Introduction to Statistics

Prof.ssa Chiara Seghieri, Dott.ssa Costanza Tortù

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

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

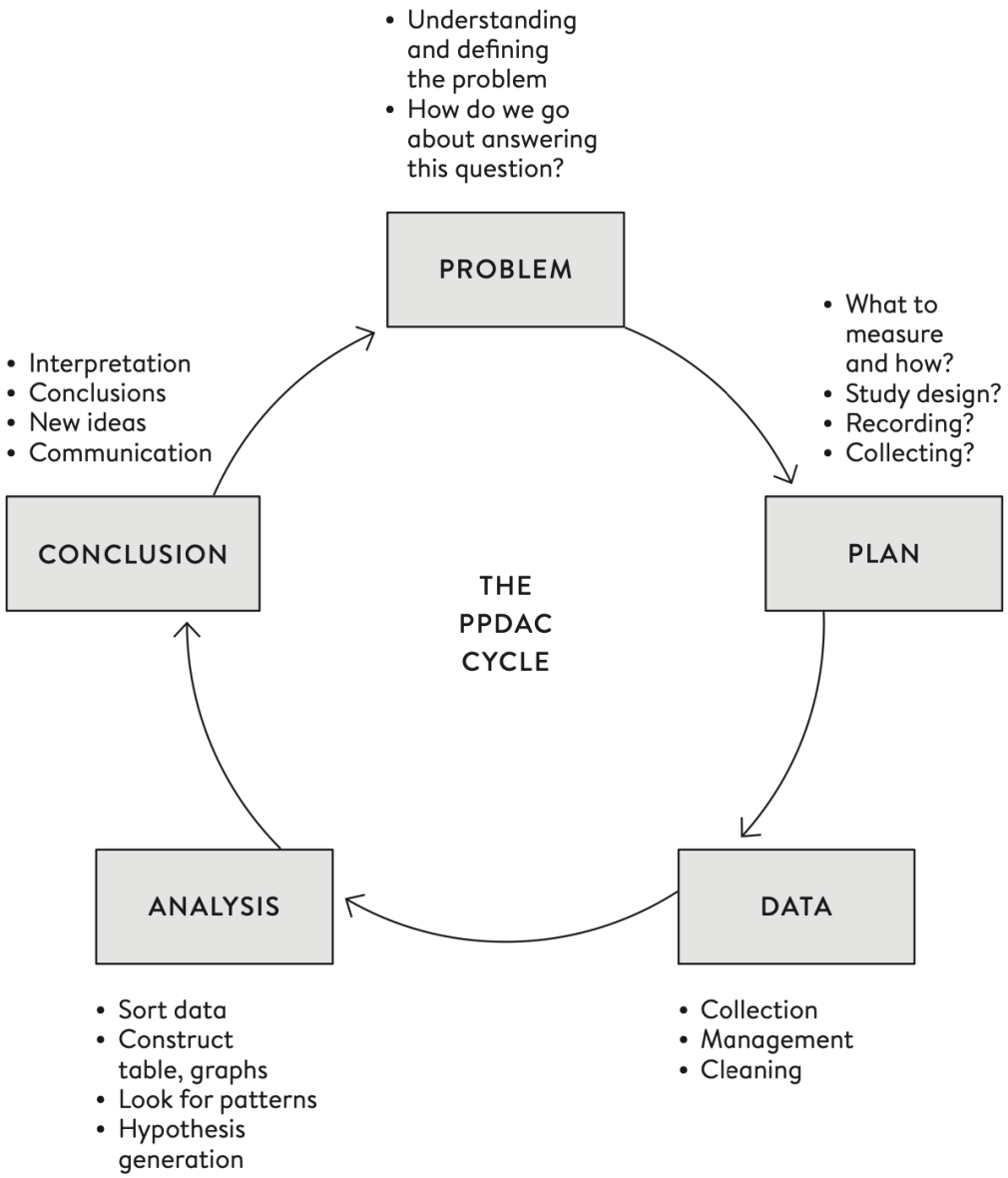
1 - What is statistics

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.

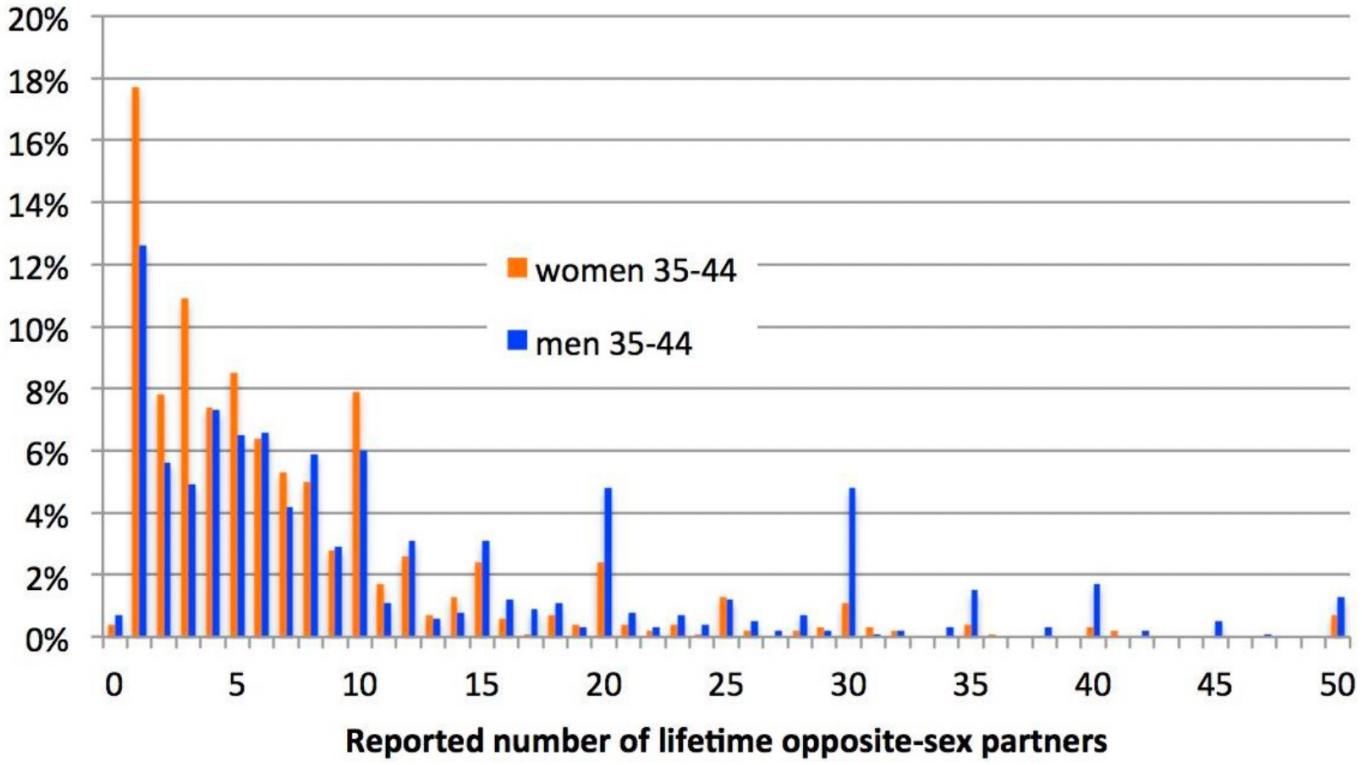
The data Cycle. Statistics helps answer real questions and support decision making



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

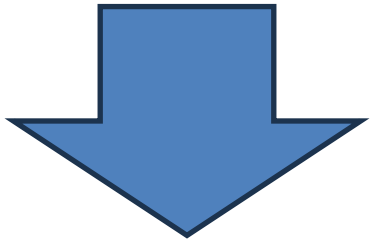
- **Problem**: cannot know this as a fact
- **Plan**: survey in which people are carefully asked about the sexual activity (Natsal)
- **Data**: reports of numbers of partners
- **Analysis**: plotting and summary statistics

How many sexual partners do people report?



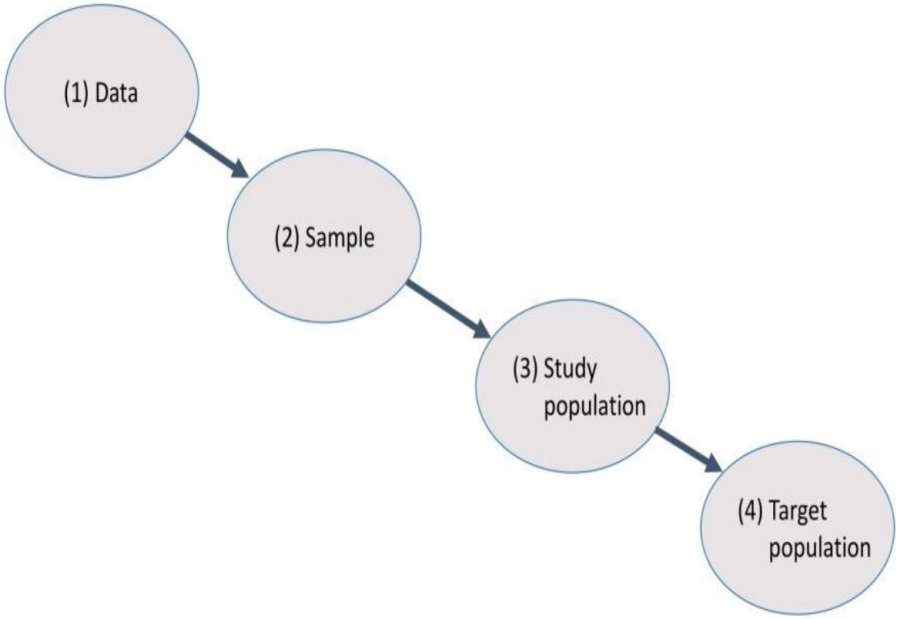
Reported number of sexual partners in lifetime	Men aged 35–44	Women aged 35–44
Mean	14.3	8.5
Median	8	5
Mode	1	1
Range	0 to 500	0 to 550
Inter-quartile range	4 to 18	3 to 10
Standard deviation	24.2	19.7

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**

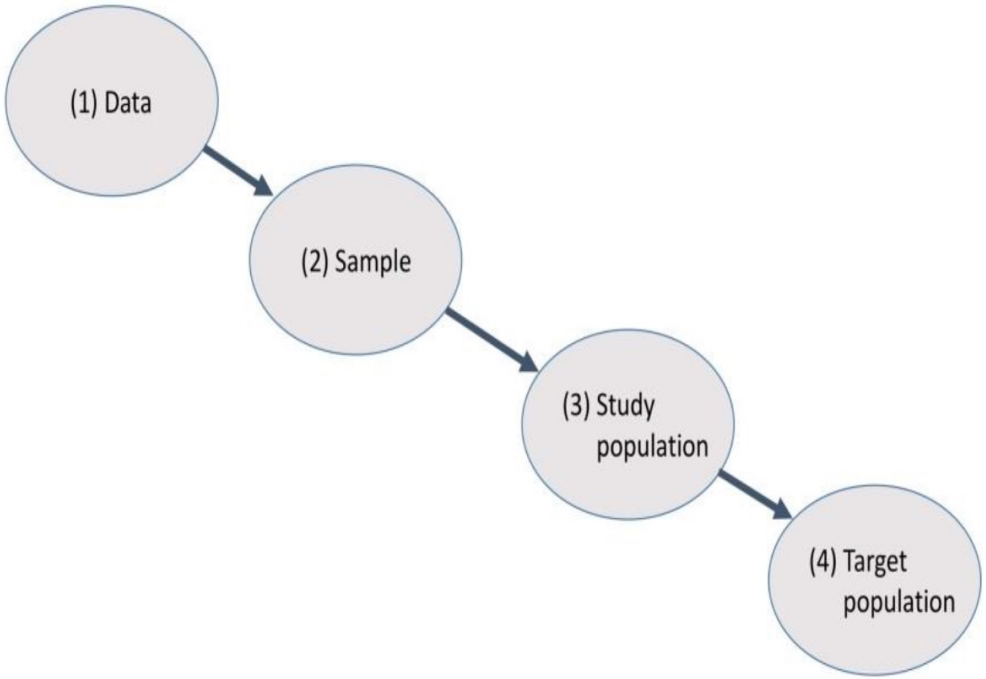
Inference and bias

*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?????

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*

2 – Types of studies

Type of Studies

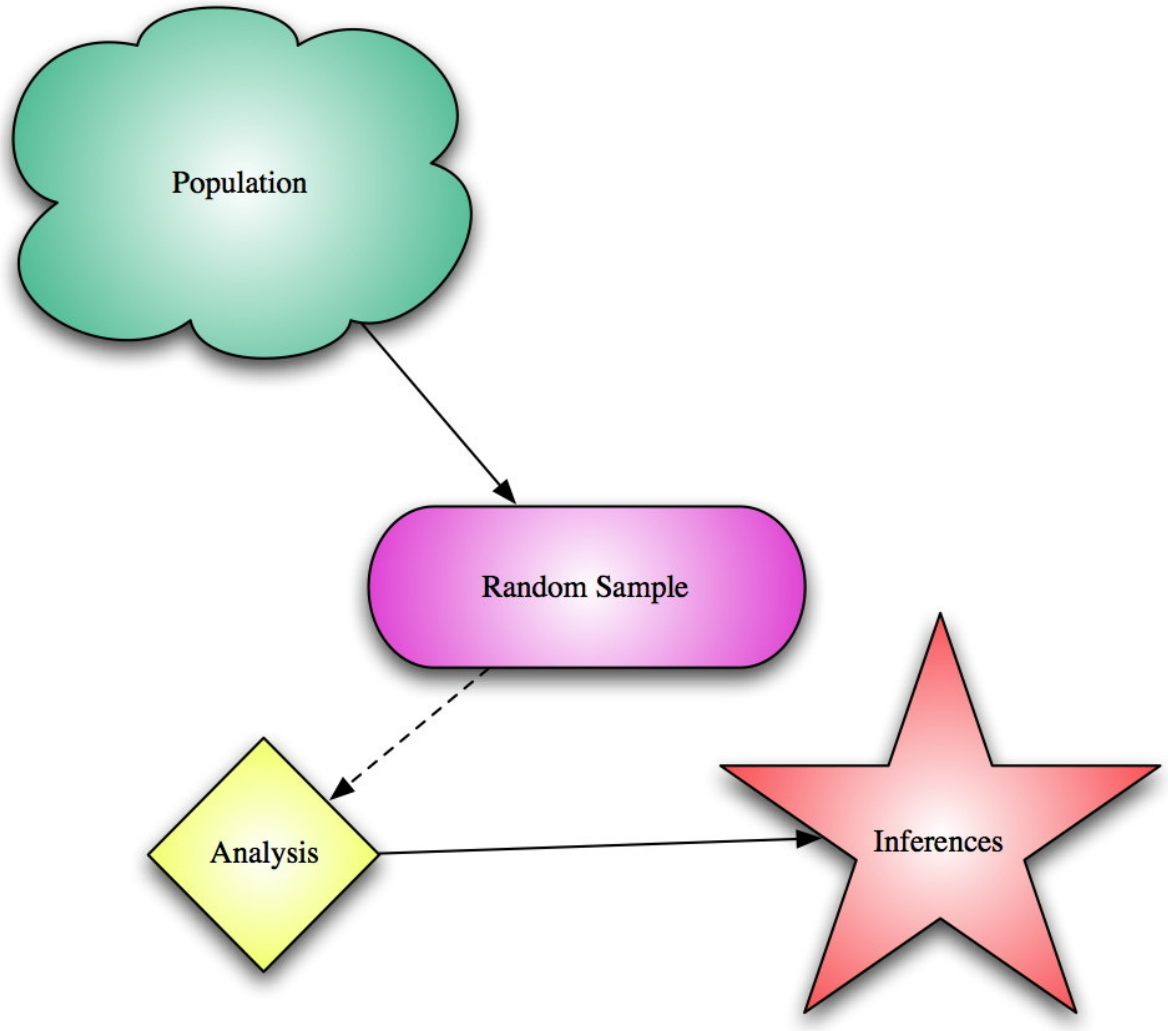
There are two primary types of data collection: **observational studies** and **experiments**.

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, without having any other information on individuals.

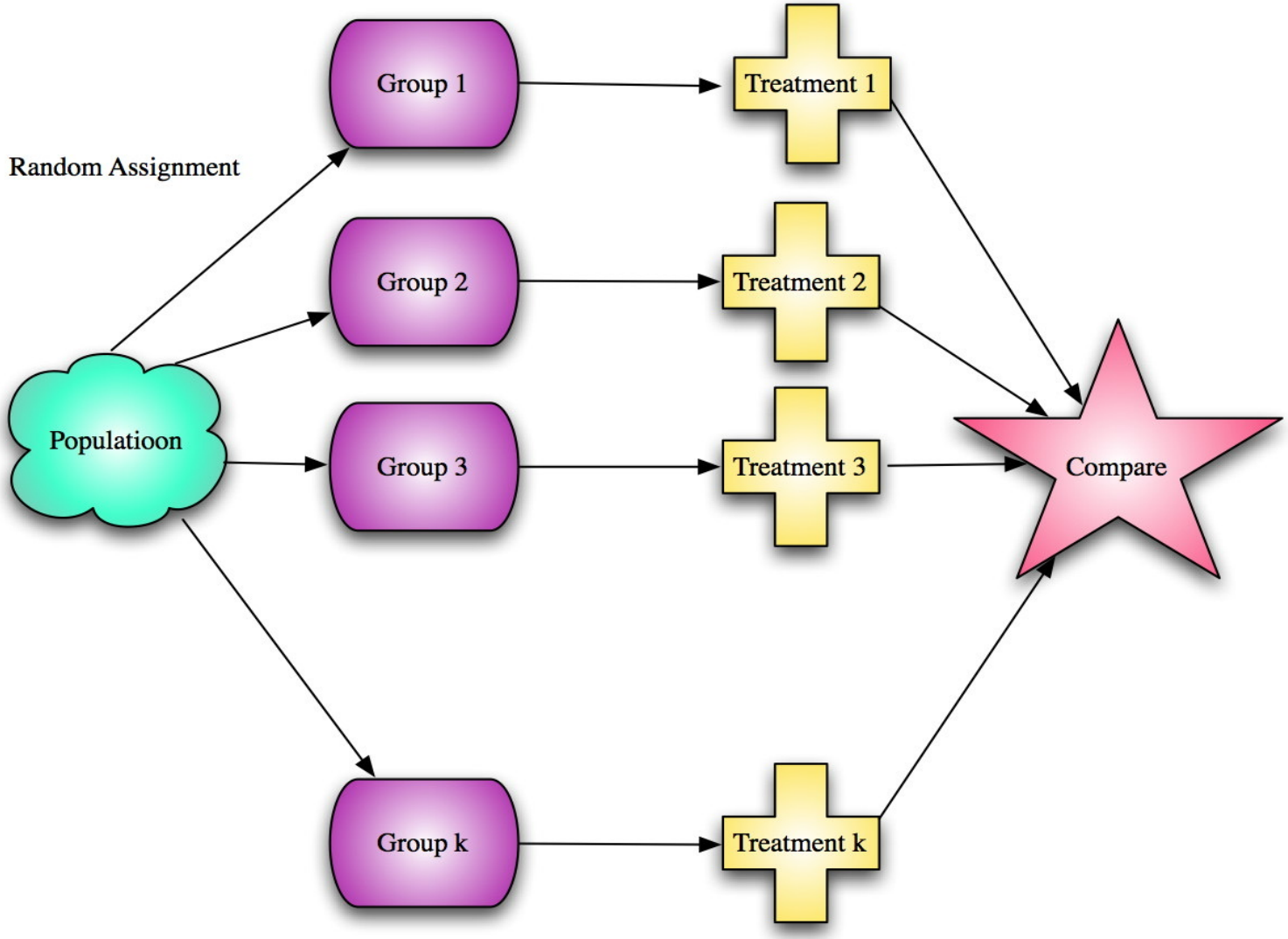
When researchers want to establish a causal connection, they try conduct an experiment – if it is possible! - .

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 randomization reasonably guarantees that treated and untreated patients have similar characteristics

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.

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

There are many other important questions that can be studied only using observational studies.

Does smoking cause lung cancer? Is a new medication for treating migraine headaches more effective than the current treatment that doctors most often prescribe?

To extract a causal information from observational studies researchers must rely on pre - treatment characteristics and try to make their framework **as good as random**

3 – Introduction to sampling theory

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.

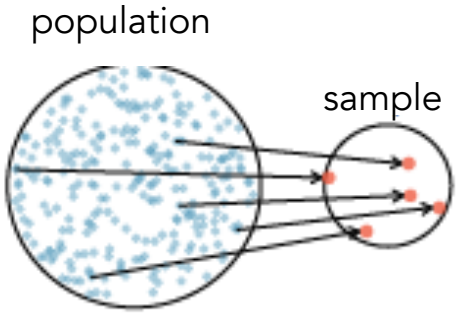
✓The only guarantee for that comes from the method used to select the sample (**sampling method**)→**probability sampling**

✓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.

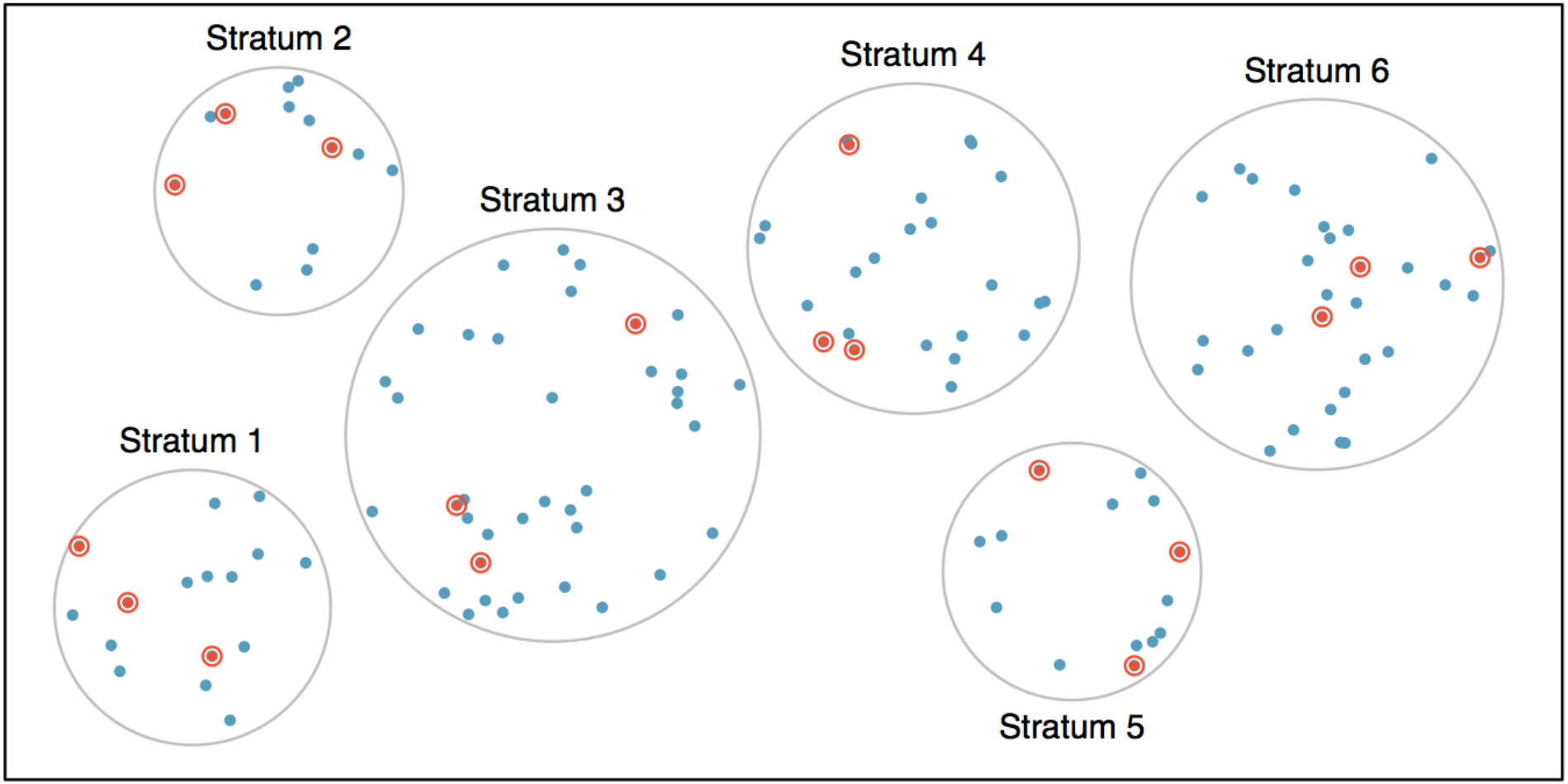
Stratified random sampling

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

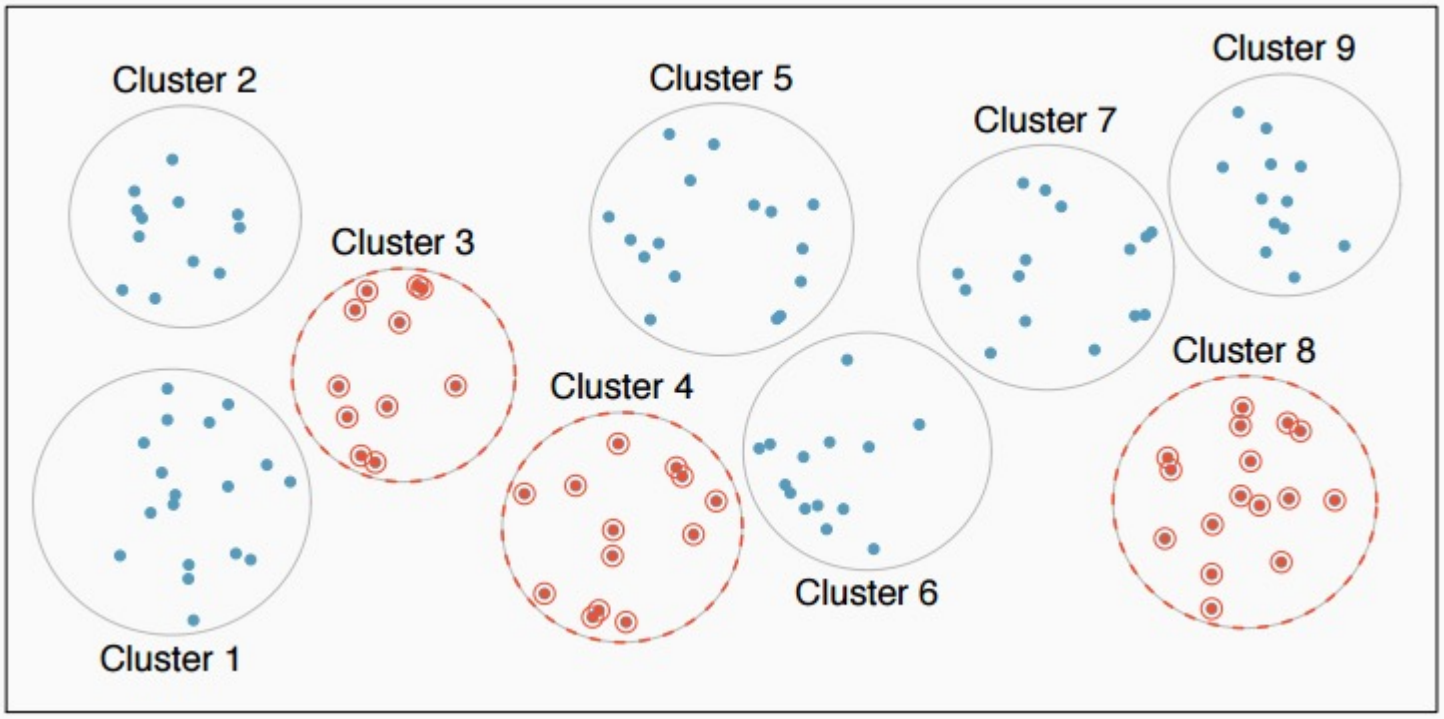
we break up the population into many groups (usually naturally occurring groups like municipalities, classes, hospitals,…), called clusters.
Then we sample a fixed number of clusters and include all observations from each of those clusters in the sample.

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.

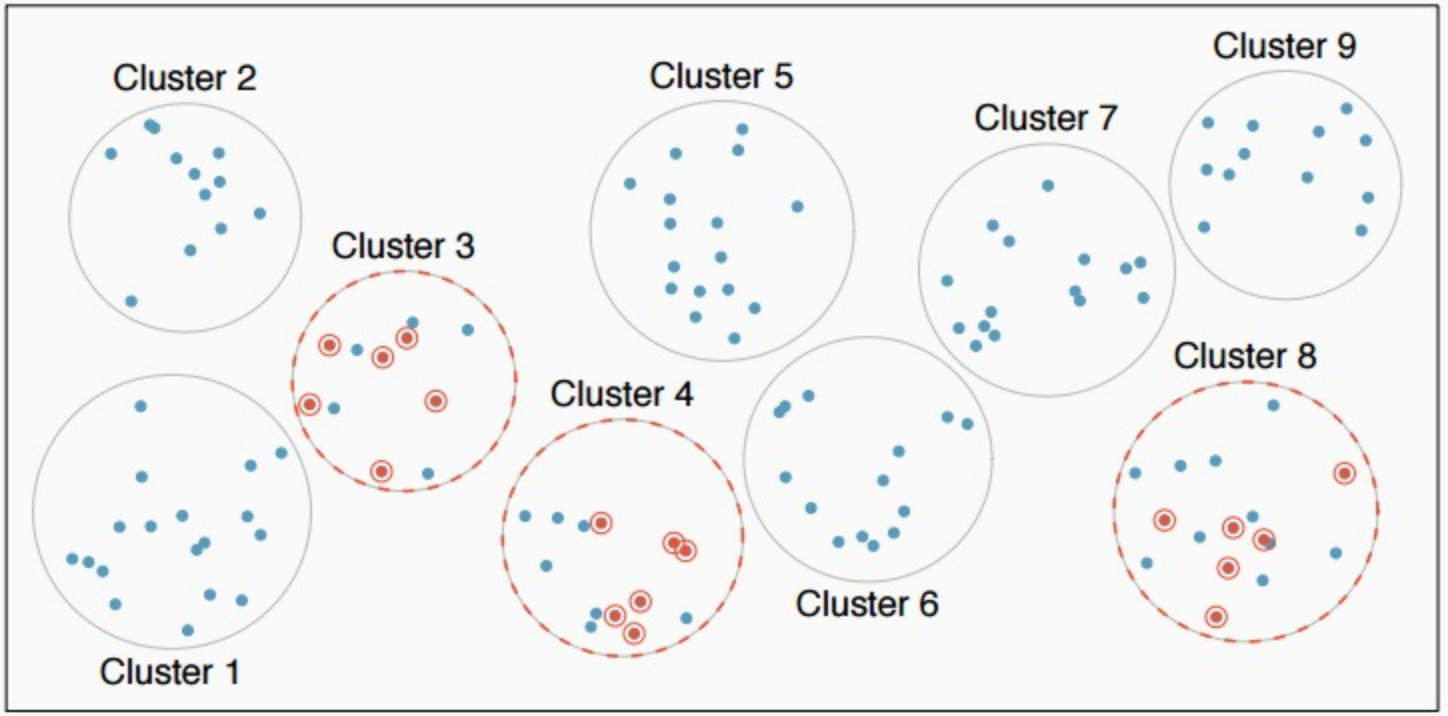
Cluster Sample

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:

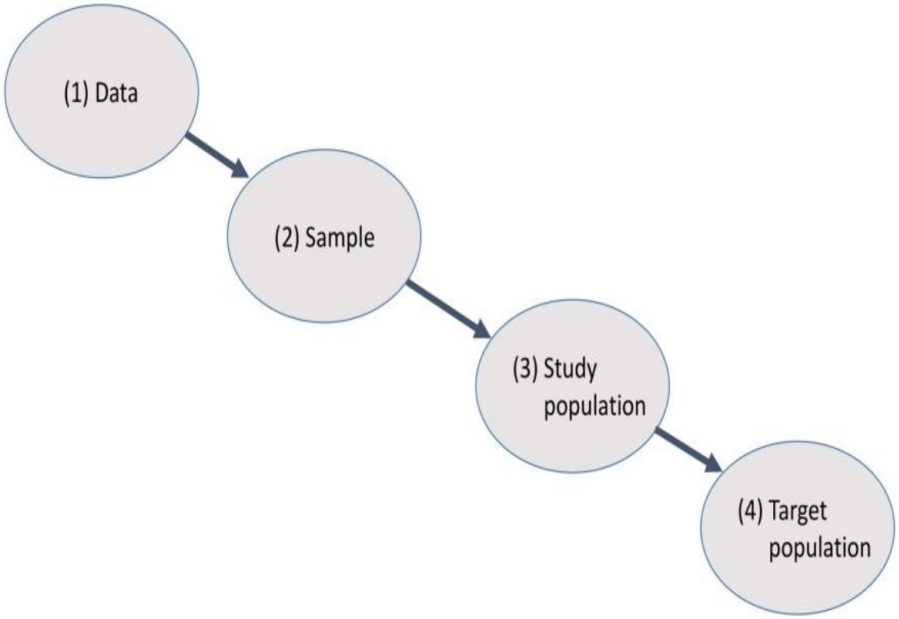
we are interested in estimating the malaria rate in a rural area of India. There are 60 villages in that area each more or less similar to the next. Our goal is to test 300 individuals for malaria. What sampling method should be employed?

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.

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.

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.

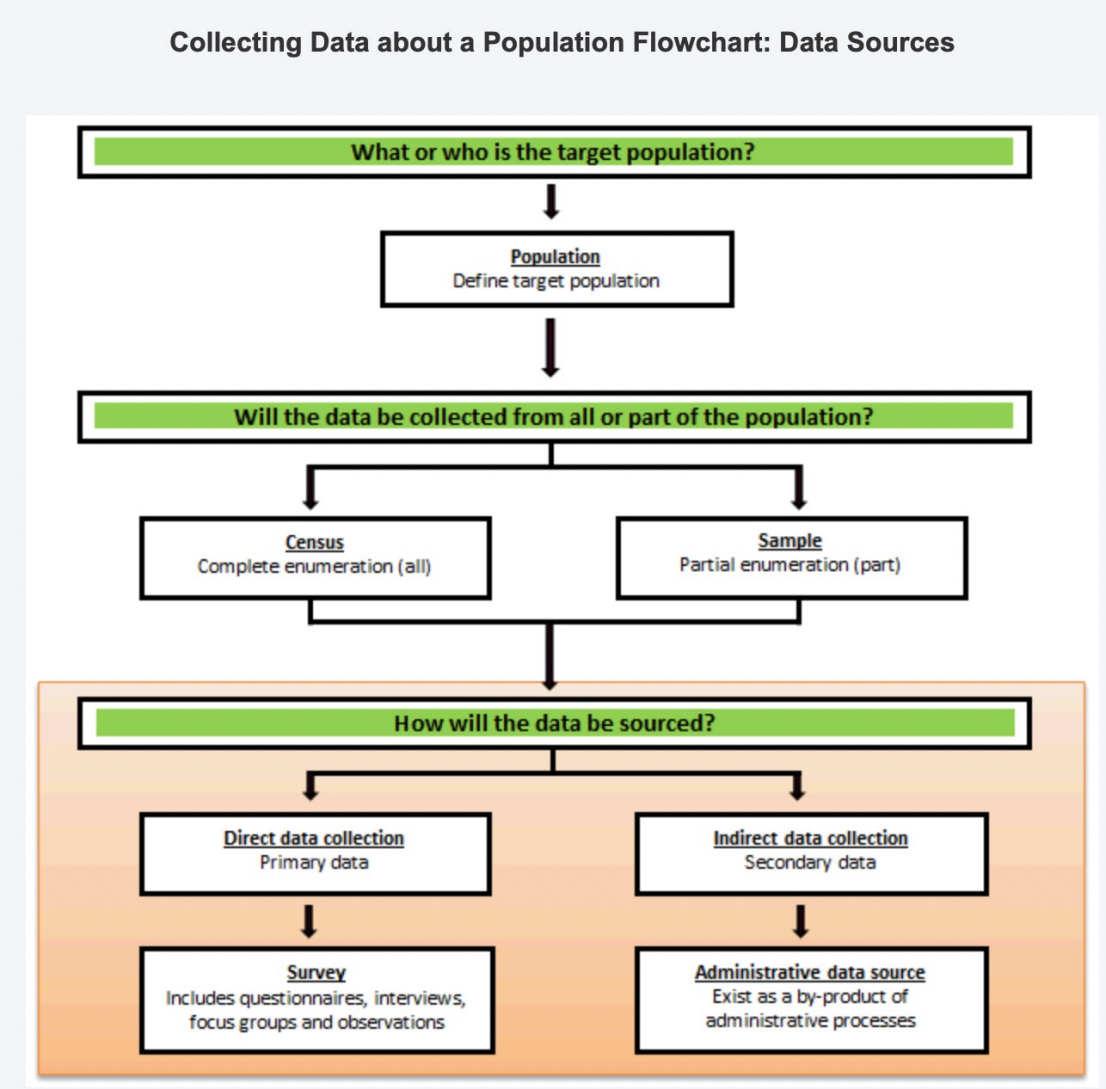
Category	Excellent/Very Good/ Good	Excellent	Very Good	Good	Fair	Ok
Overall Experience	92%	43%	35%	14%	4%	4%
Boarding	86%	39%	30%	17%	7%	7%
Crew Friendliness	95%	55%	29%	11%	3%	2%
Service onboard	93%	45%	32%	16%	4%	3%
Food & Drink Range	82%	24%	26%	32%	10%	8%

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'.

Let's concentrate on observational studies



Australian Bureau of Statistics

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.

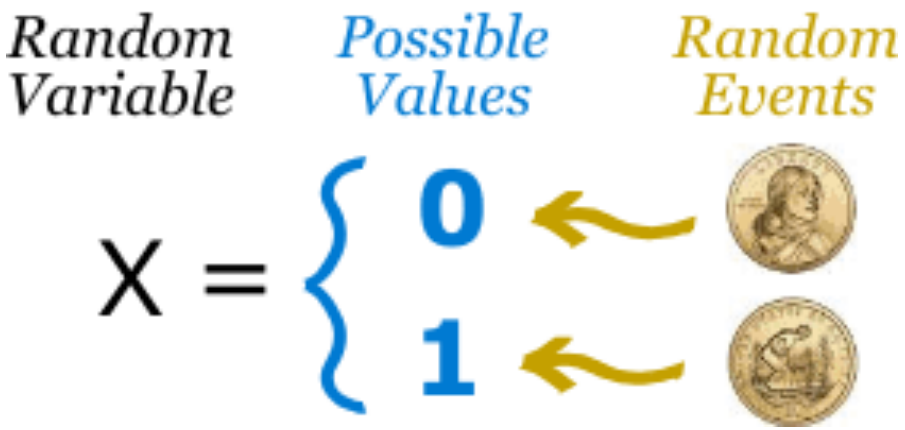
5 - Probability and population

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.

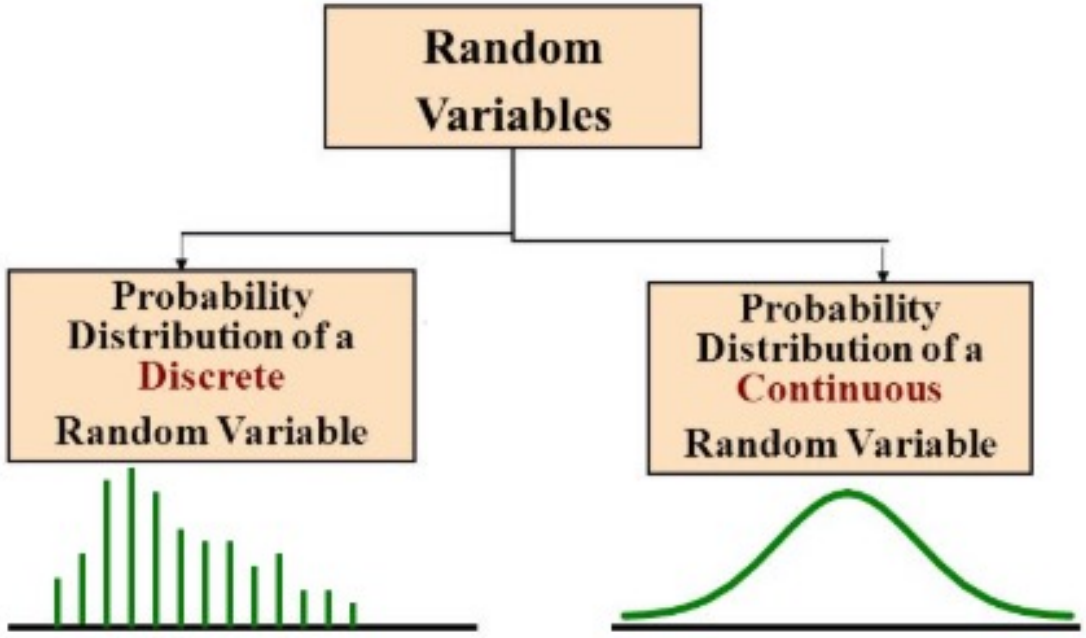
Continuous random variables have an infinite continuum of possible values.

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

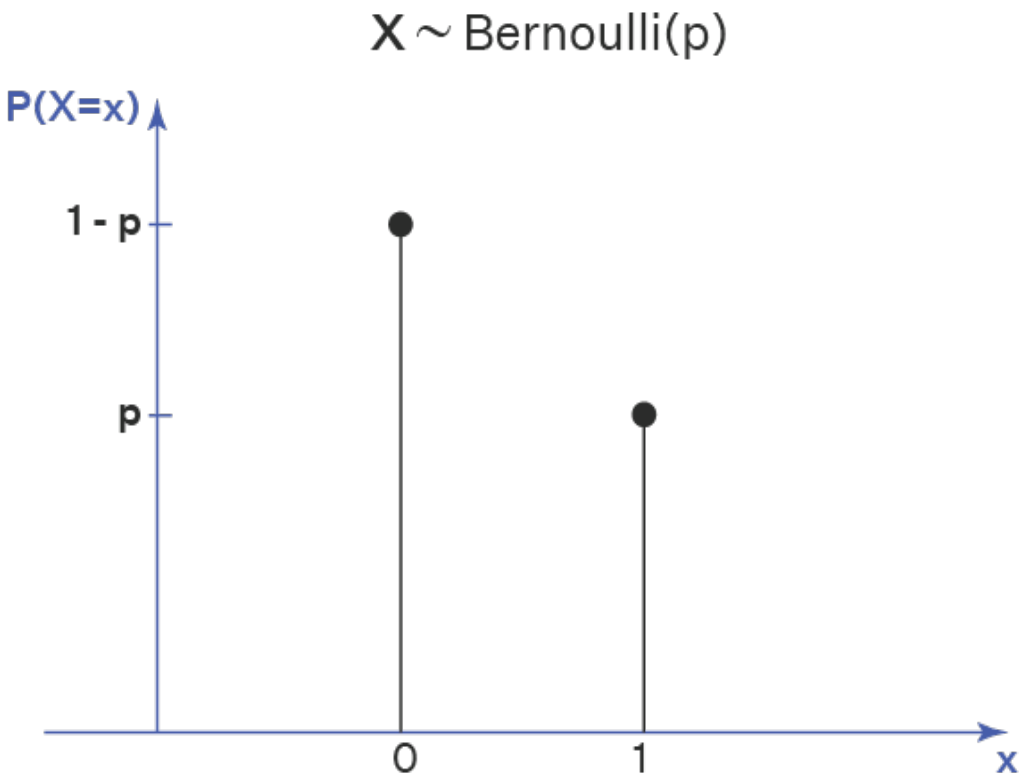
X	2	3	4	5	6	7
P(X)	0.413	0.236	0.211	0.090	0.032	0.018

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.



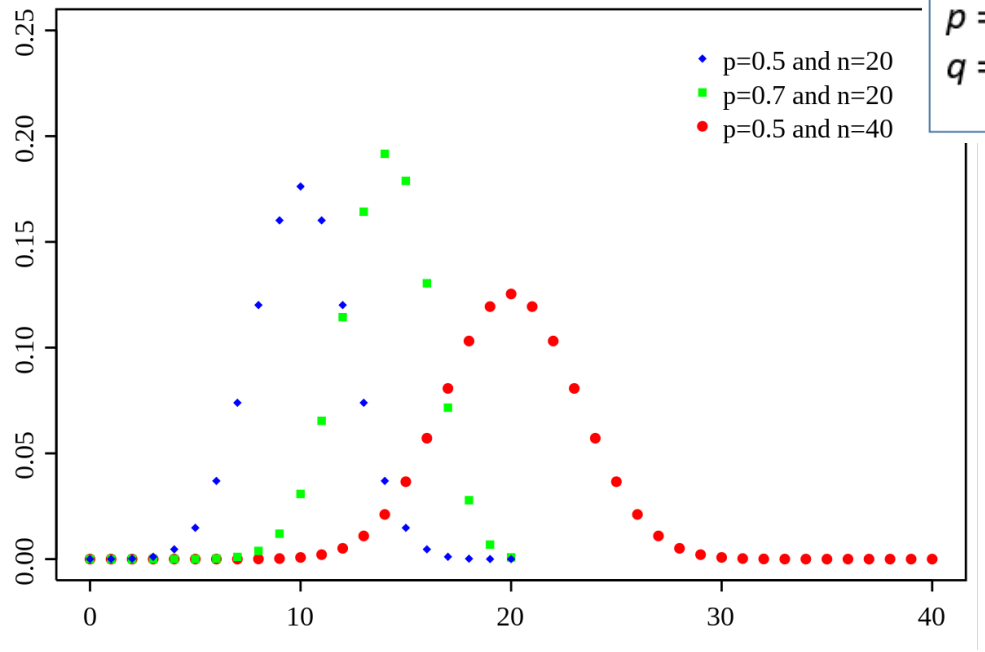
B – BINOMIAL DISTRIBUTION

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

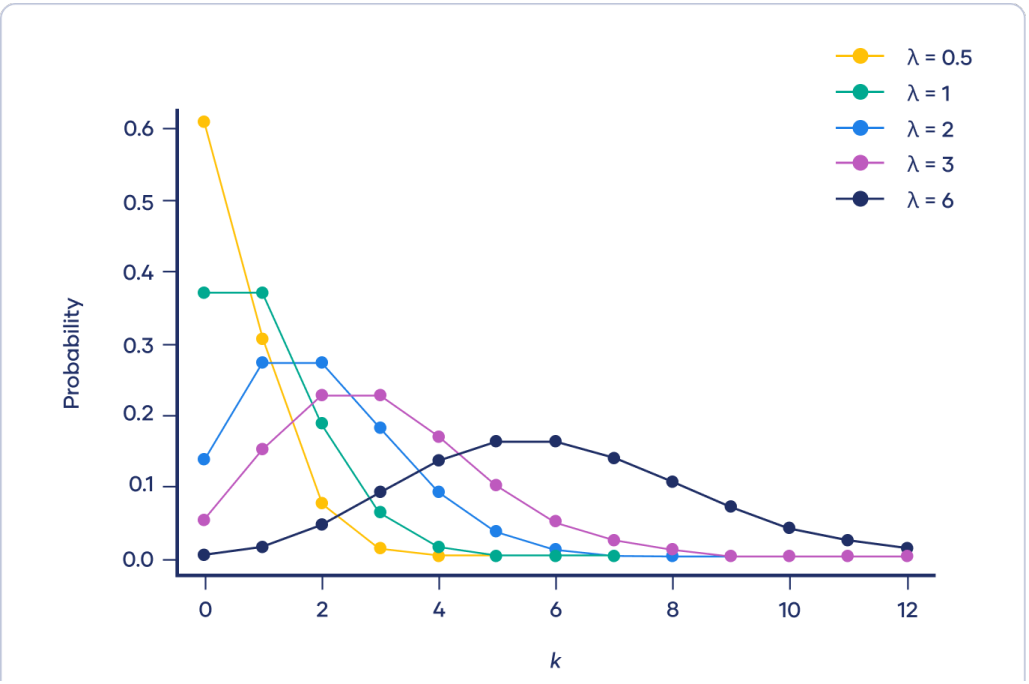
$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

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

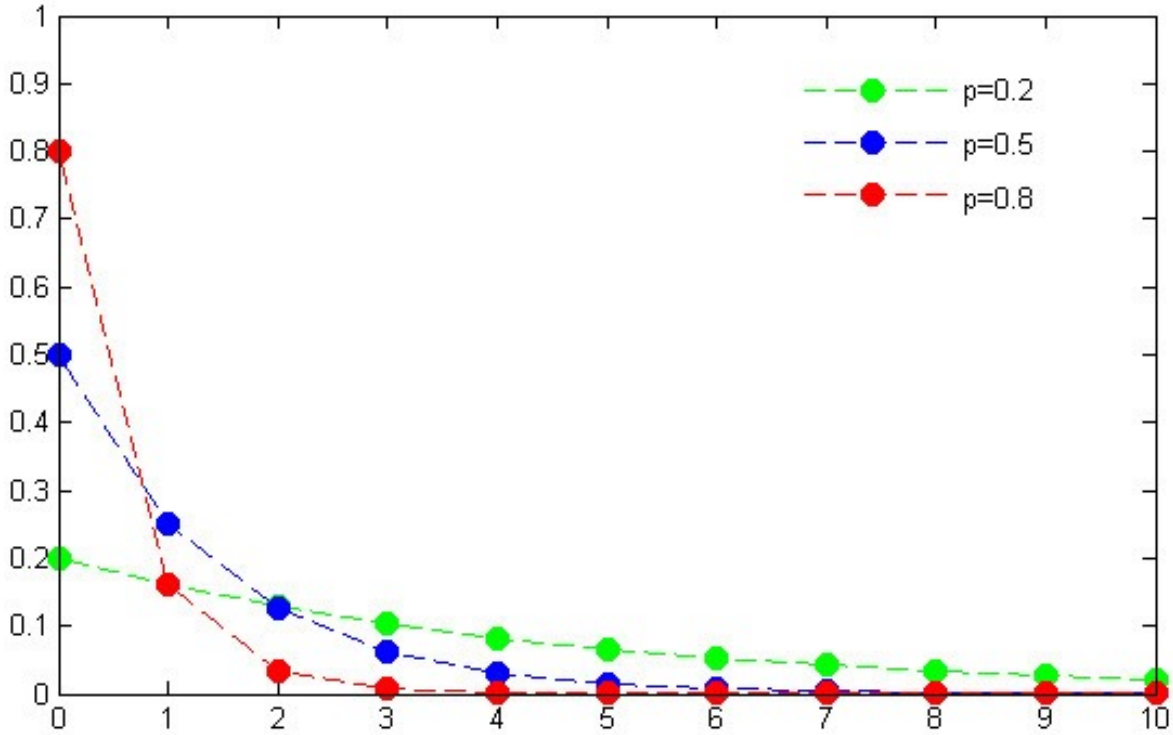
$$P(X = x) = (1-p)^{x-1}p$$

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.

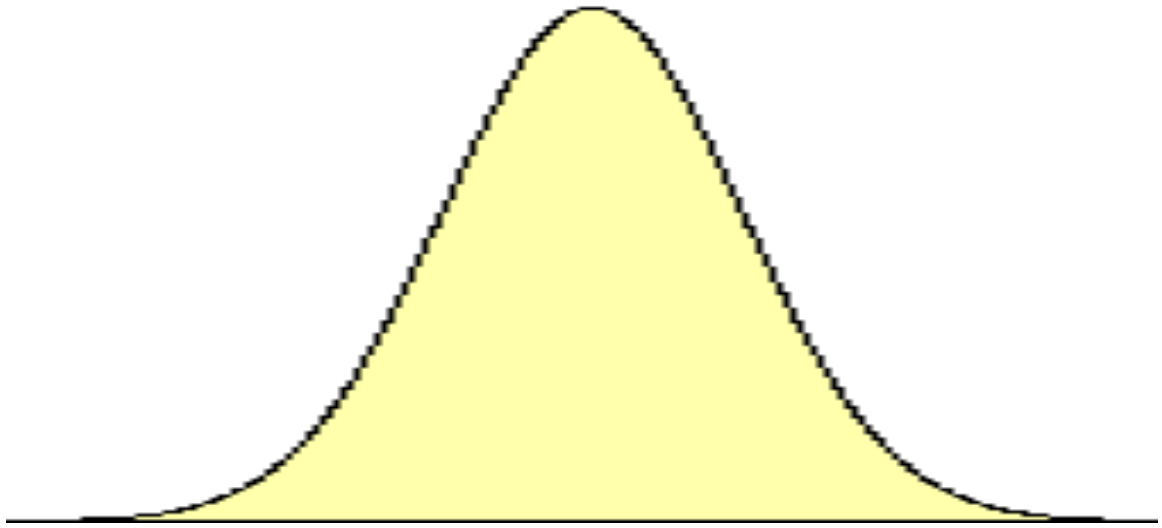
The parameter p measures the probability of success of a single attempt.

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:

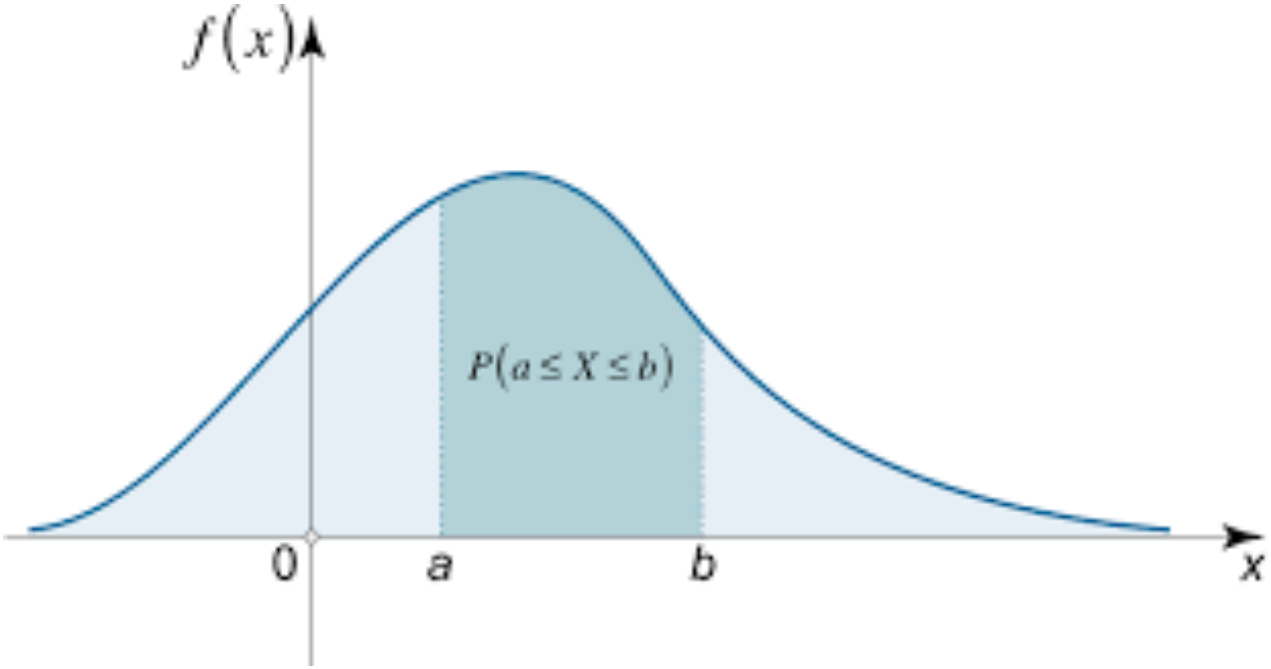


Continuous case

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%).

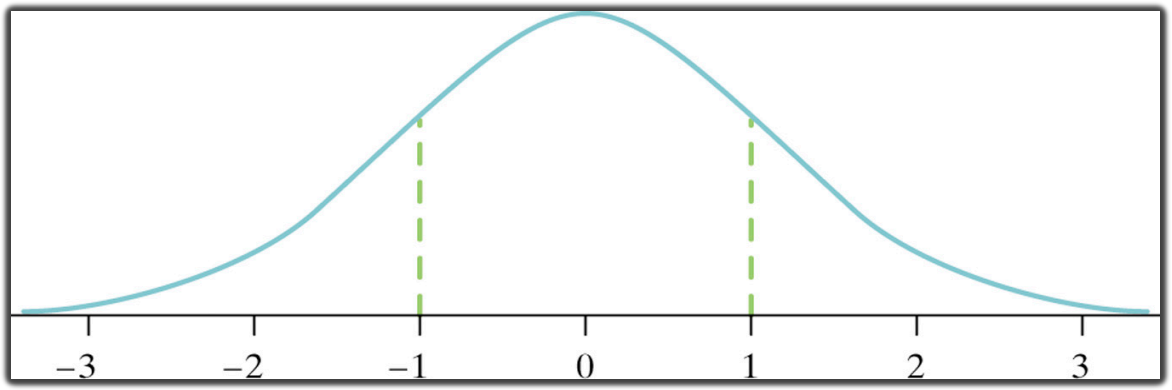
The probabilities associated with continuous functions are just areas under the curve (integrals!).



A - The Standard Normal Distribution

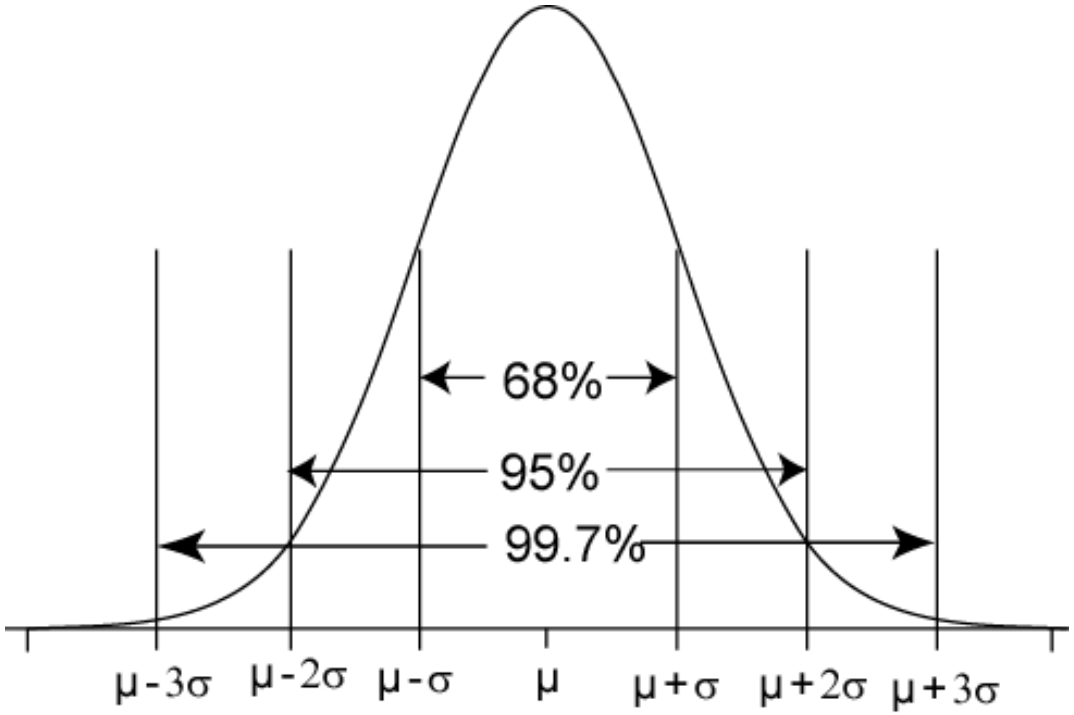
- The **standard Normal distribution** is the Normal distribution with mean 0 and standard deviation 1.
- Shown as ***N***(0,1)
- If a variable *x* has any Normal distribution $N(\mu, \sigma)$, with mean μ and standard deviation σ, then the standardized variable

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



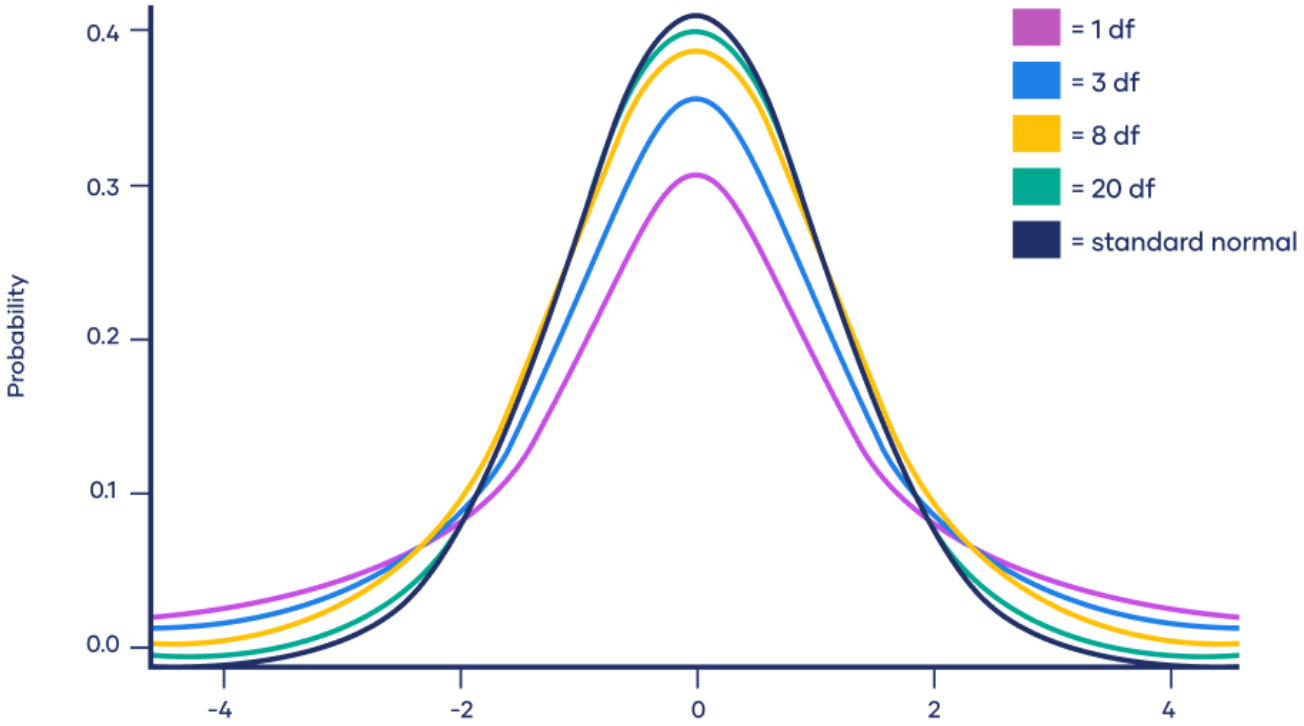
68-95-99.7 Rule for Normal Distributions

- 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

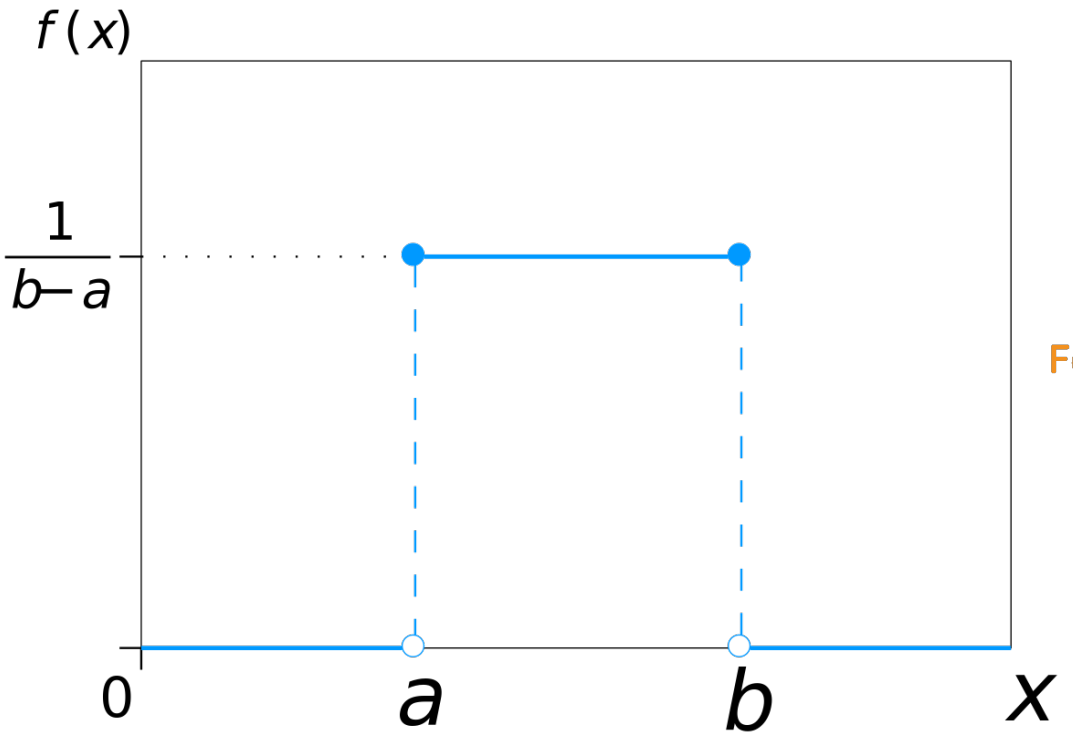
The **T-Student distribution** is particularly used for making inferences about population means when the sample size is small or when the population standard deviation is unknown.



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.



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

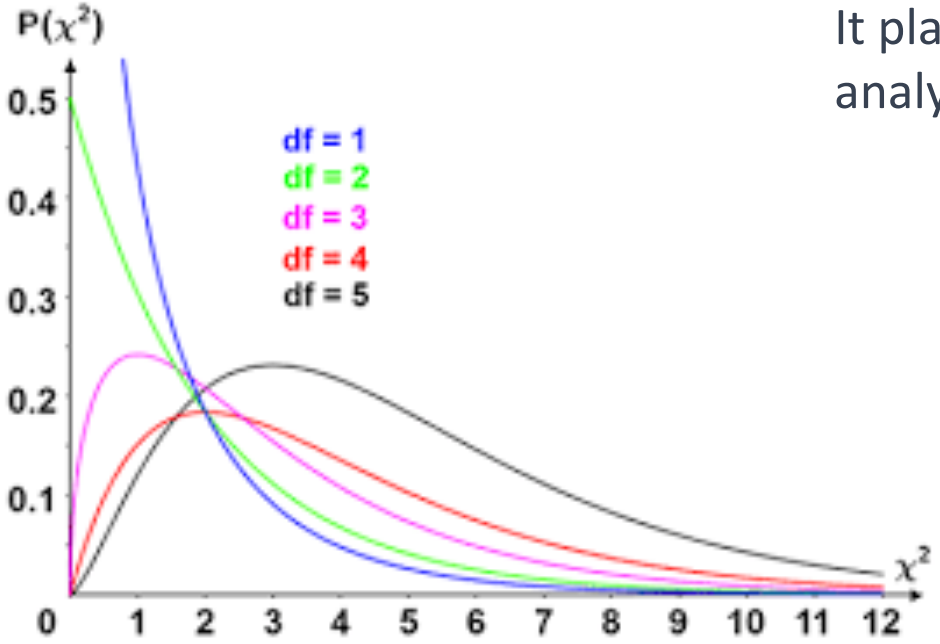
$$\text{Mean} = \frac{(a + b)}{2}$$

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

D – CHI-SQUARE DISTRIBUTION

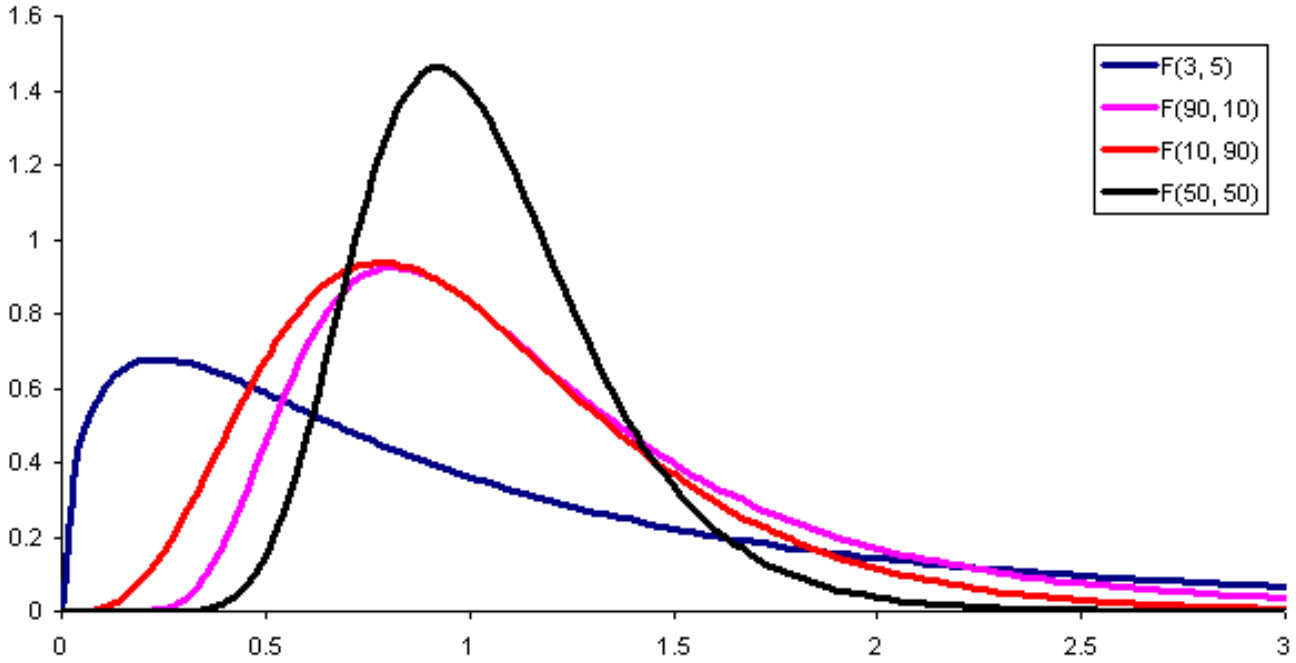
Common applications of the **Chi-squared distribution** include **testing the fit** of a model to observed data, **comparing observed and expected frequencies** in contingency tables, and evaluating the variability in a sample.

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



E – FISHER DISTRIBUTION

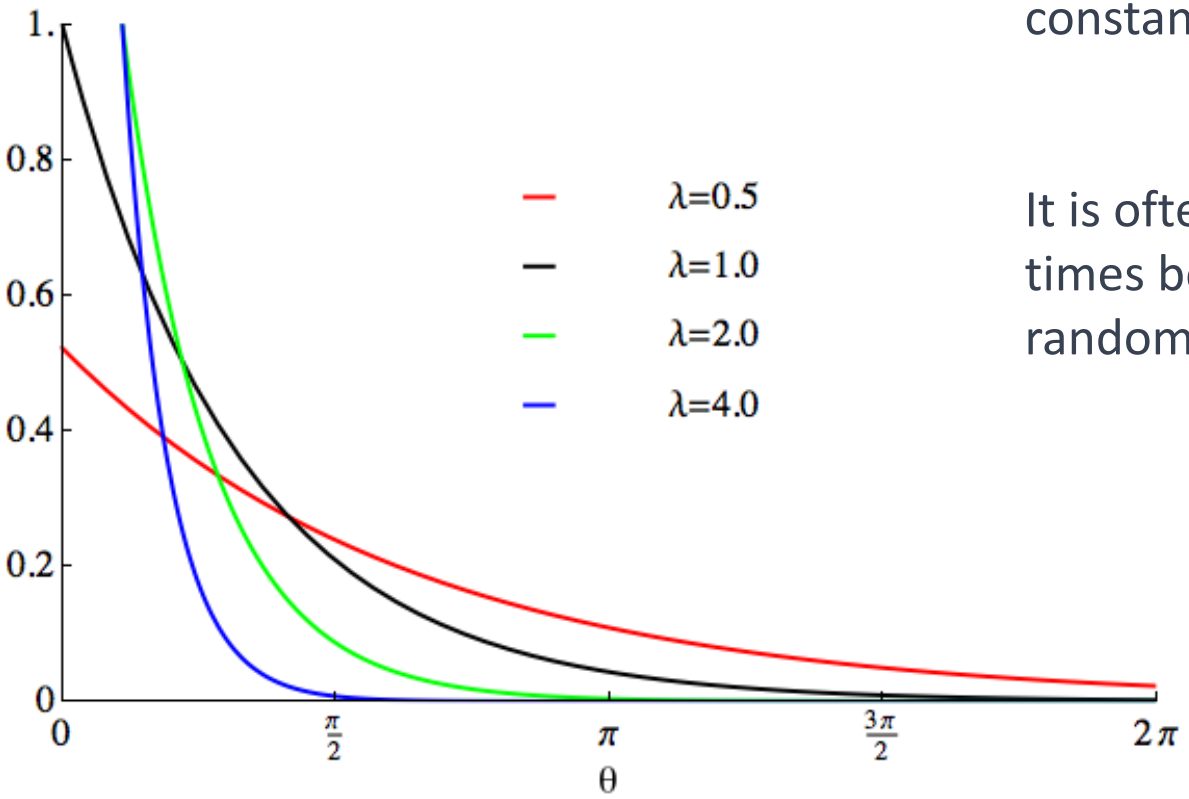
The **Fisher distribution** has two parameters:
- Degrees of Freedom (df1): This parameter, represents the degrees of freedom associated with one sample or population.
- Degrees of Freedom (df2): The second parameter represents the degrees of freedom associated with another sample or population.



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

F – EXPONENTIAL DISTRIBUTION

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

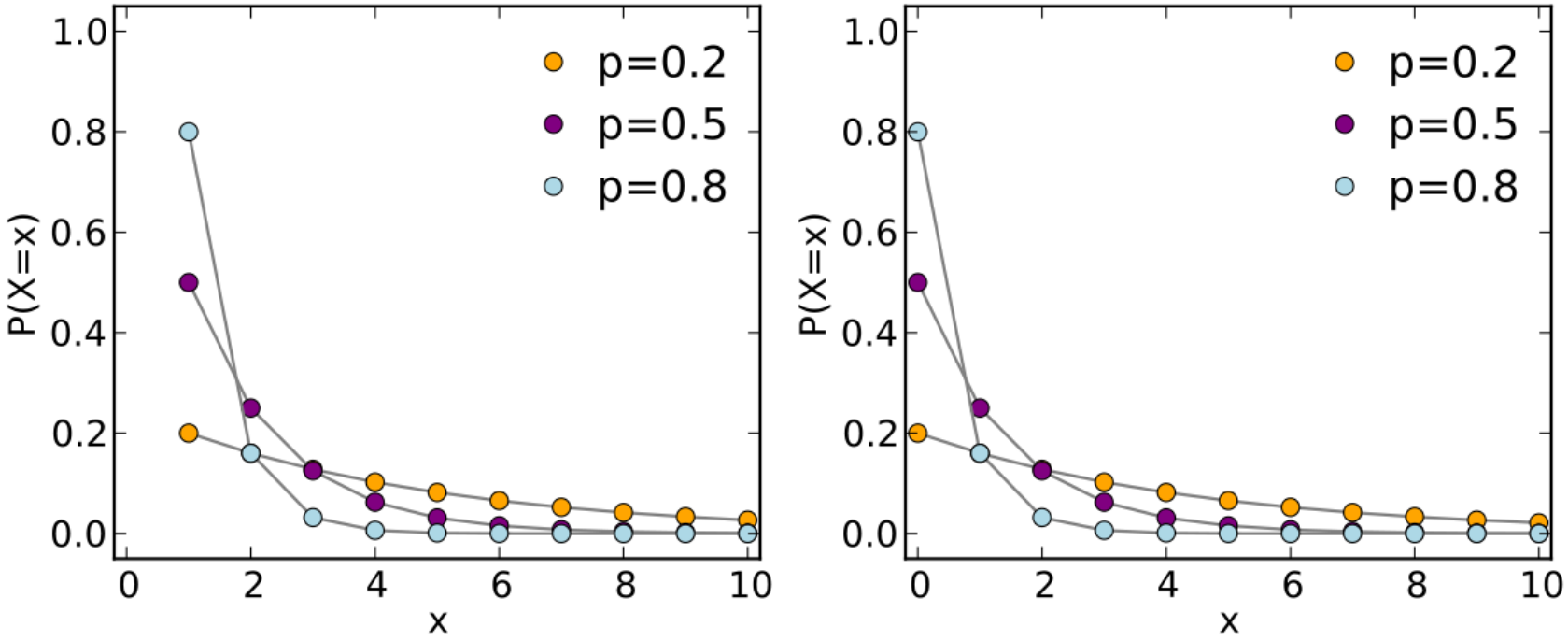


The **Exponential distribution** models the time between events in a process where events occur continuously and independently at a constant average rate.

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.



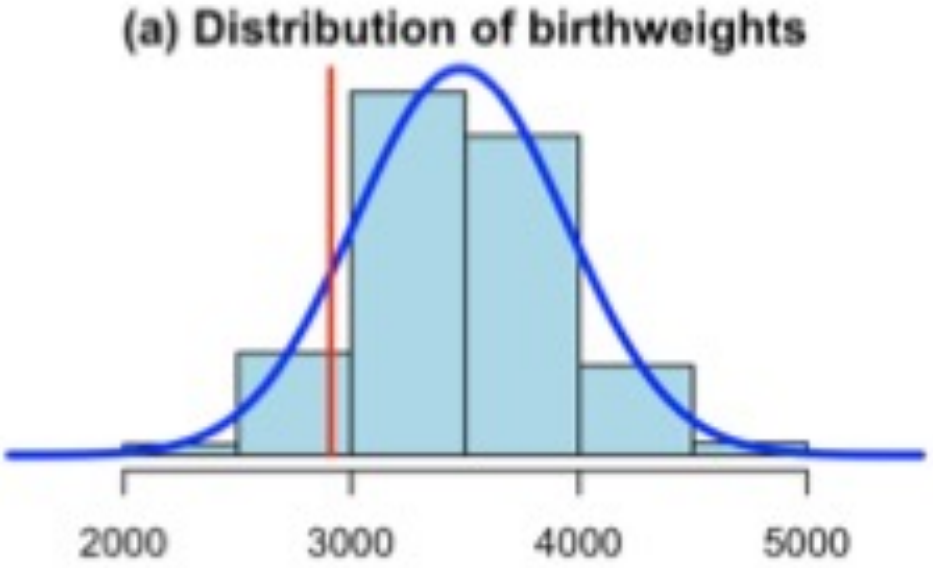
Probability and Statistics

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.

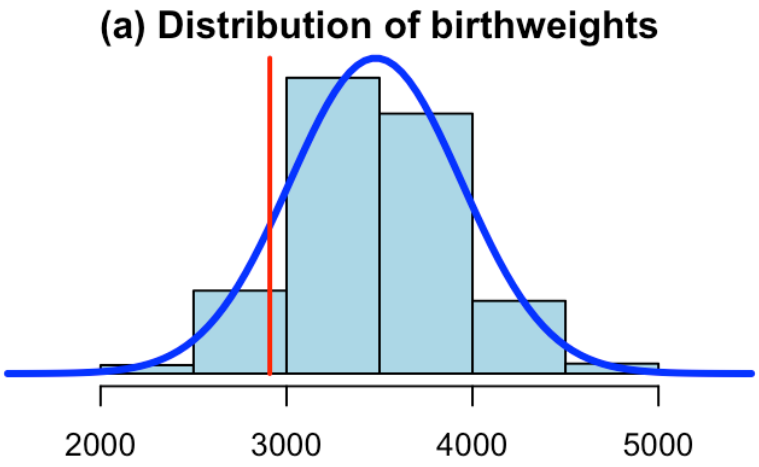
Many social phenomena show a notable regularity in their global trend (while individual events might be completely unpredictable). There might be a good correspondence between empirical data and mathematical probability distribution, the data behave as if a known random mechanism had generated them.

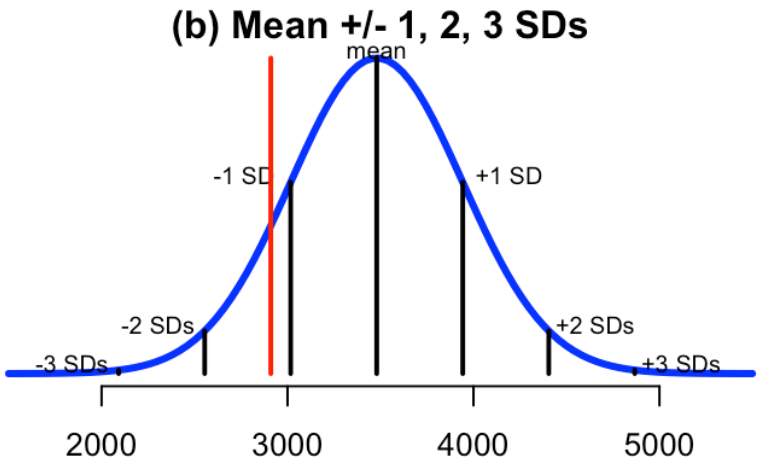
US vital statistics:1,096,277 full-term births to non-hispanic white women in the United States for 2013



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

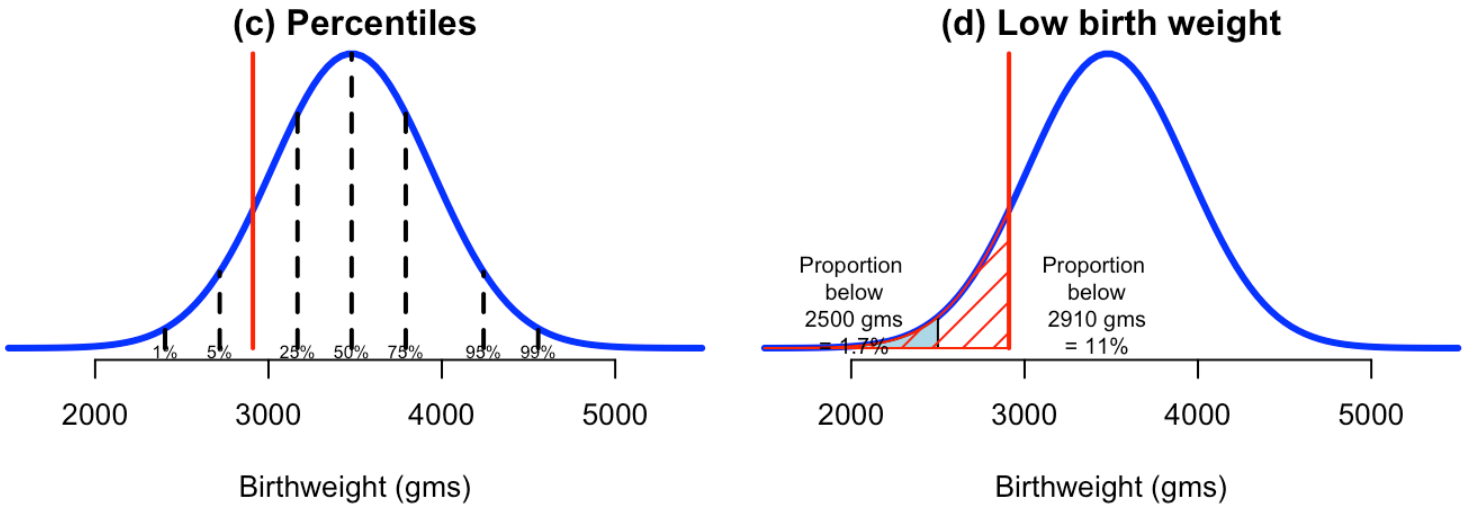


Figure (d) The proportion of low-birth-weight babies (dark shaded area), and babies less than 2,910g (light shaded area). It then represents:
- % of population of babies with a low birthweight
- probability that a randomly chosen baby (born in 2013…) weights less than 2.5 kg

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

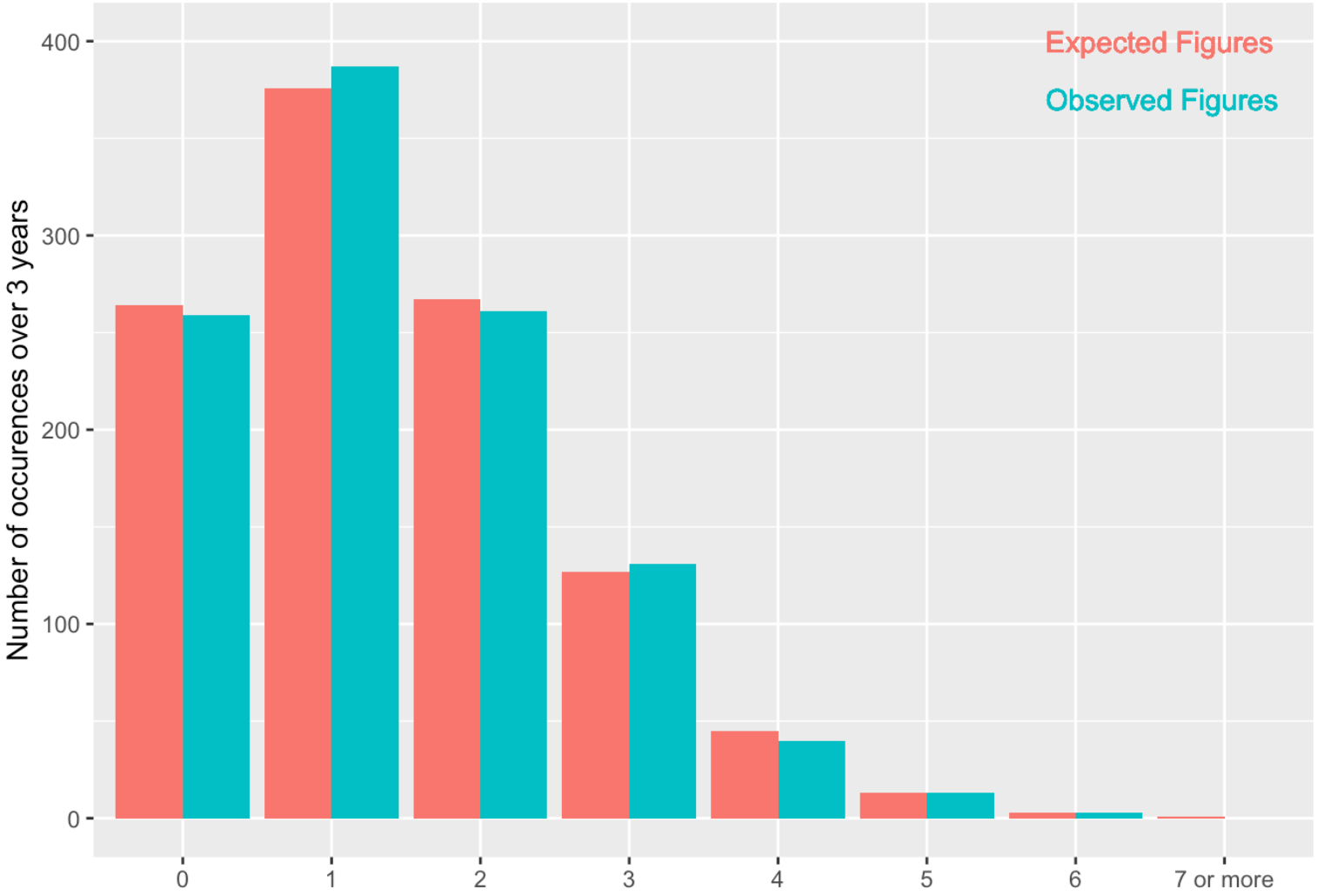
In this example we know the population distribution and parameters (mean, standard deviation, percentiles) however, in practice, we generally do not know the population, we therefore apply the induction process.

How often we do expect 7 or more homicides in a day in England and Wales?

We collect the daily data from March 2013 to April 2016 in England and Wales and we
Found on average 1.41 a day. In this period there were no episodes of 7 or more homicides but can be sure that this might not happen?

We should hypothesise a probability distribution for the daily number of homicides→we could imagine that every day we have a large population of individuals each with a very small probability of being killed→Poisson distribution which depends only on μ.

If we use 1.41 as estimate of μ, we can calculate the expected number of homicides based on the hypothesis that the phenomenon follows a Poisson distribution.



Observed and expected (assuming a Poisson distribution) daily number of recorded homicide incidents, 2014 to 2016, England and Wales (ONS data).

The probability of observing 7+ homicides in a day is 0.07% (a day like this on average every 1535 days→ not highly probable but neither impossible!)

CONNECTIONS AMONG DISTRIBUTIONS