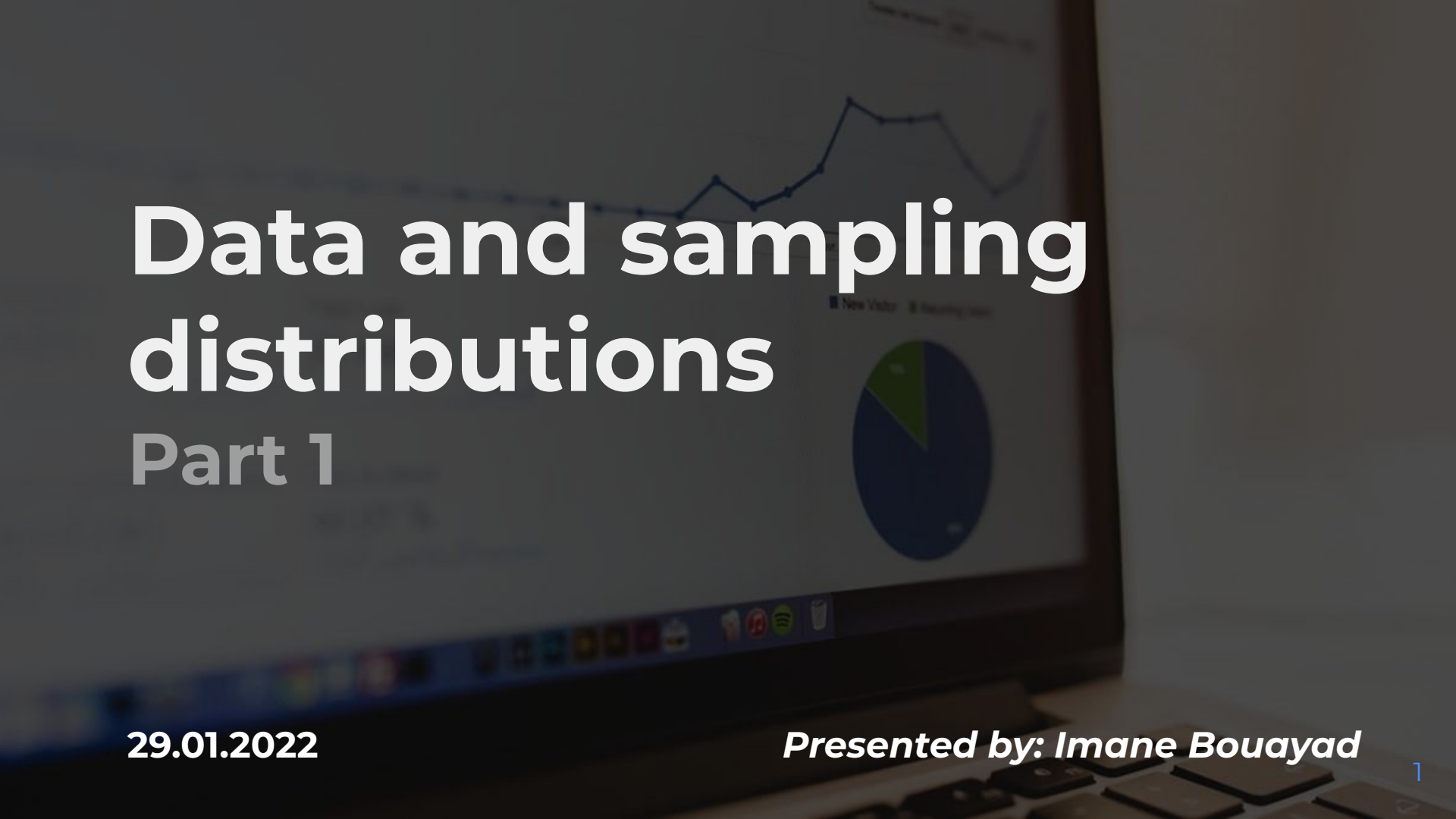


Data and sampling distributions

Part 1

29.01.2022

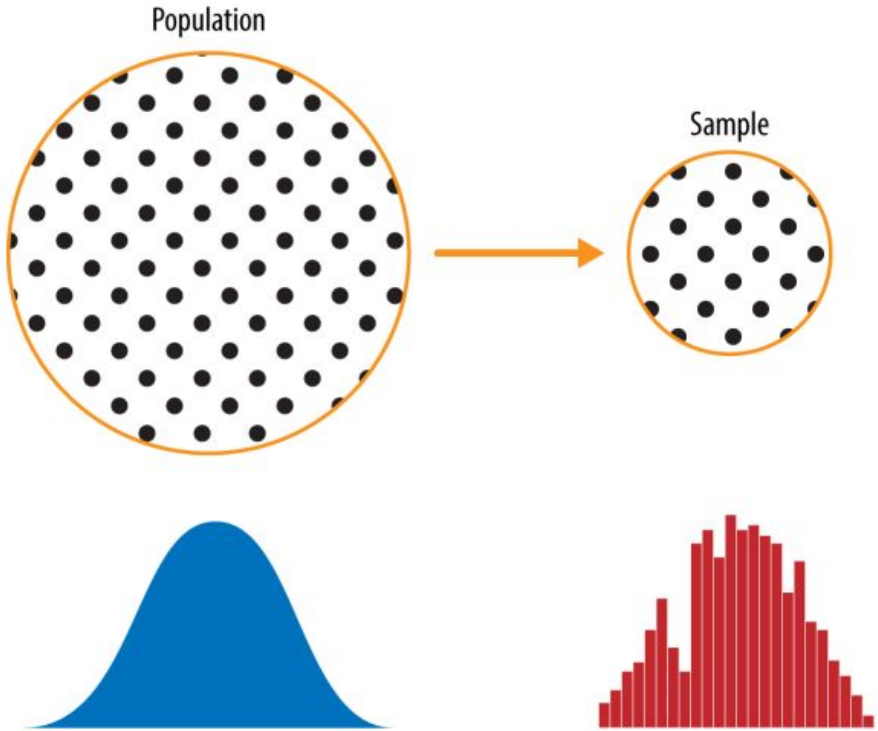
Presented by: Imane Bouayad


Chapter 6:
Statistical Machine
Learning
Chapter 7:
Unsupervised Learning
Chapter 5:
Classification
Practical Statistics
for Data Scientists
Chapter 1: Exploratory
Data Analysis
Chapter 4:
Regression and
Prediction
Chapter 3:
Statistical Experiments
and Significance
Testing
Chapter 2:
Data and Sampling
Distributions
We are here !

Presentation outline

- Random sampling and sample bias
- Selection bias
- Sampling distribution of statistics
- The bootstrap
- Confidence interval
- Normal distribution

Sampling



Obtaining good samples

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

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

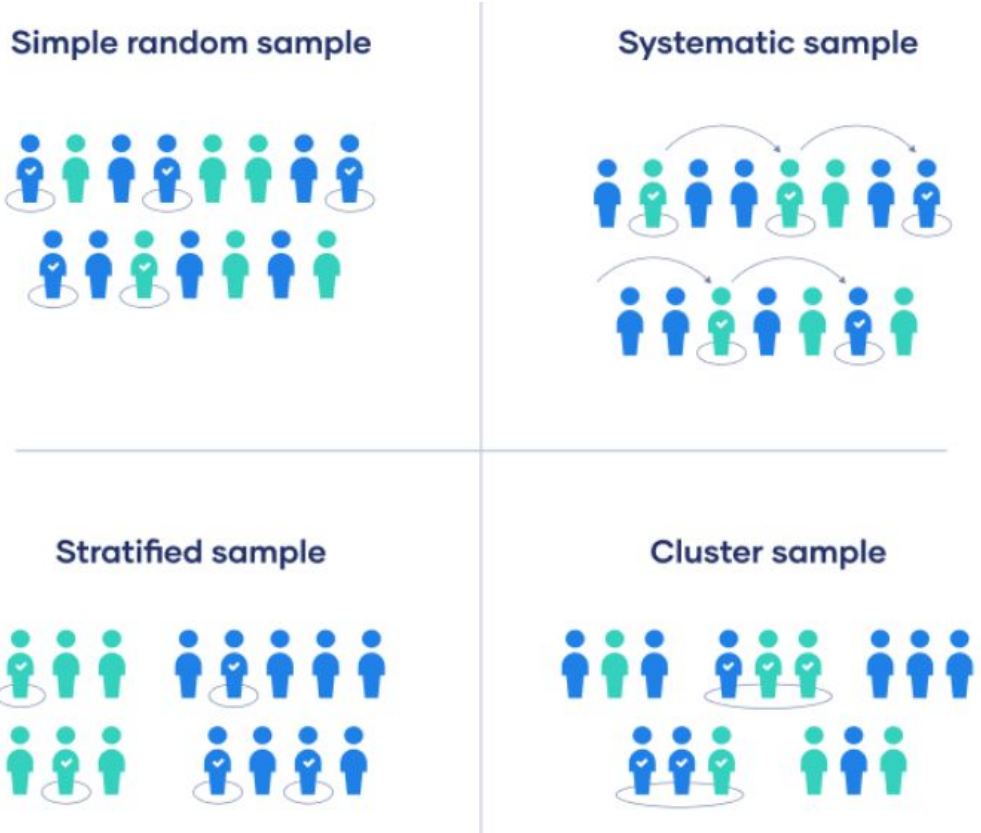
Random sampling

Random sampling is a process in which each available member of the population being sampled has an equal chance of being chosen for the sample at each draw.

Two types of random sampling:

- With replacement
- Without replacement

Random sampling



Source: Scribbr[7]

Random sampling

Mr. Marino has compiled a list of 1,348 students in his high school. He has selected a sample of 42 students by choosing every 14th student on this list.

Which type of sampling is he using?

1. Simple random sampling
2. Stratified random sampling
3. Cluster sampling
4. Systematic random sampling

Random sampling

Mr. Marino has compiled a list of 1,348 students in his high school. He has selected a sample of 42 students by choosing every 14th student on this list.

Which type of sampling is he using?

1. Simple random sampling
2. Stratified random sampling
3. Cluster sampling
4. **Systematic random sampling**

Random sampling

An analyst is investigating teachers' attitudes toward year-round schooling. She is particularly interested in describing the attitudes of teachers from small, medium, and large schools.

Which sampling procedure should be used to ensure her sample is representative of these types of schools?

1. Simple random sampling
2. Stratified random sampling
3. Cluster sampling
4. Systematic random sampling

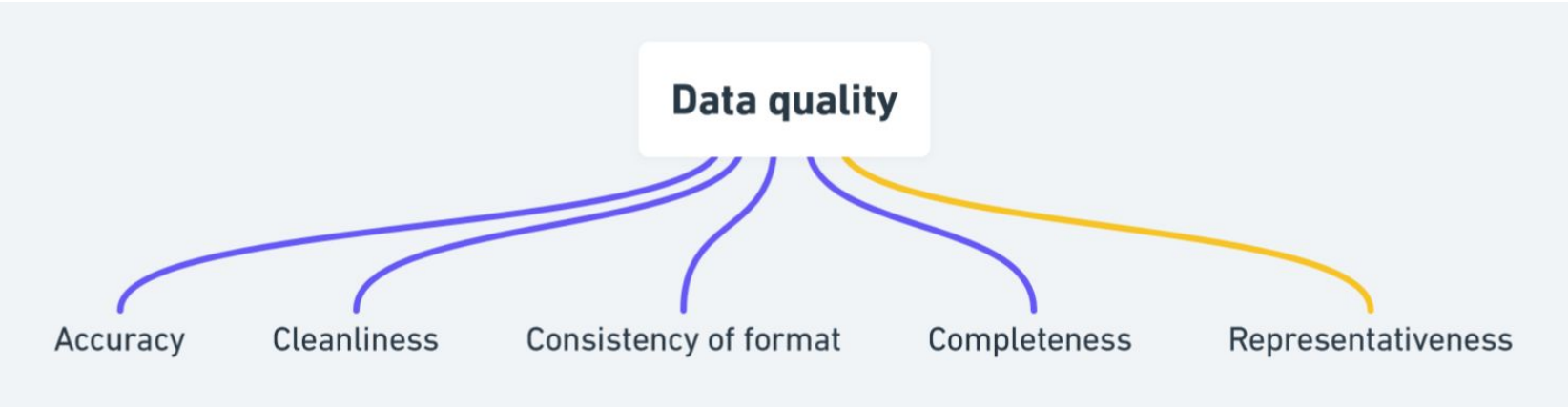
Random sampling

An analyst is investigating teachers' attitudes toward year-round schooling. She is particularly interested in describing the attitudes of teachers from small, medium, and large schools.

Which sampling procedure should be used to ensure her sample is representative of these types of schools?

1. Simple random sampling
2. **Stratified random sampling**
3. Cluster sampling
4. Systematic random sampling

Data quality and random sampling



Statistics adds the notion of **representativeness**.

In order to account for representativeness of our data, we have to deal with bias.

Selection bias

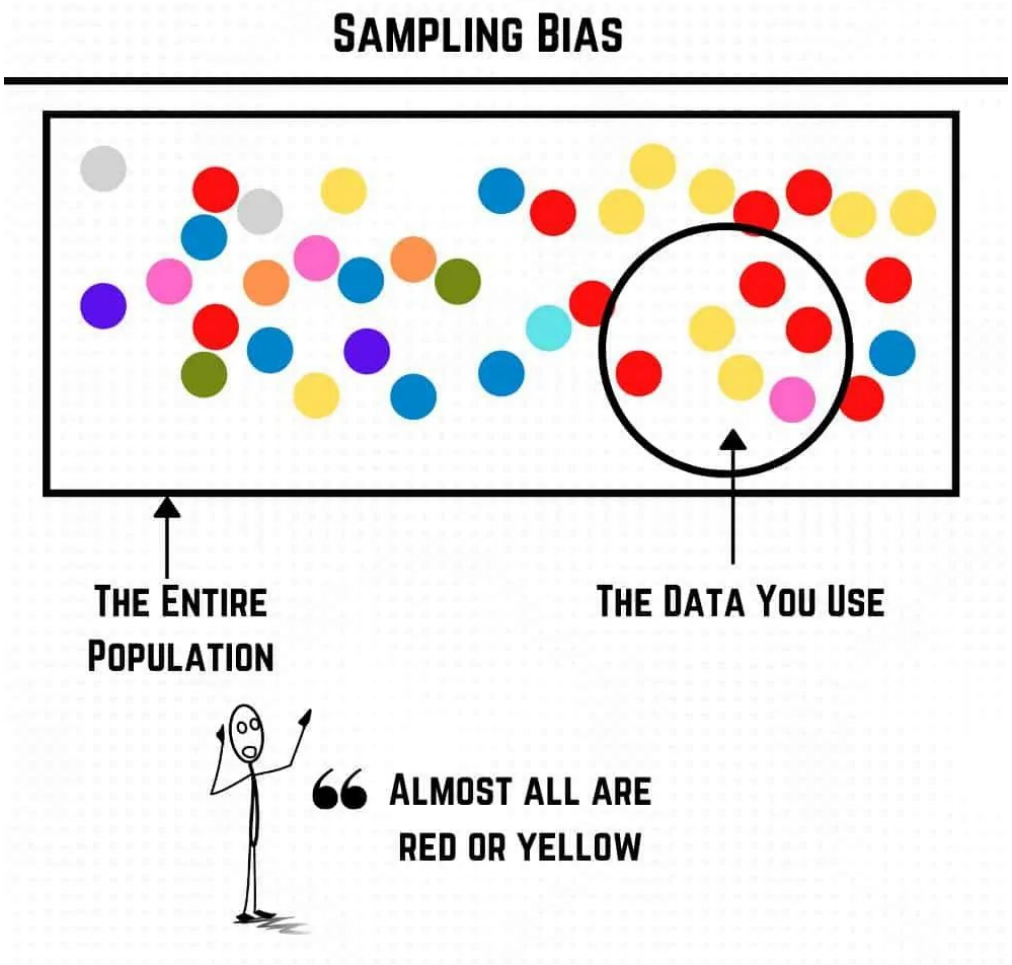
"Selection bias refers to the practice of selectively choosing data—consciously or unconsciously—in a way that leads to a conclusion that is misleading or ephemeral."

Examples of selection bias:

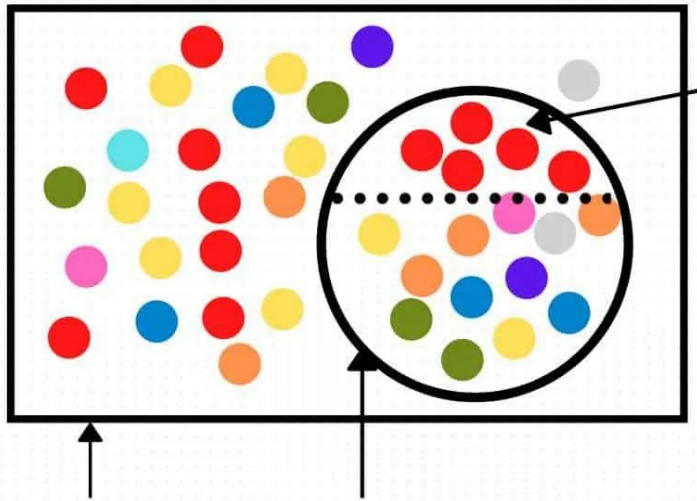
- Sampling bias
- Early termination
- Cherry picked data

Selection bias

- **Non-response:** If only a small fraction of the randomly sampled people choose to respond to a survey, the sample may no longer be representative of the population.
- **Voluntary response:** Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue. Such a sample will also not be representative of the population.
- **Convenience sample:** Individuals who are easily accessible are more likely to be included in the sample.


SAMPLING BIAS
THE ENTIRE POPULATION
THE DATA YOU USE
“ALMOST ALL ARE RED OR YELLOW

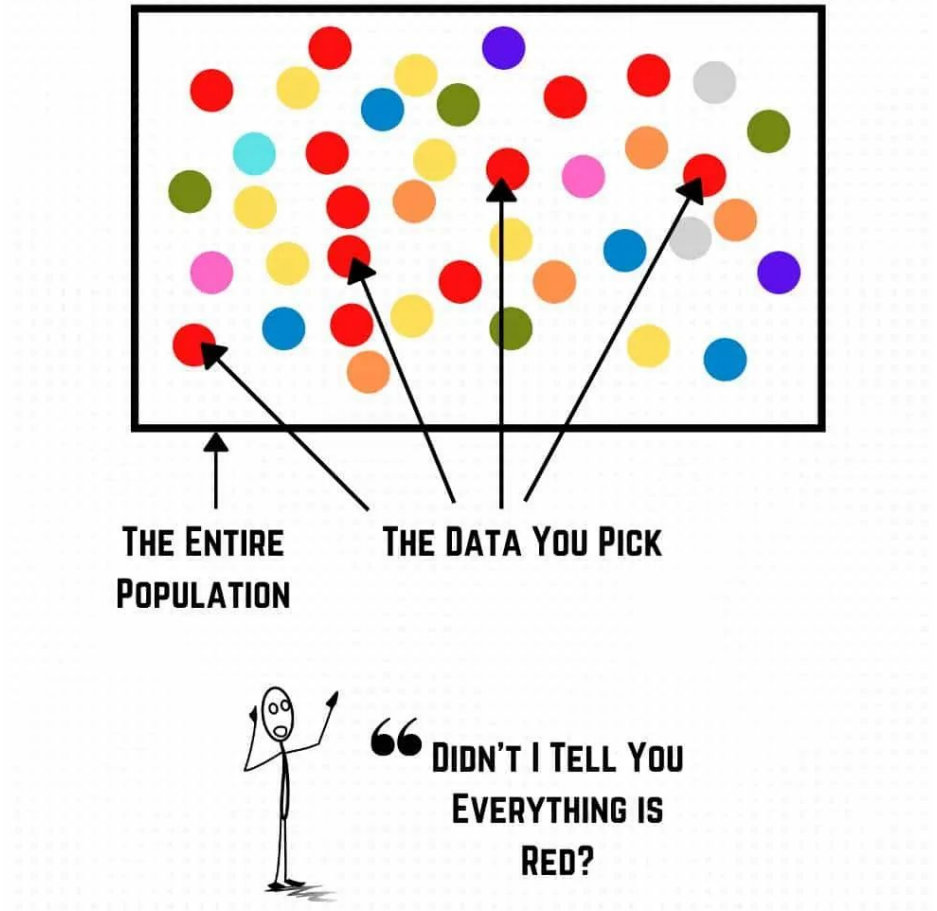
Early Termination



After First 5 Tests

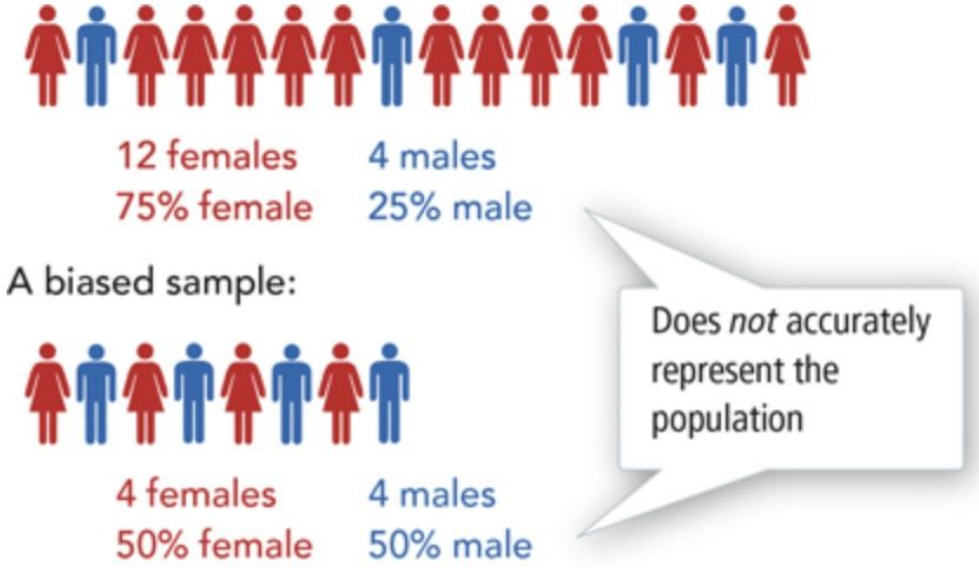
"Everything is Red

DATA BIAS

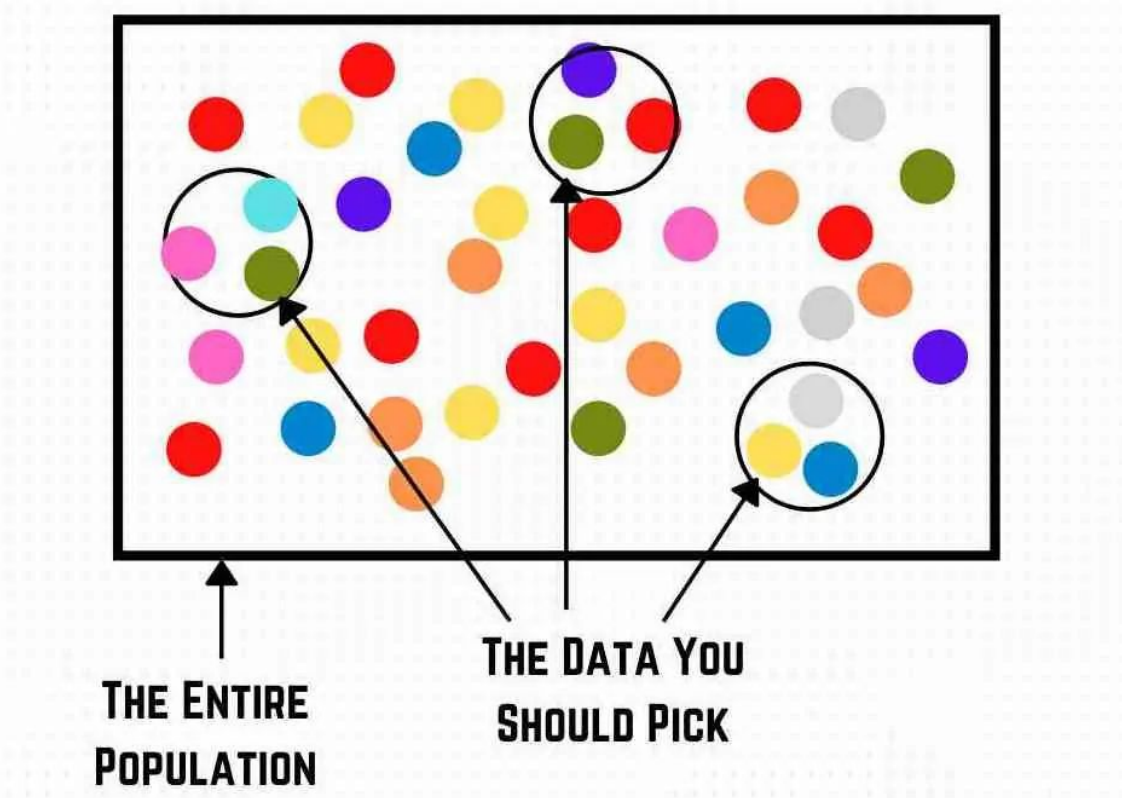

THE ENTIRE POPULATION
THE DATA YOU PICK
"DIDN'T I TELL YOU EVERYTHING IS RED?

Sample bias

"Sampling bias occurs when some members of a population are systematically more likely to be selected in a sample than others."



How to avoid selection bias



School district is considering whether it will no longer allow high school students to park at school after two recent accidents where students were severely injured. As a first step, they survey parents by mail, asking them whether or not the parents would object to this policy change. Of ***6,000*** surveys that go out, ***1,200*** are returned. Of these 1,200 surveys that were completed, ***960*** agreed with the policy change and ***240*** disagreed.

Which of the following statements are true?

1. Some of the mailings may have never reached the parents.
2. The school district has strong support from parents to move forward with the policy approval.
3. It is possible that majority of the parents of high school students disagree with the policy change.
4. The survey results are unlikely to be biased because all parents were mailed a survey.

(a) Only 1 (b) 1 and 2 (c) 1 and 3 (d) 3 and 4 (e) Only 5

School district is considering whether it will no longer allow high school students to park at school after two recent accidents where students were severely injured. As a first step, they survey parents by mail, asking them whether or not the parents would object to this policy change. Of ***6,000*** surveys that go out, ***1,200*** are returned. Of these 1,200 surveys that were completed, ***960*** agreed with the policy change and ***240*** disagreed.

Which of the following statements are true?

1. Some of the mailings may have never reached the parents.
2. The school district has strong support from parents to move forward with the policy approval.
3. It is possible that majority of the parents of high school students disagree with the policy change.
4. The survey results are unlikely to be biased because all parents were mailed a survey.

(a) Only 1 (b) 1 and 2 (c) 1 and 3 (d) 3 and 4 (e) Only 5

Regression to the mean

“The notion of **regression to the mean** was first worked out by Sir Francis Galton. The rule goes that, in any series with complex phenomena that are dependent on many variables, where chance is involved, **extreme outcomes tend to be followed by more moderate ones**.”

Regression to the mean

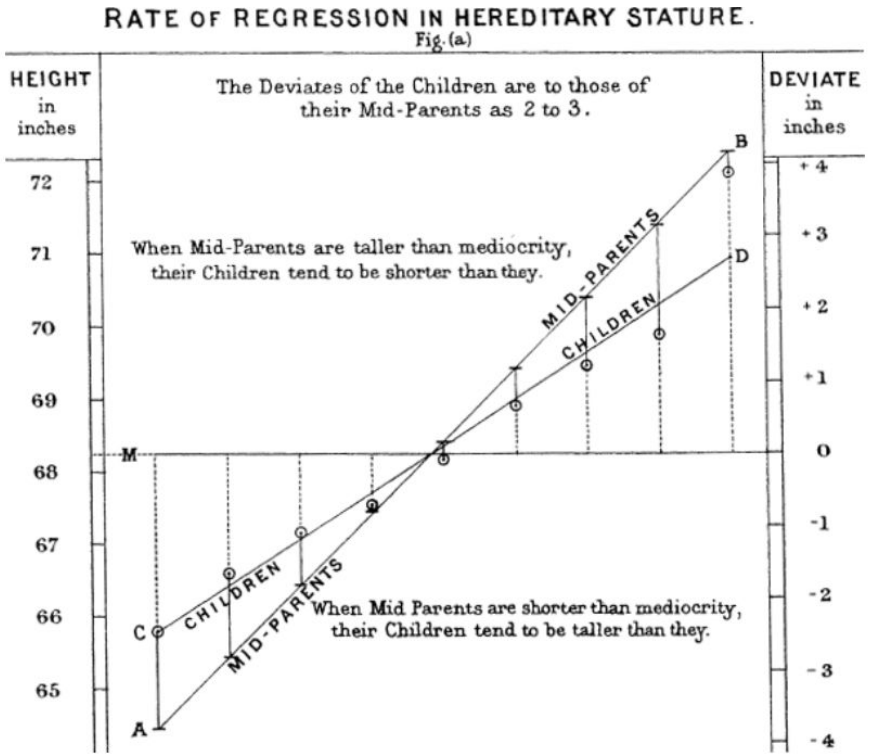
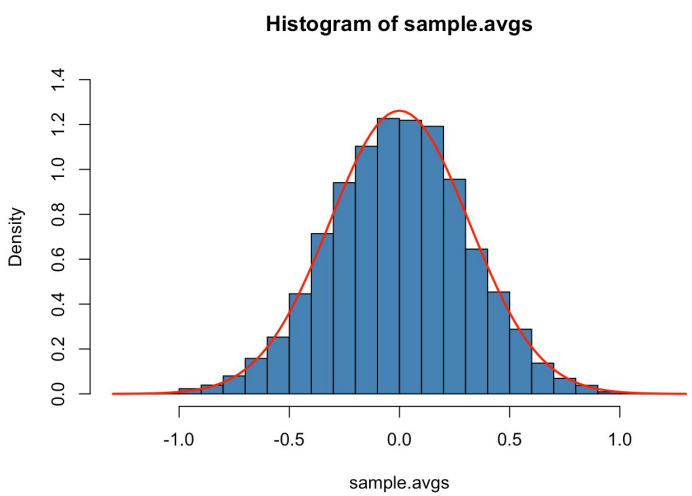
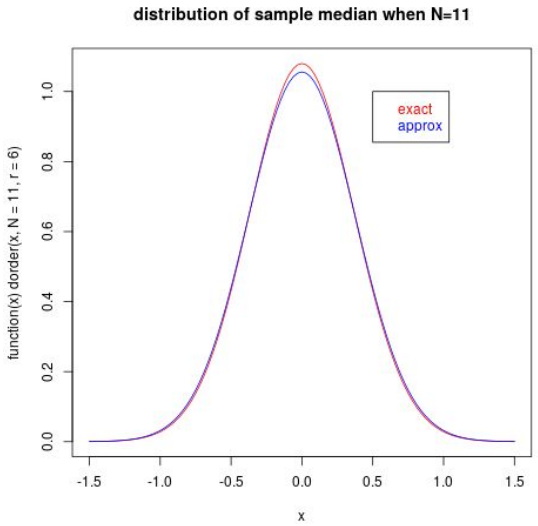


Figure 2-5. Galton's study that identified the phenomenon of regression to the mean

Sampling distribution of a statistic

The sampling distribution of a given population is the distribution of frequencies of a range of different outcomes that could possibly occur for a **statistic** of a population.





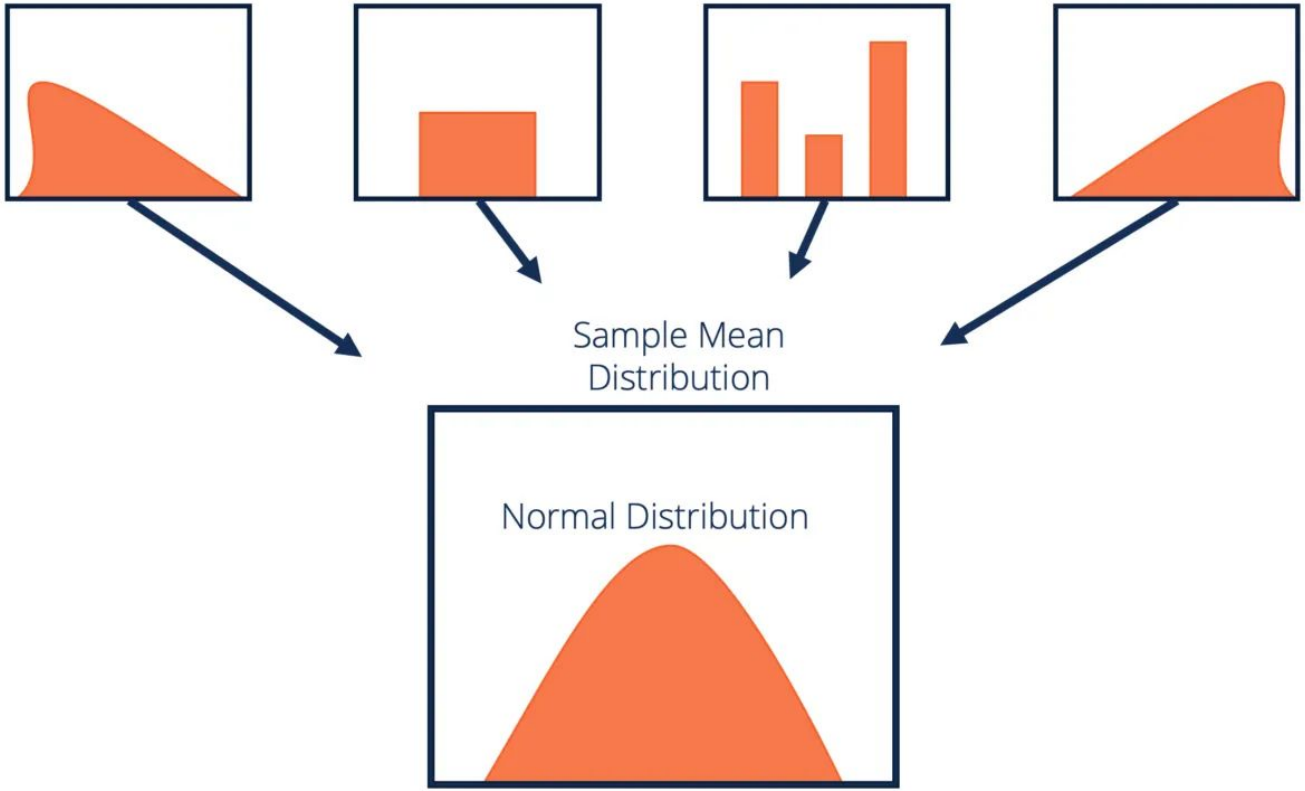
Central limit theorem

According to the Central Limit Theorem, the sampling distribution of the mean:

- is normally distributed.
- has a mean equal to the population mean.
- has standard deviation (also called standard error) equal to the population standard deviation divided by the square root of the sample size.

$$\text{standard error} = \frac{\text{Population standard deviation}}{\sqrt{\text{sample size}}}$$

Central limit theorem



Standard error and sample size

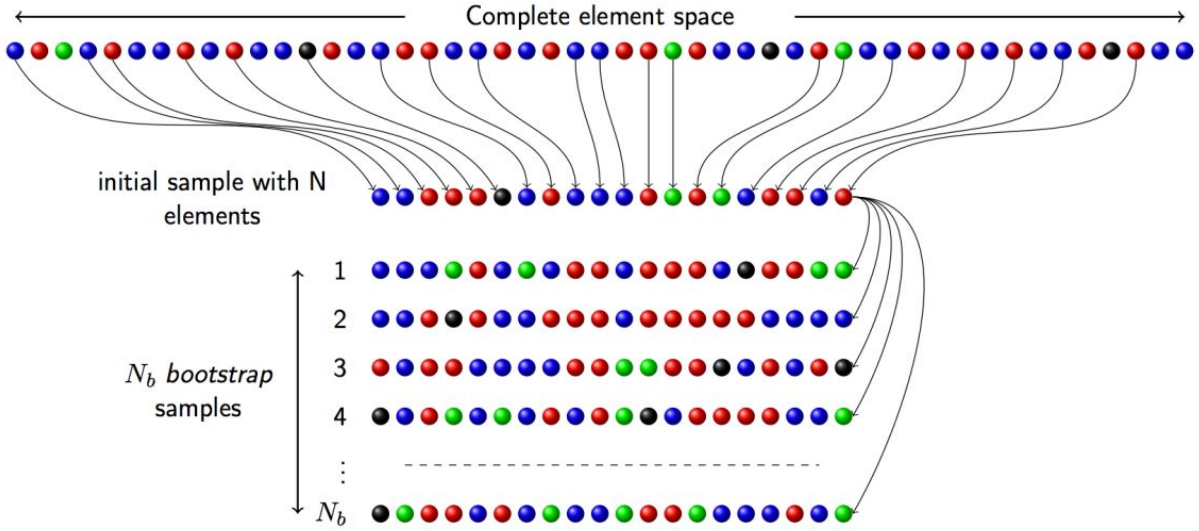
When you increase the sample size, the standard error of the mean decreases.

This can be seen from the formula:

$$\text{standard error} = \frac{\text{Population standard deviation}}{\sqrt{\text{sample size}}}$$

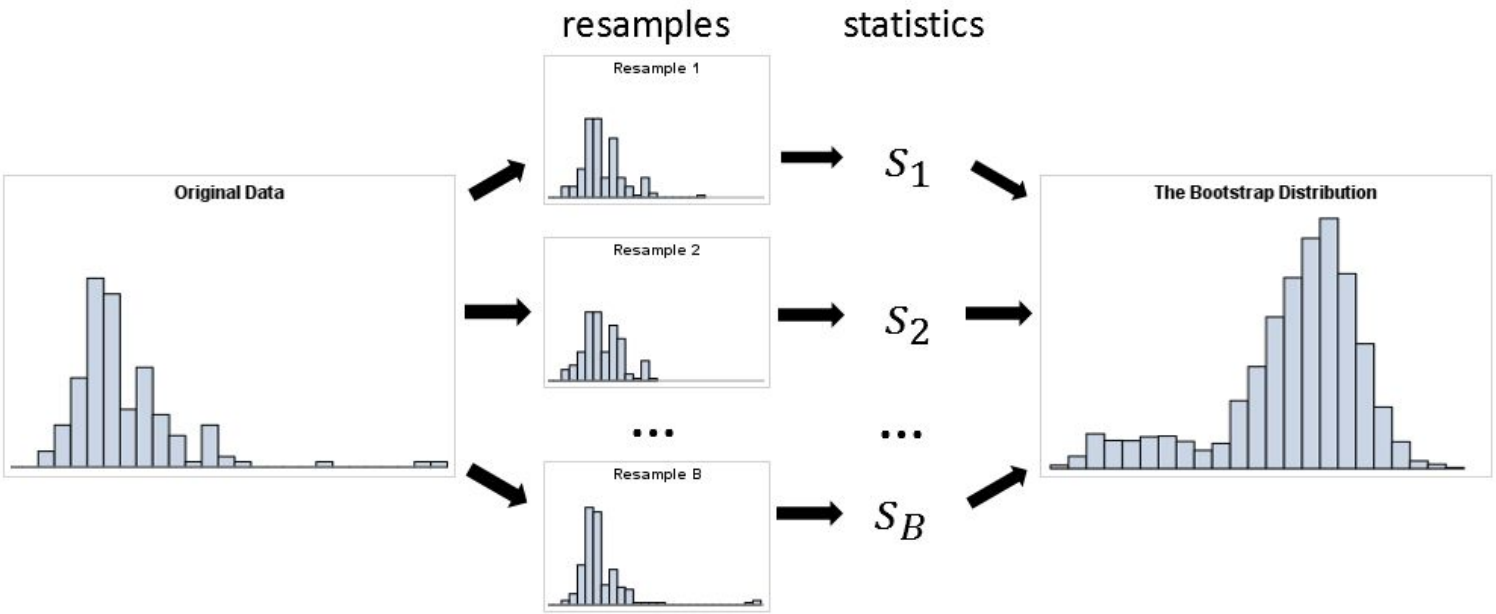
The bootstrap

Bootstrapping is any test or metric that uses random sampling with replacement, and falls under the broader class of resampling methods.

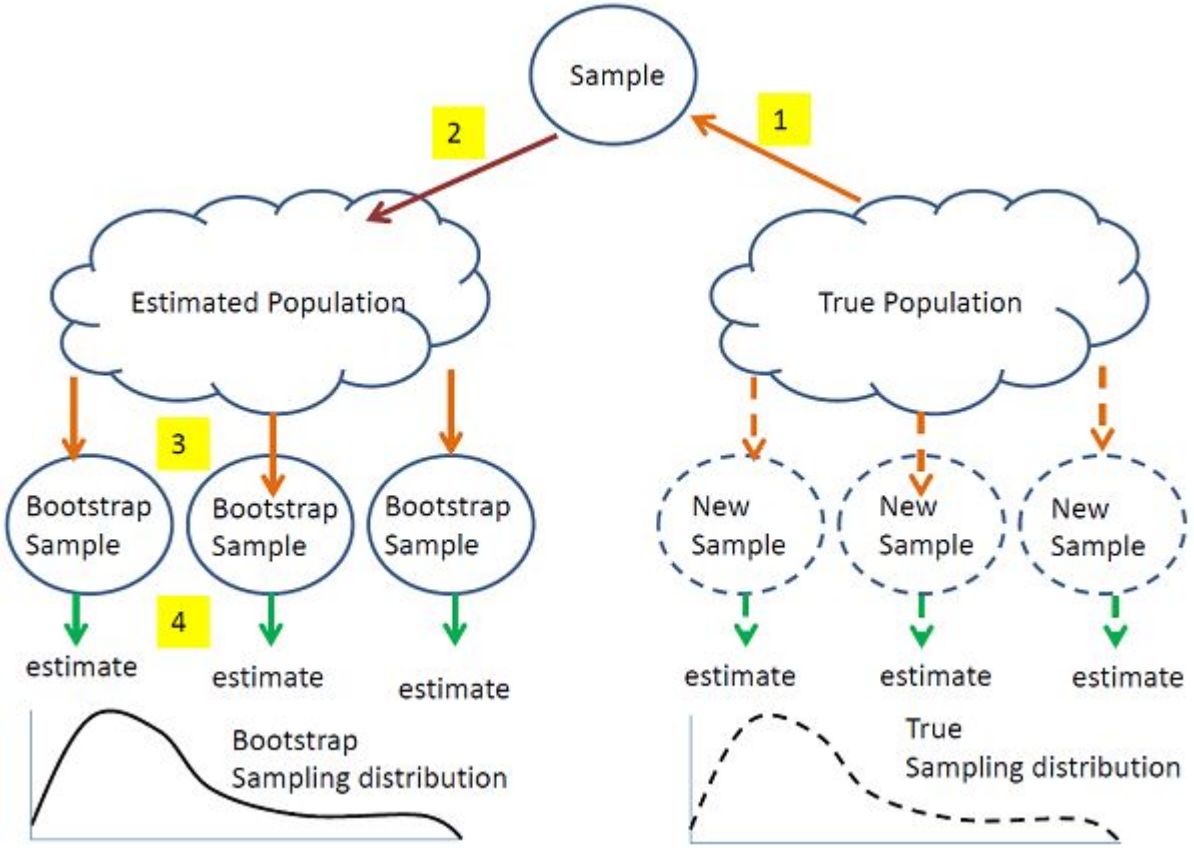


Resampling

Resampling is the method that consists of drawing repeated samples from the original data samples.

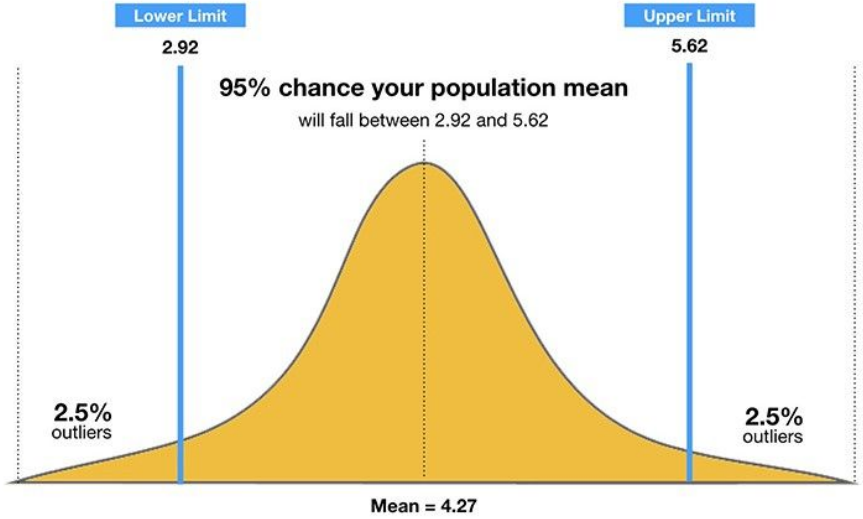


Bootstrapping vs resampling



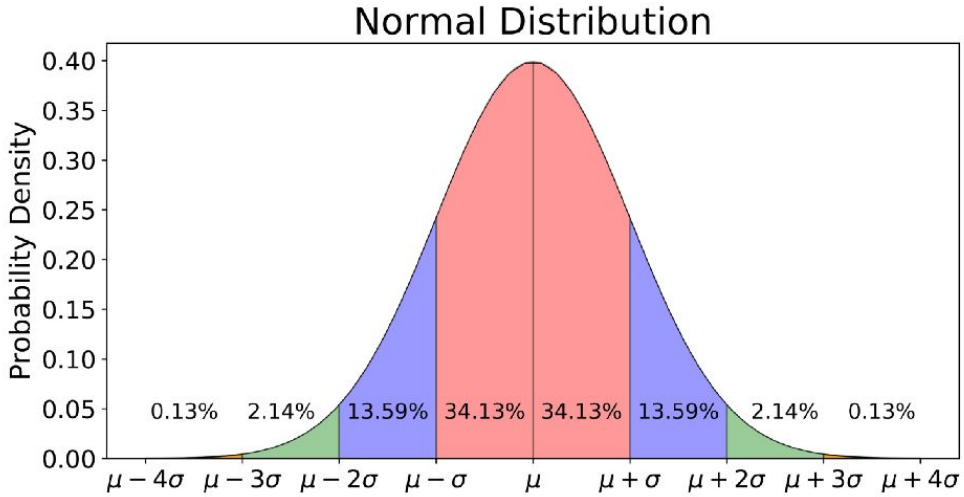
Confidence interval

- Measure the degree of uncertainty or certainty in a sampling method.
- A confidence interval can take any number of probabilities, with the most common being a 95% or 99% confidence level.



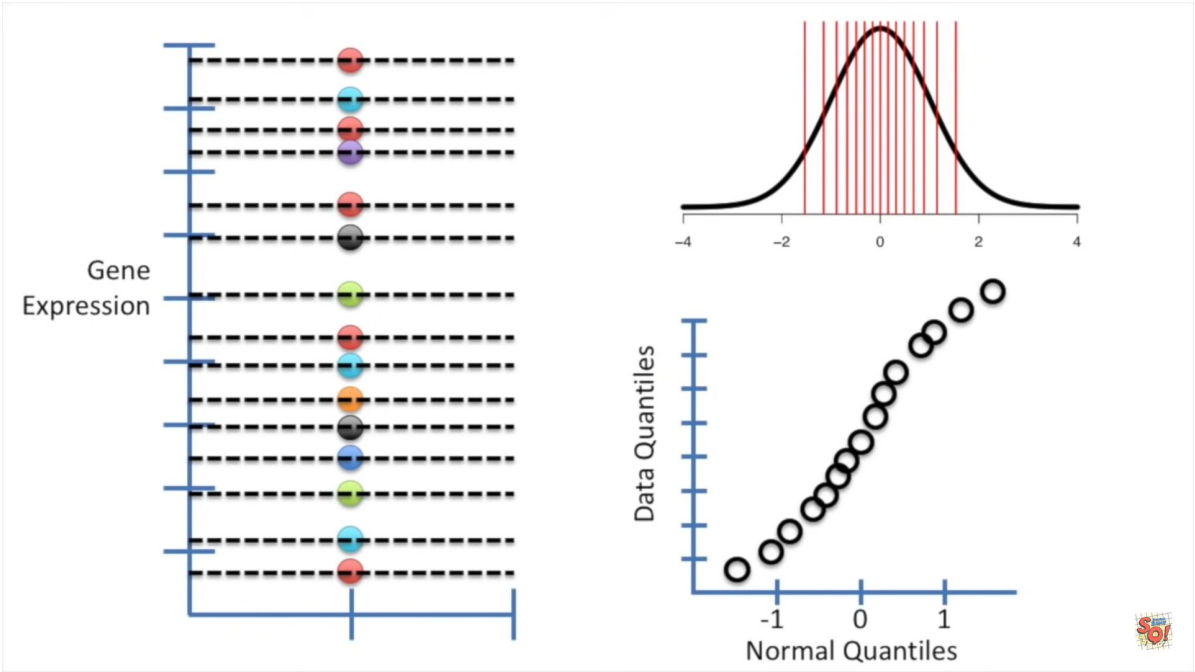
Normal distribution

Normal distribution also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.



QQ plots

QQ plot is a plot of the quantiles of the first data set against the quantiles of the second data set.



QQ plots

Have you been to an ophthalmologist ?

- QQ plots are like glass's filters.

- They show you how your data quantiles distribution compare to other known quantile distribution.

References

https://www.scribbr.com/methodology/sampling-methods/

Statistics course - Master of data science by Hajar Moussanif

https://studylib.net/doc/10304353/math-225-quiz-1--name-_____

https://fs.blog/regression-to-the-mean/

https://www.investopedia.com/

https://online.stat.psu.edu/stat555/node/119/

https://productiveclub.com/selection-bias/#3_Indirect_causes