linked to other sources
- Data may not be representative of the whole population

Surveillance systems

- Can provide detailed data on a single condition or from selected sites
- Data may be useful for correction of over- or under-reporting
- Not always representative of population
- Some systems may collect little information relevant to equity stratifiers
- Possible improvements:
  - Include individual or small-area identifiers
  - Integrate surveillance functionality into larger health information systems with full coverage

EXTRA INFORMATION

Types of surveillance systems

- **Outbreak disease surveillance**
  - aims to track cases of epidemic-prone diseases as well as their risk factors
  - often relies on frequent reporting by health facilities, such as laboratories
- **Sentinel surveillance**
  - uses a sample of clinics for intensified monitoring
  - is used by disease programmes such as HIV and malaria
- **Risk factor surveillance**
  - describes data collection and analysis in noncommunicable disease monitoring
  - often focuses on data obtained through surveys
- **Demographic surveillance**
  - found in many low- and middle-income countries
  - sites have a longitudinal birth and death registration system for a local population to collect information about cause of death and other health-related data

Data source mapping: step 1

List of data sources by type (partial table)

Data source type	Data source	Year(s) of data collection	Notes
Census	National census	1990, 2000, 2010	
Administrative	Immunization records	2000–2006	Annual collection
Household survey	Standard DHS	1994, 1999, 2004, 2009	
…			

Note: DHS = Demographic and Health Survey

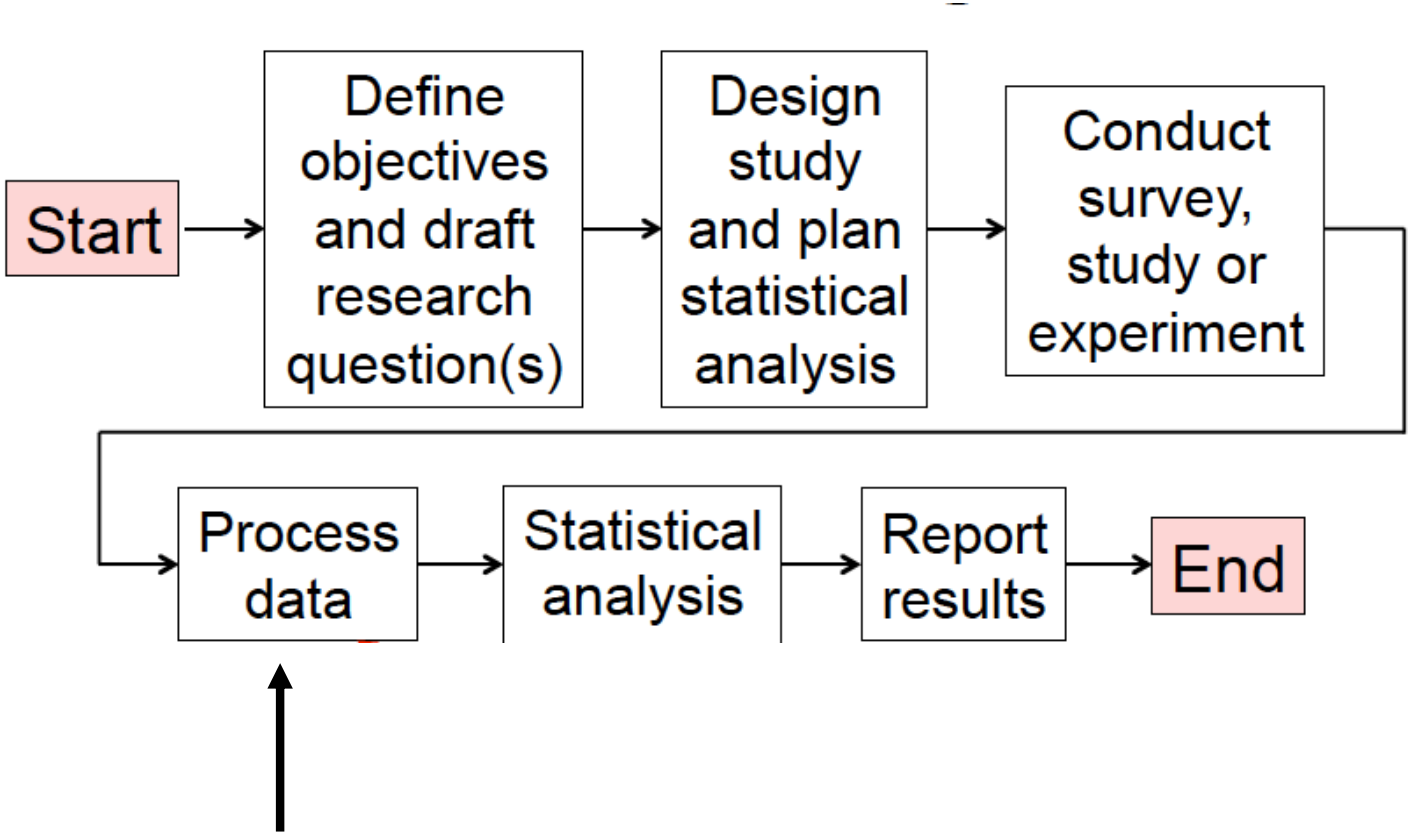
Data source mapping: step 2

List of data sources and equity stratifiers (partial table)

No.	Data source and year	Equity stratifier				Notes
		Sex	Wealth	Place of residence	Province or region	
1	Immunization records 2000–2006				√	
2	DHS 2009	√	√	√	√	17 provinces
3	DHS 2004	√		√	√	13 provinces
	…					

DATA & DESCRIPTIVES

The research study process



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

Ensures you collect 'good' data!

Allows you to draw valid conclusions and answer your research question(s),

- Reduces potential bias
- Reduce variability in your data
- Enables you to see the big picture
- Improves accuracy (precision) of results
- Reduces amount of data needed
- Reduces cost (time or money)
- Surveys or observational studies cannot identify causes and effects
- Designed experiments can!

Data Matrices (nxp)

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 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

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

Relationships between variables

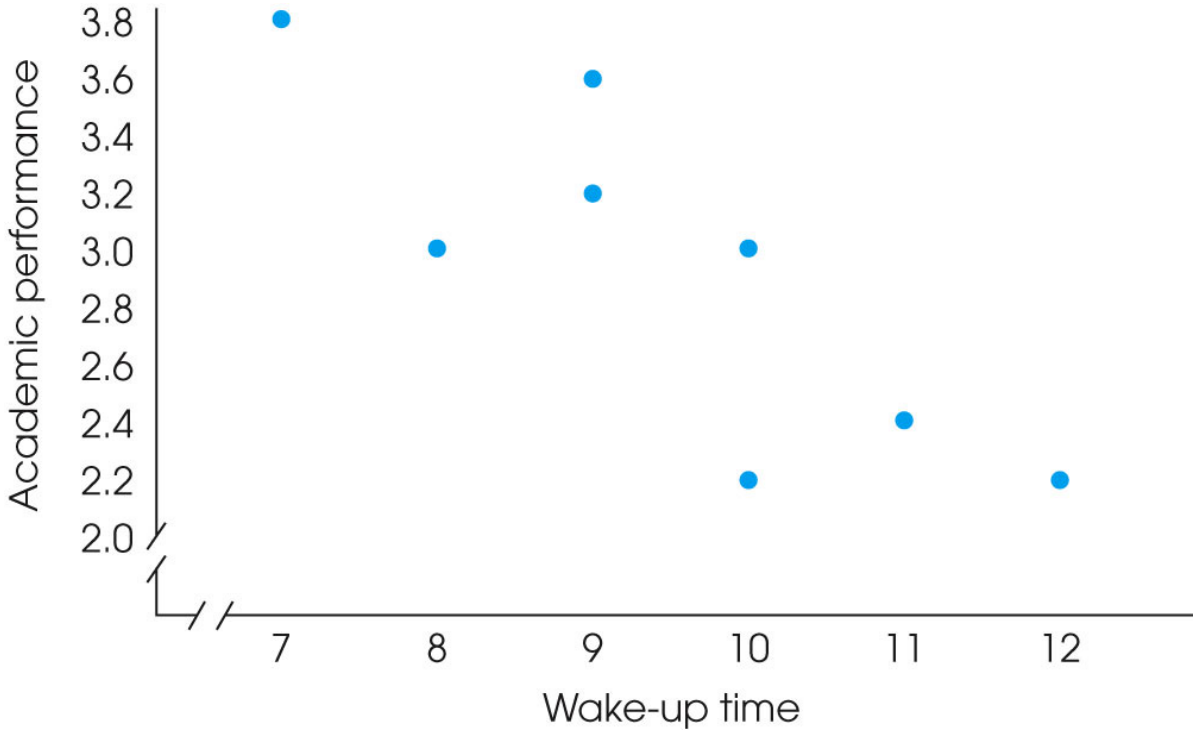
Many analyses are motivated by a researcher looking for a relationship between two or more variables. A social scientist may like to answer some of the following questions:

(1) Does a higher than average increase in county population tend to correspond to counties with higher or lower median household incomes?
(2) How useful a predictor is median education level for the median household income for US counties?
(3) Is lower wake-up time associated with lower academic performance?

To answer these questions, data must be collected, such as the urban data set. Examining summary statistics could provide insights for the above mentioned questions. Additionally, graphs can be used to visually explore data.

Scatterplots are one type of graph used to study the relationship between two **numerical variables.**

Child	Wake-up Time	Academic Performance
A	11	2.4
B	9	3.6
C	9	3.2
D	12	2.2
E	7	3.8
F	10	2.2
G	10	3.0
H	8	3.0



EXPLANATORY AND RESPONSE VARIABLES

When we ask questions about the relationship between two variables, we sometimes also want to determine if the change in one variable causes a change in the other.

Consider the following rephrasing of an earlier question:
If there is an increase in the median household income in a county, does this drive an increase in its population?
In this question, we are asking whether one variable affects another.
If this is our underlying belief, then median household income is the **explanatory** variable and the population change is the **response variable** in the hypothesized relationship.

EXPLANATORY AND RESPONSE VARIABLES

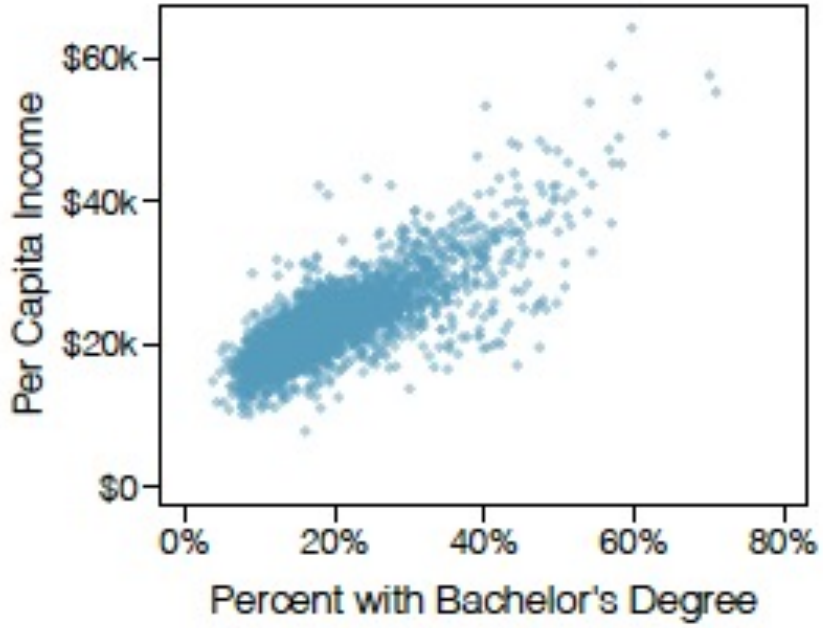
When we suspect one variable might causally affect another, we label the first variable the explanatory variable and the second the response variable. Explanatory might affect variable response variable.

For many pairs of variables, there is no hypothesized relationship, and these labels would not be applied to either variable in such cases.

Bear in mind that the act of labeling the variables in this way does nothing to guarantee that a causal relationship exists. A formal evaluation to check whether one variable causes a change in another requires an experiment.

Income and education in US counties. The scatterplot below shows the relationship between per capita income (in thousands of dollars) and percent of population with a bachelor's degree in 3,143 counties in the US in 2010.

(a) What are the explanatory and response variables?
(b) Describe the relationship between the two variables.
(c) What Can we conclude?

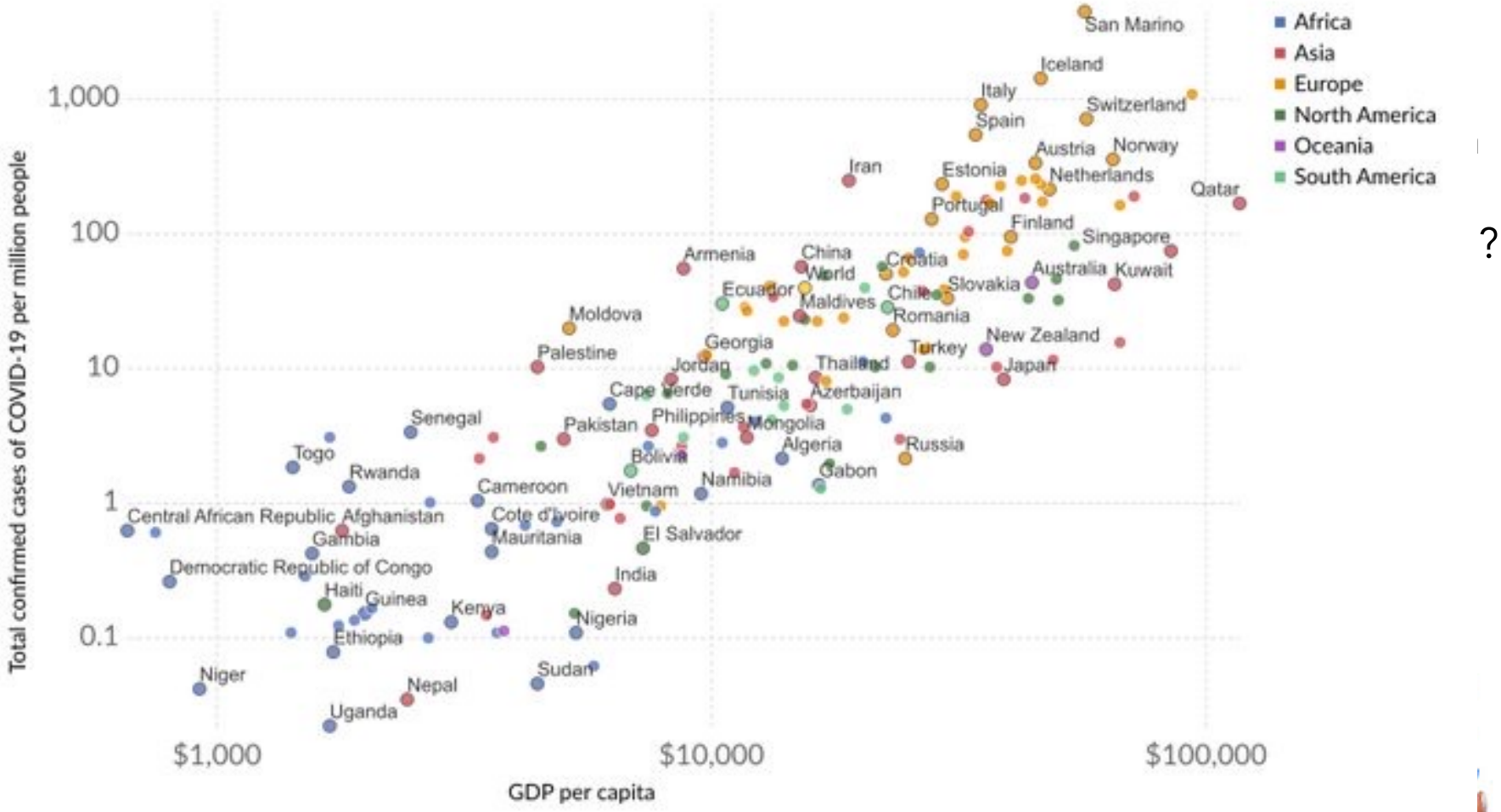


Total confirmed cases per million people vs. GDP per capita, Mar 22, 2020

The number of confirmed cases of COVID-19 is lower than the number of total cases. The main reason for this is limited testing.

GDP per capita is adjusted for price differences between countries (it is expressed in international dollars).

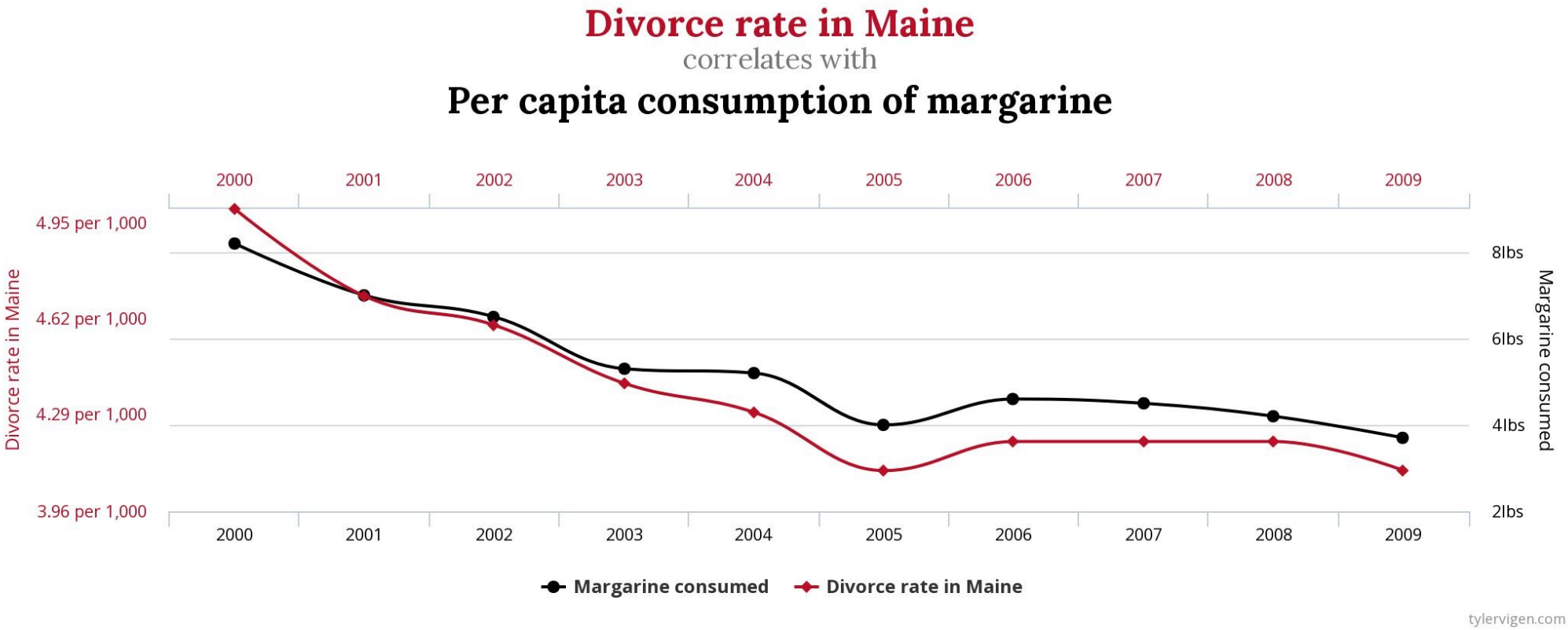
Our World in Data



Source: European CDC – Latest Situation Update Worldwide, World Bank

CC BY

?



http://www.tylervigen.com/

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

Data Matrix: aggregated data

rate

Average years of ed.

% change.

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 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

Proportions, Rates & Ratios – aggregated indexes

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.

Example: If 20 of 100 clients are less than 35 years of age, what is the proportion of young clients in the clinic?
20/100 = 1/5
A way to express a proportion is the percentage (proportion multiplied by 100)

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

A proportion is always a ratio
A rate is always a ratio
A rate may or may not be a proportion

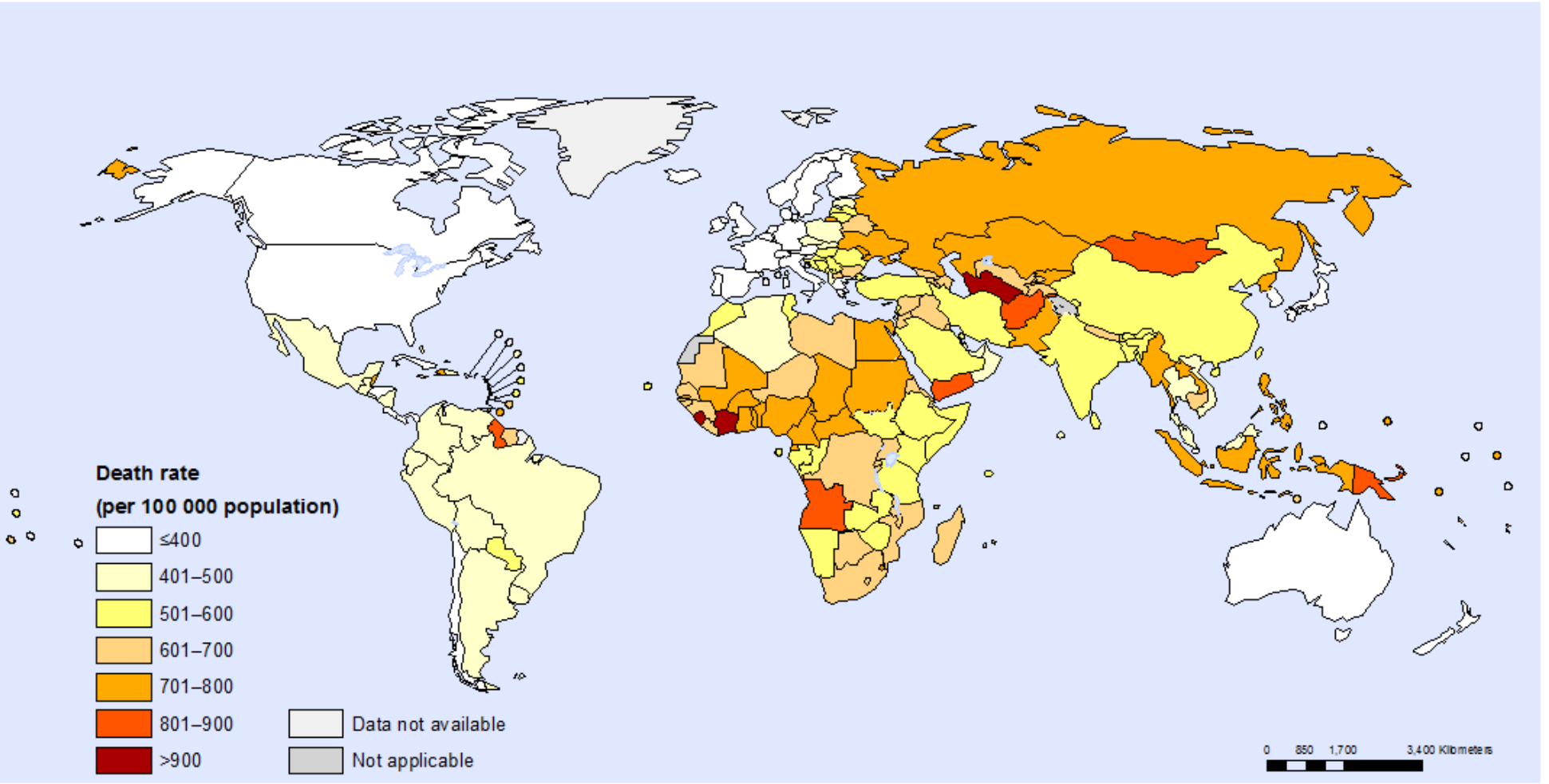
Death rate (all causes), crude (per 1,000 people)

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

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

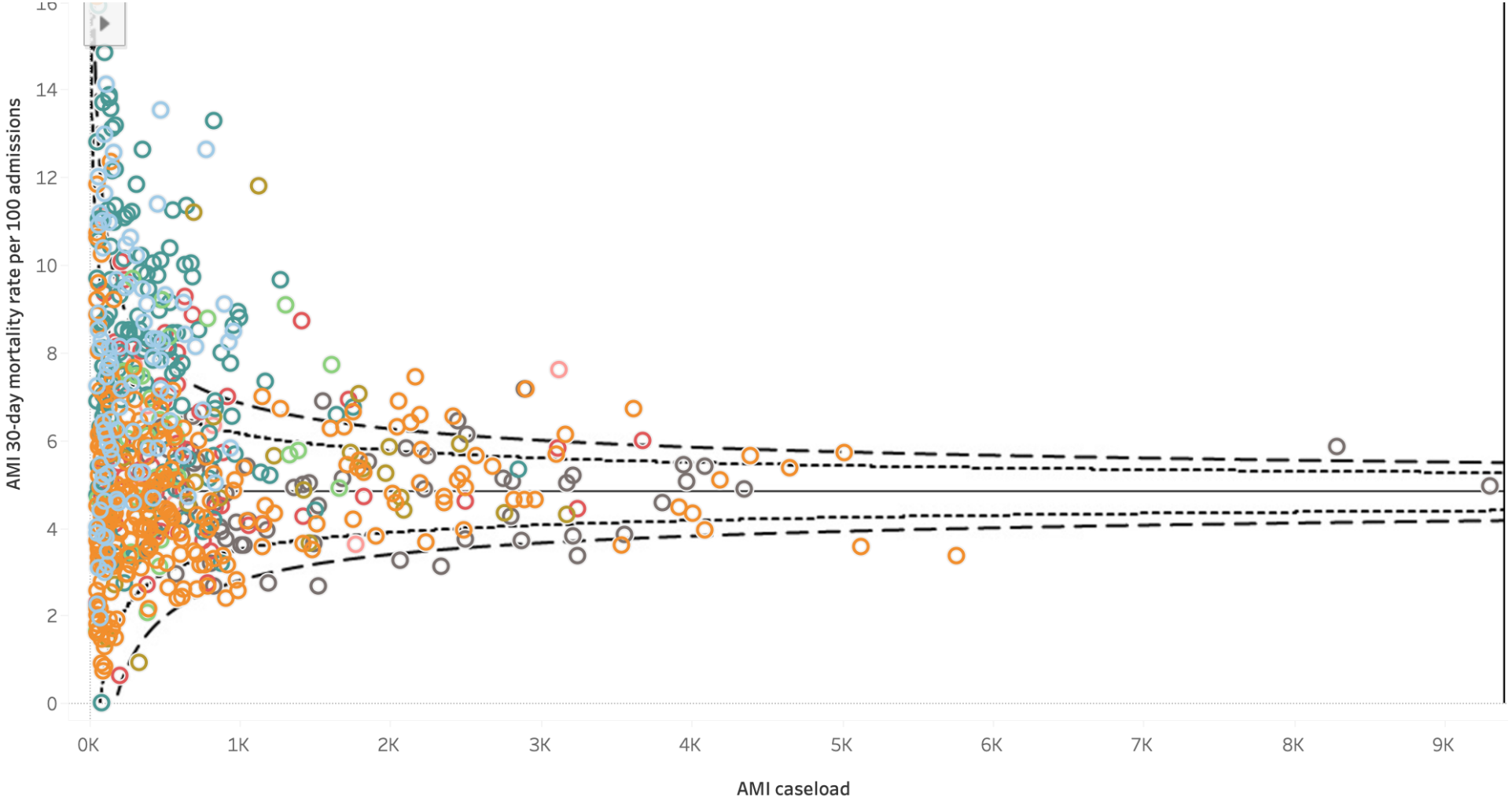


The boundaries and names shown and the designations used on this map do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate border lines for which there may not yet be full agreement.

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

World Health Organization

© WHO 2017. All rights reserved.



	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

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)

Assessing Change in Two Rates:

Absolute (arithmetic) change = rate2–rate1
Relative change = rate2/rate1
Proportional (percent) change = (rate2–rate1)/rate1

Example

1989: rate1 = 1,153 – 1996: rate2 = 307

307–1,153 = –846 or an absolute decrease in the rate of 846 cases per 100,000 persons
– (would = 0 if no change)

307 / 1153 = 0.27 or 1–0.27=0.73 or 73% relative decrease in rate – (would = 1 if no change)

Proportional change – (307–1,153)/1,153 = –0.73 or 73% relative decrease in rate.
– (would = 0 if no change)

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)

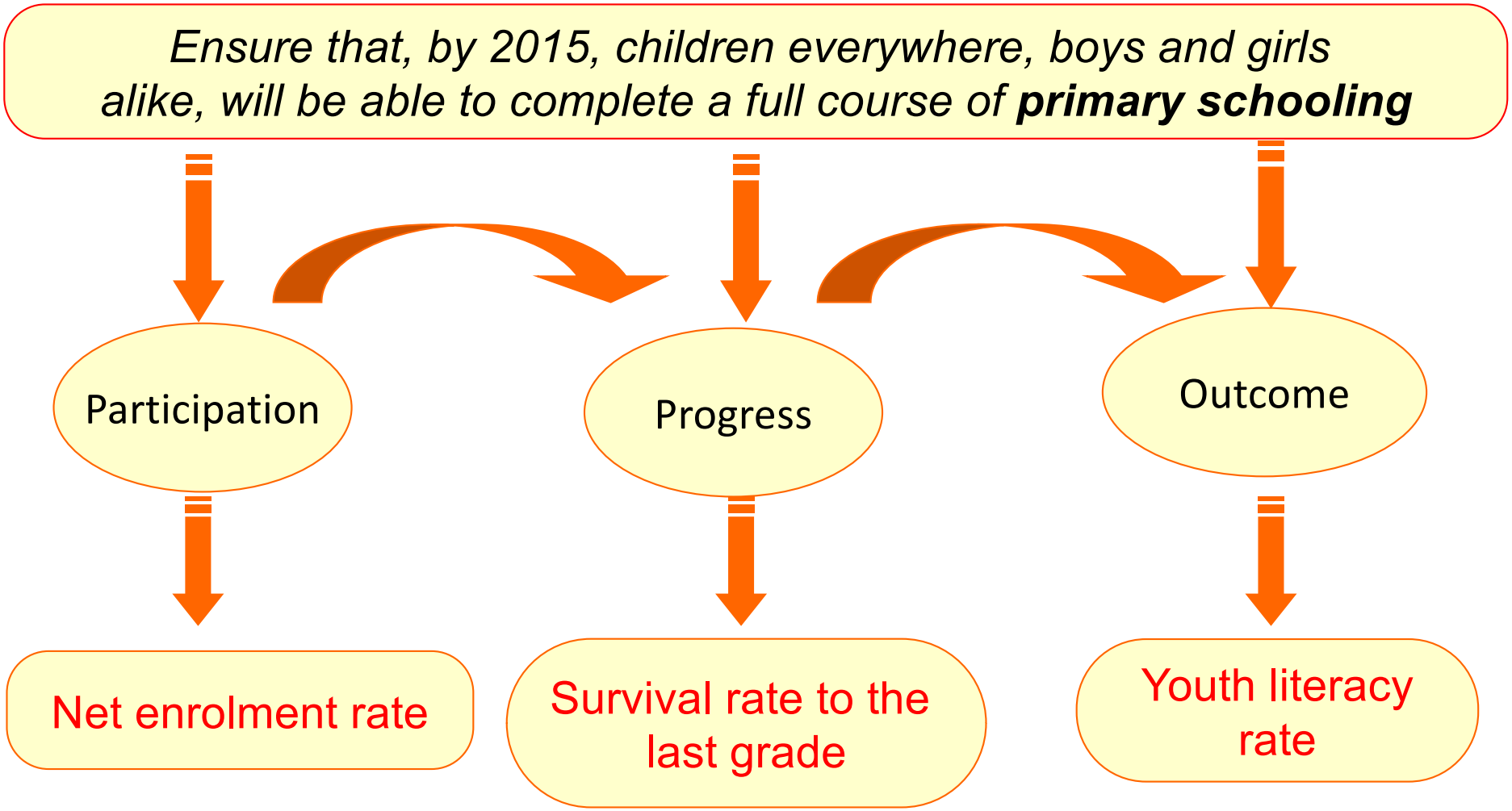
9. Ratio of girls to boys in primary, secondary and tertiary education
10. Ratio of literate females to males of 15-to-24-year-olds

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

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.

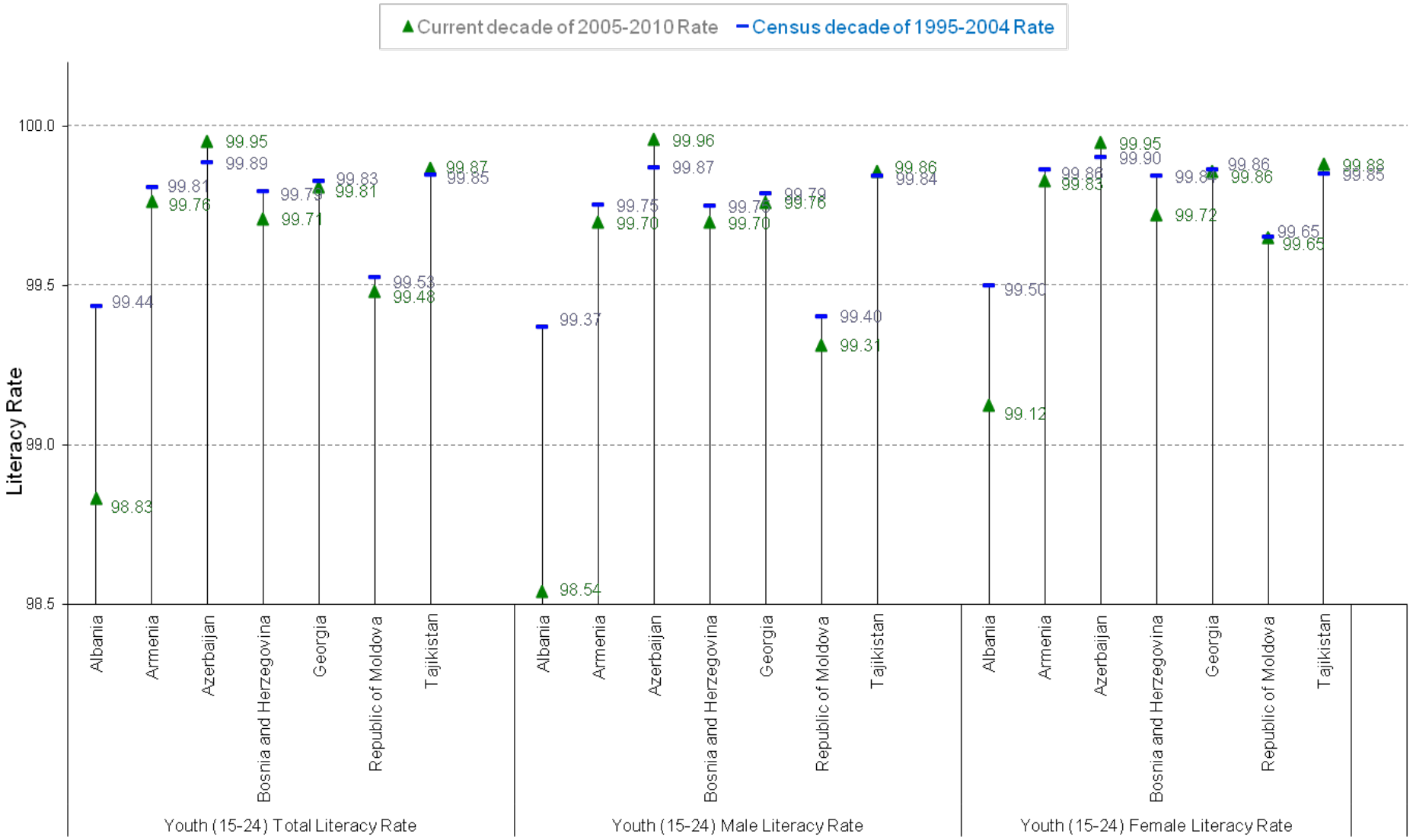
The youth literacy rate reflects the outcomes of the primary education system over the previous 10 years, and is often seen as a proxy measure of social progress and economic achievement

Youth literacy rate (15-24 years)

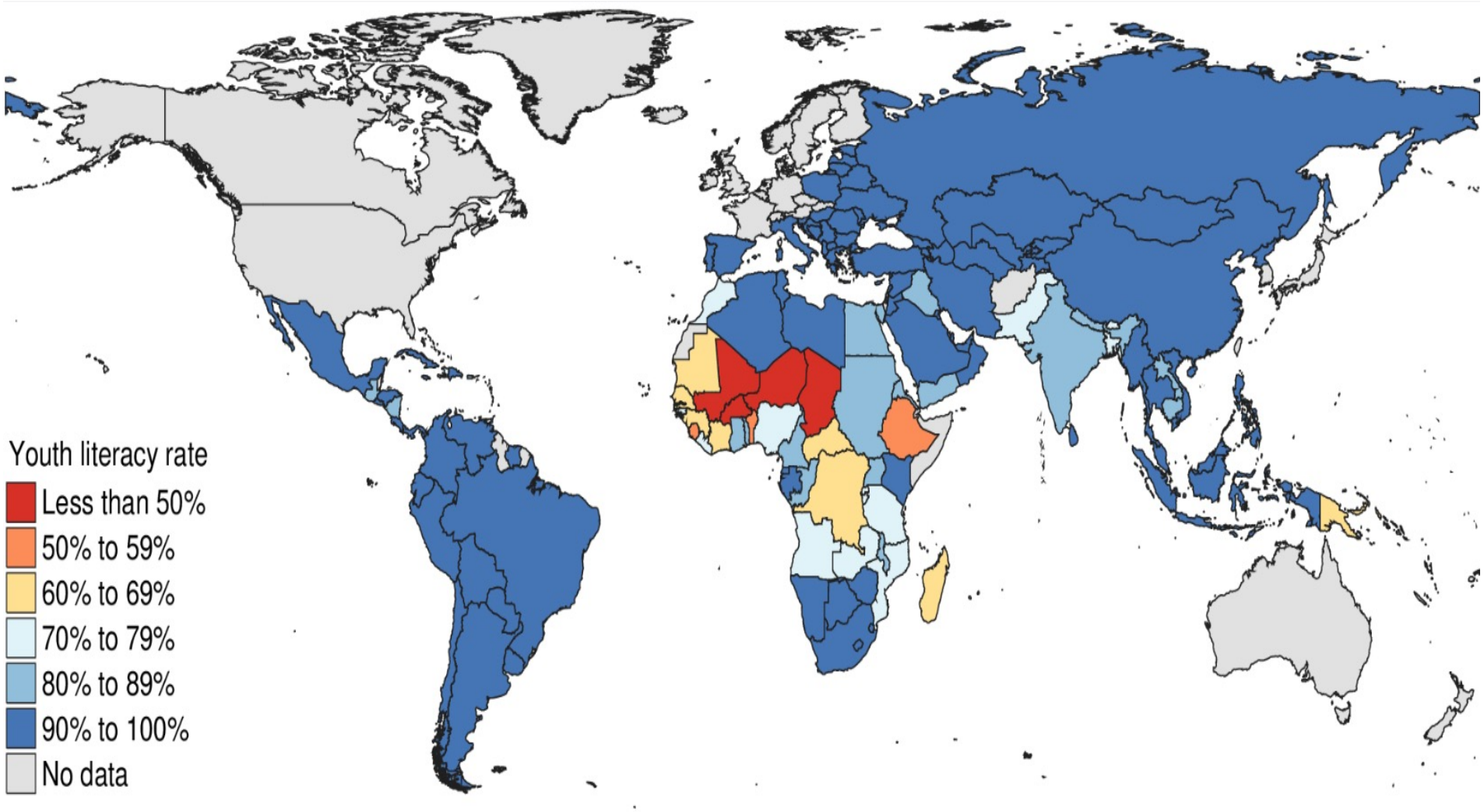
Interpretation: The indicator ranges from 0% (all youth are illiterate) to 100% (all youth are literate). Literacy rates below 100 per cent indicate the need to increase school participation and education quality.

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.

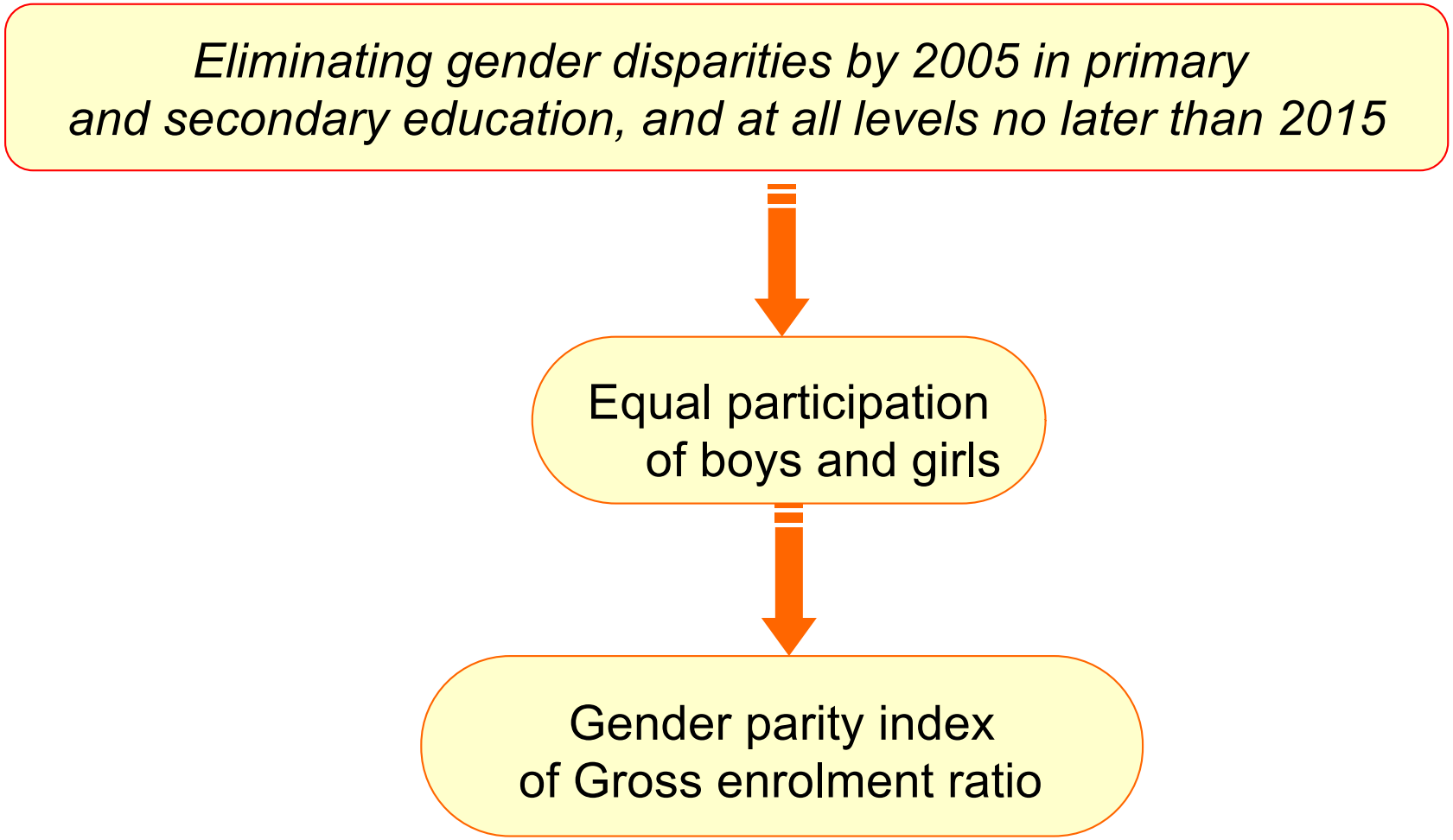
Practices for identifying literates and illiterates during actual census enumeration may also vary. Errors in literacy self-declaration can affect the reliability of the statistics.

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:



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:

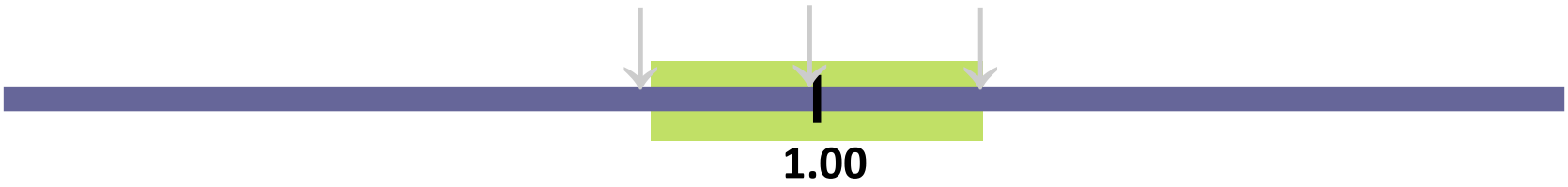
GPI_{GER} Gender parity index for Gross enrolment ratio

GER_{Female} Gross enrolment ratio for female

GER_{Male} Gross enrolment ratio for male

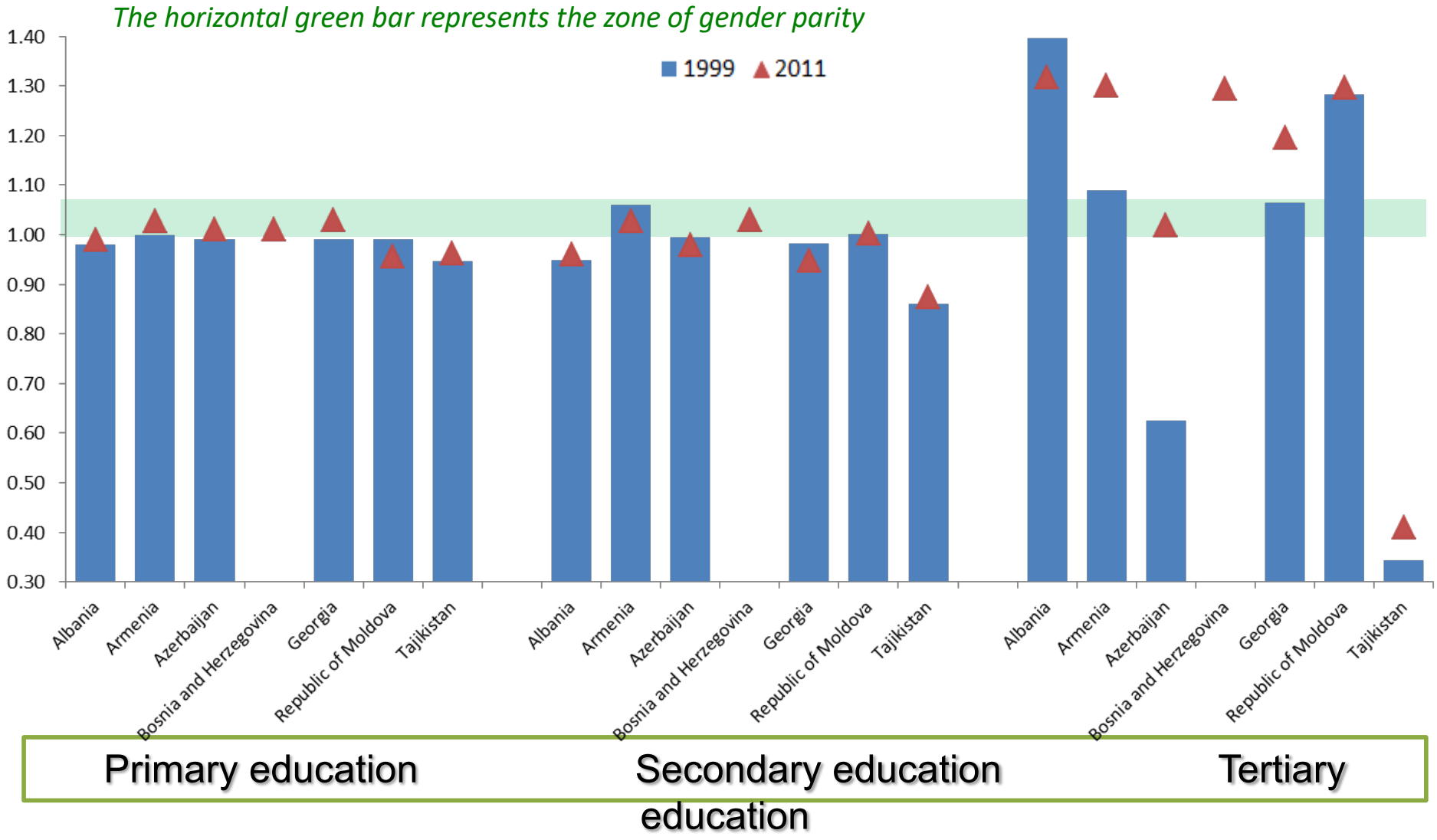
Measuring gender parity
GER for Primary education, 2011

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



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…

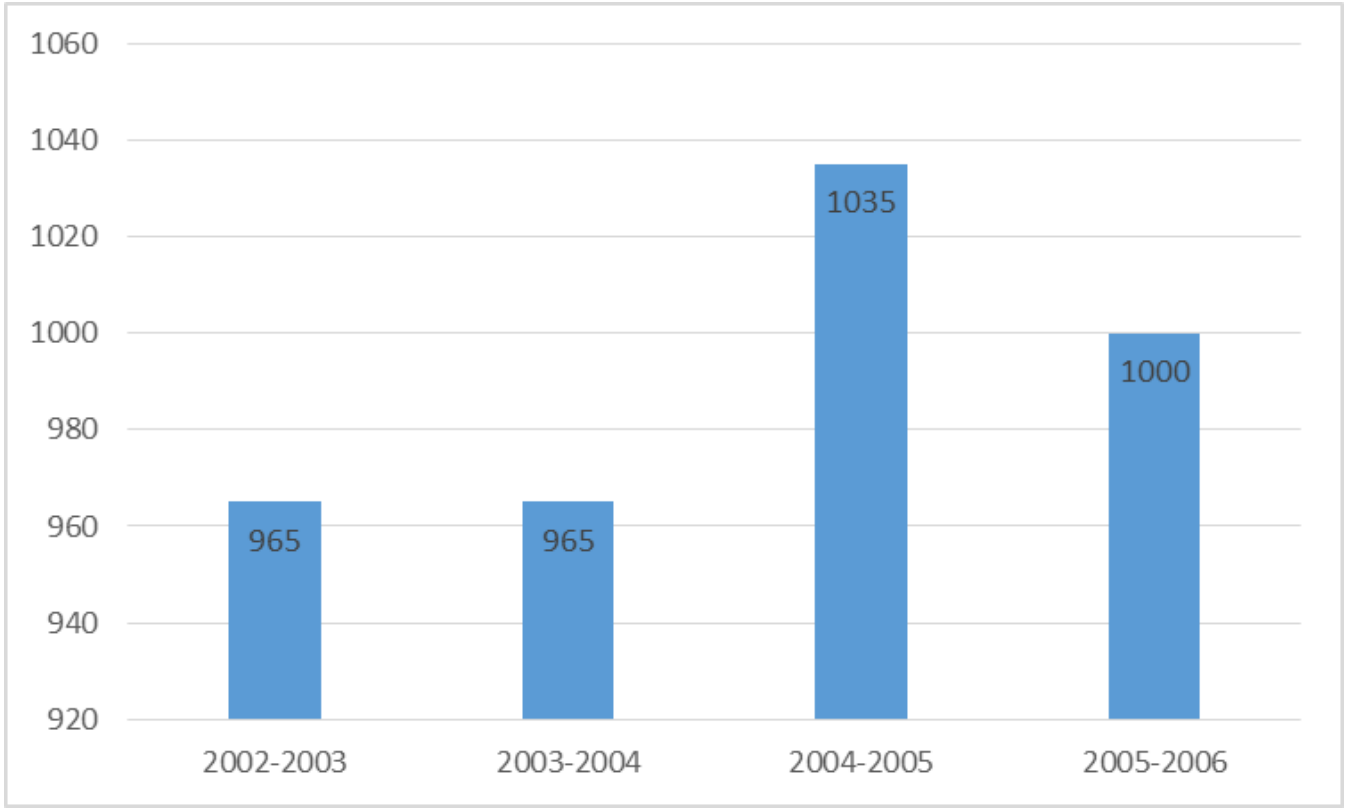
1. Weighted sum
2. Principal component/factor analysis
3. Regression
4. Structural equation modeling
5. ….

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

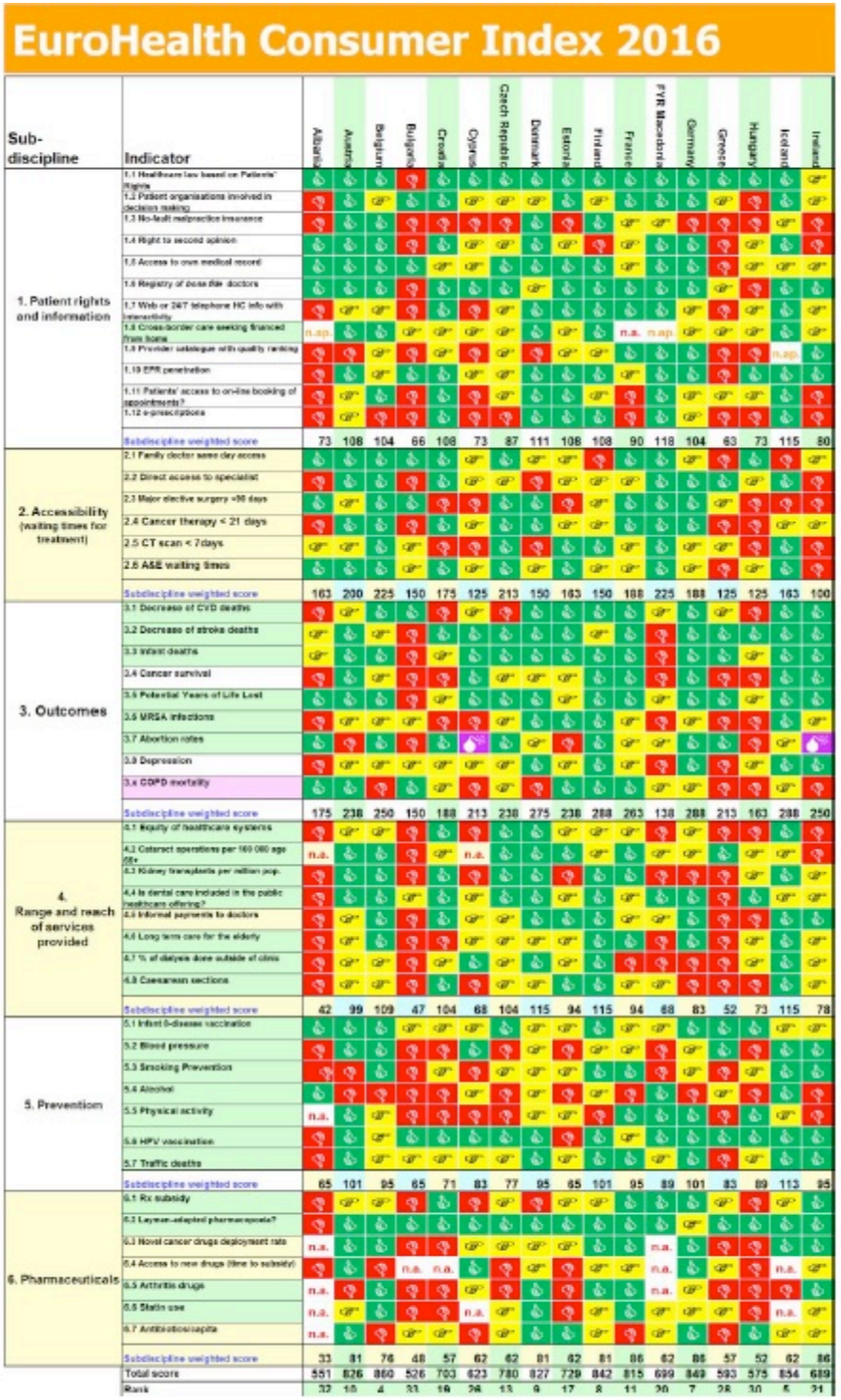
The aim has been to select a limited number of indicators, within a definite number of evaluation areas, which in combination can present a telling tale of how the healthcare consumer is being served by the respective systems.

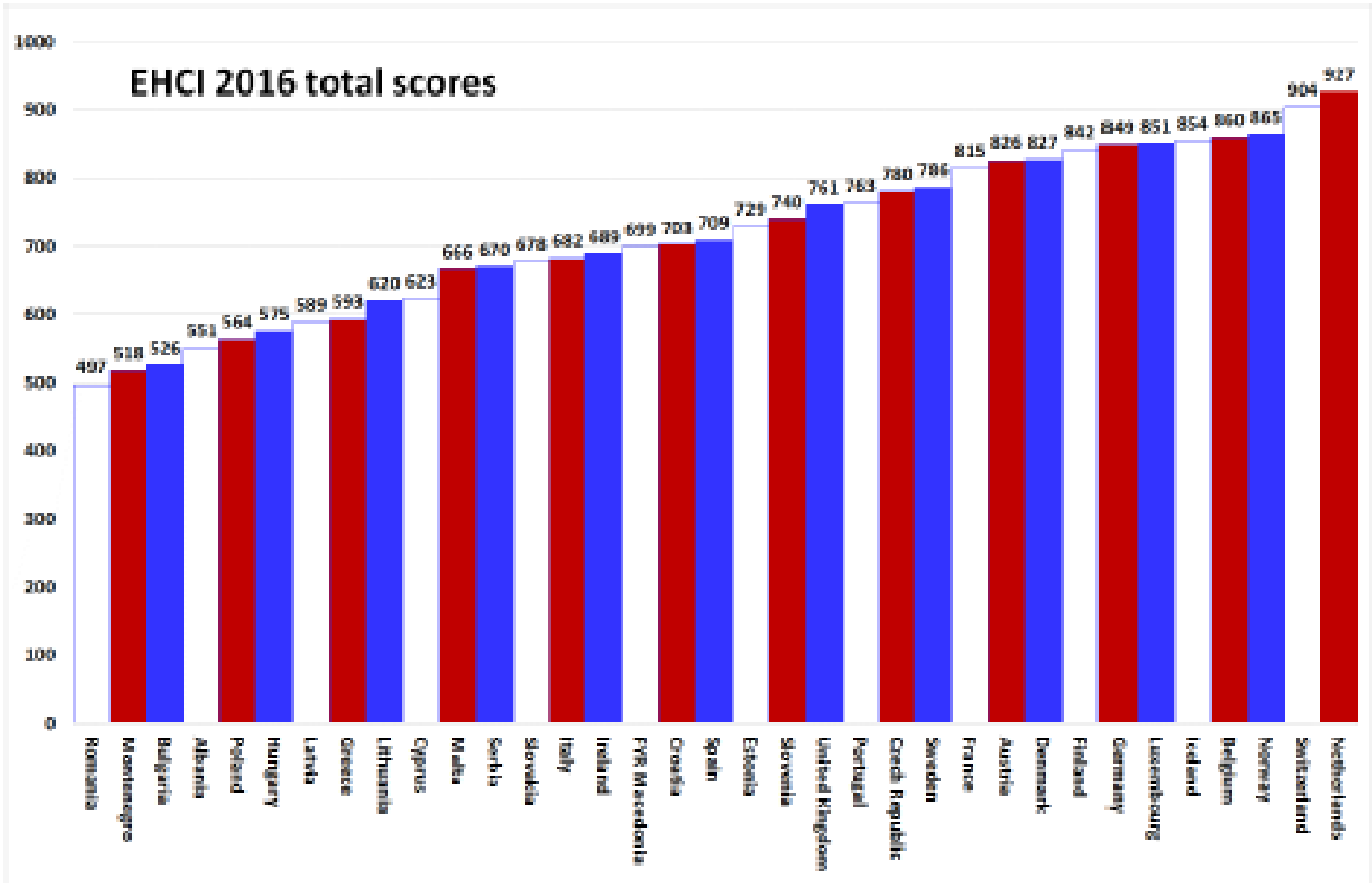
Comparing healthcare systems performance in 35 countries.

http://www.healthpowerhouse.com/

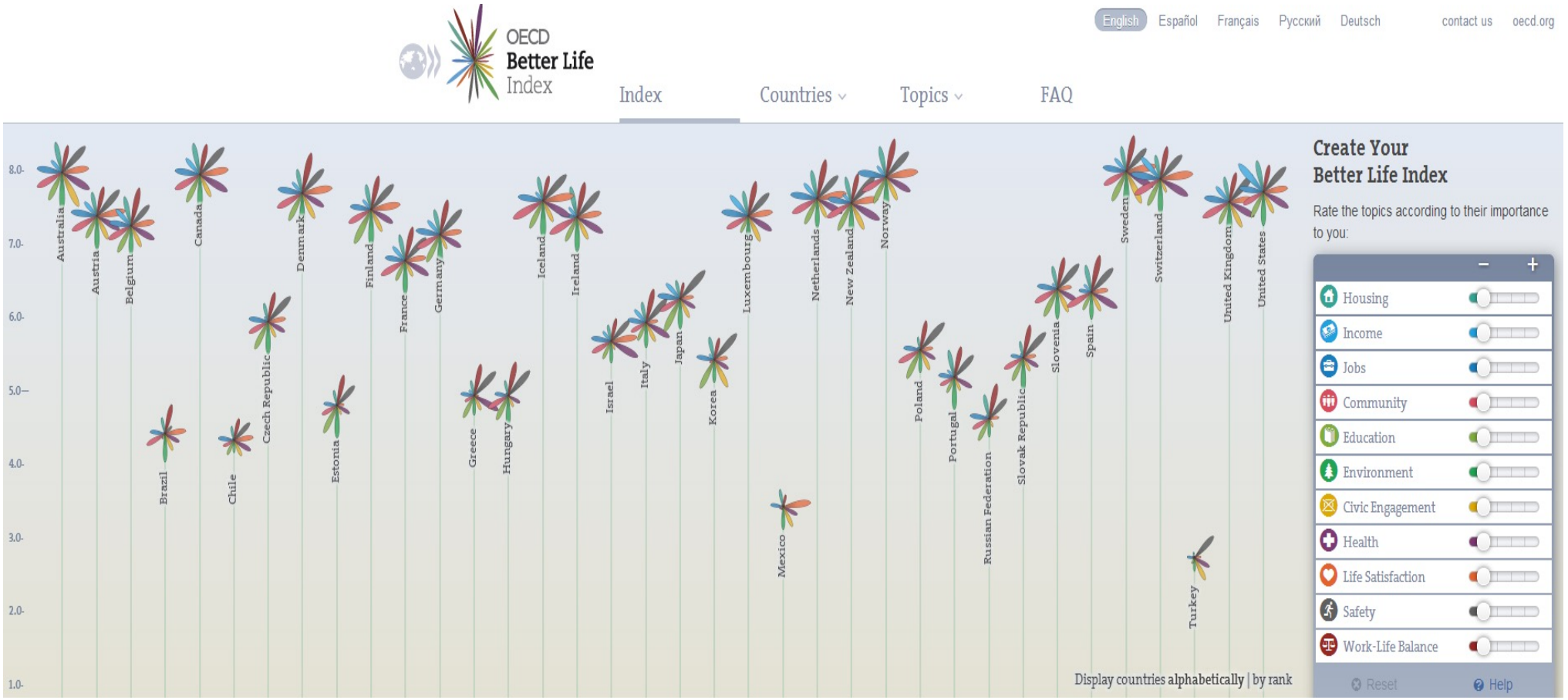
Sub-discipline	Relative weight ("All Green" score contribution to total maximum score of 1000)
1.Patient rights, information and e-Health	125
2.Accessibility (Waiting time for treatment)	225
3.Outcomes	300
4.Range and reach of services ("Generosity")	125
5.Prevention	125
6.Pharmaceuticals	100
Total sum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



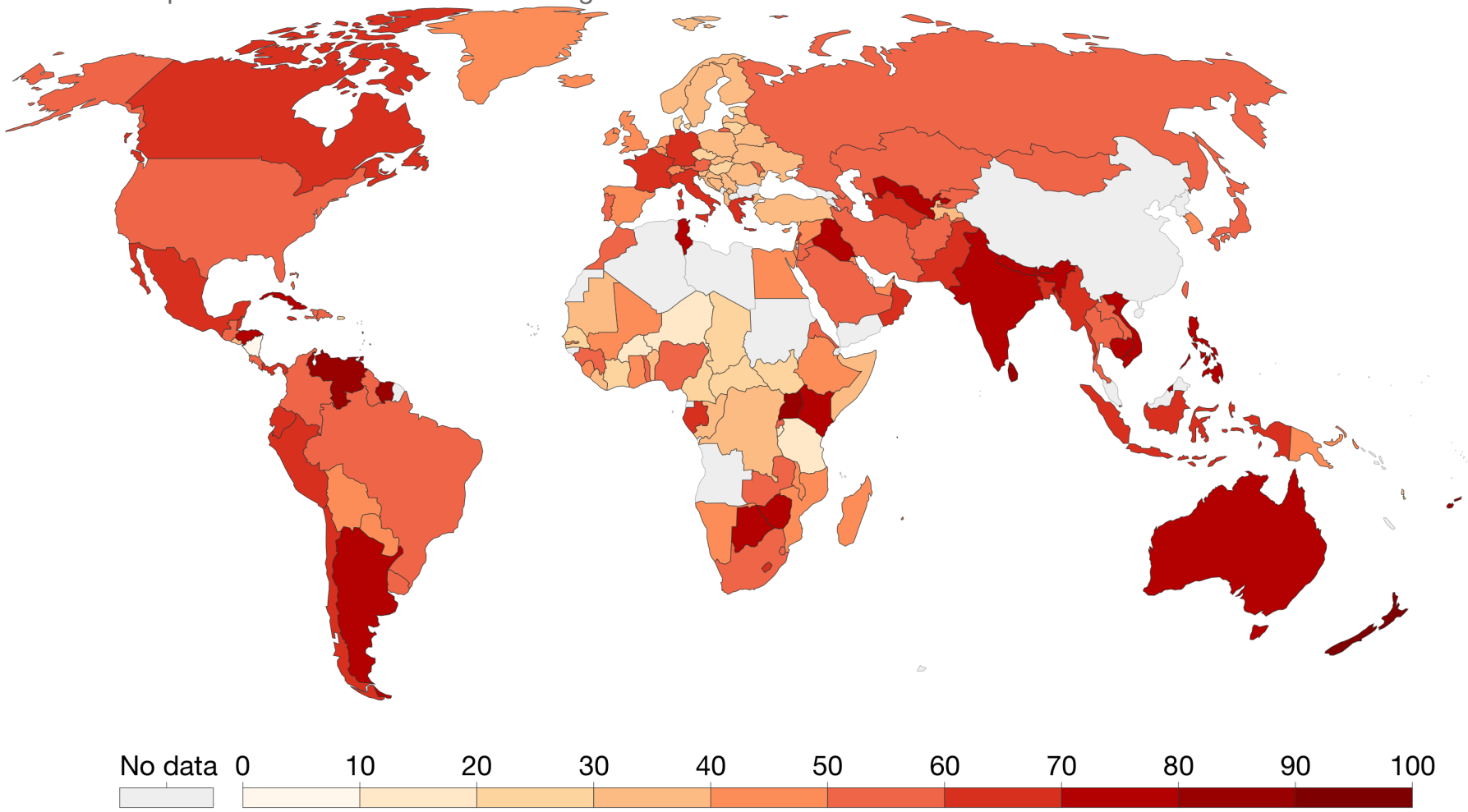

EHCI 2016 total scores
1000
900
800
700
600
500
400
300
200
100
0
497
518
526
551
564
575
589
593
620
623
666
670
678
682
689
699
703
709
729
740
761
763
780
786
815
826
827
843
849
851
854
860
865
904
927
Romania
Montenegro
Bulgaria
Albania
Poland
Hungary
Latvia
Greece
Lithuania
Cyprus
Malta
Serbia
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



COVID-19: Stringency Index

This is a composite measure based on nine response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest). If policies vary at the subnational level, the index is shown as the response level of the strictest sub-region.



Source: Hale, Angrist, Goldszmidt, Kira, Petherick, Phillips, Webster, Cameron-Blake, Hallas, Majumdar, and Tatlow (2021). "A global panel database of pandemic policies (Oxford COVID-19 Government ResponseTracker)." Nature Human Behaviour. – Last updated 6 September 2021, 10:50 (London time)

OurWorldInData.org/coronavirus • CC BY

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.

Practice

Data were collected about students in a statistics course. Three variables were recorded for each student:

1. number of siblings,
2. student height,
3. Whether or not the student had previously taken a statistics course.

Classify each of the variables.

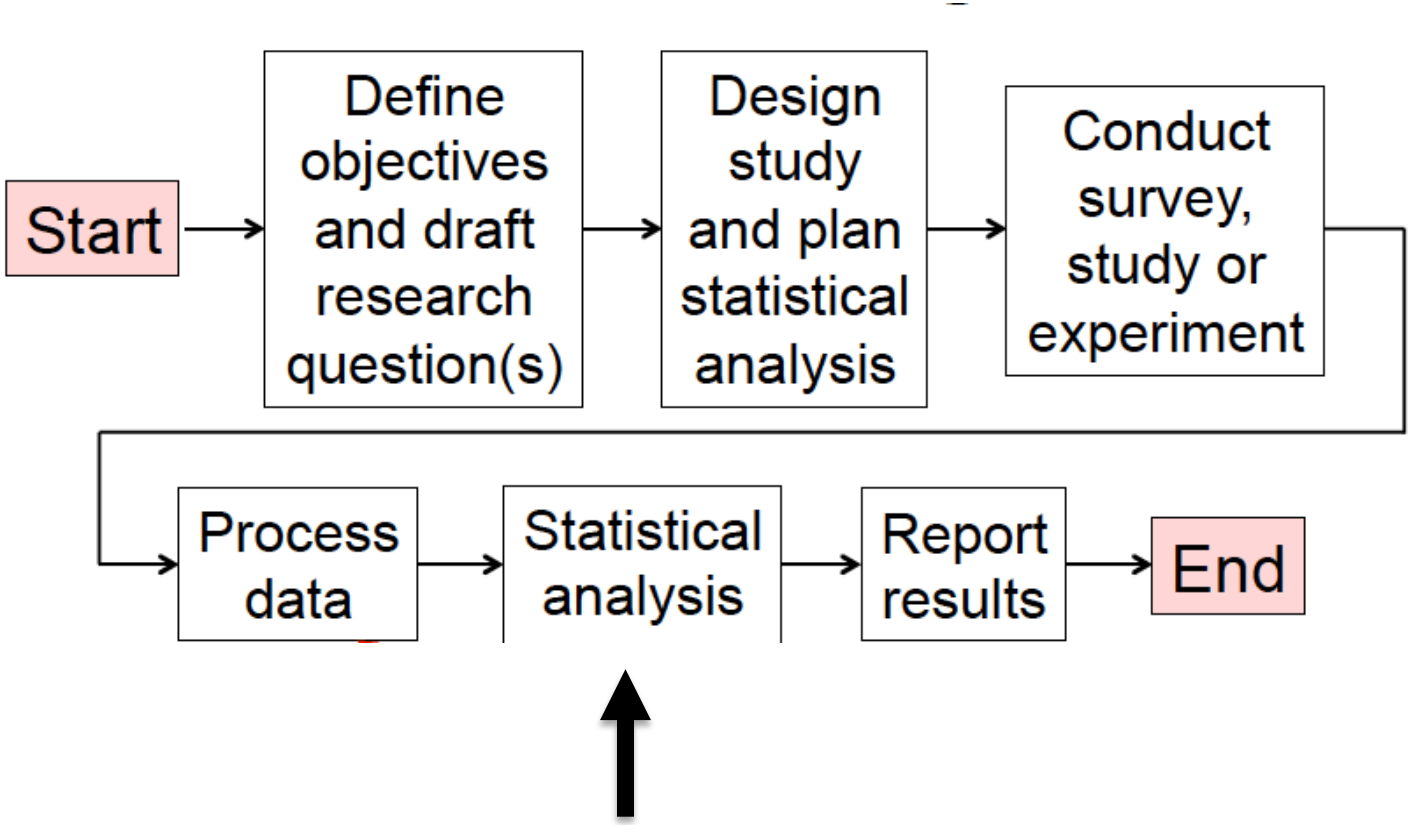
Smoking habits of UK residents. A survey was conducted to study the smoking habits of UK residents. Below is a data matrix displaying a portion of the data collected in this survey. \cig" stands for cigarettes, and \N/A" refers to a missing component of the data.

	sex	age	marital	grossIncome	smoke	amtWeekends	amtWeekdays
1	Female	42	Single	Under £2,600	Yes	12 cig/day	12 cig/day
2	Male	44	Single	£10,400 to £15,600	No	N/A	N/A
3	Male	53	Married	Above £36,400	Yes	6 cig/day	6 cig/day
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1691	Male	40	Single	£2,600 to £5,200	Yes	8 cig/day	8 cig/day

(a) What does each row of the data matrix represent?
(b) How many participants were included in the survey?
(c) Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

EXPLORING AND SUMMARIZING DATA

The research study process



Summarize Data to Reveal Meaningful Information, Patterns, and Relationships

How you do this depends on the nature of the data, e.g., nominal, ordinal, etc.

Analyze:

-One variable at a time (***Univariate Analysis***).
Example question: How many of the students in the freshman class are female?

-Two variables at a time (***Bivariate Analysis***)
Example question: Is there a relationship between the number of females in Computer Programming and their scores in Mathematics?

- Multiple variables at a time (***Multivariate Analysis)***
Example question: which are the determinants of voting behaviour?

Summarize Data to Reveal Meaningful Information, Patterns, and Relationships

How to summarize information:

Two stages:
reduce the data to a single relatively compact *table* (**frequency table**, crosstabulation, control table, etc.) or corresponding ***chart*** (bar graph, histogram, dot chart, box chart, scattergram, etc.)

reduce it further, if possible and depending on the nature of the variable, to one or several ***summary statistical measures*** (measures of central tendency, dispersion, correlation, etc.).

We first look at the process of summarizing *data* down to *frequency tables*, *bar graphs*, and *histograms*.
Then (univariate) measures of central tendency and dispersion.

Frequency tables

After collecting data, the first task for a researcher is to organize and simplify the data so that it is possible to get a general overview of the results. This is the goal of descriptive statistical techniques.
One method for simplifying and organizing data is to construct a **frequency table**.

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

n =1000 individuals (sample size) on the rows
p=number of variables on the colums

Notation

$X =$ variable

$n =$ sample size

$k =$ num of values of X

$x_i =$ value i of X

$n_i =$ absolute frequency of xi

$f_i =$ relative frequency of xi

X	n_i	$f_i = n_i / n$
x_1	n_1	f_1
x_2	n_2	f_2
$\vdots$	$\vdots$	$\vdots$
x_i	n_i	f_i
$\vdots$	$\vdots$	$\vdots$
x_k	n_k	f_k
Totale	n	1

tab sex

	n: absolute freq	f: relative freq (%)	cumulative freq (%)
sex of individual	Freq.	Percent	Cum.
male	**457**	**45.70**	**45.70**
female	**543**	**54.30**	**100.00**
Total	**1,000**	**100.00**	

tab pd005

marital status	Freq.	Percent	Cum.	
married	**584**	**58.40**	**58.40**	=f1
separated	**20**	**2.00**	**60.40**	=f1+f2
divorced	**17**	**1.70**	**62.10**	=f1+f2+f3
widowed	**80**	**8.00**	**70.10**	=……
never married	**299**	**29.90**	**100.00**	
Total	**1,000**	**100.00**		

tab ph001

health in general	Freq.	Percent	Cum.
very good	**335**	**33.67**	**33.67**
good	**388**	**38.99**	**72.66**
fair	**196**	**19.70**	**92.36**
bad	**60**	**6.03**	**98.39**
very bad	**16**	**1.61**	**100.00**
Total	**995**	**100.00**	

tab ph001, m

health in general	Freq.	Percent	Cum.
very good	**335**	**33.50**	**33.50**
good	**388**	**38.80**	**72.30**
fair	**196**	**19.60**	**91.90**
bad	**60**	**6.00**	**97.90**
very bad	**16**	**1.60**	**99.50**
.	**5**	**0.50**	**100.00**
Total	**1,000**	**100.00**	

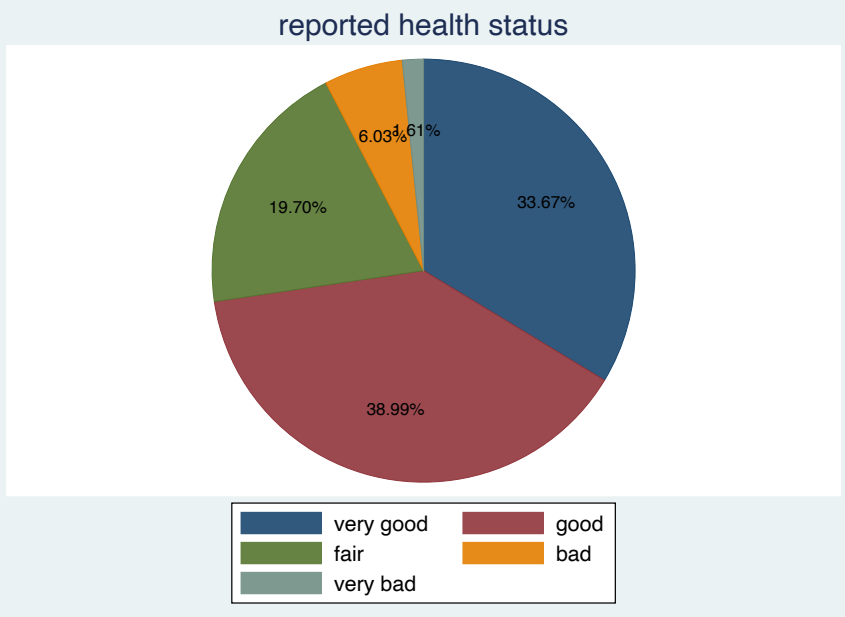
Age is a continuous variable, we need to create a new variable in which age is divided in classes!

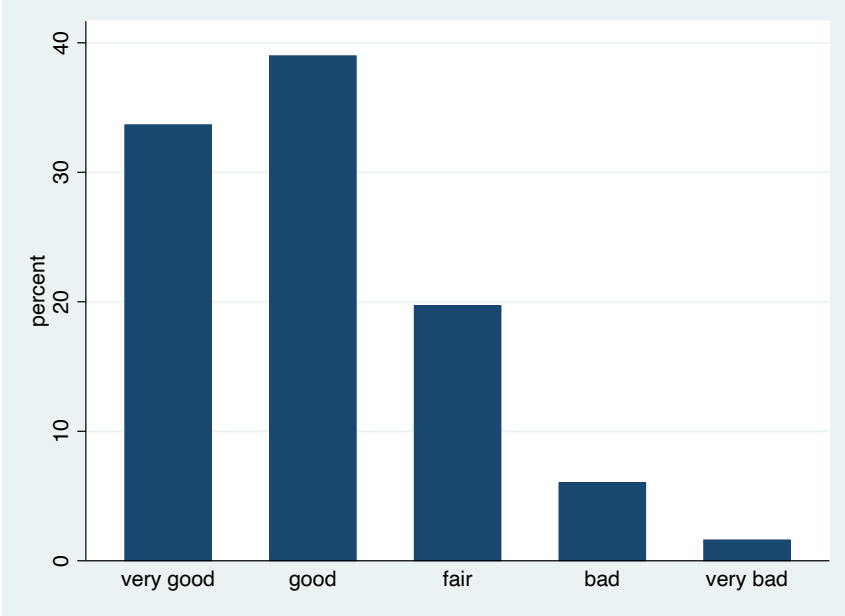
tab agegr

RECODE of age (age of individual)	Freq.	Percent	Cum.
up_to_25_years	**185**	**18.50**	**18.50**
from_26_to_35	**175**	**17.50**	**36.00**
from_36_to_45	**204**	**20.40**	**56.40**
from_46_to_55	**161**	**16.10**	**72.50**
from_56_to_65	**111**	**11.10**	**83.60**
from_66_to_75	**91**	**9.10**	**92.70**
over_76_years	**73**	**7.30**	**100.00**
Total	**1,000**	**100.00**	

Pie Charts and Bar Charts

health in general	Freq.	Percent	Cum.
very good	335	33.67	33.67
good	388	38.99	72.66
fair	196	19.70	92.36
bad	60	6.03	98.39
very bad	16	1.61	100.00
Total	995	100.00	





for qualitative variables!

the frequencies are on the vertical axis and are proportional to the heights of the bars

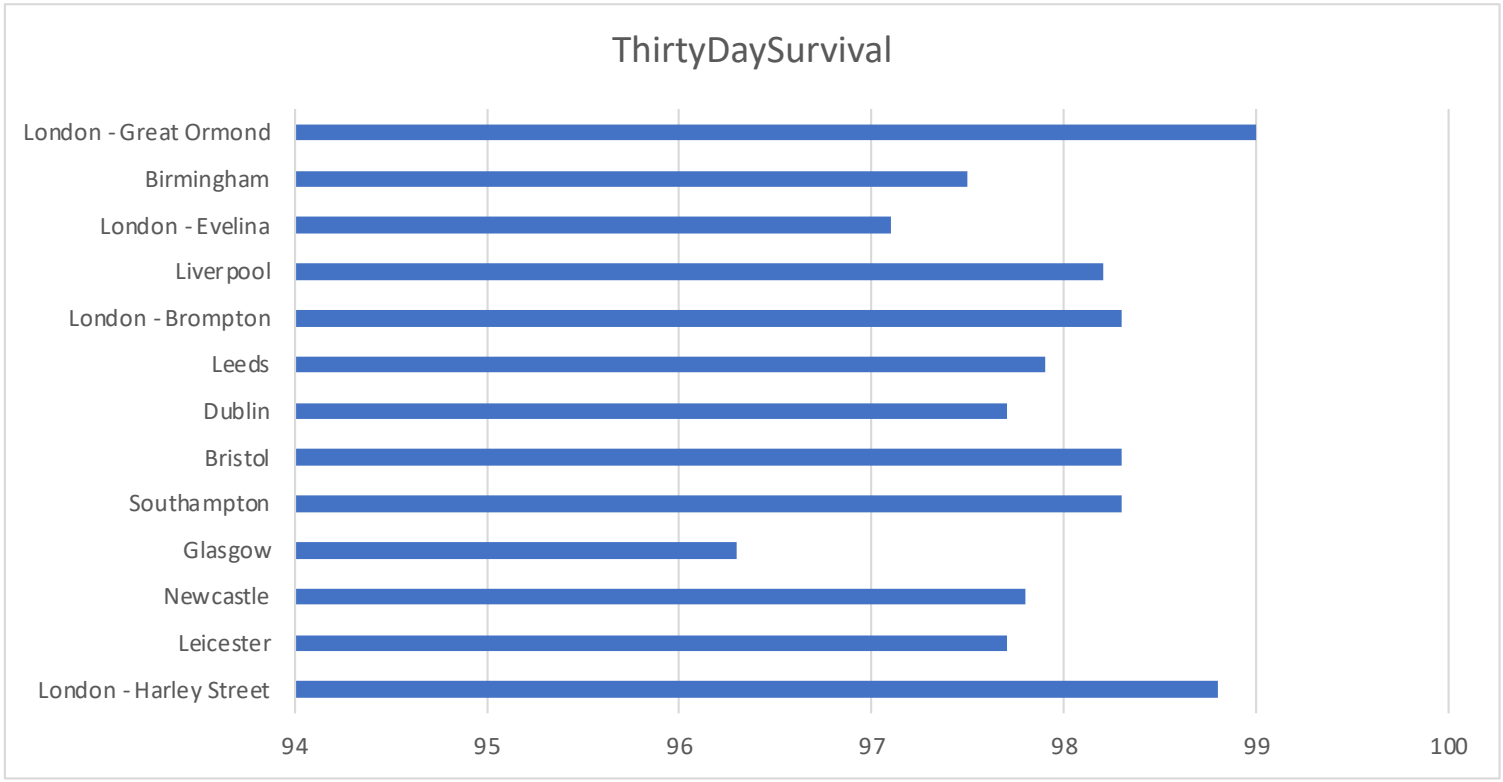
The Bristol heart scandal occurred in England during the 1990s

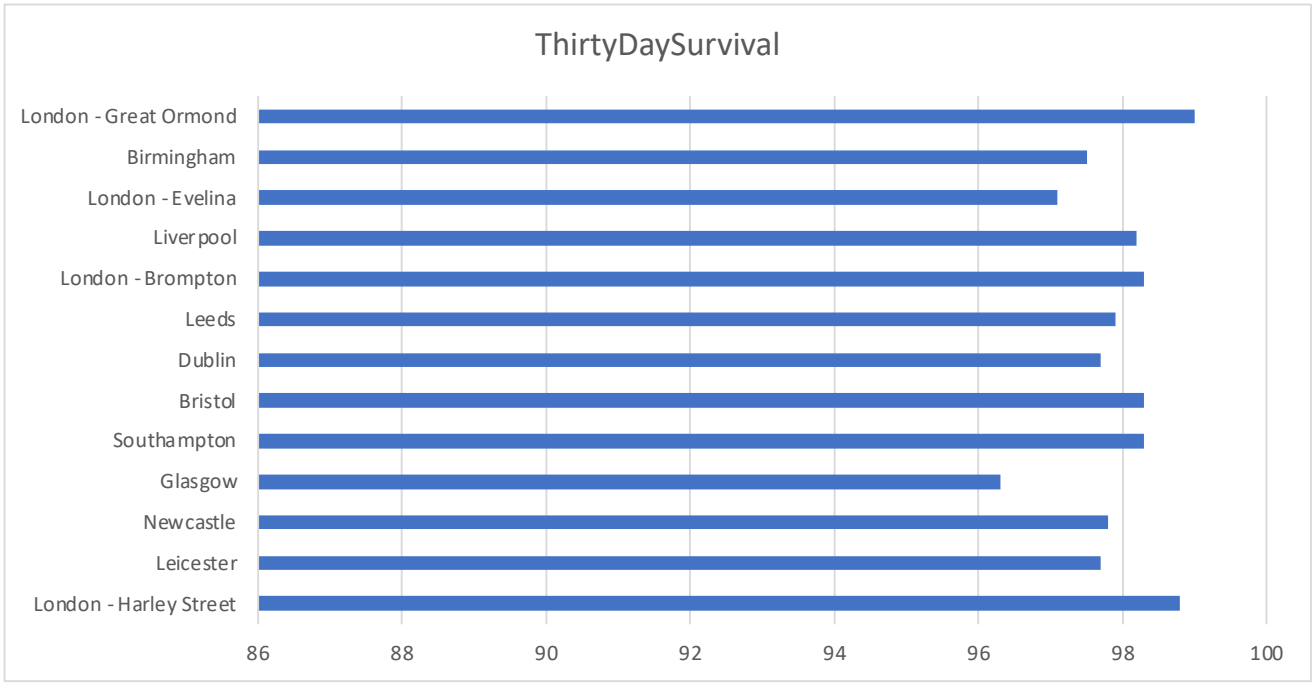
For children aged less than 90 days had a mortality nearly four times higher (4.0) than that elsewhere in England

2012-2015

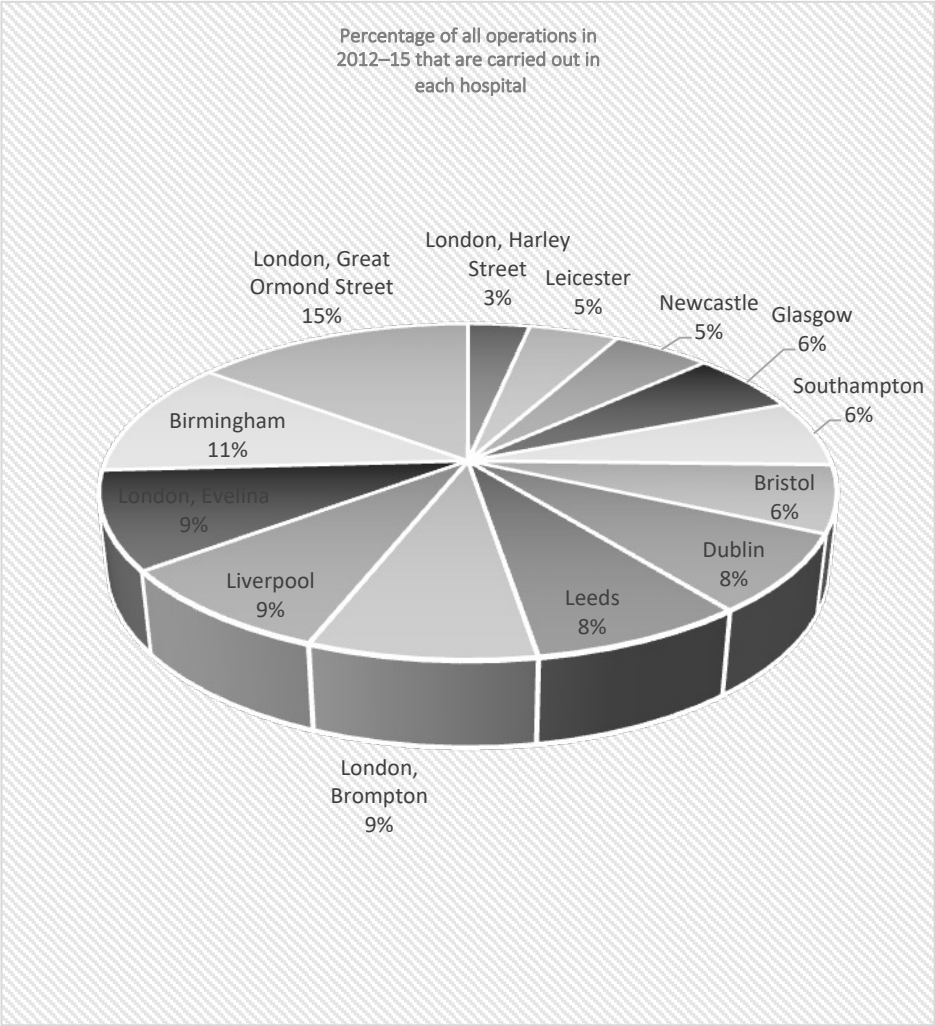
Hospital	Operations	Survivors	Deaths	ThirtyDaySurvival	PercentageDying
London - Harley Street	418	413	5	98.8	1.2
Leicester	607	593	14	97.7	2.3
Newcastle	668	653	15	97.8	2.2
Glasgow	760	733	27	96.3	3.7
Southampton	829	815	14	98.3	1.7
Bristol	835	821	14	98.3	1.7
Dublin	983	960	23	97.7	2.3
Leeds	1038	1016	22	97.9	2.1
London - Brompton	1094	1075	19	98.3	1.7
Liverpool	1132	1112	20	98.2	1.8
London - Evelina	1220	1185	35	97.1	2.9
Birmingham	1457	1421	36	97.5	2.5
London - Great Ormond	1892	1873	19	99	1

https://www.childrensheartsurgery.info


ThirtyDaySurvival
London - Great Ormond
Birmingham
London - Evelina
Liverpool
London - Brompton
Leeds
Dublin
Bristol
Southampton
Glasgow
Newcastle
Leicester
London - Harley Street
94
95
96
97
98
99
100


ThirtyDaySurvival
London - Great Ormond
Birmingham
London - Evelina
Liverpool
London - Brompton
Leeds
Dublin
Bristol
Southampton
Glasgow
Newcastle
Leicester
London - Harley Street
86
88
90
92
94
96
98
100

	% ops carried out in each hospital	Number of operations	Number of deaths	Number of survivors
London, Harley Street	3.2	418	5	413
Leicester	4.7	607	14	593
Newcastle	5.2	668	15	653
Glasgow	5.9	760	27	733
Southampton	6.4	829	14	815
Bristol	6.5	835	14	821
Dublin	7.6	983	23	960
Leeds	8.0	1038	22	1016
London, Brompton	8.5	1094	19	1075
Liverpool	8.8	1132	20	1112
London, Evelina	9.4	1220	35	1185
Birmingham	11.3	1457	36	1421
London, Great Ormond Street	14.6	1892	19	1873
	100.0	12933	263	12670



The proportion of all child heart operations being carried out in each hospital, displayed in a 3D pie chart from Excel. This deeply unpleasant chart makes categories near the front look bigger, and so makes it impossible to make visual comparisons between hospitals.

Two-way frequency tables – bivariate analysis

Also known as *contingency tables*, crosstabs help you to analyze the relationship between two or more categorical variables

tab ph001 sex, row col

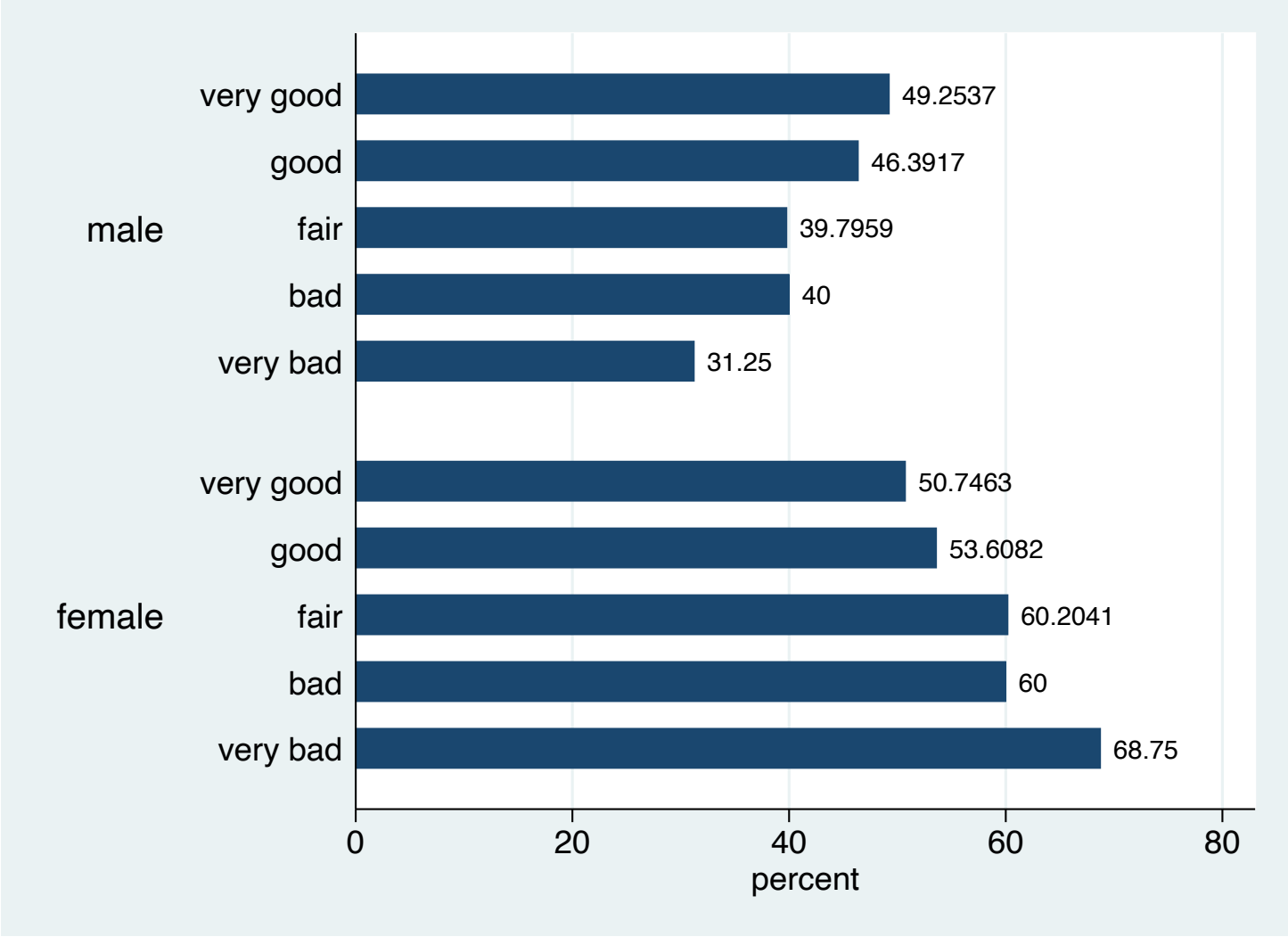
health in general	sex of individual male	female	Total
very good	165	170	335
	49.25	50.75	100.00
	36.50	31.31	33.67
good	180	208	388
	46.39	53.61	100.00
	39.82	38.31	38.99
fair	78	118	196
	39.80	60.20	100.00
	17.26	21.73	19.70
bad	24	36	60
	40.00	60.00	100.00
	5.31	6.63	6.03
very bad	5	11	16
	31.25	68.75	100.00
	1.11	2.03	1.61
Total	452	543	995
	45.43	54.57	100.00
	100.00	100.00	100.00

The first value in a cell: the number of observations for each xtab. In this case, 165 respondents are 'male' and reported to be in a 'very good' health status, 170 are 'female' and and reported to be in a 'very good' status.

The second value in a cell: row percentages for the first variable in the xtab. Out of those who report to be in 'very good' health status, 49.25% are males and 50.75% are females.

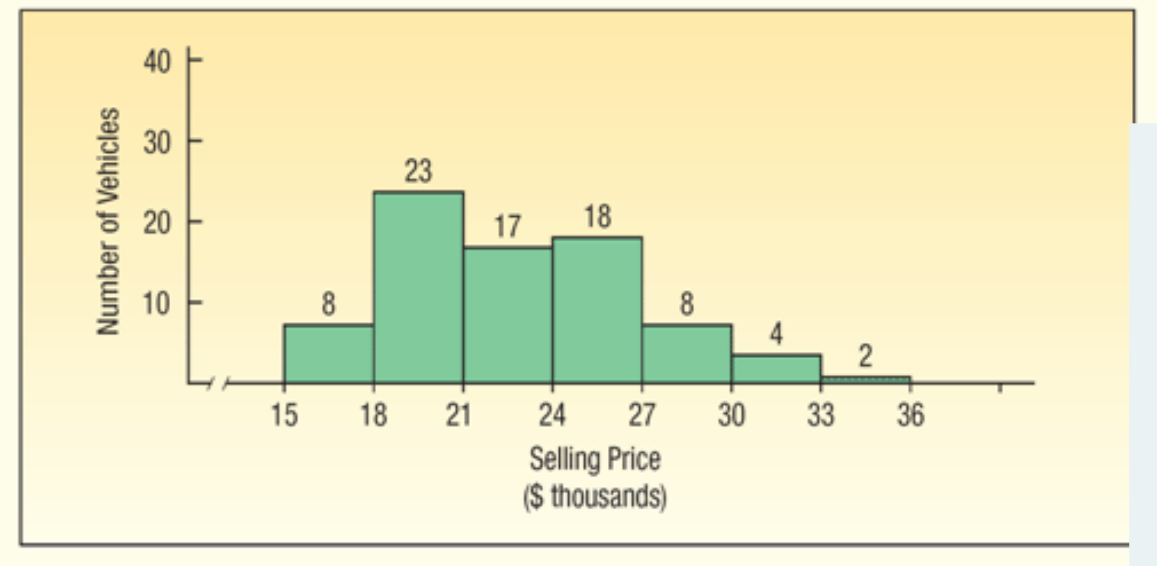
The third value in a cell: column percentages for the second variable in the xtab. Among males, 36.50% report a 'very good' health while 31.31% of females report a 'very good' health

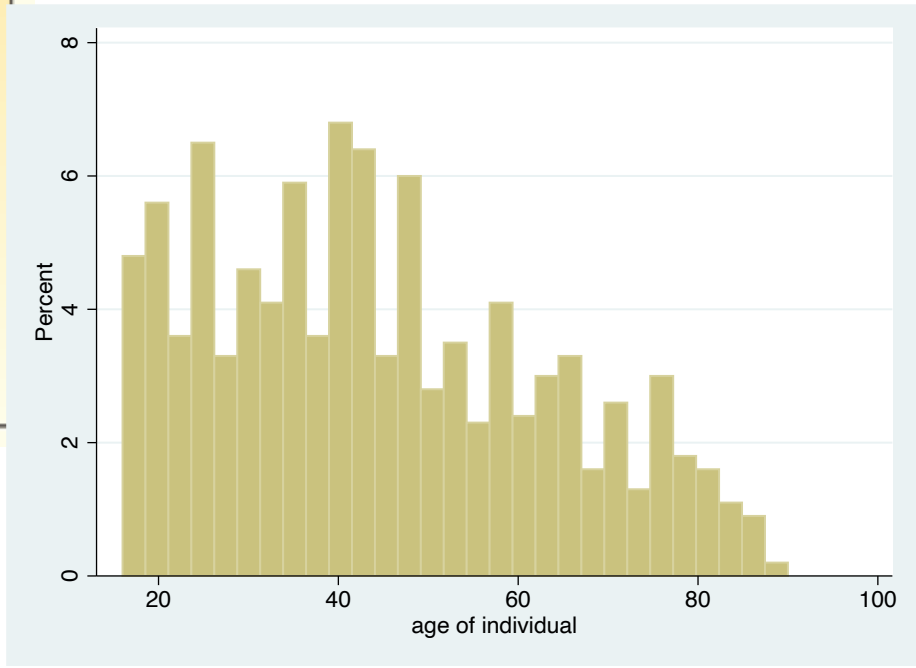
catplot ph001 sex, percent(ph001) blabel(bar)



Histogram

A graph in which the classes are marked on the horizontal axis and the class frequencies on the vertical axis. The class frequencies are represented by the heights of the bars and the bars are drawn adjacent to each other.



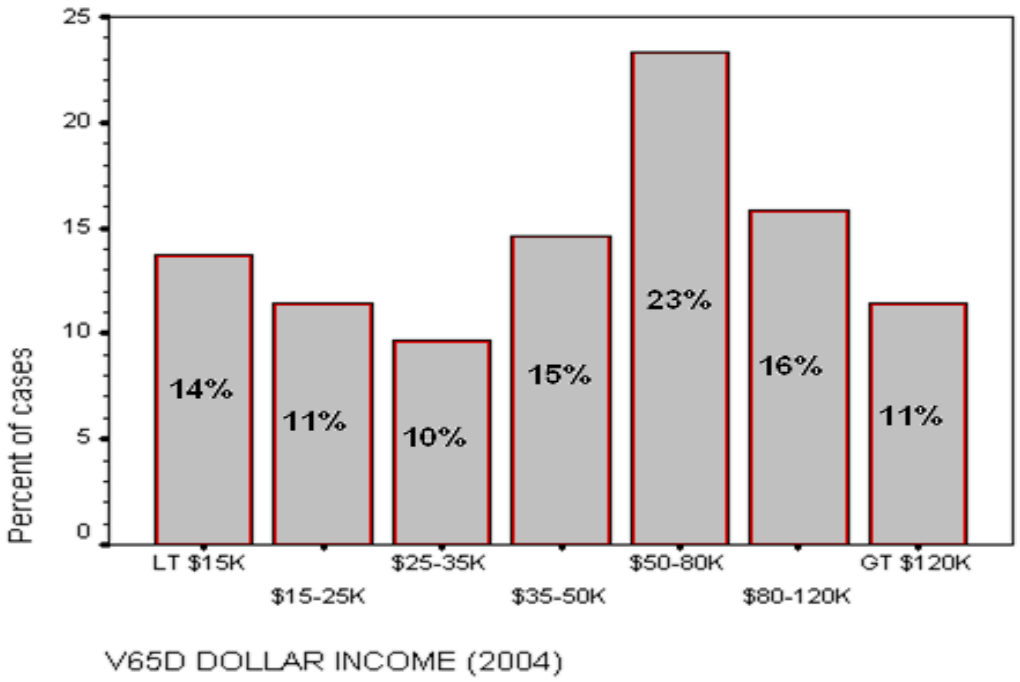


Frequency table for income

		Freq.	Percent	Valid %	Cum.
%					
Less than $15,000	145	12.0	13.7	13.7	
$15,000 to $25,000	121	10.0	11.4	25.2	
$25,000 to $35,000	102	8.4	9.7	34.9	
$35,000 to $50,000	154	12.7	14.6	49.5	
$50,000 to $80,000	246	20.3	23.3	72.8	
$80,000 to $120,000	167	13.8	15.8	88.6	
More than $120,000	120	9.9	11.4	100.0	
Total	1055	87.0	100.0		
Missing NA	157	13.0			
Total	1212	100.0			

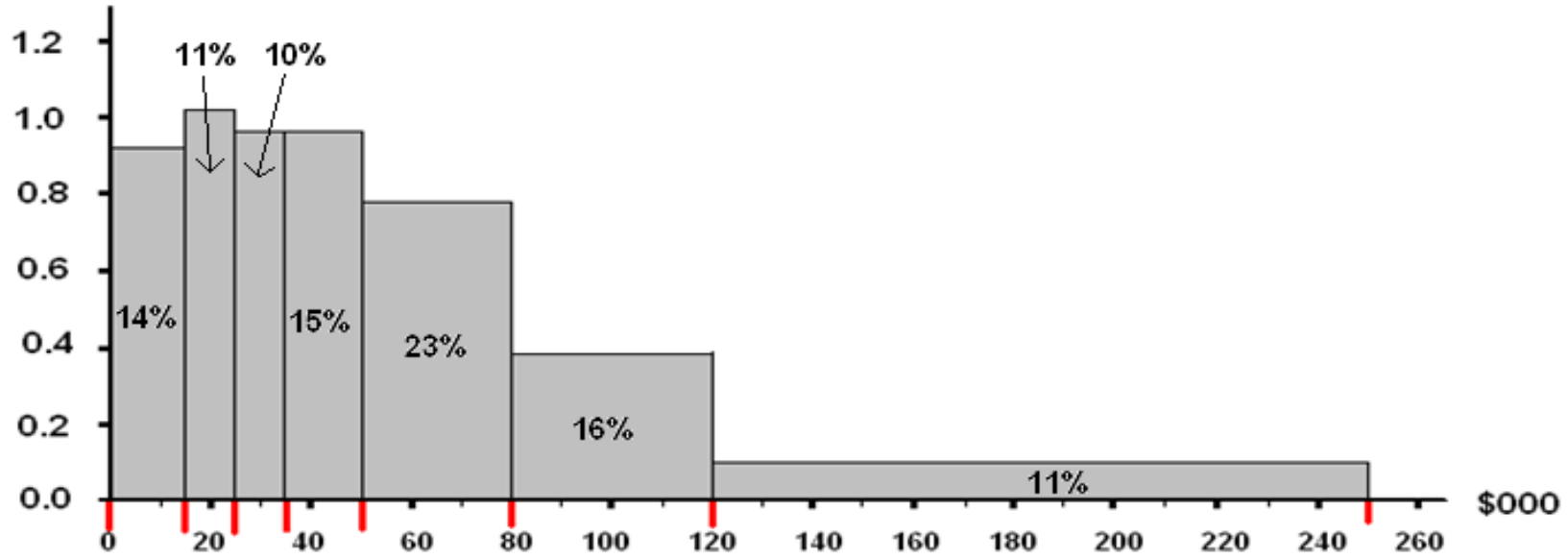
example from SPSS

Bar Chart



The bar chart appears to display a distribution of income that is approxi-mately "uniform" – that is, all bars are approximately the same height, except for a distinctive peak (or "mode") in the third highest income category.
Indeed, the impression the bar graph conveys to the eye is that there are more well-off than not-so-well-off people.
However, this impression is quite misleading, as you can begin to under-stand when you look more closely at the income class intervals and notice that they are not of equal width.
Here is the histogram of the same INCOME data =>

Histogram



The fundamental difference between a bar graph and a histogram:
in a bar graph, *frequency is represented by the height of the bars* (all of which have the same width);
in a histogram, *frequency is represented by the area of the "bars"* (which may have different widths, reflecting the different "widths" of the class intervals).
With equal class intervals, the area of a bar depends only on its height, so Histogram ≈ Frequency Bar Chart
But with unequal class intervals, the area of a bar depends on both its height and its width, so Histogram ≠ Frequency Bar Chart

Histogram (cont.)

the *area* [*not* height] of each rectangle is proportional to the frequency associated with that class interval.
How tall should each rectangle be?
The width of each rectangle is the width of the class interval, and you should remember that:

Area = *Height* × *Width* so *Height* = *Area* / *Width*

Since *Area* here represents *Frequency*, we have the formula:

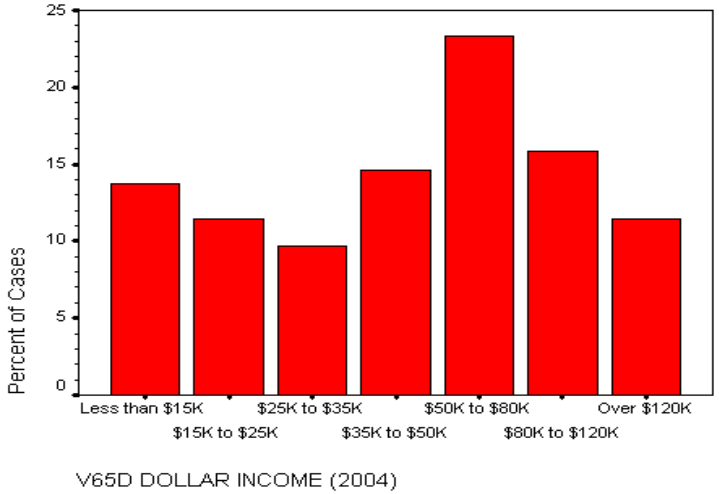
Height = *Frequency* / *Width*,
where *Width* is the width of the class interval.

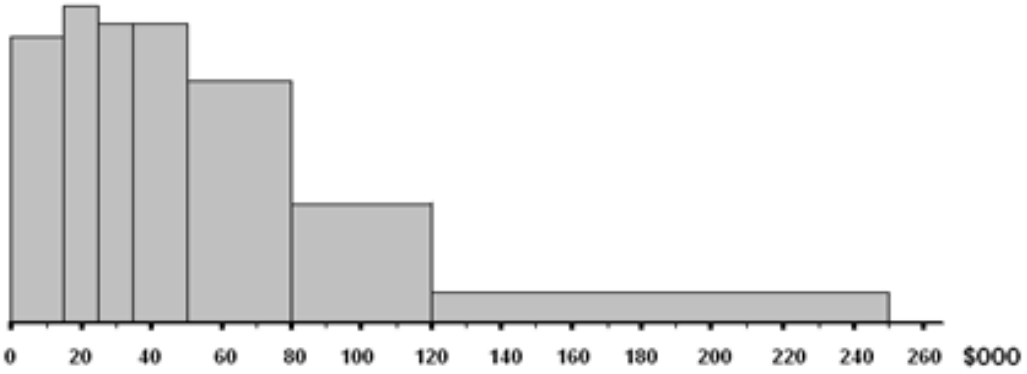
Histogram (cont.)

Now we can calculate the following (relative) heights of all the bars/rectangles. (Since only relative magnitudes matter, we can ignore the $000 = $K in INCOME values.)

Class Interval	*Width*	*Freq.*	Freq/Width	*Height*
0-15	15	13.7	13.7 / 15	= 0.913
15-25	10	11.4	11.4 / 10	= 1.140
25-35	10	9.7	9.7 / 10	= 0.970
35-50	15	14.6	14.6 / 15	= 0.973
50-80	30	23.3	23.3 / 30	= 0.777
80-120	40	15.8	15.8 / 40	= 0.395
120-250	130	11.4	11.4/130	= 0.088

Now we can draw the appropriate scale on the vertical axis.
The tallest rectangle has a height of about 1.14





Given that height in a histogram does *not* represent frequency, what does it represent?

The answer is that height represents *density* — that is, *how densly observed values of cases are "packed into" each class interval*.

Note that the class interval $50-80K includes about twice as many cases (23.3%) as the interval $15-25K (11.4%).

This fact is reflected in the *bar graph* in Figure 1 by the fact that the bar on the $50-80K interval is about *twice as high* as the bar over the $15-25K interval.

It is reflected in the histogram in Figure 2 by the fact that the "bar" (rectangle) on the $50-80K interval has about *twice the area* of the bar on the $15-25K interval.

But the 23.3% of the cases in the $50-80K interval are spread over an income interval that is three times as wide as the interval into which the 11.4% of the cases in the $15-25K interval are packed, so the height of the former (wide) bar is actually less than the height of the latter (thin) bar.

Histogram vs. Frequency Bar Graph

If *all class intervals all have the same width*, then the histogram is essentially no different from a bar chart

Otherwise (i.e., if the class intervals are not all of equal width), a bar chart and a histogram of the same data may look quite different,
in which event the bar chart presents a misleading picture of the data,
while the histogram presents a more accurate picture.
The histogram, unlike the bar chart takes account of the *interval* property of the variable.

The INCOME histogram was based on a small number of (rather wide) class intervals and a modest number of cases (n = 1212).

Remember that INCOME is continuous.

Suppose we have INCOME data that is recorded very *precisely*, e.g., to the near dollar or even cent.

Suppose also we have a huge --- approaching infinite --- number of cases.

We could then refine INCOME into narrower and narrow (i.e., *more precise*) class intervals, redrawing the histogram accordingly.

If we pushed this process to the limit, we would end up with what would be an essentially *continuous* (and probably fairly smooth) *density curve*

This is illustrated in the following series of charts using a symmetric ("normal') distribution and equal class intervals that get narrower and narrower (i.e., more precise).