ASM1:
Introduction to Statistics

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

About this course…

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

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

Encouraging critical thinking!

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

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

About this course…

The knowledge and skills you will gain in this course highly depend on the level of your participation in class learning activities.

The texts are not the class! A great part of the work of the class comes from:

- our discussions in class. You are encourage to actively take part to the class with questions and observations;

- and from practicing **by yourself** with the data .

Final Assignment

Students will be asked to work with assigned dataset to answer to specific research questions.

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

In practice:

- Divide in groups (max 2 students per group)
- Perform statistical analysis to answer to the research question that will be assigned to you by using the corresponding dataset.
- Present your findings in about 10 slides in a scientific manner

Outline

1. What is statistics
2. Types of studies
3. Introduction to sampling theory
4. Sources of error in statistical data
5. Probability and population

1 - What is statistics

Statistics is…

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

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

Statistics in the context of a general process of investigation:

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

That is, statistics has three primary components:

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

A well-written statistical question refers to:

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

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

News Article

Map Shows Best Cities To Live in Right Now

PUBLISHED
JUN 22, 2025 AT 08:15 AM EDT

Economist Intelligence Unit's (EIU) Global Liveability Index, evaluates cities across 30 indicators grouped into five categories, which include stability, health care, culture and environment, education, and infrastructure.

Servizio | Lavoro

Istat: a settembre +67mila occupati rispetto ad agosto. Tasso di disoccupazione sale al 6,1%

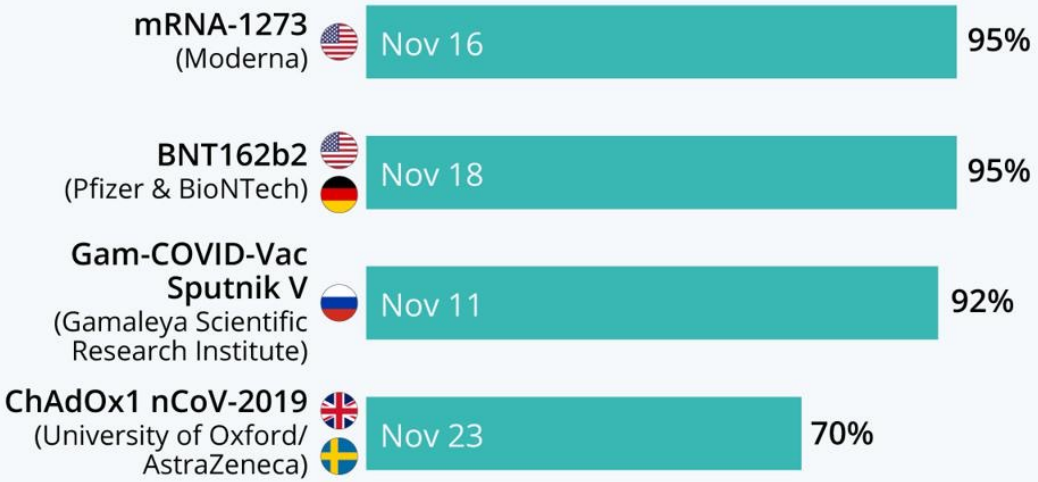
Il tasso di disoccupazione giovanile al 20,6% (+0,9 punti)

di Redazione Roma

30 ottobre 2025

How Effective Are The Covid-19 Vaccine Candidates?

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

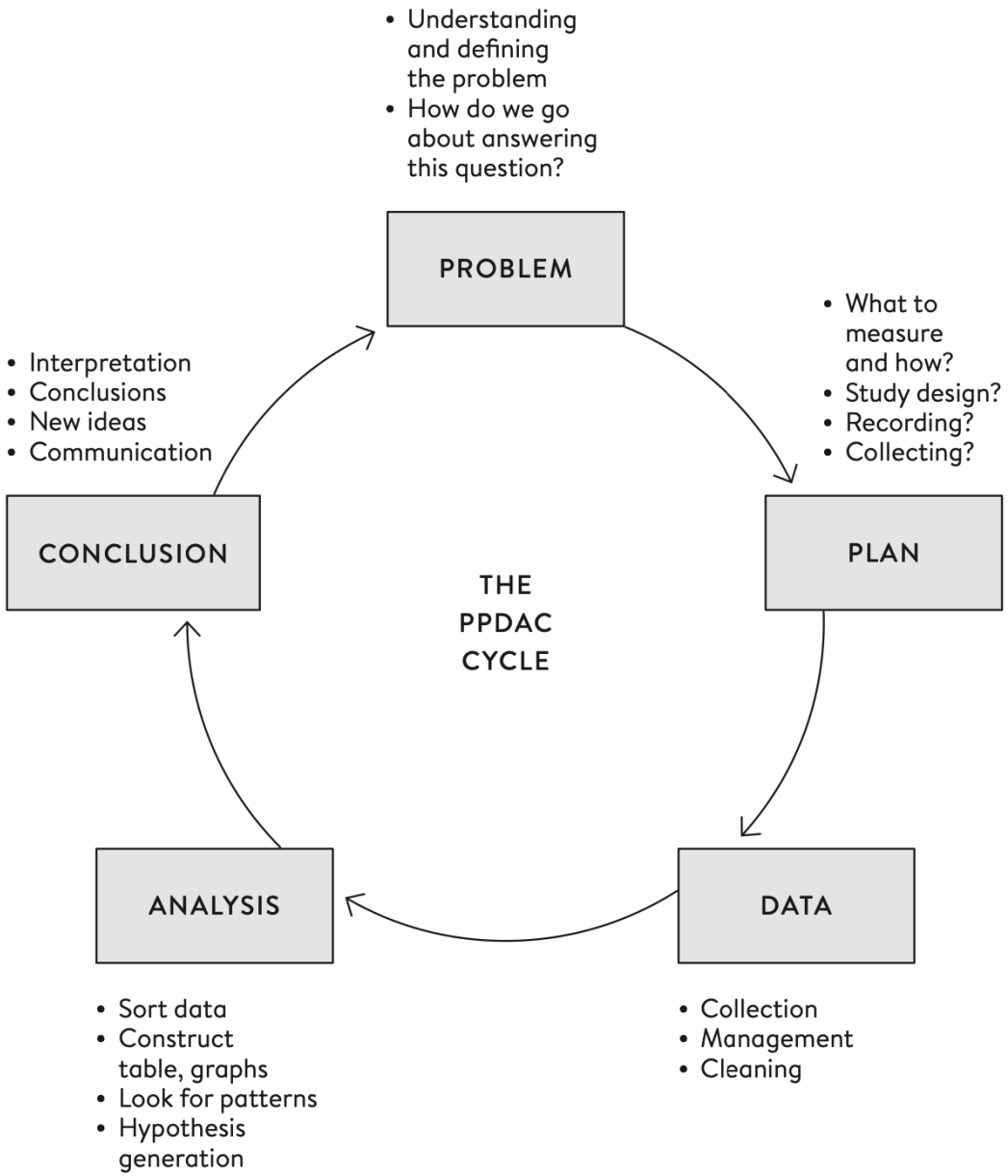


* As of Nov 23, 2020. Phase III trials for BNT162b2 are complete.
Other trials are ongoing and findings have not been peer-reviewed.

Sources: Respective companies, Russian health ministry

Forbes statista

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



2 – Types of studies

Type of Studies

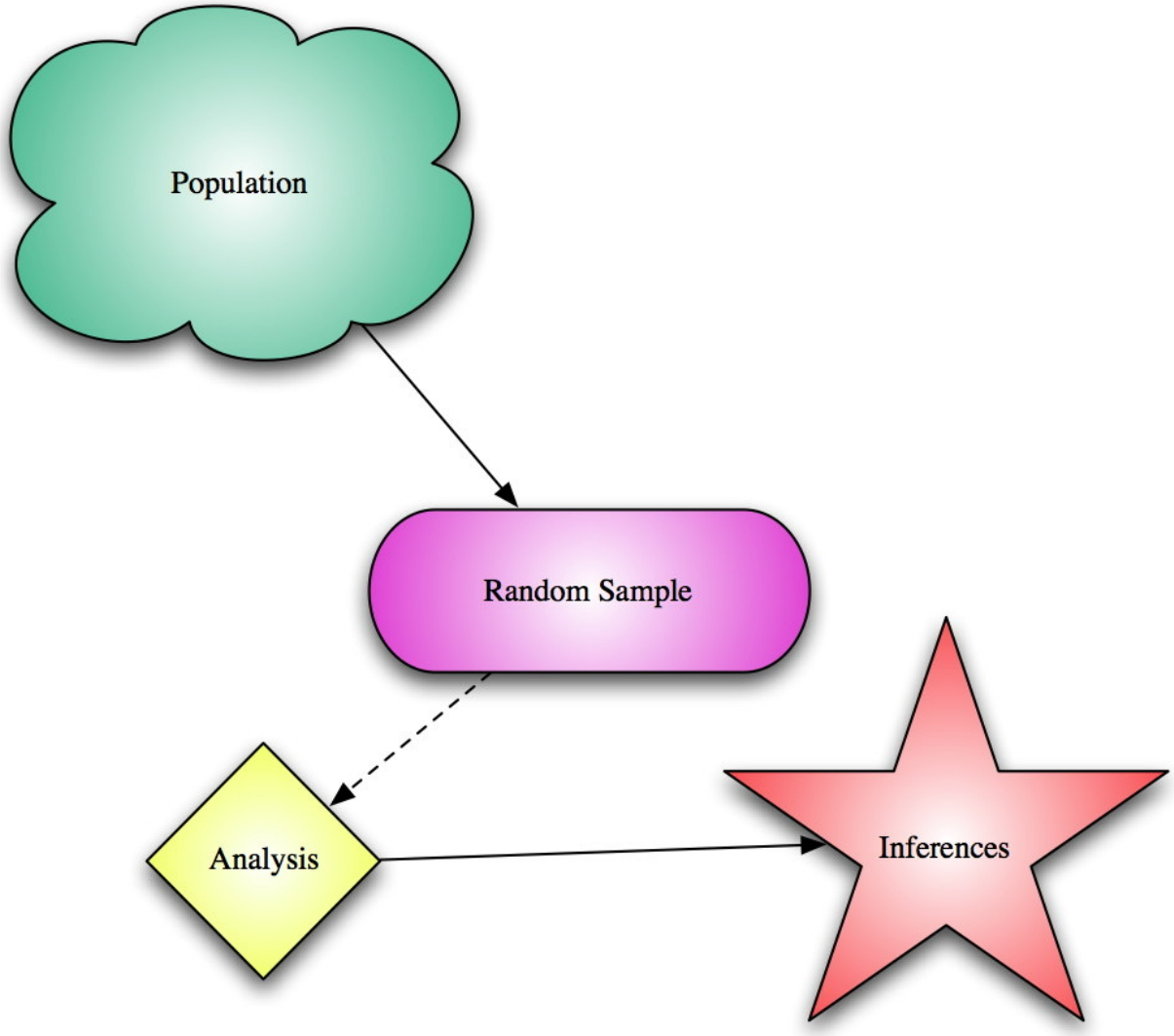
There are two primary types of data collection: **observational studies** and **experiments**.

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

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

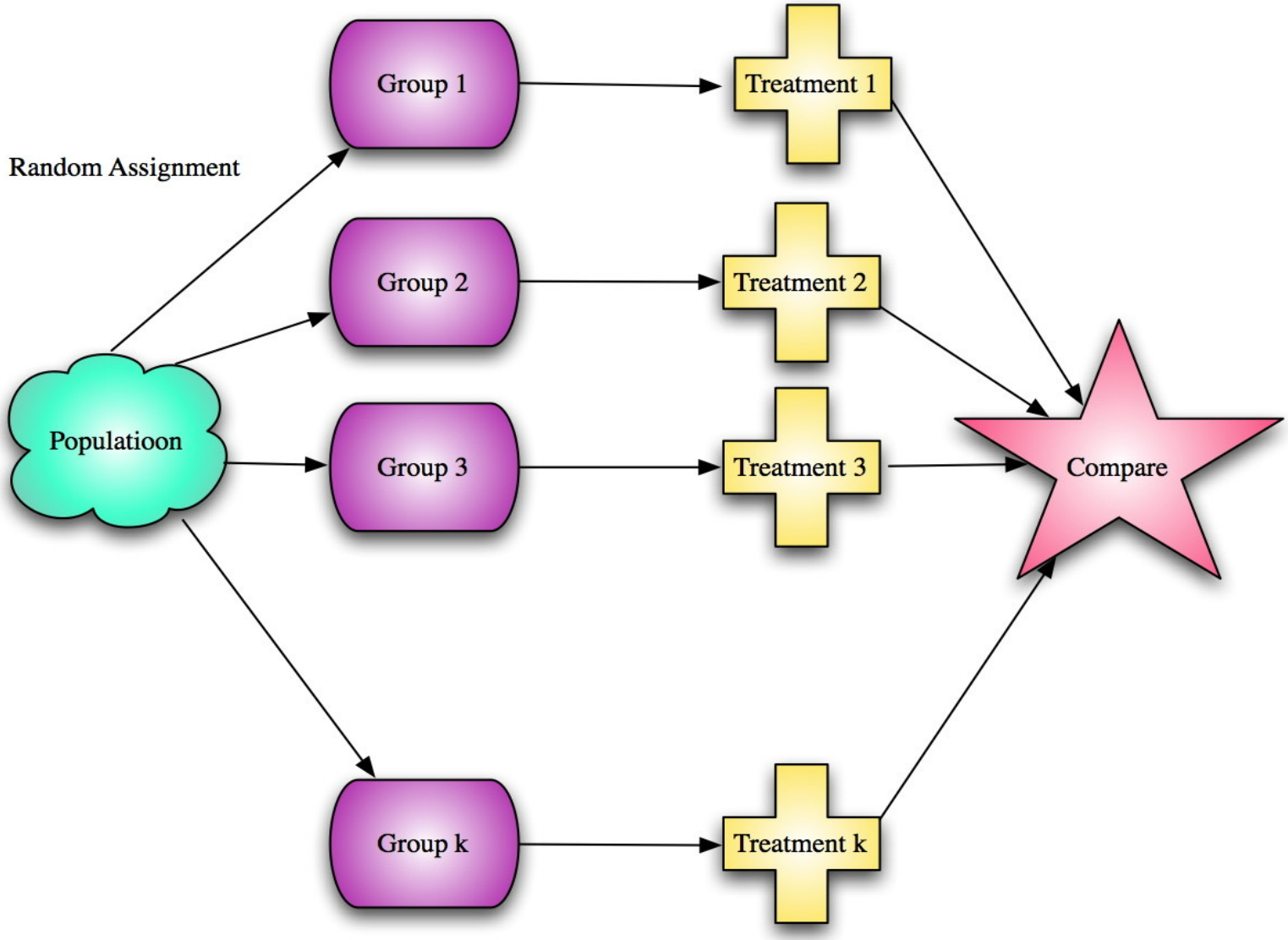
When researchers want to establish a causal connection, they conduct an experiment.

Observational studies



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

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


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

Experiments

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

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

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

Natural Experiments

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

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

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.

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.

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.

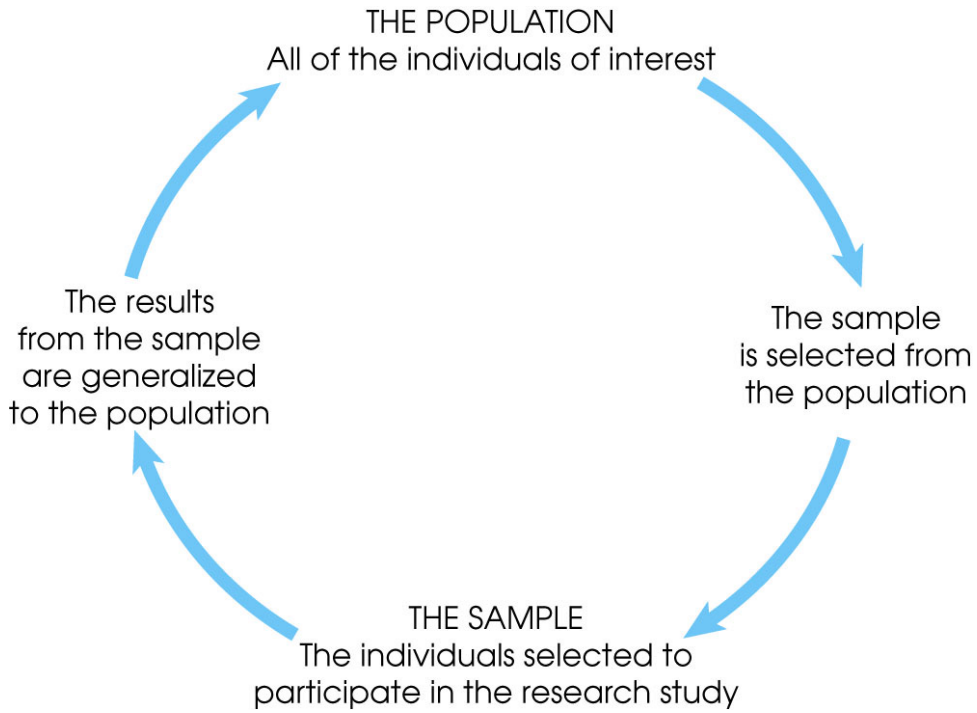
Height, age, income, country of birth, grades obtained at school are all examples of variables. Variables may be classified into various categories.

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.

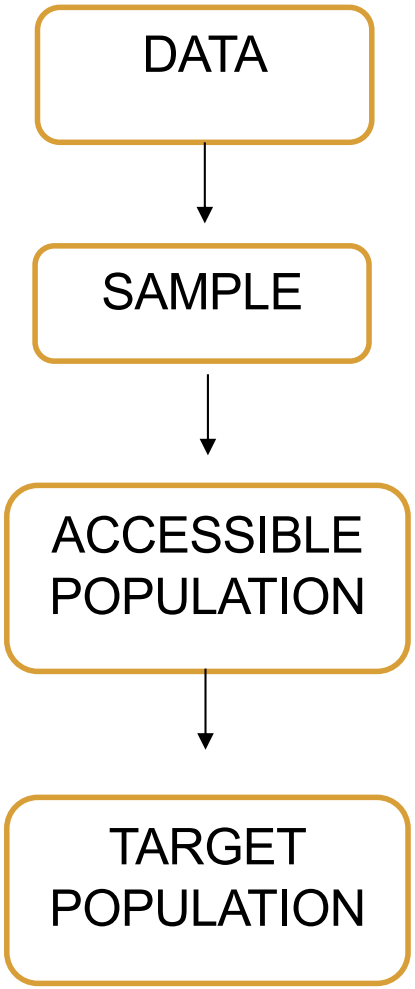
STATISTICS

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

Obtaining good samples

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

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

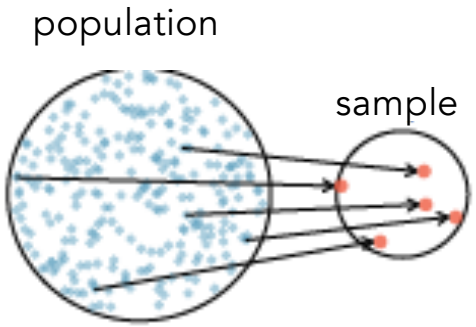
✓The only guarantee for that comes from the method used to select the sample (sampling method)→**probability sampling**

✓There are several sampling methods that guarantee representativeness.

Obtaining good samples

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

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



Simple random sampling

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

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

Procedure:

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

Simple random sampling

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

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

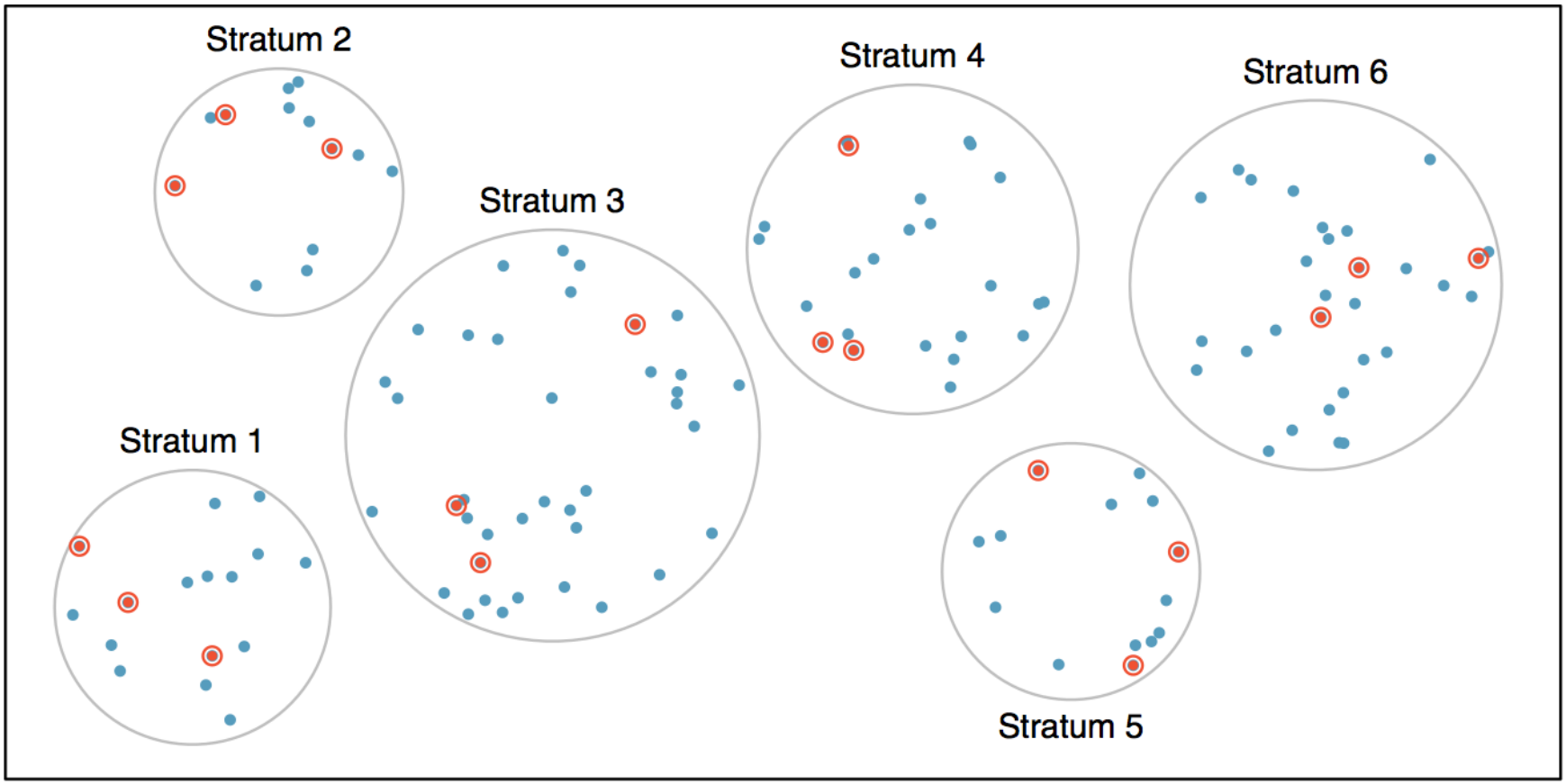
Stratified random sampling

The population is divided into groups called strata.

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

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

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



As example:

If a random selection of students were selected and you wanted to be sure that students majoring in psychology and in business and in others were included with those from all other groups, you could separate the population into three and randomly select members from each.

Cluster and multistage random sampling

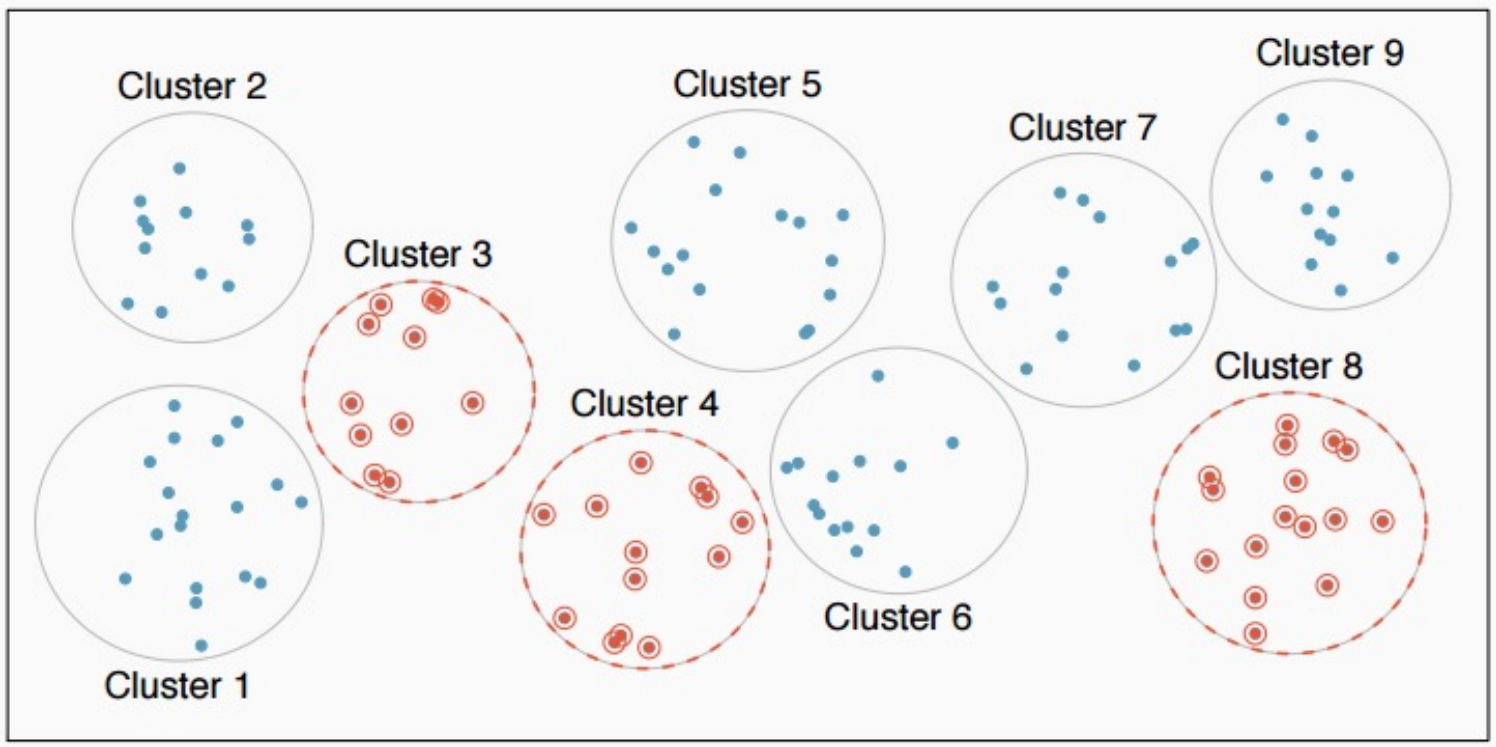
we break up the population into many groups (usually naturally occurring groups like municipalities, classes, hospitals,…), called clusters.
Then we sample a fixed number of clusters and include all observations from each of those clusters in the sample.

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

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

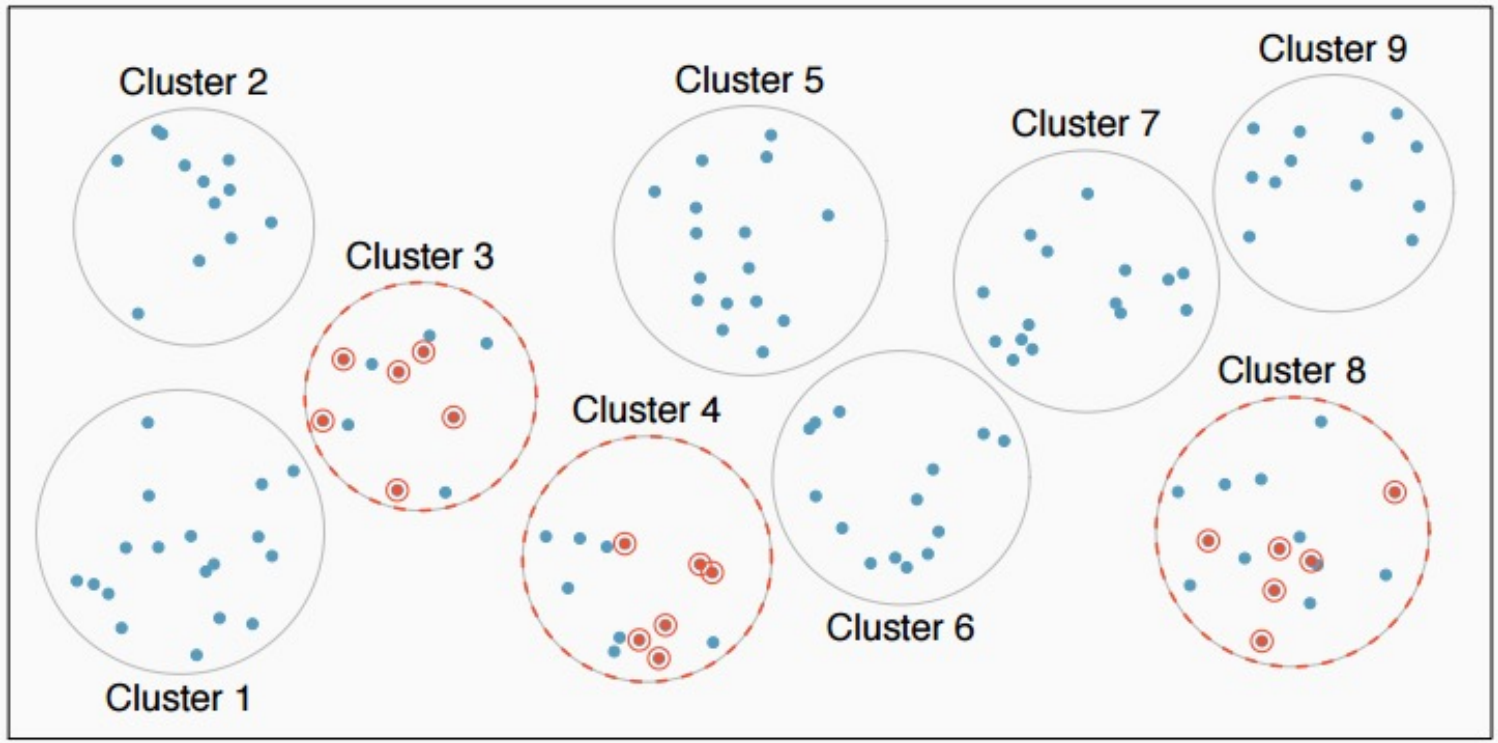
Cluster Sample

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



Multistage Sample

We take a simple random sample of clusters, and then take a simple random sample of observations from the sampled clusters



Cluster and multistage random sampling

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

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

Example:

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

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

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

NON PROBABILITY SAMPLES:

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

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

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

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

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

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

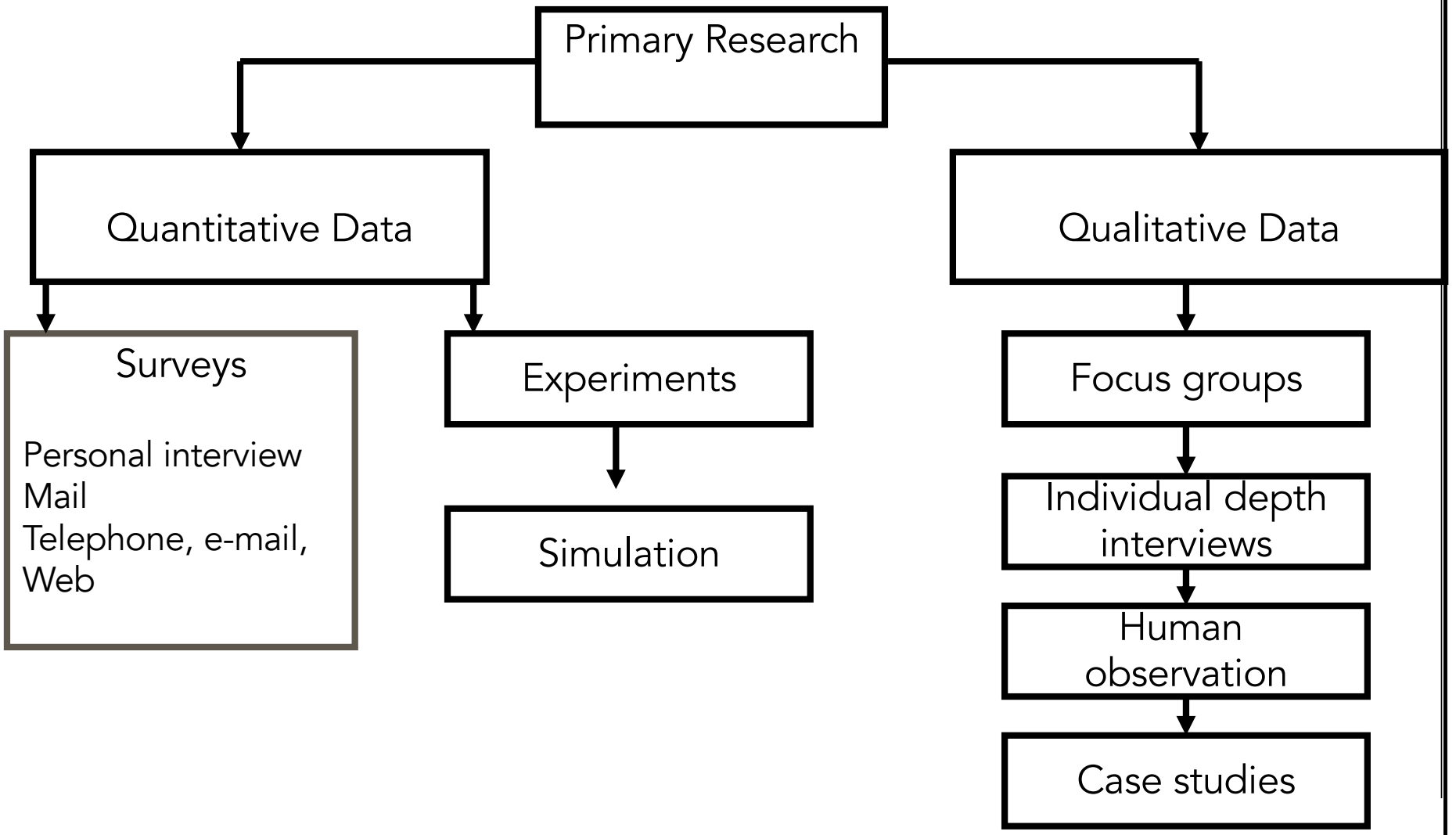
- Primary:

  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

Ensures you collect 'good' data!

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

- Reduces potential bias
- “Control” 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!

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

Data Matrix: individual level data

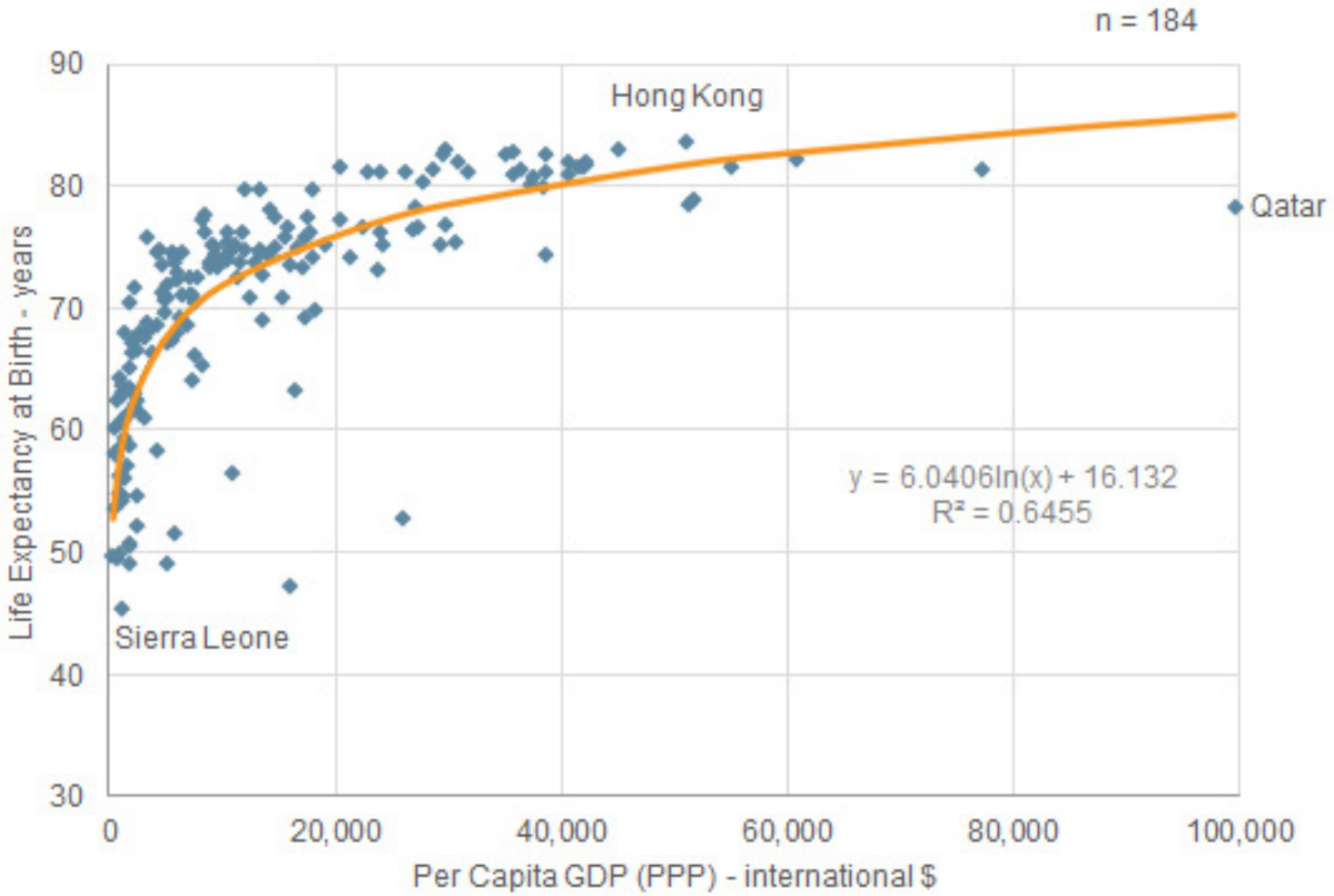
wave	country	hid	pid	pd001	age	sex	maritalstatu	pe001	personalincome	healthstatus
w2 surve	spain	6068101	60681101	1948	47	male	married	paid emp	2400695	good
w6 surve	denmark	5445702	54457103	1974	25	female	married	paid emp	129000	very goo
w3 surve	spain	5882101	58821101	1934	62	male	married	paid emp	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

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

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

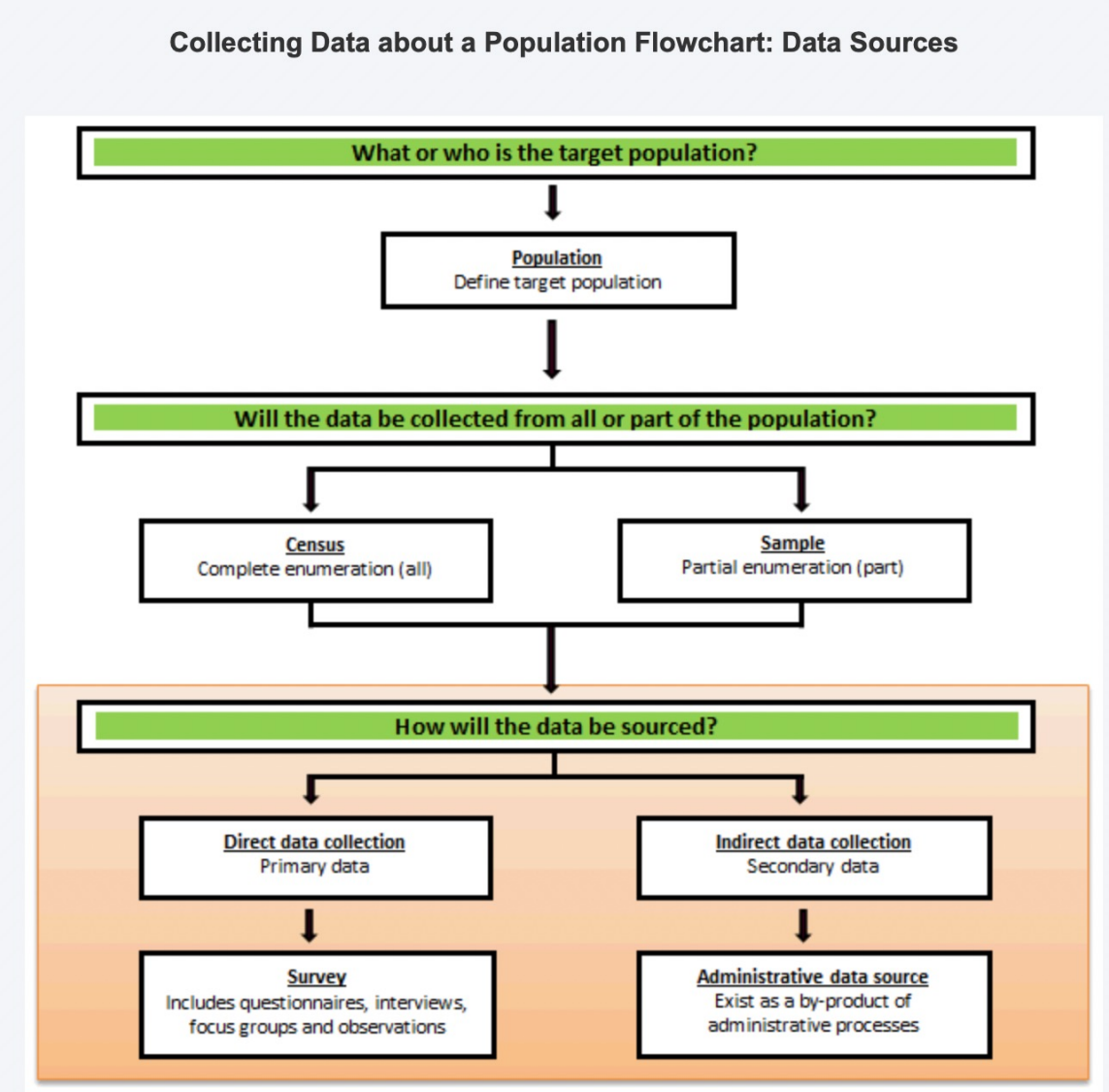
- Trustable

- 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

Let's concentrate on observational studies and start analysing data

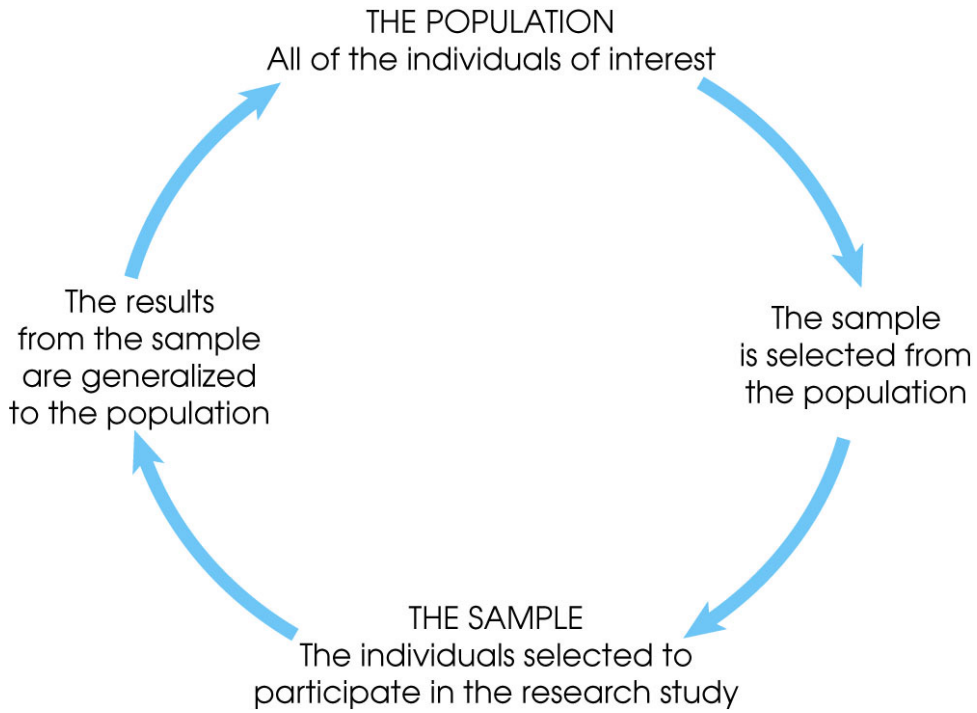


Australian Bureau of Statistics

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**



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

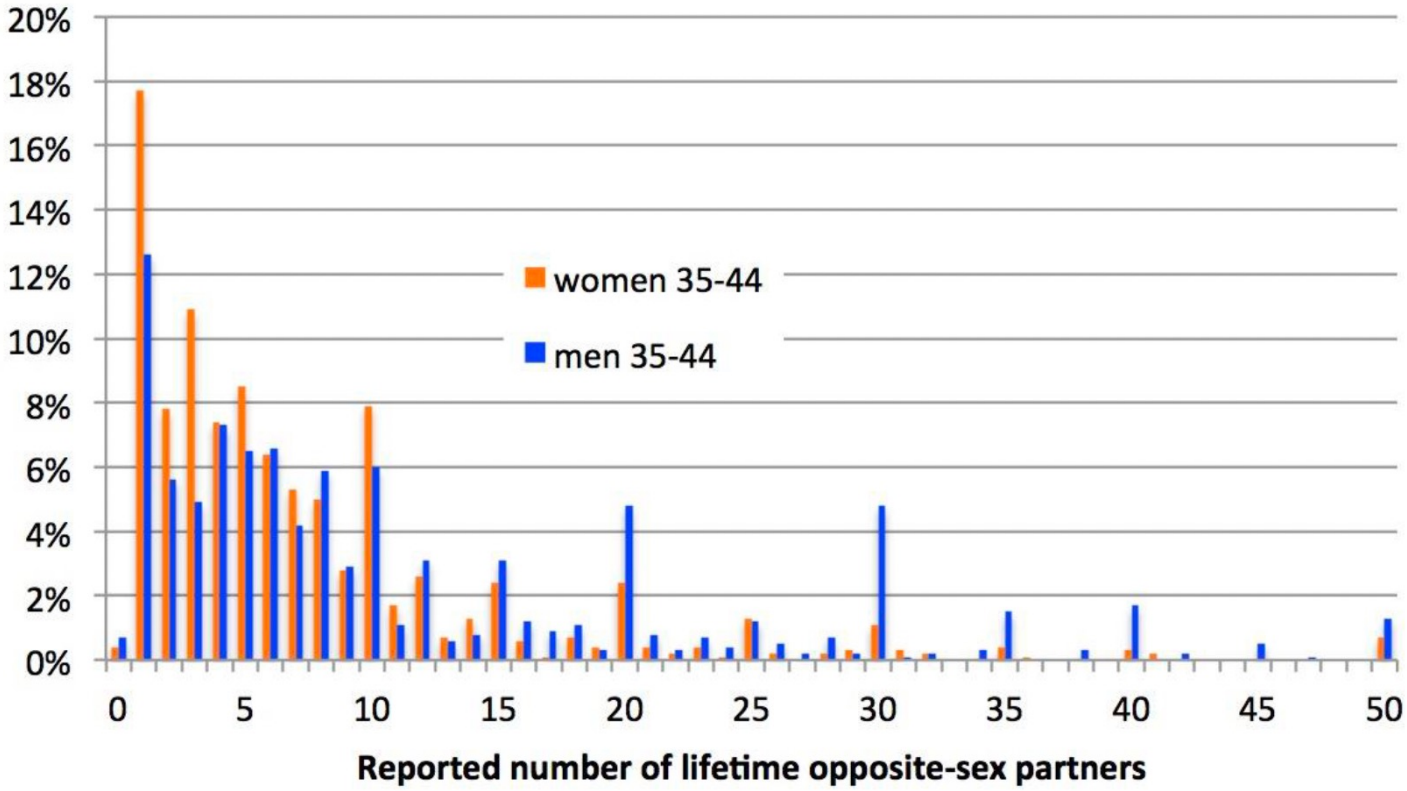
✓**Statistics:** known number that summarize a characteristics of the sample. A statistic is often used to point estimate the parameter in the population.

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

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

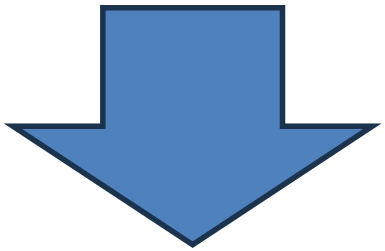
- **Problem**: cannot know this as a fact
- **Plan**: survey in which people are carefully asked about the sexual activity (Natsal)
- **Data**: reports of numbers of partners
- **Analysis**: plotting and summary statistics

How many sexual partners do people report?



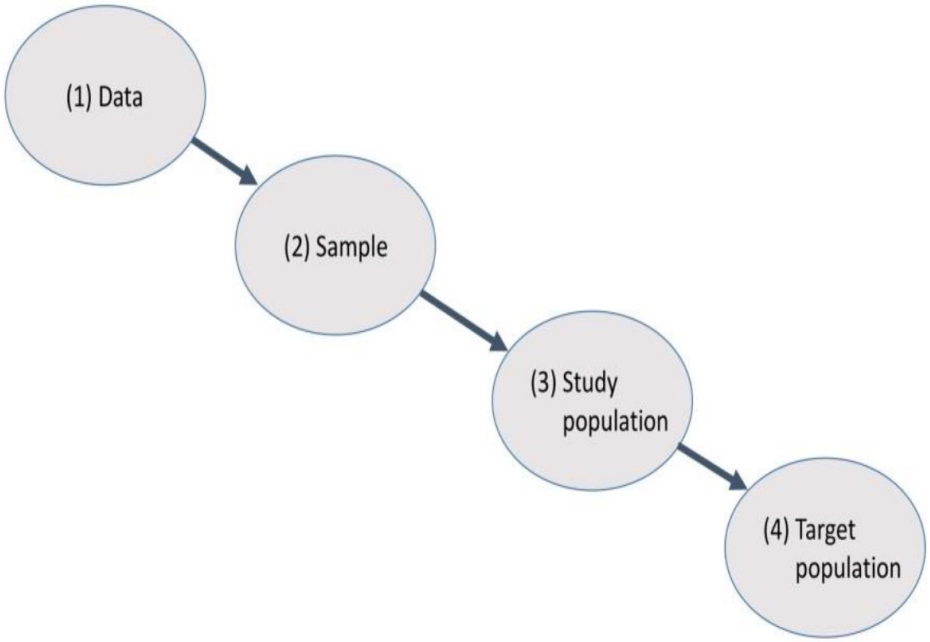
Reported number of sexual partners in lifetime	Men aged 35–44	Women aged 35–44
Mean	14.3	8.5
Median	8	5
Mode	1	1
Range	0 to 500	0 to 550
Inter-quartile range	4 to 18	3 to 10
Standard deviation	24.2	19.7

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



INFERENCE

INDUCTION PROCESS



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

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

From 3 to 4: **external validity**

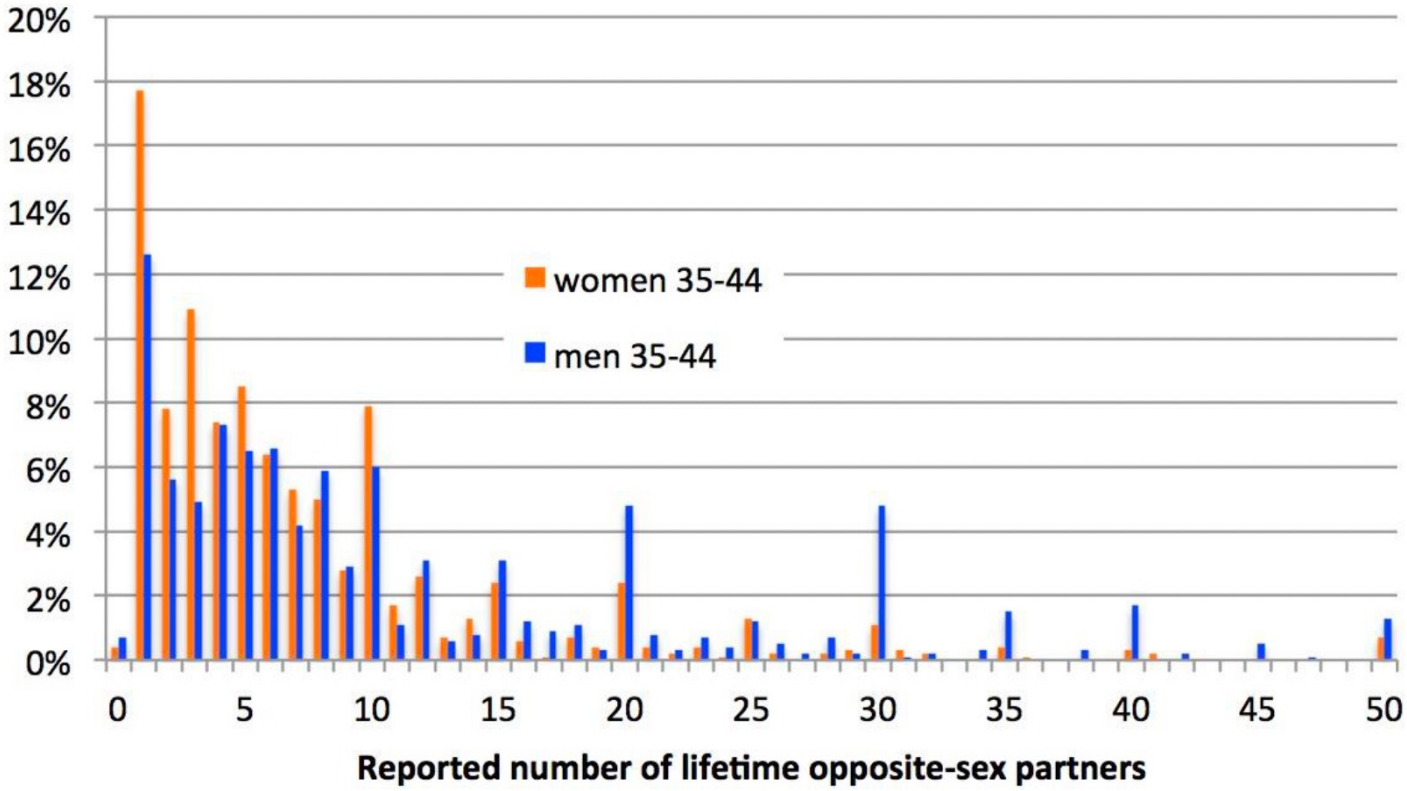
Inference and bias

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

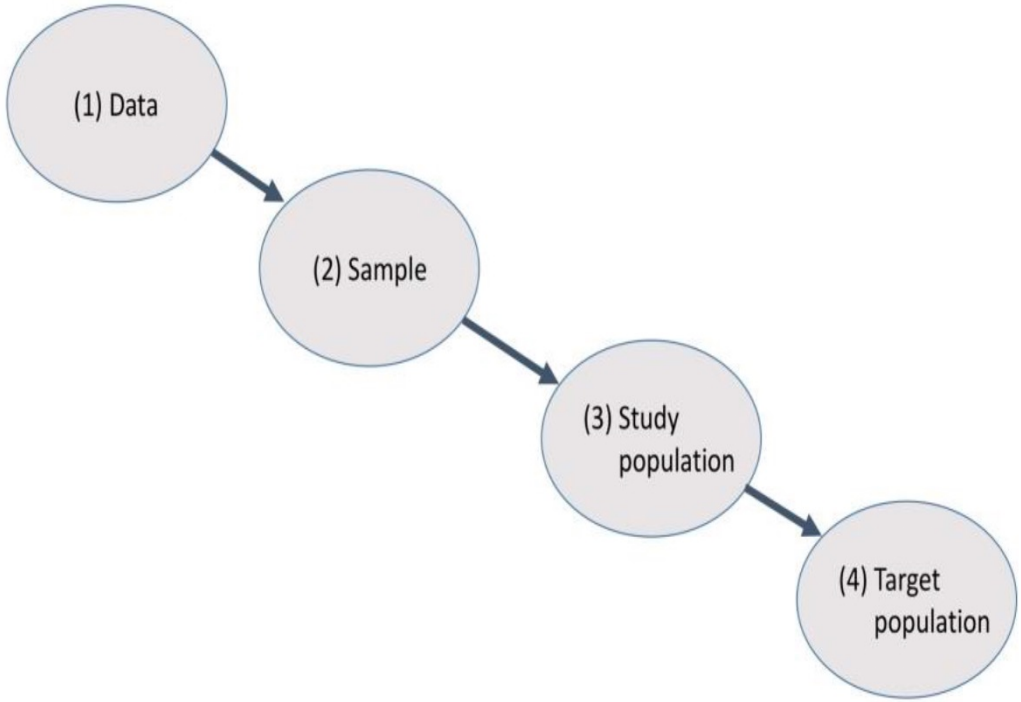
Reported number of sexual partners in lifetime	Men aged 35–44	Women aged 35–44
Mean	14.3	8.5
Median	8	5
Mode	1	1
Range	0 to 500	0 to 550
Inter-quartile range	4 to 18	3 to 10
Standard deviation	24.2	19.7

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

How many sexual partners do people report?



Induction: the stages in generalising from data



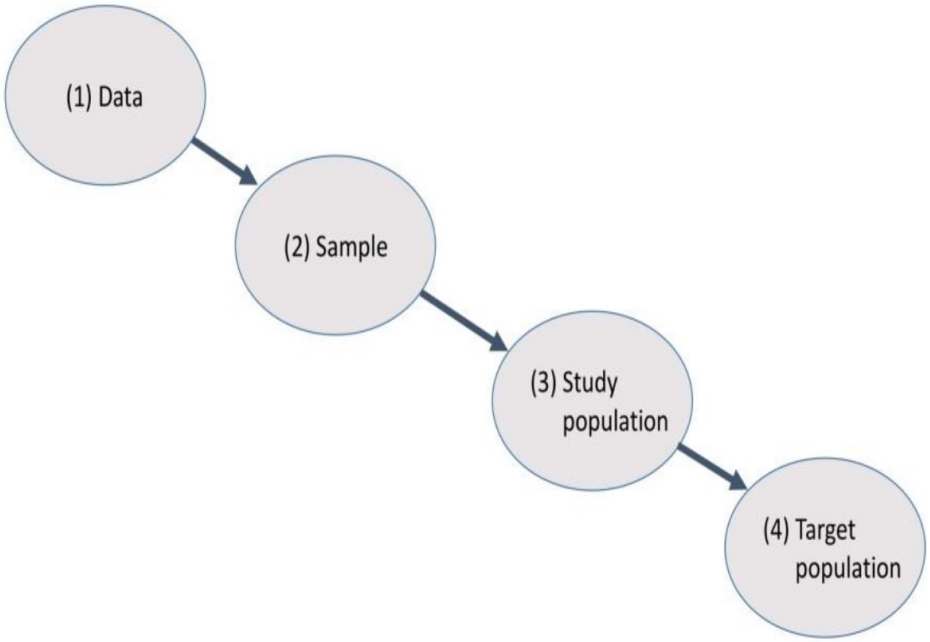
- ***1 to 2.*** How reliable are the reports?
- *Poor memory, social acceptability bias etc*

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

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

4 - Types and sources of error in statistical data

INDUCTION PROCESS



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

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

From 3 to 4: external validity

What is sampling error? (1/2)

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

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

What is sampling error? (1/2)

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

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

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

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

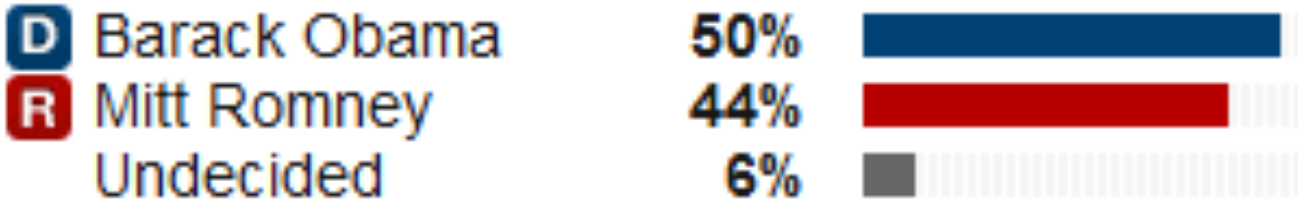
The discrepancy (natural difference that exist by chance) between a sample statistic and its population parameter is called **sampling error**.

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

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? (2/3)

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

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

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

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

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

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

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

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

Biased survey questions: positive (negative) framing

92% Of Ryanair Customers Satisfied With Flight Experience

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

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

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

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

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

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

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

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

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

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

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

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

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

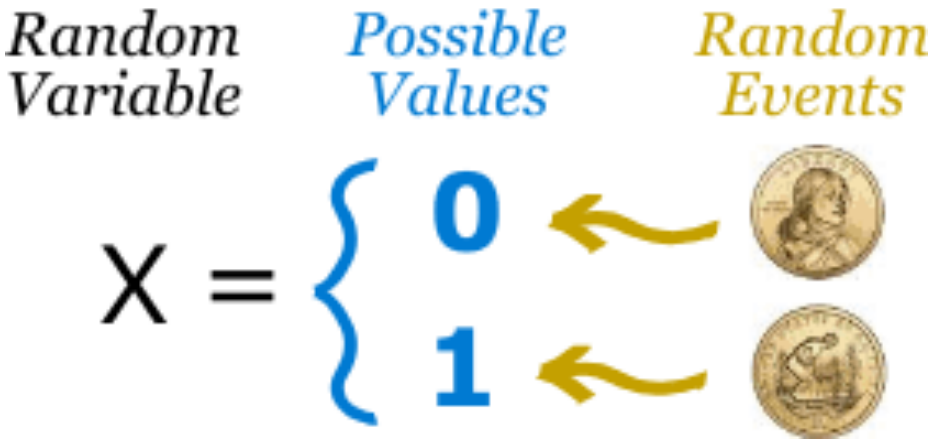
5 - Probability and population

What is a random variable

Random Variable X: function that maps outcomes of a random process to real values.

A function from the sample space S to the set R of all real numbers.

Examples: tossing a coin, or you want to know how many sixes you get if you roll the dice a certain number of times. Your random variable, X could be equal to 1 if you get a six and 0 if you get any other number.



Random Variables: continuous & discrete

Random variables can be **discrete** or **continuous**

Discrete random variables have a countable (or finite) number of outcomes.

Examples: Dead/alive, satisfied/not satisfied, etc.

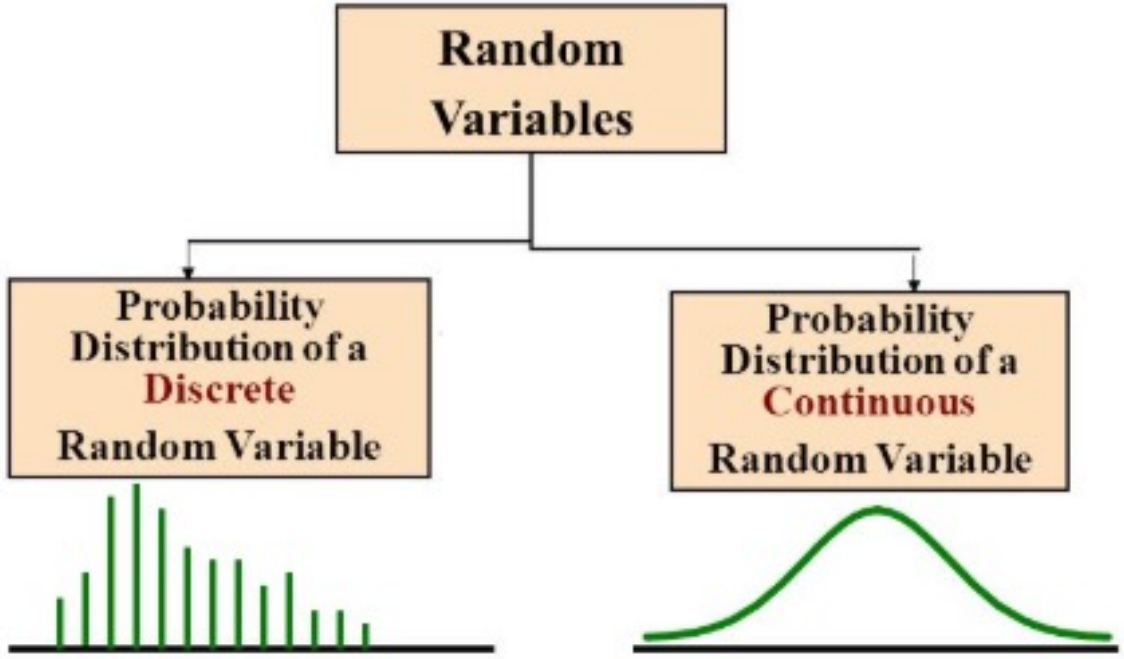
Continuous random variables have an infinite continuum of possible values.

Examples: blood pressure, weight, the speed of a car, QI.

Random Variable and Probability Distribution

The **probability distribution** of a random variable is the collection of possible outcomes along with their probabilities:

- Discrete case: $\Pr(X = x) = p_\theta(x)$
- Continuous case: $\Pr(a \le X \le b) = \int_a^b p_\theta(x)dx$



Discrete Random Variables

- Discrete random variables can be summarized by listing all values along with the probabilities
  - Called a **probability distribution**
- Example: number of members in US families

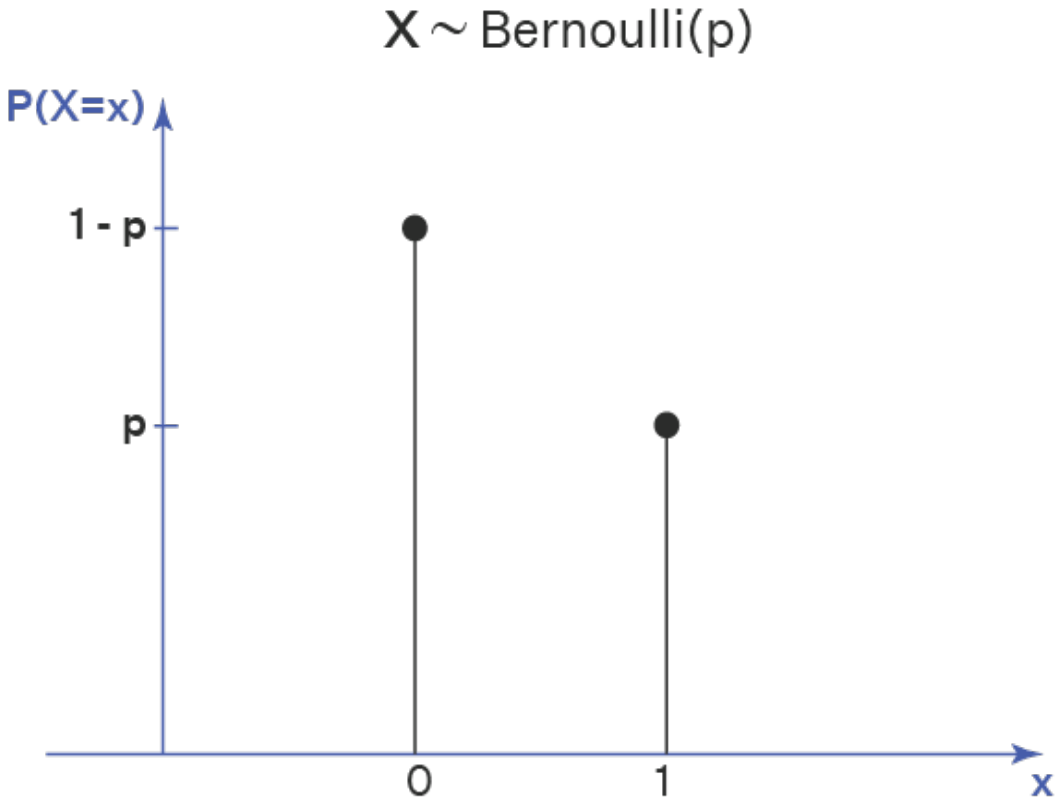
X	2	3	4	5	6	7
P(X)	0.413	0.236	0.211	0.090	0.032	0.018

A – BERNOULLI DISTRIBUTION

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

The **Bernoulli distribution** models a single trial with two possible outcomes: success and failure.

Example: modeling of binary events such as the outcome of a coin flip (heads or tails), success or failure of a medical treatment, or the occurrence of an event in a single trial.



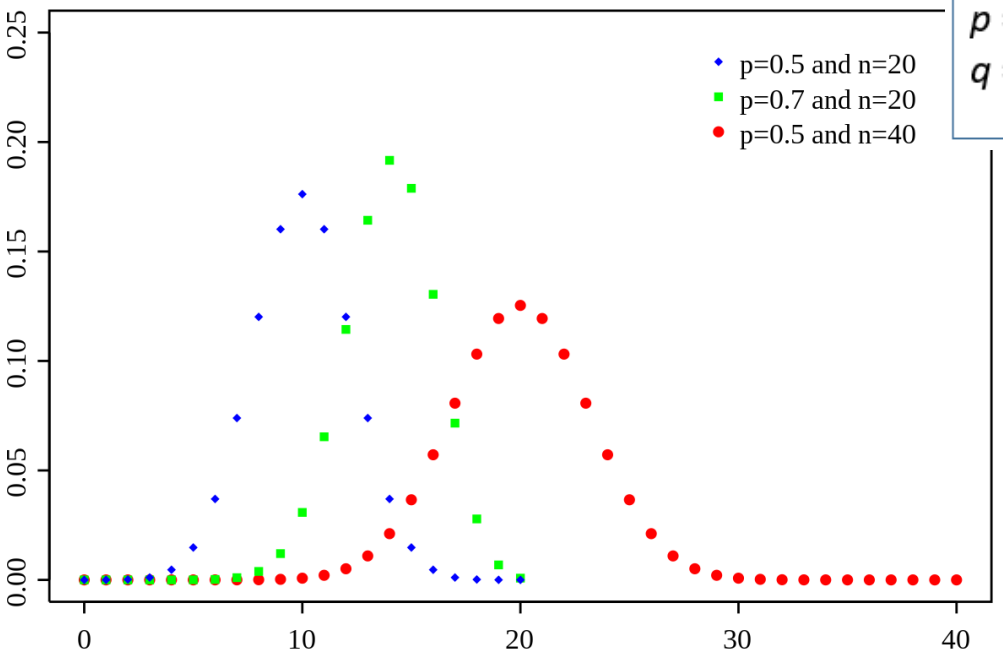
B – BINOMIAL DISTRIBUTION

The **Binomial distribution** models the number of successes in a fixed number of independent and identically distributed Bernoulli trials

Binomial Distribution Formula

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!\,x!} p^x q^{n-x}$$

where
n = the number of trials (or the number being sampled)
x = the number of successes desired
p = probability of getting a success in one trial
$q = 1 - p$ = the probability of getting a failure in one trial



Example: you want to model the number of defective items in a batch of manufactured products

C – POISSON DISTRIBUTION

The **Poisson distribution** models the number of occurrences that happen within a fixed interval of time or space, when these events occur with a known average rate and are independent of the time since the last event.

Poisson Distribution Formula

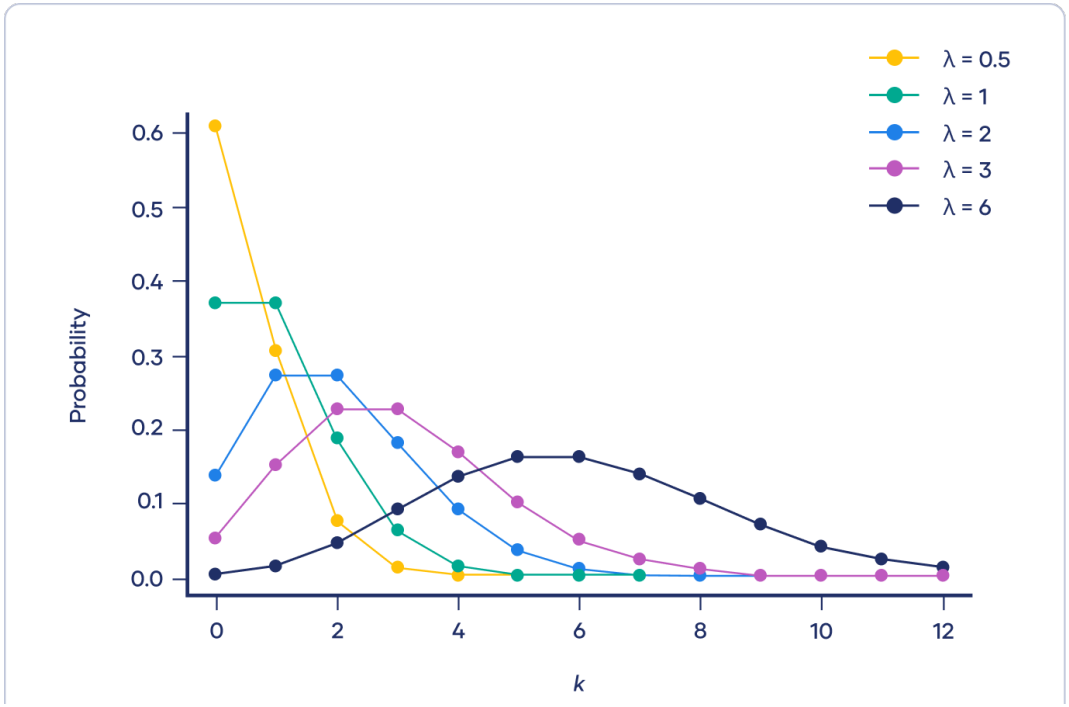
$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

where

x = 0, 1, 2, 3, ...

λ = mean number of occurrences in the interval

e = Euler's constant ≈ 2.71828



Example: modeling the number of phone calls at a call center in a given hour

D – GEOMETRIC DISTRIBUTION

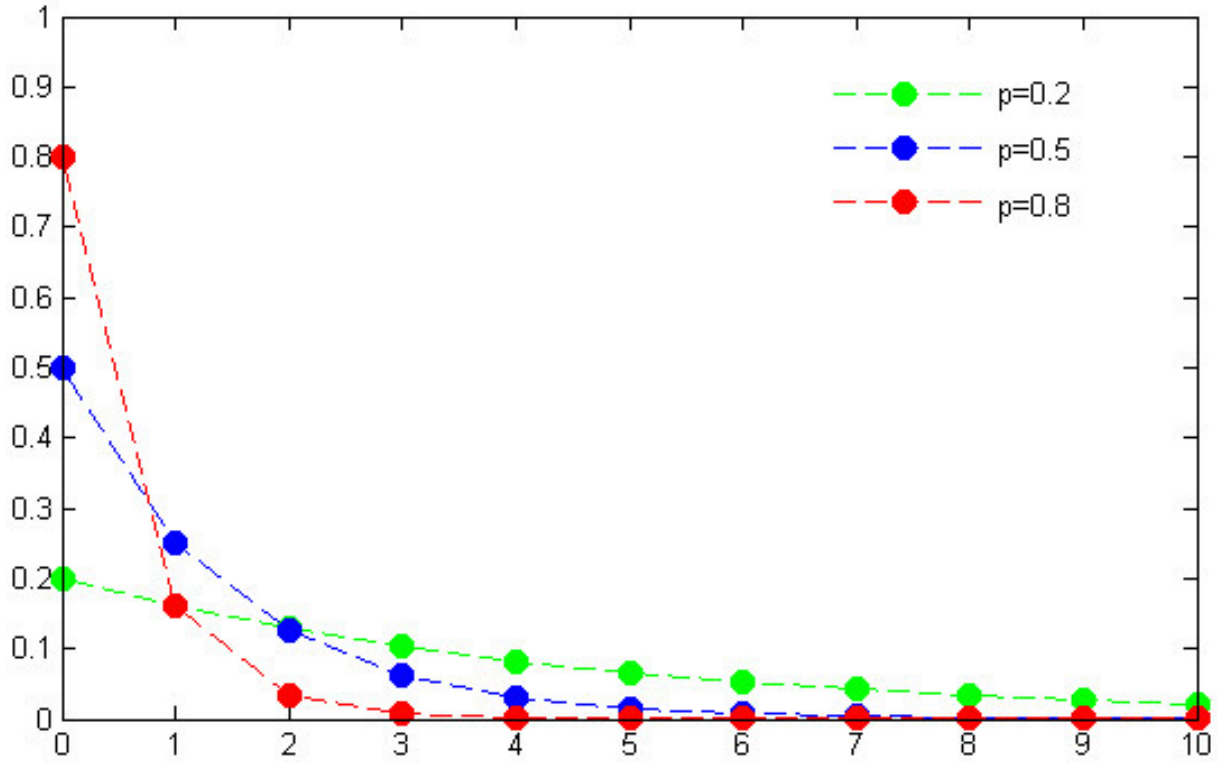
$$P(X = x) = (1-p)^{x-1}p$$

probability of getting the first success in the x-th trial

The **geometric distribution** describes the number of Bernoulli trials required for a success to occur.

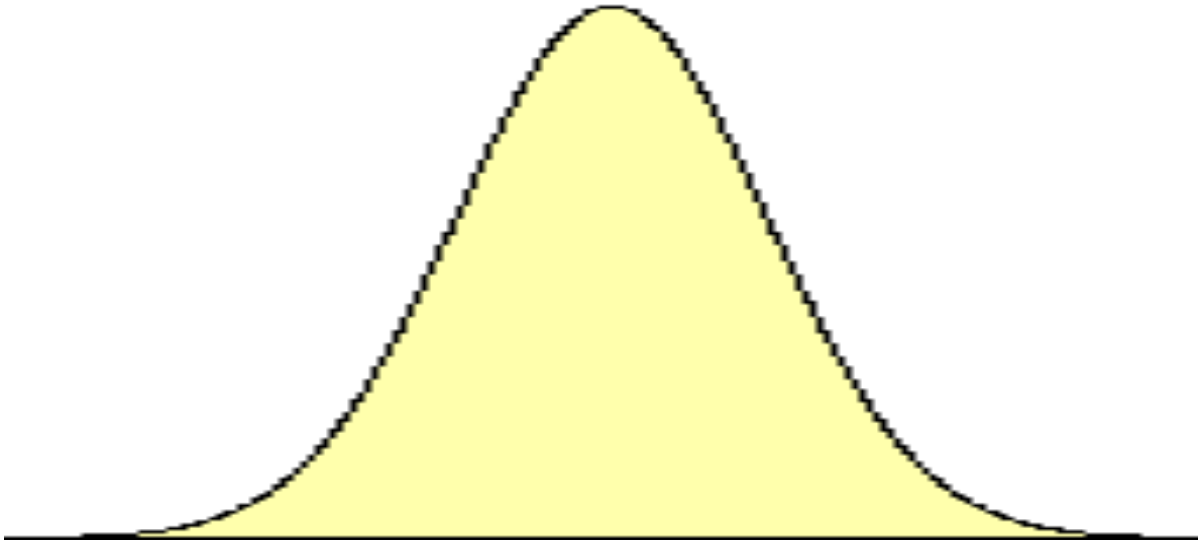
The parameter p measures the probability of success of a single attempt.

Commonly used in scenarios such as modeling the number of trials needed for a coin to come up heads



Continuous Random Variables

- Continuous random variables have a **non-countable** number of values
- Can't list the entire probability distribution, so we use a **density curve** instead of a histogram
- Eg. Normal density curve:

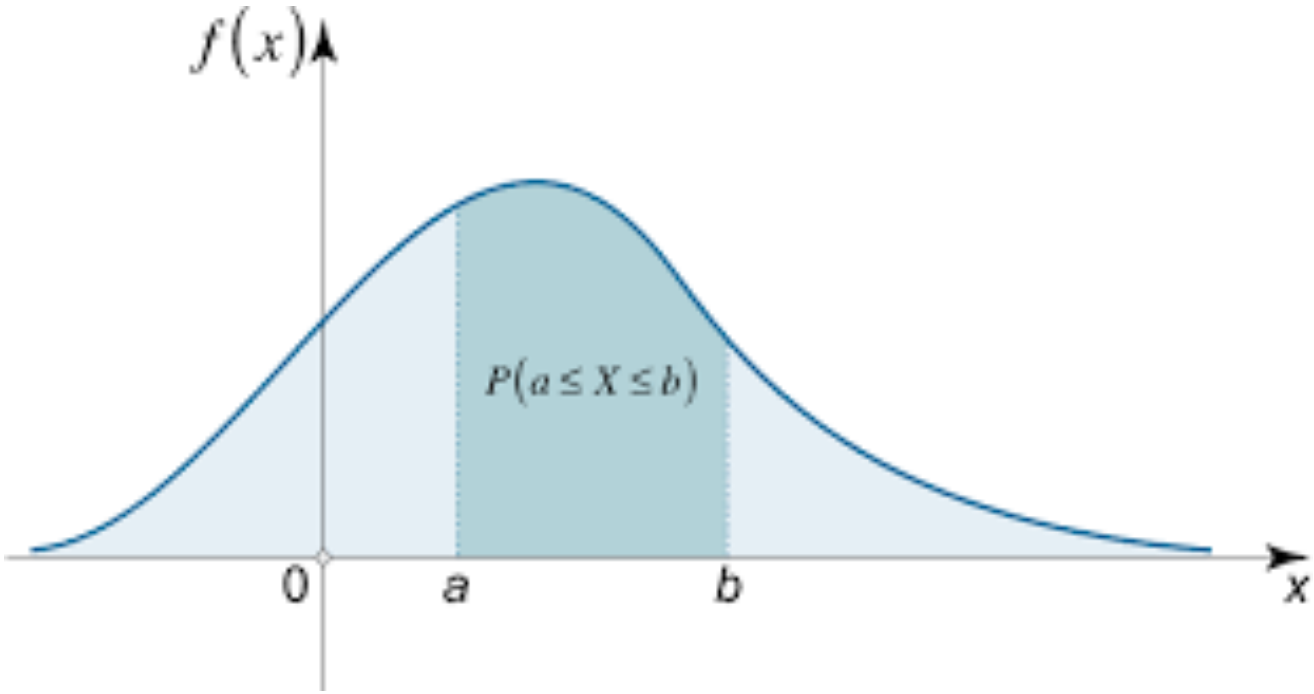


Continuous case

The **probability function** that accompanies a continuous random variable is a continuous mathematical function that integrates to 1.

Probabilities are given for a range of values, rather than a particular value (e.g., the probability of getting a math score between 29 and 30 is 2%).

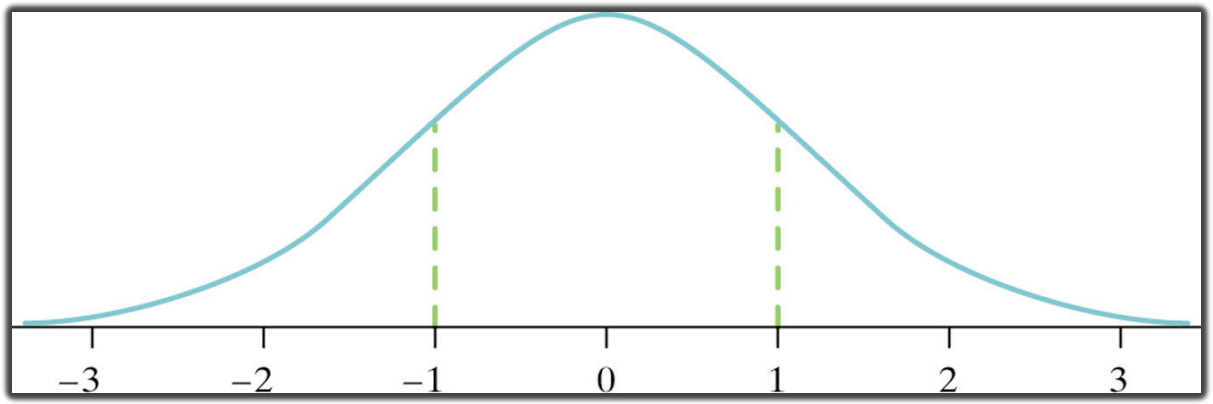
The probabilities associated with continuous functions are just areas under the curve (integrals!).



A - The Standard Normal Distribution

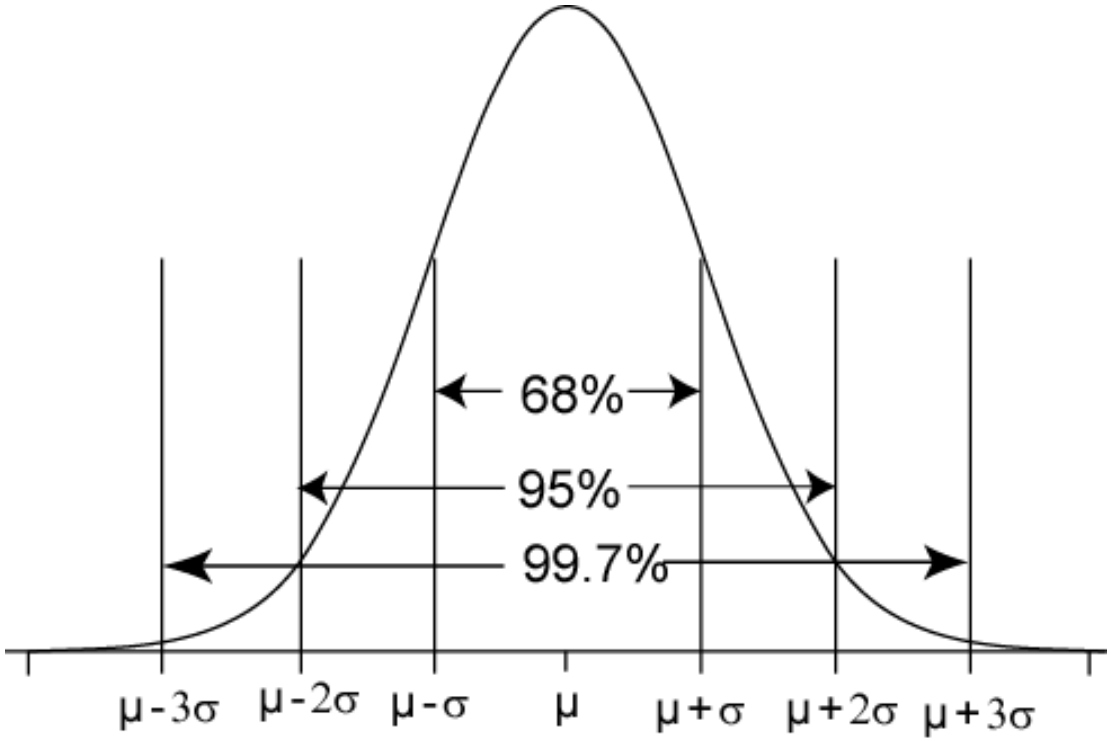
- The **standard Normal distribution** is the Normal distribution with mean 0 and standard deviation 1.
- Shown as ***N***(0,1)
- If a variable x has any Normal distribution $N(\mu, \sigma)$, with mean μ and standard deviation σ, then the standardized variable

$$z = \frac{x - \mu}{\sigma}$$



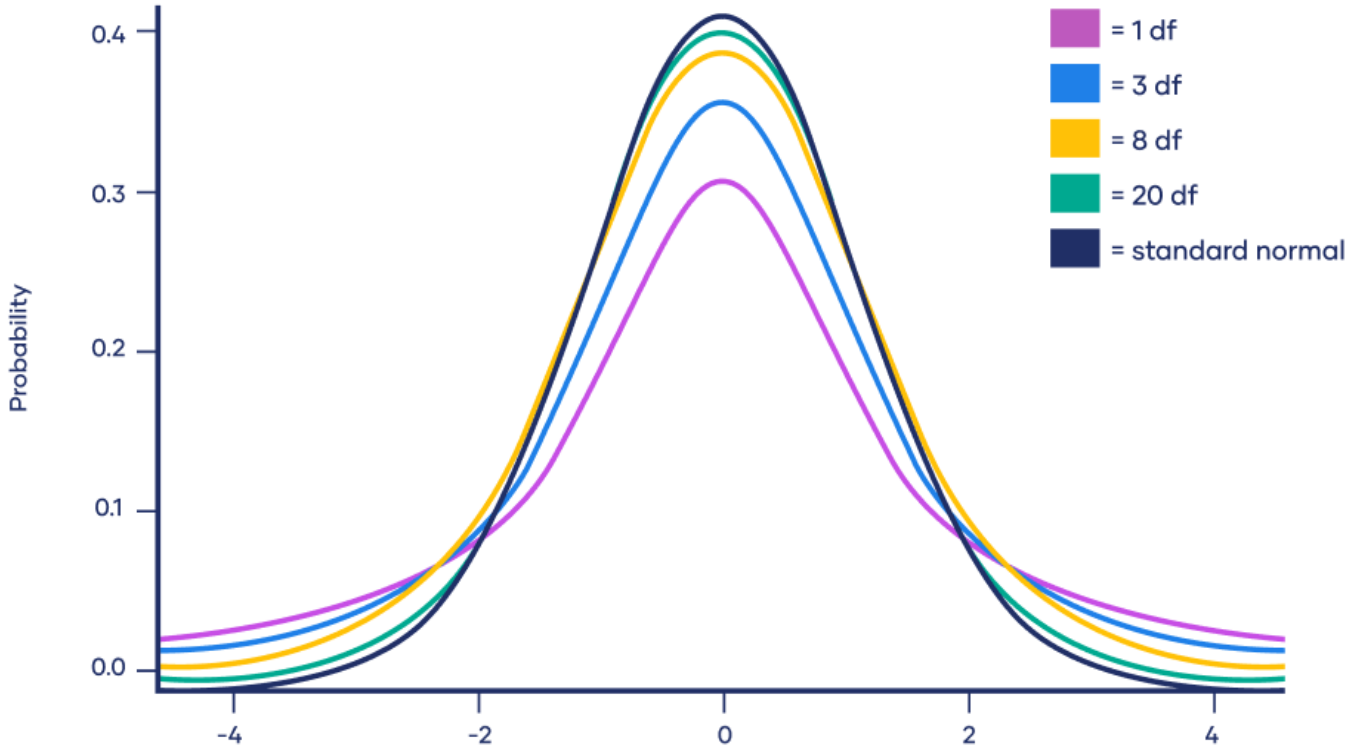
68-95-99.7 Rule for Normal Distributions

- 68% of the AUC within ±1**σ** of μ
- 95% of the AUC within ±2**σ** of μ
- 99.7% of the AUC within ±3**σ** of μ



B – T-STUDENT DITRIBUTION

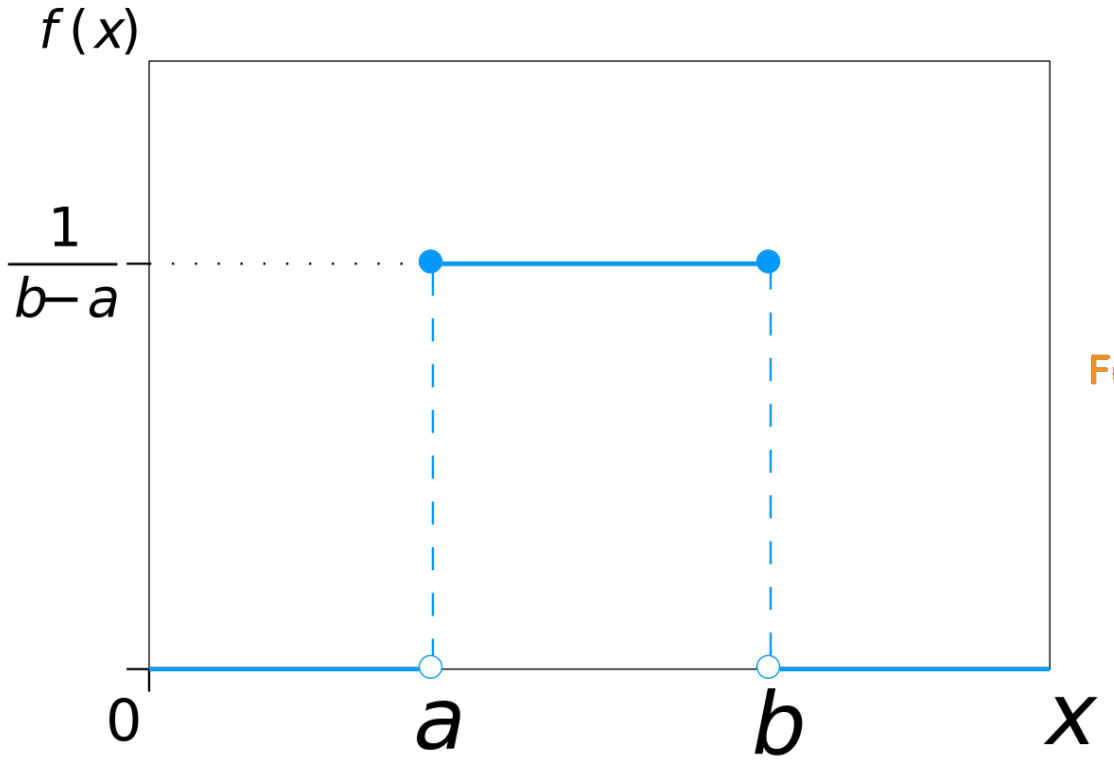
The **T-Student distribution** is particularly used for making inferences about population means when the sample size is small or when the population standard deviation is unknown.



C – UNIFORM DISTRIBUTION

The **Uniform distribution** models a continuous random variable with a constant and equal probability of taking any value within a specified interval.

The uniform distribution is defined by two parameters:
a: The lower bound of the interval.
b: The upper bound of the interval, where a < b.



$$F(x) = \frac{1}{(b - a)}$$

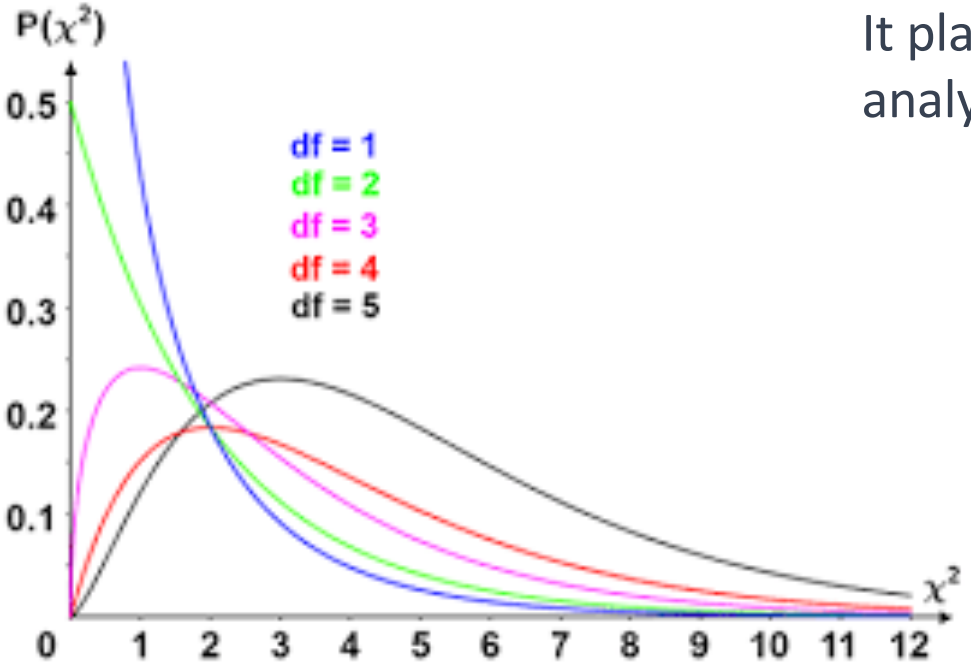
$$\text{Mean} = \frac{(a + b)}{2}$$

$$\sigma = \sqrt{\frac{(b - a)^2}{12}}$$

D – CHI-SQUARE DISTRIBUTION

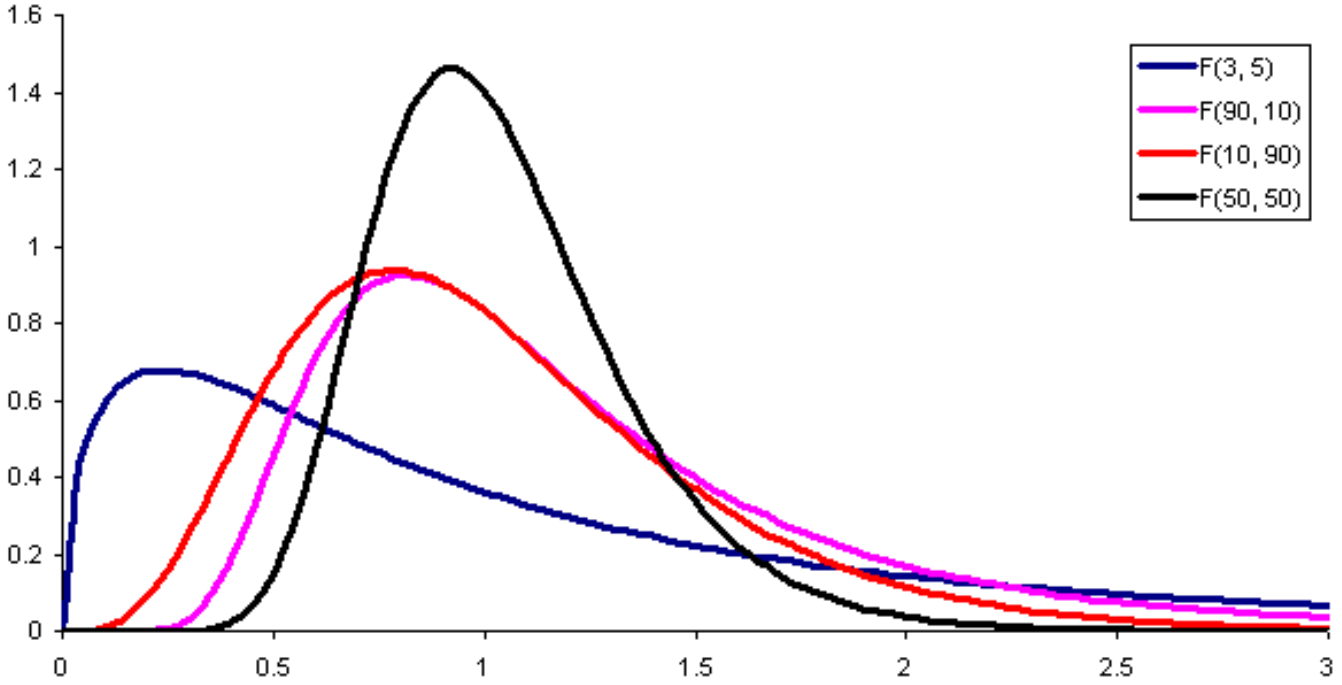
Common applications of the **Chi-squared distribution** include **testing the fit** of a model to observed data, **comparing observed and expected frequencies** in contingency tables, and evaluating the variability in a sample.

It plays a vital role in various statistical analyses and hypothesis tests



E – FISHER DISTRIBUTION

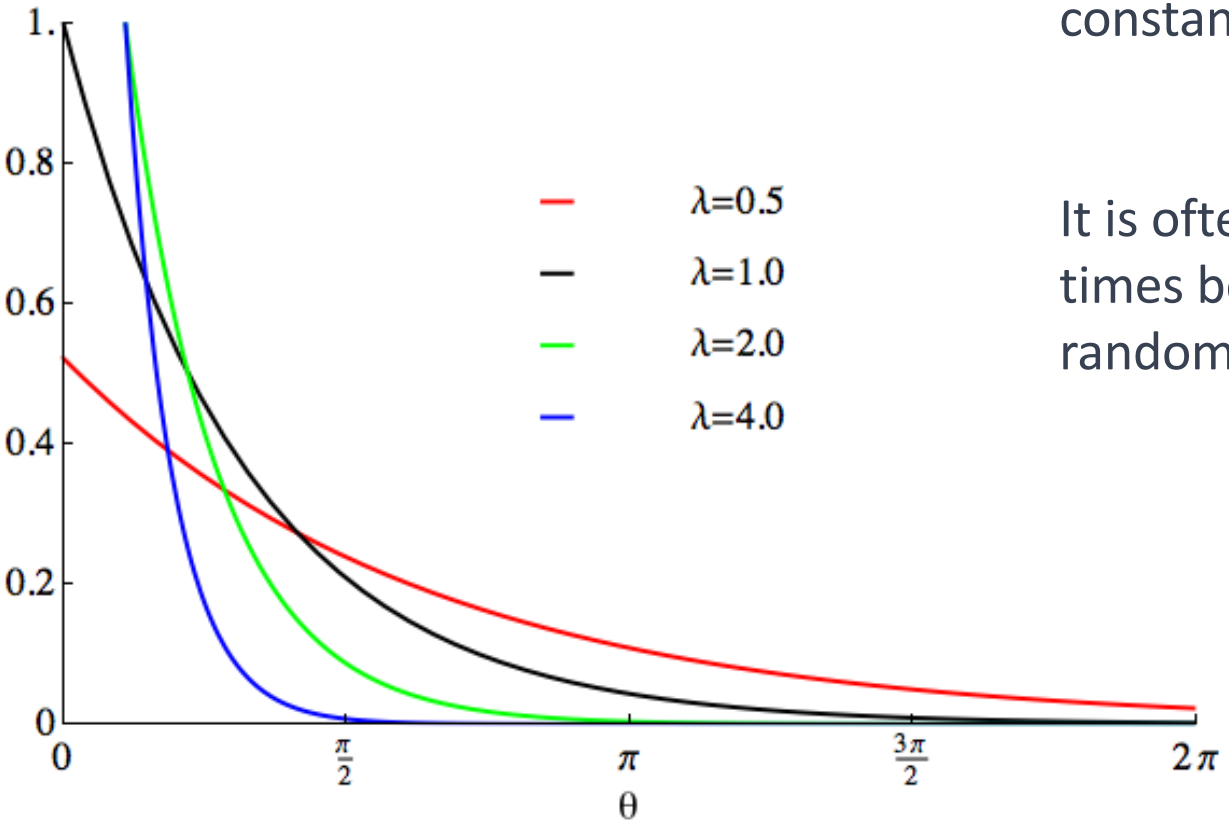
The **Fisher distribution** has two parameters:
- Degrees of Freedom (df1): This parameter, represents the degrees of freedom associated with one sample or population.
- Degrees of Freedom (df2): The second parameter represents the degrees of freedom associated with another sample or population.



The **F distribution** is widely used to compare variances of two samples.

F – EXPONENTIAL DISTRIBUTION

$$f(x) = \begin{cases} 0, & x < 0 \\ \lambda e^{-\lambda x}, & x \geq 0 \end{cases}$$



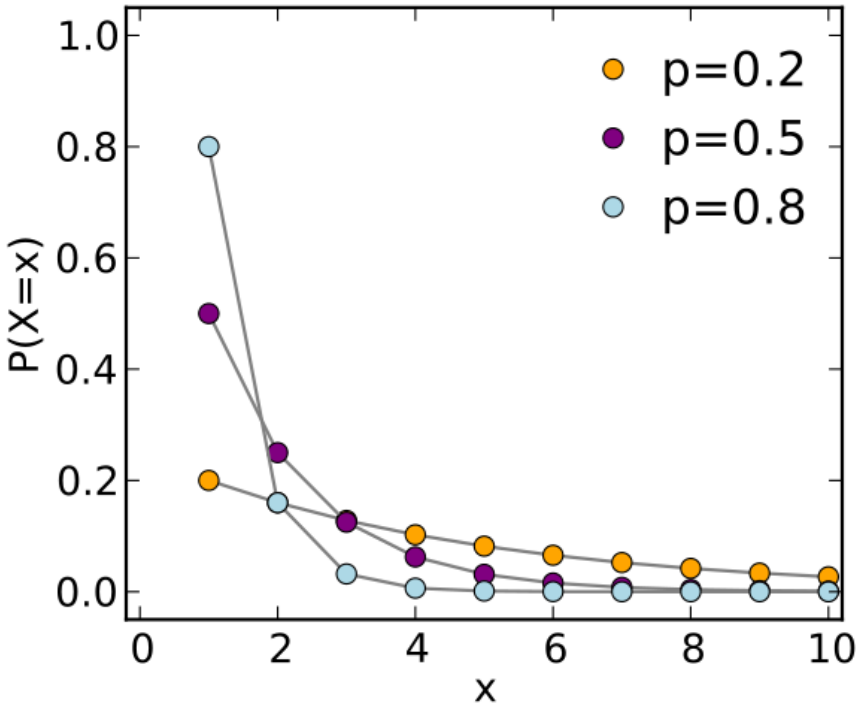
The **Exponential distribution** models the time between events in a process where events occur continuously and independently at a constant average rate.

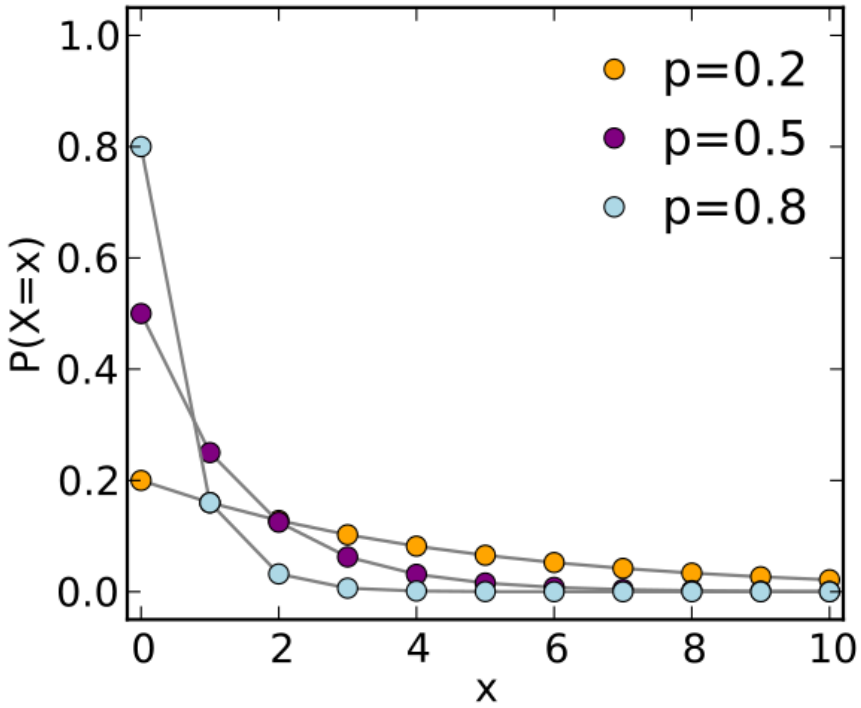
It is often used to model the waiting times between occurrences of random events.

The **Exponential distribution** is the continuous counterpart of the Geometric distribution!

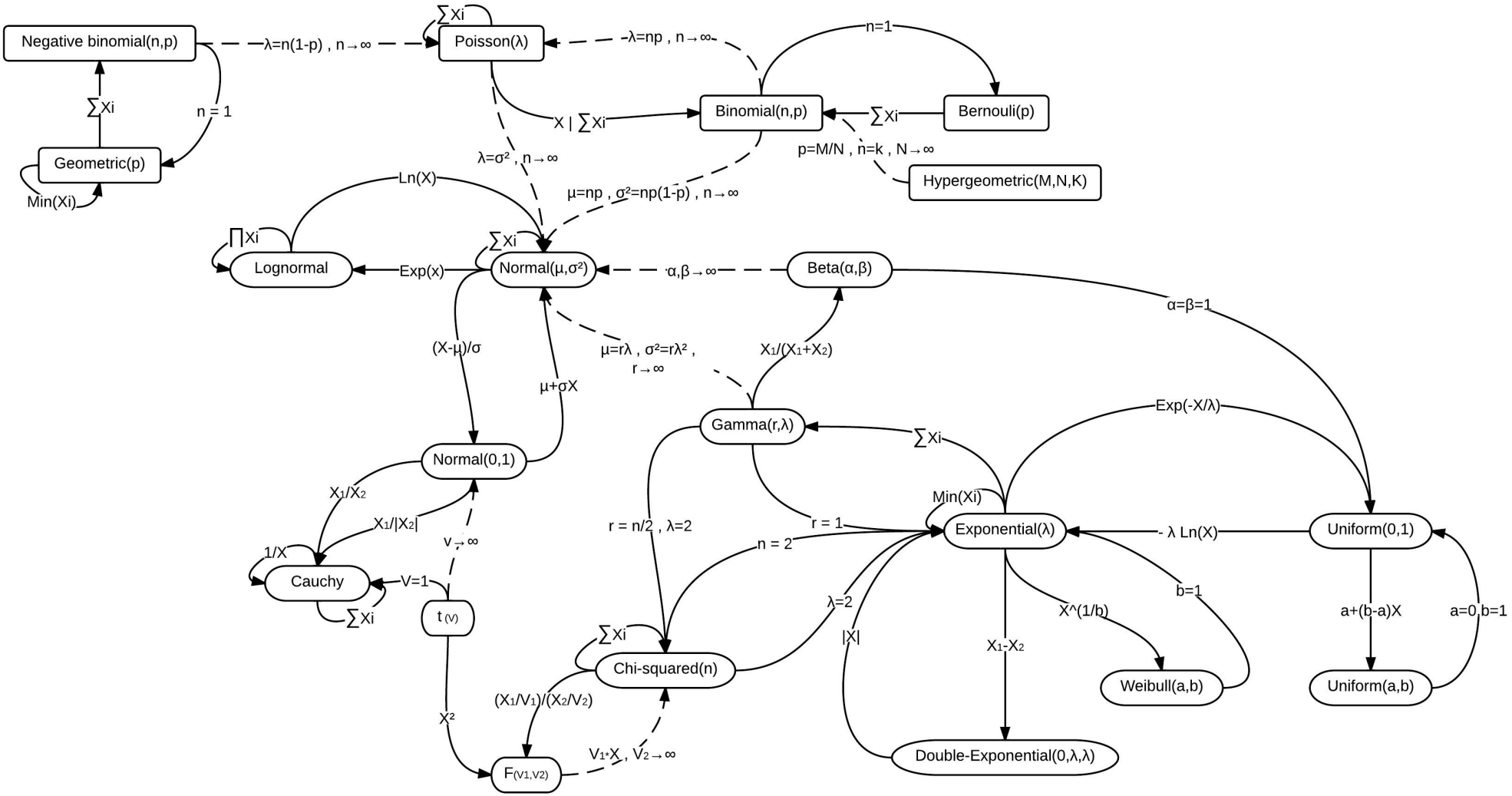
The key connection between these distributions is their **memoryless property**. The probability of an event happening in the future is independent of how much time or how many trials have already passed.

- In the exponential distribution, it's the time between events, and in the geometric distribution, it's the number of trials until the first success.





CONNECTIONS AMONG DISTRIBUTIONS



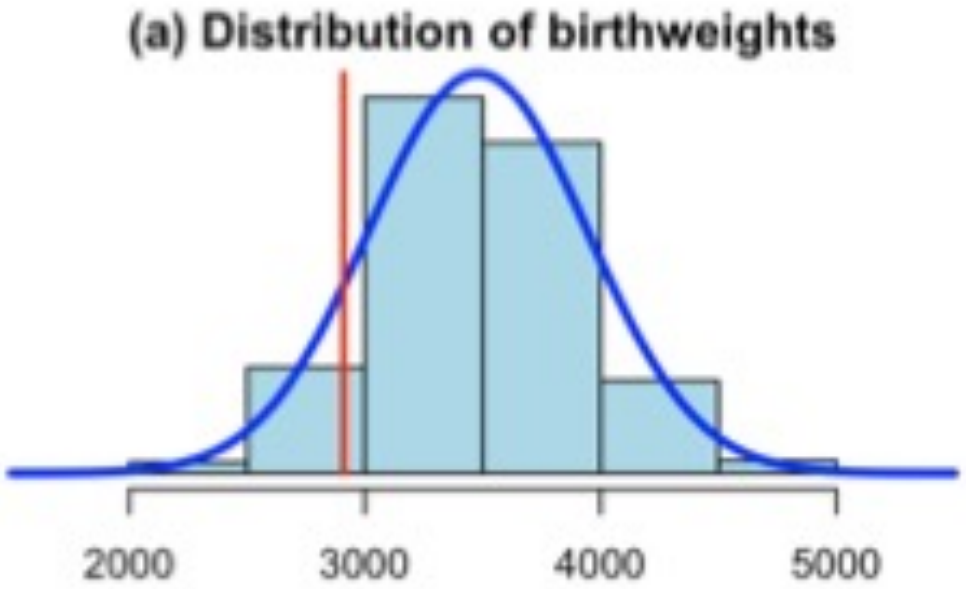
Probability and Statistics

In probability, we start with a **model** that describes how likely a random event is going to happen. We then predict the likelihood of the event happening.

In statistics, we are given data and asked what kind of model is likely to have generated it. We infer the truth or the model based on the actual data observed.

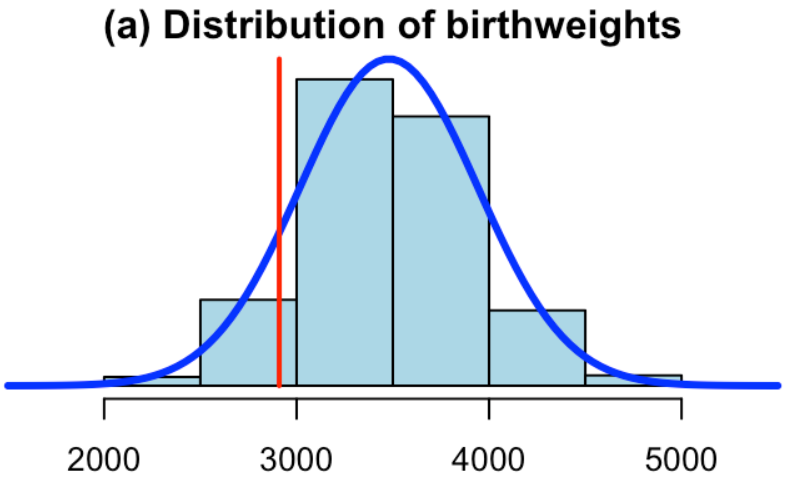
Many social phenomena show a notable regularity in their global trend (while individual events might be completely unpredictable). There might be a good correspondence between empirical data and mathematical probability distribution, the data behave as if a known random mechanism had generated them.

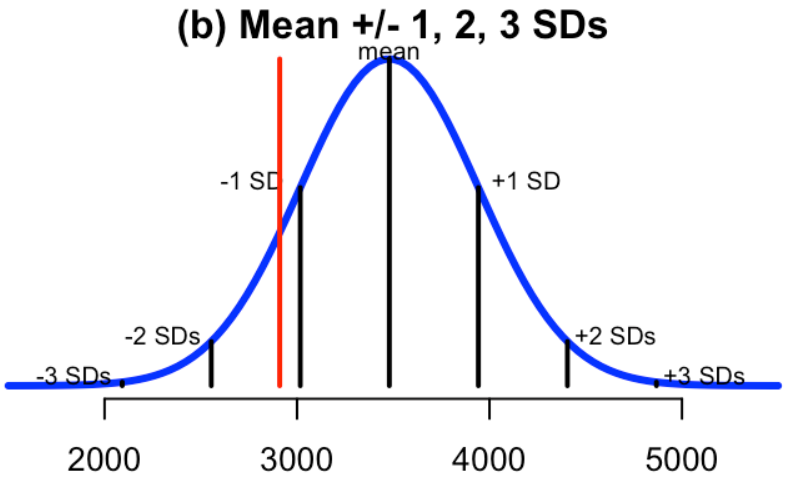
US vital statistics:1,096,277 full-term births to non-hispanic white women in the United States for 2013



This is not the complete population but the size is so large that we can approximate to the population.

We consider an American woman who had a baby of 2.91 kg (red line). This can be interpret as a sample of size 1.


(a) Distribution of birthweights
2000
3000
4000
5000


(b) Mean +/- 1, 2, 3 SDs
mean
-1 SD
+1 SD
-2 SDs
+2 SDs
-3 SDs
+3 SDs
2000
3000
4000
5000

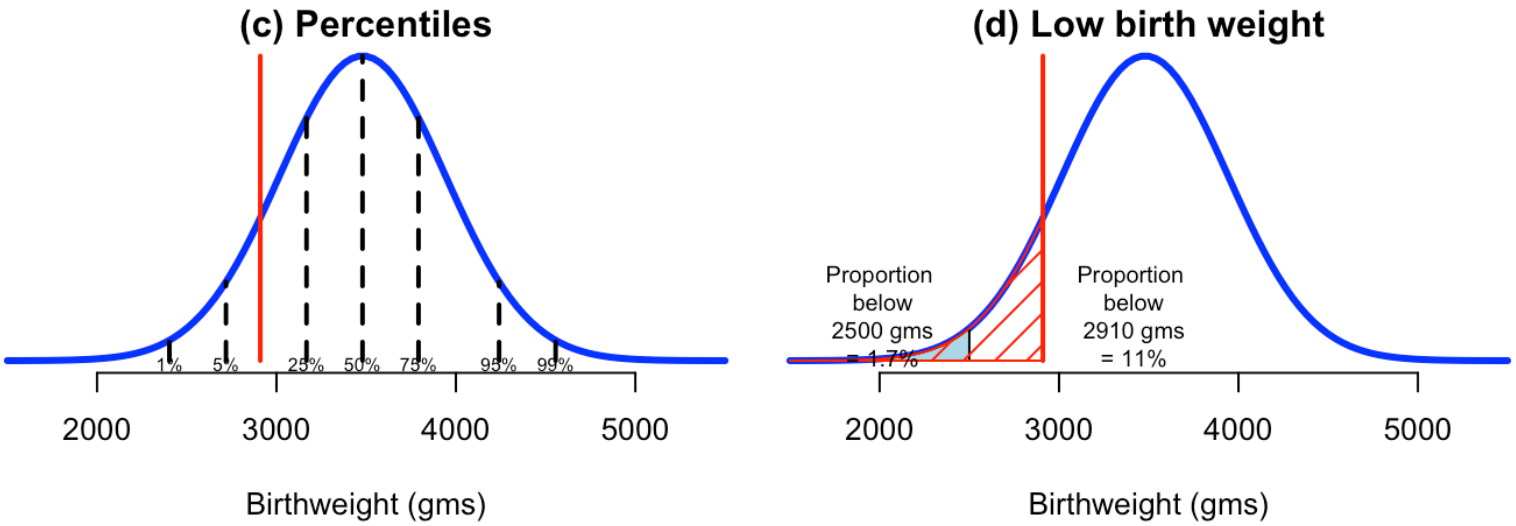
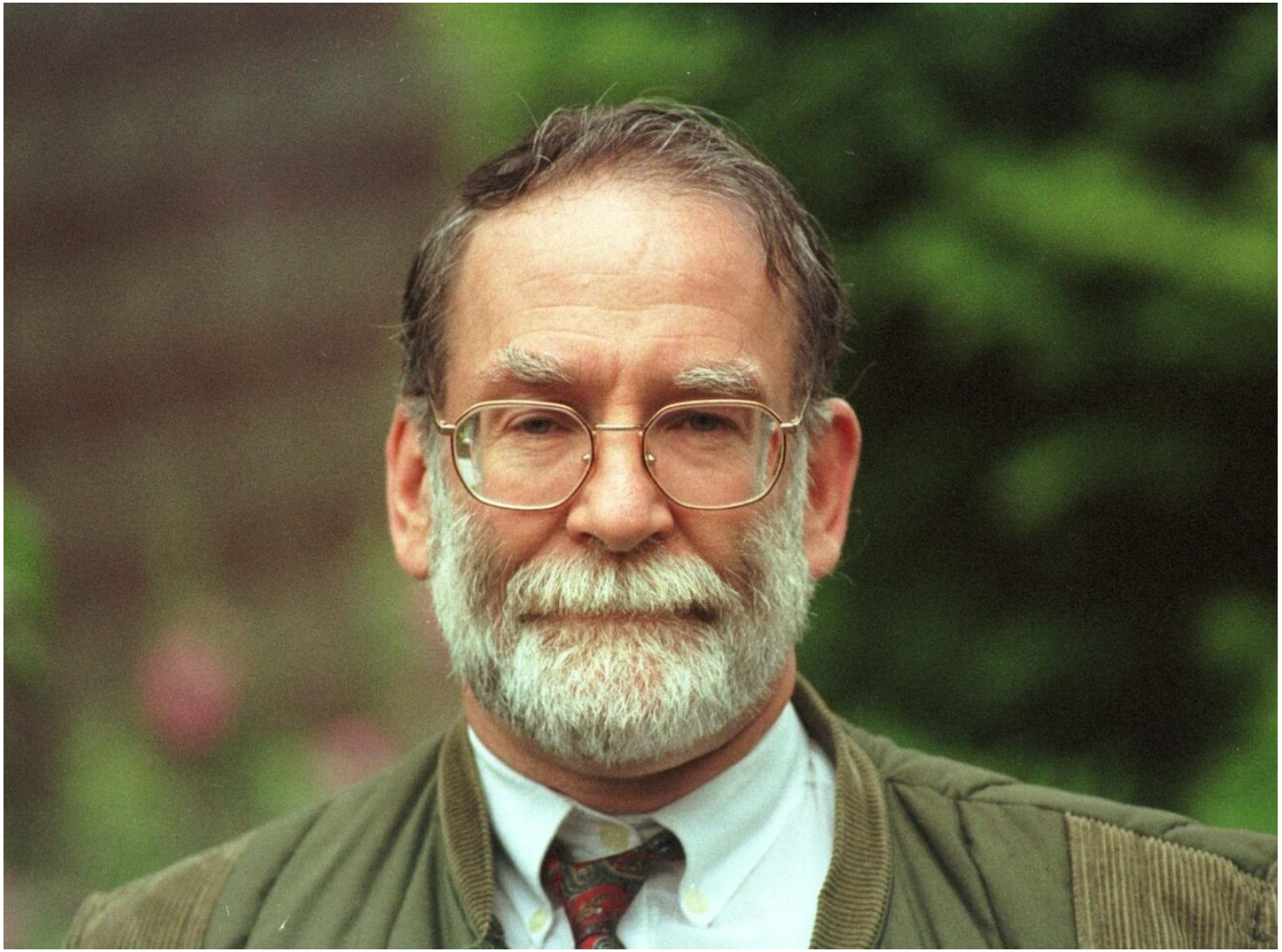


Figure (d) The proportion of low-birth-weight babies (dark shaded area), and babies less than 2,910g (light shaded area). It then represents:
- % of population of babies with a low birthweight
- probability that a randomly chosen baby (born in 2013…) weights less than 2.5 kg

The distribution is the collection of individuals but it is also a probability distribution for a randomly chosen observation!

In this example we know the population distribution and parameters (mean, standard deviation, percentiles) however, in practice, we generally do not know the population, we therefore apply the induction process.

Do you know him?



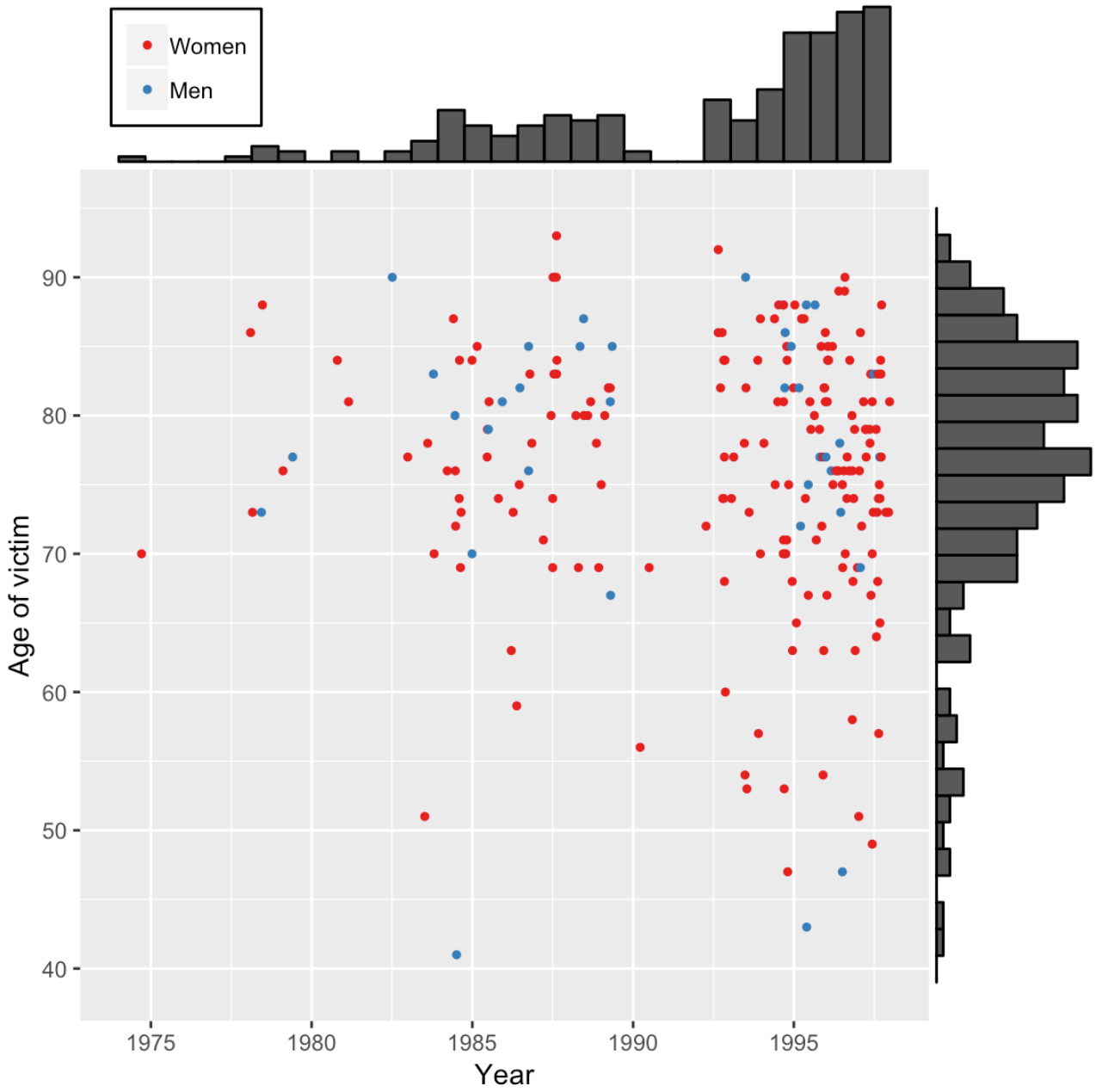
Harold Shipman

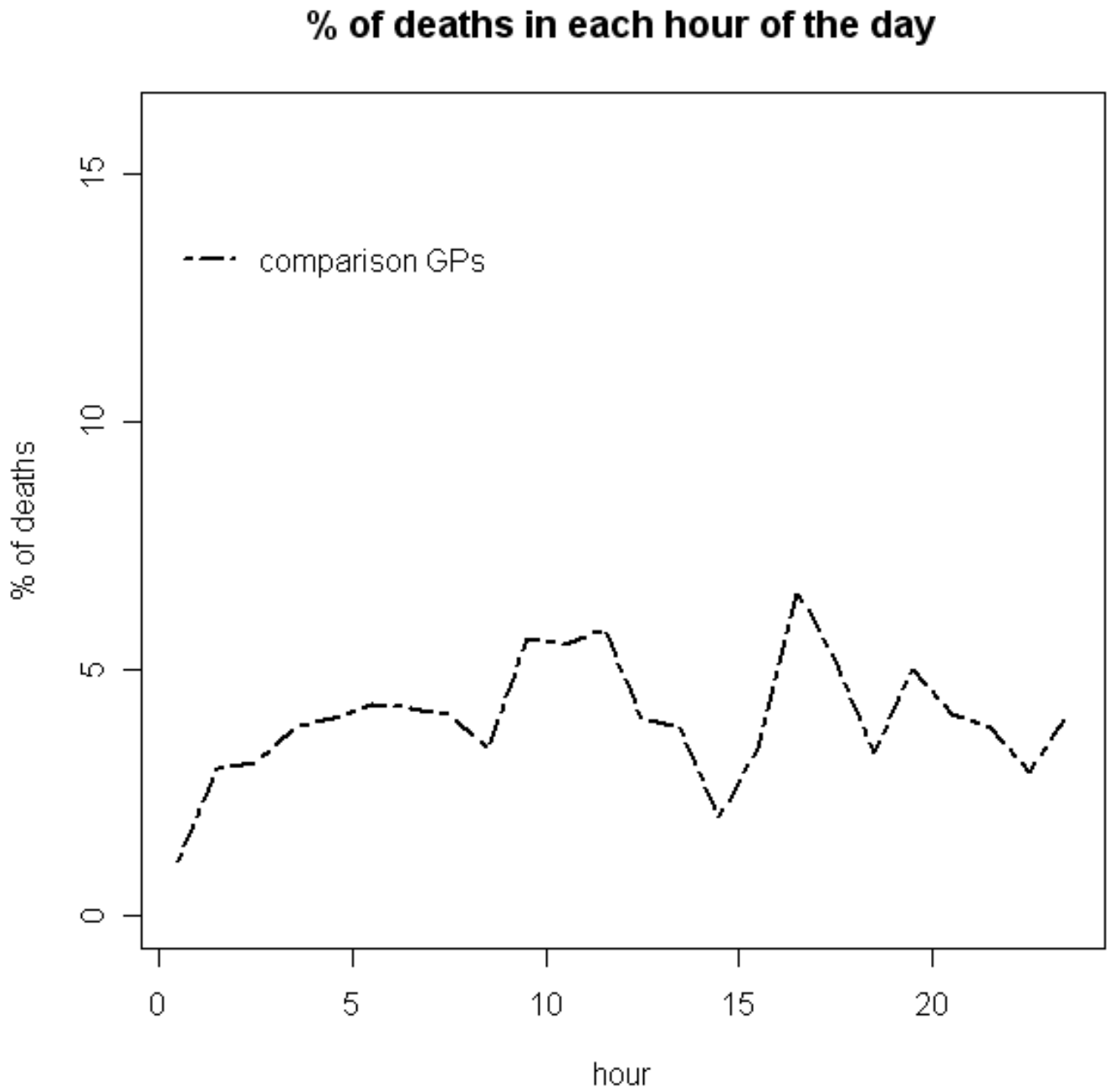
The Shipman Inquiry

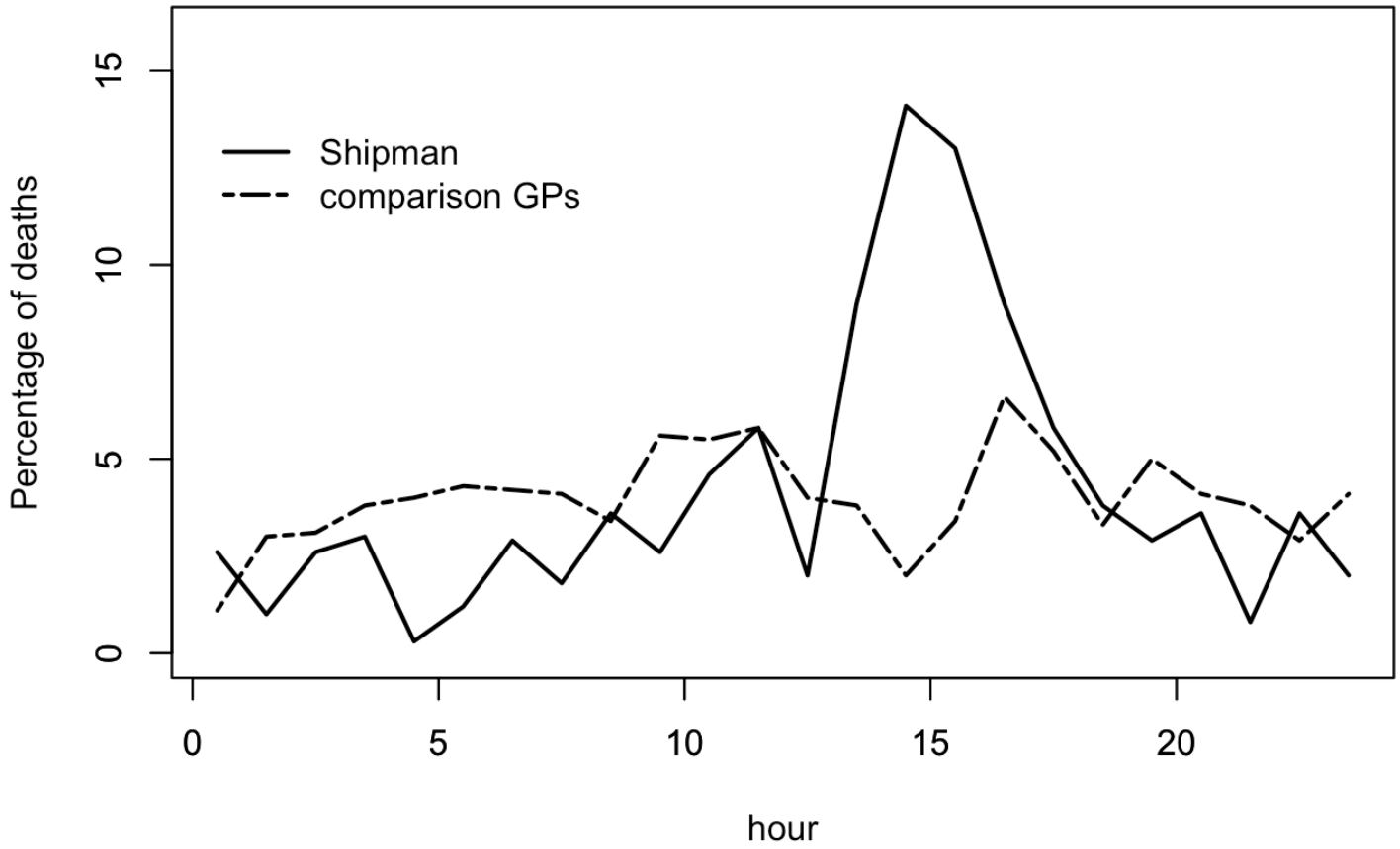
Harold Shipman had been convicted on 31 January 2000 of the murder of 15 of his patients while he was a GP.

The inquiry in July 2002 established that he had killed at least 215 people, and may have killed as many as 260, although the true number could be even higher. The analysed approximately 270,000 pages of evidence (including birth certificates).

Looking at data:
What was the pattern of Harold Shipman's murders?

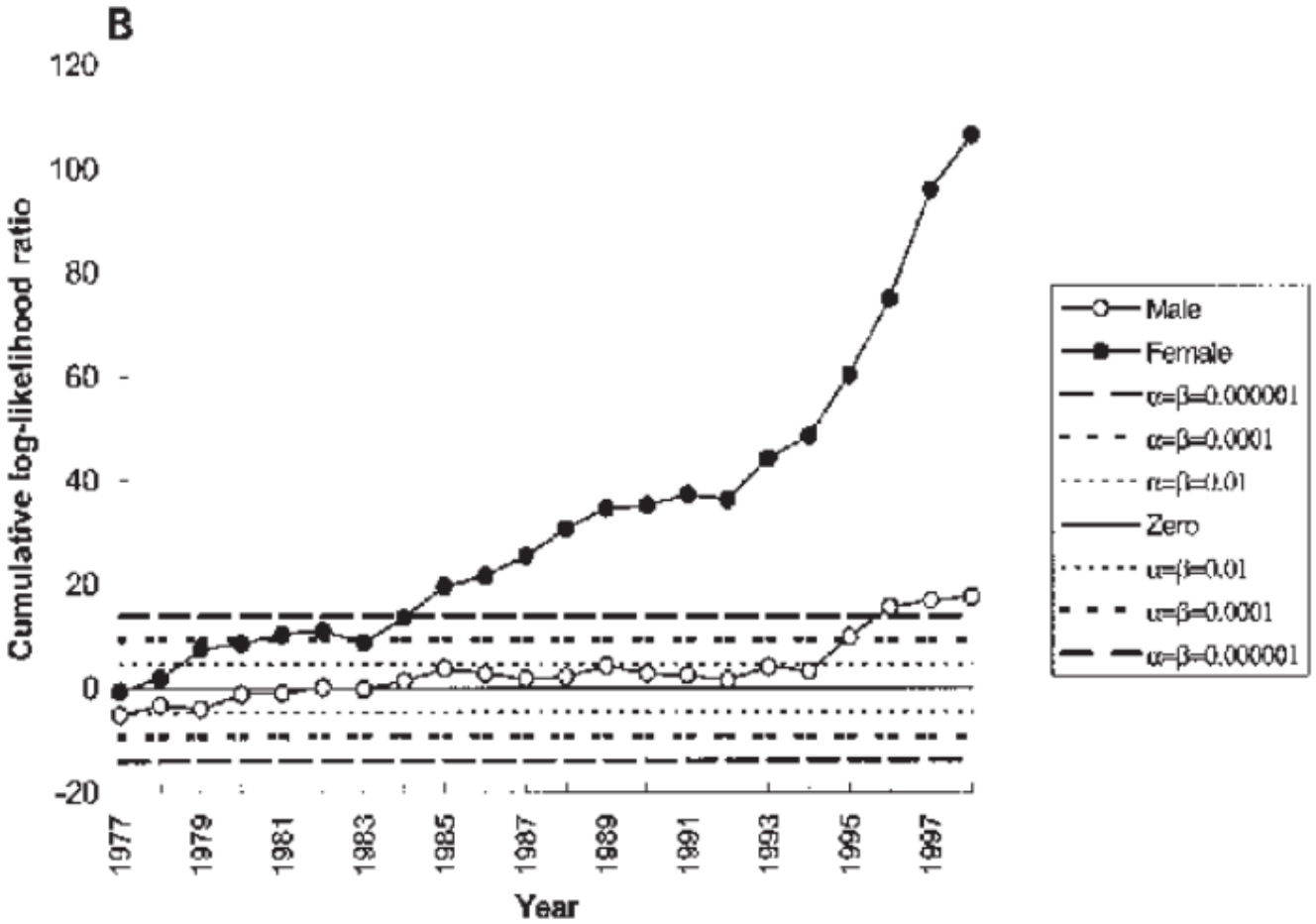

Women
Men
Age of victim
90
80
70
60
50
40
1975
1980
1985
1990
1995
Year


% of deaths in each hour of the day
comparison GPs
% of deaths
0
5
10
15
0
5
10
15
20
hour


Shipman
comparison GPs
Percentage of deaths
0
5
10
15
0
5
10
15
20
hour

Shipman Inquiry & statistical process control

FIGURE 1 - uploaded by William H. Woodall



Sequential Probability Ratio Test (SPRT) for Detection of a Doubling in Mortality Risk: Age >64 Years and Death in Home/Practice for Dr. Harold Shipman. (Figure 2(B) of Spiegelhalter et al. (2003)). Reproduced by permission of the Oxford University Press.