Introduction to Bioinformatics and Biostatistics

Marta R. Hidalgo
Bioinformatics and Biostatistics Unit, CIPF

October 24th, 2018

Unidad de Bioinformática y Bioestadística

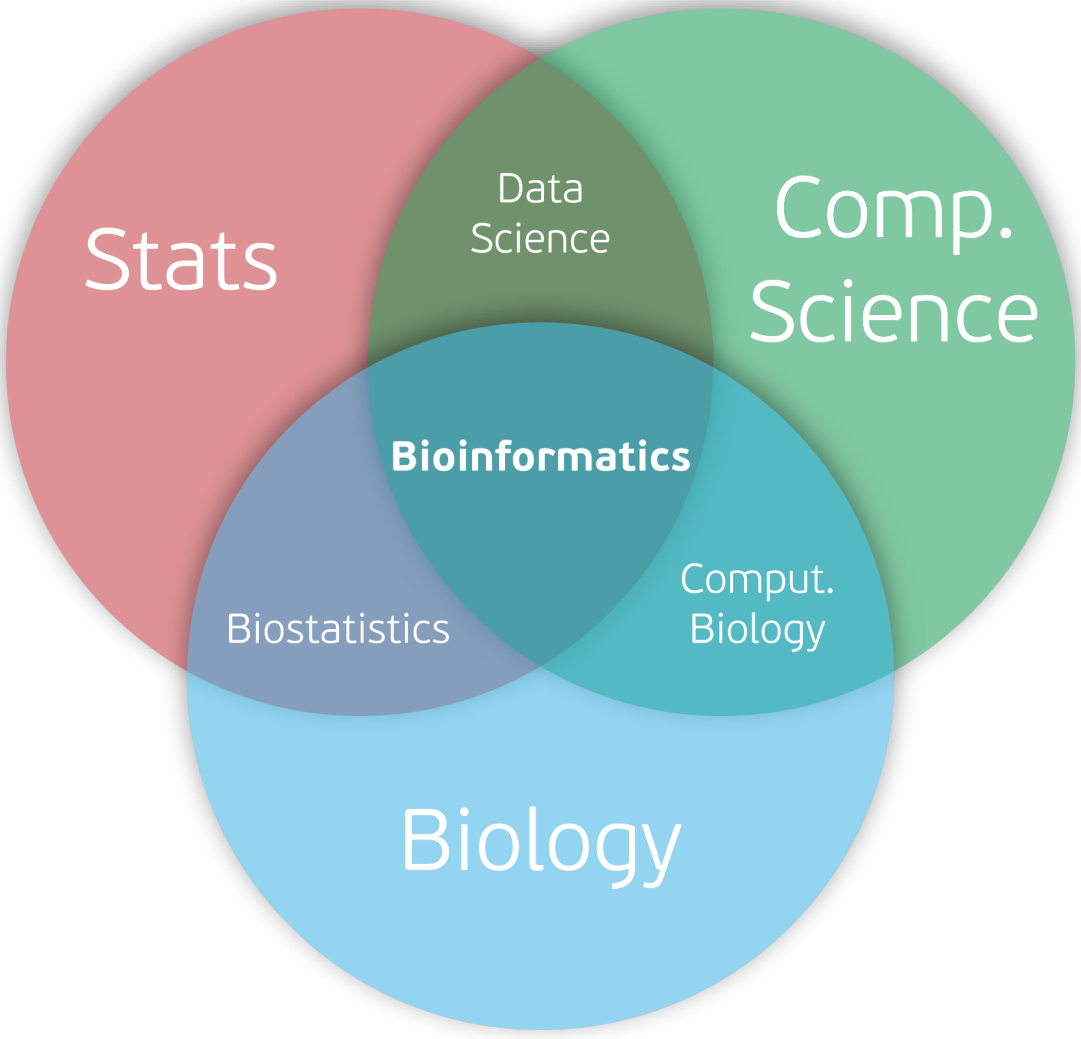
PRINCIPE FELIPE
CENTRO DE INVESTIGACION

WODA
WEB-BASED OMICS DATA ANALYSIS

Outline

Outline

What is Bioinformatics?



What is Bioinformatics?

Big Data

DNA contains 3.200M of bases
Each genome contains ~20.000 genes
Studies include up to 100.000 patients

Bioinformatics Database Building

Management of massive biological resource data and development of databases

Structural Bioinformatics

Prediction of the structure of a protein
Creation of new drugs

Bioinformatics analysis

Sequencing techniques:
- Disease mechanisms
- Phylogenetics
- Population genetics

What is Bioinformatics?

Big Data

DNA contains 3.200M of bases
Each genome contains ~20.000 genes
Studies include up to 100.000 patients

Bioinformatics Database Building

Management of massive biological resource data and development of databases

Structural Bioinformatics

Prediction of the structure of a protein
Creation of new drugs

OMICS

Bioinformatics analysis

Sequencing techniques:
- Disease mechanisms
- Phylogenetics
- Population genetics

What is Bioinformatics?

- Genomics
- Metabolomics
- Proteomics
- Transcriptomics

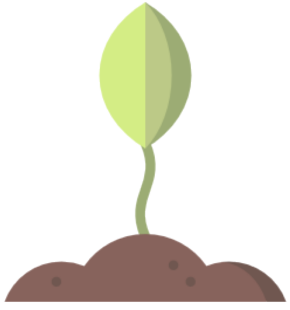
Data bases

What for?

- experiment data (raw / processed)
- variants
- pathways

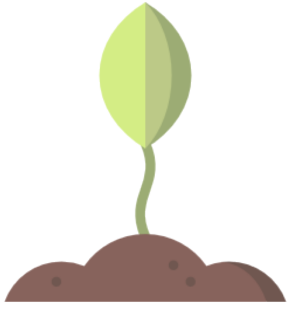
Outline

Variables



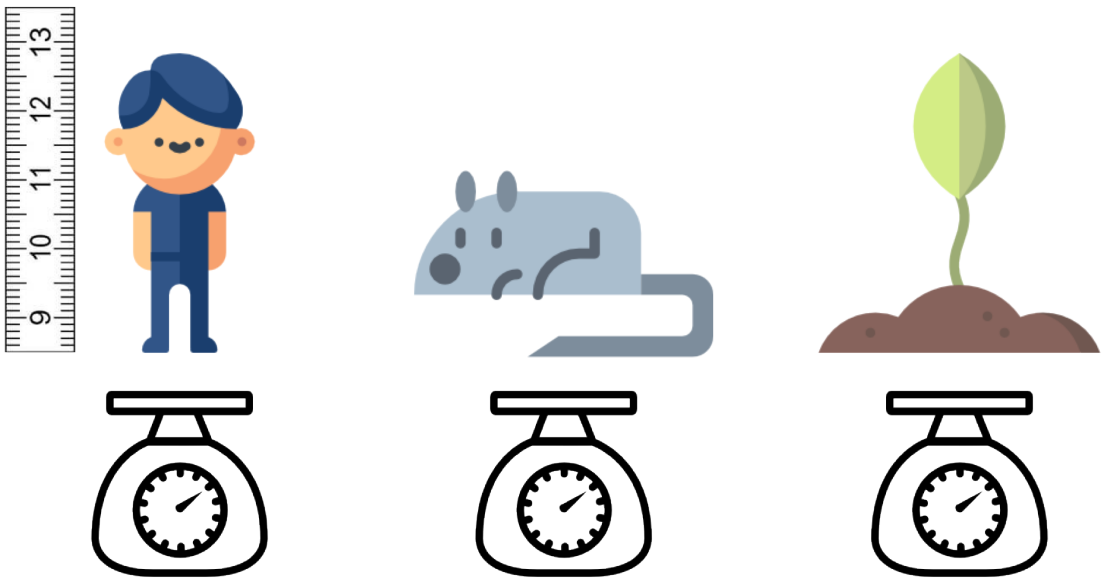
	Human	Mouse	Plant

Variables



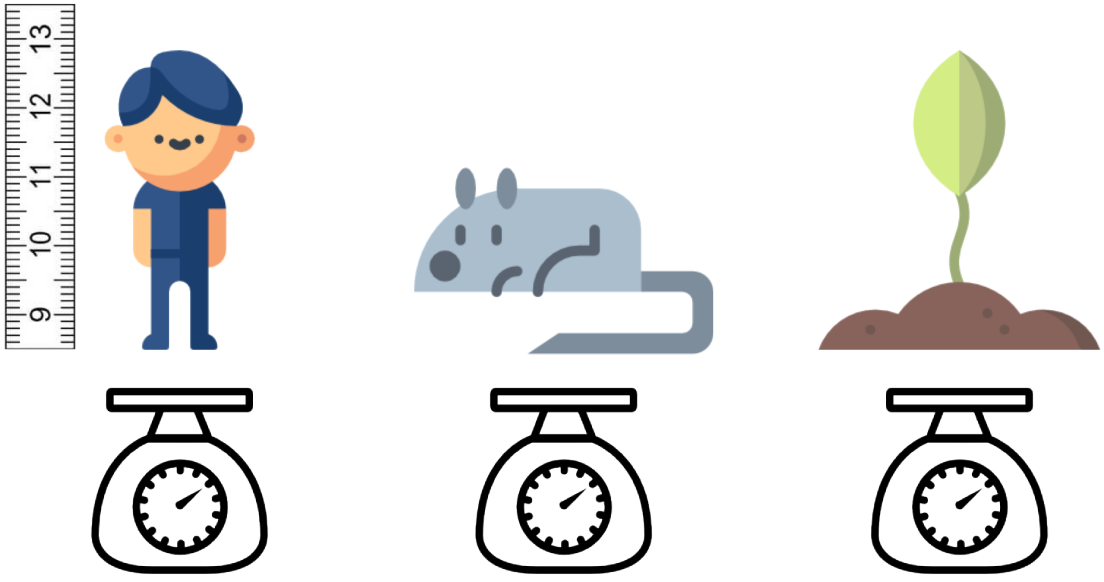
	Human	Mouse	Plant
Height	170cm	25cm	10cm

Variables



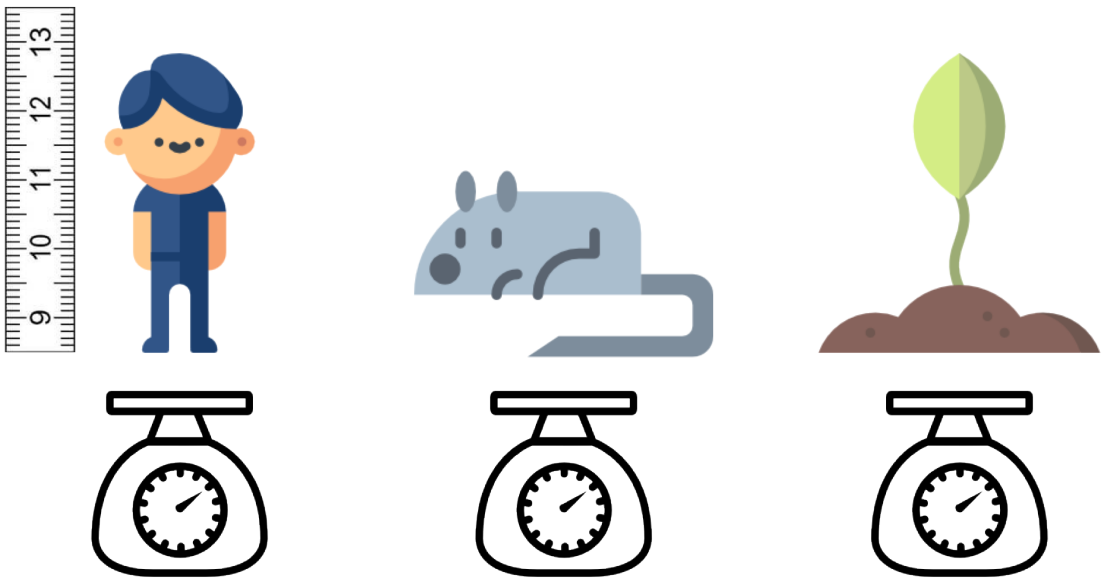
	Human	Mouse	Plant
Height	170cm	25cm	10cm
Weight	68.000g	150g	10g

Variables



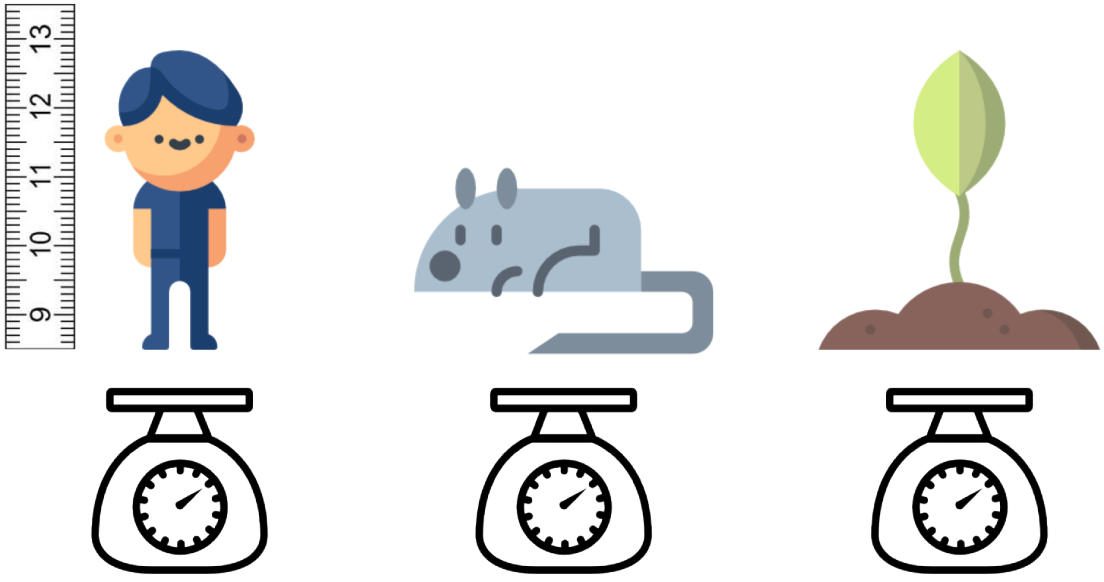
	Human	Mouse	Plant
Height	170cm	25cm	10cm
Weight	68.000g	150g	10g
Color eyes	Brown	Black	NA

Variables



	Human	Mouse	Plant
Height	170cm	25cm	10cm
Weight	68.000g	150g	10g
Color eyes	Brown	Black	NA
Color flowers	NA	NA	White

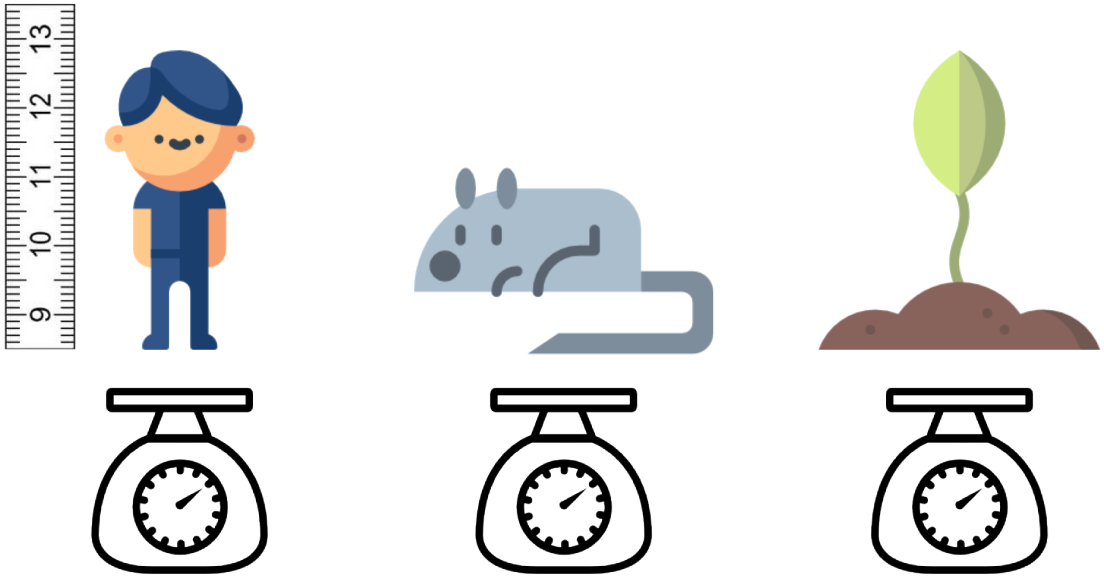
Variables



	Human	Mouse	Plant
Height	170cm	25cm	10cm
Weight	68.000g	150g	10g
Color eyes	Brown	Black	NA
Color flowers	NA	NA	White

NUMERICAL

Variables



	Human	Mouse	Plant
Height	170cm	25cm	10cm
Weight	68.000g	150g	10g
Color eyes	Brown	Black	NA
Color flowers	NA	NA	White

NUMERICAL

CATEGORICAL

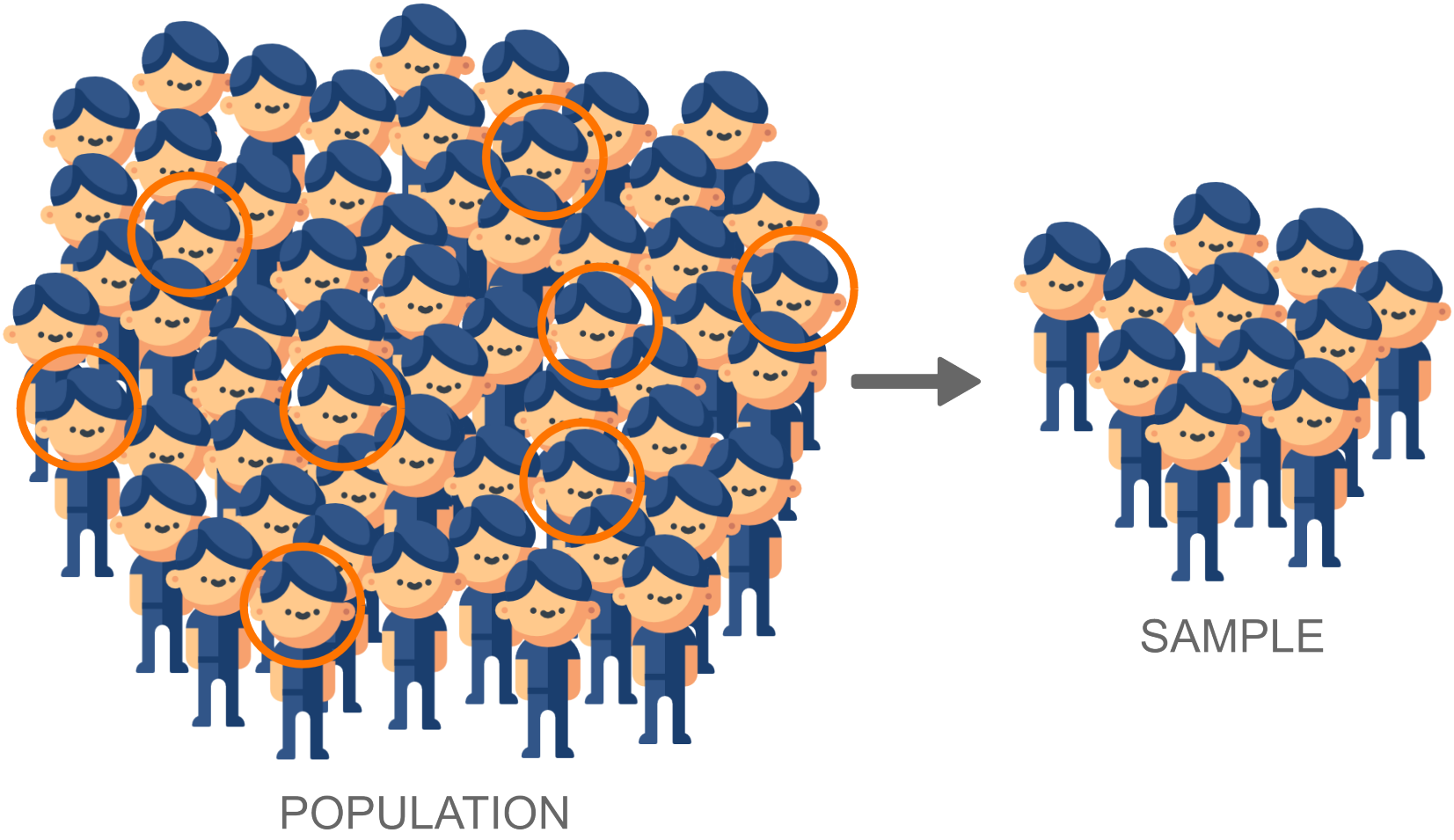
Population

POPULATION

Population

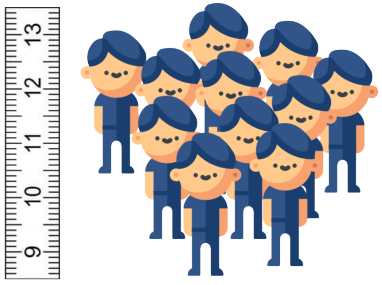
POPULATION

Population



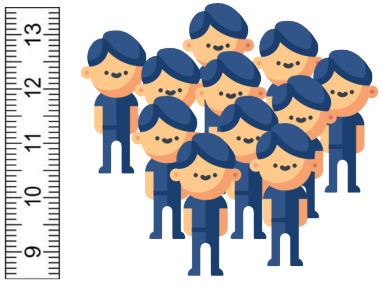
Numerical variable

Numerical variable



	Height
Human 1	1'70
Human 2	1'53
Human 3	2'01
Human 4	1'82
Human 5	1'65
Human 6	1'73
Human 7	1'91
Human 8	1.81

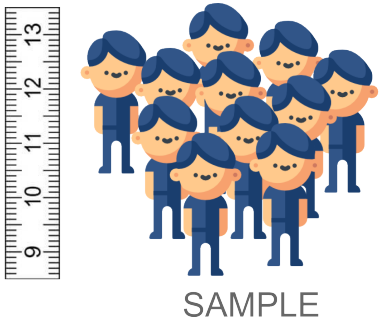
Numerical variable



	Height
Human 1	1'70
Human 2	1'53
Human 3	2'01
Human 4	1'82
Human 5	1'65
Human 6	1'73
Human 7	1'91
Human 8	1.81

SAMPLING DISTRIBUTION

Numerical variable



	Height
Human 1	1'70
Human 2	1'53
Human 3	2'01
Human 4	1'82
Human 5	1'65
Human 6	1'73
Human 7	1'91
Human 8	1.81

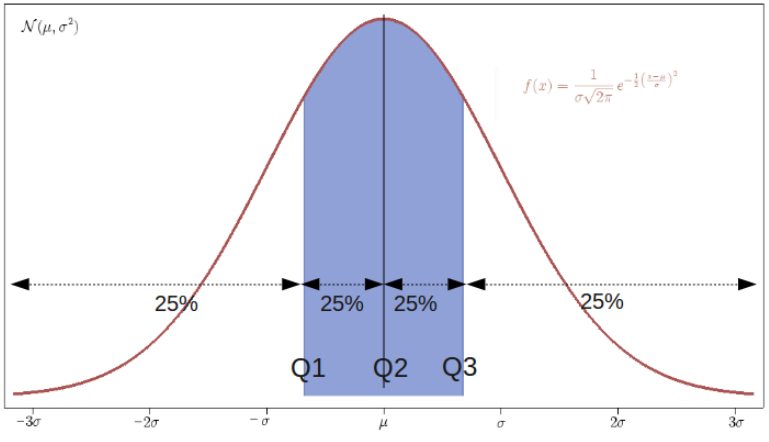
SAMPLING DISTRIBUTION

- mean
- median
- mode
- variance

Quantiles, Quartiles and Percentiles

Quantiles

Cut points dividing the observations in a sample in intervals of equal dimension.



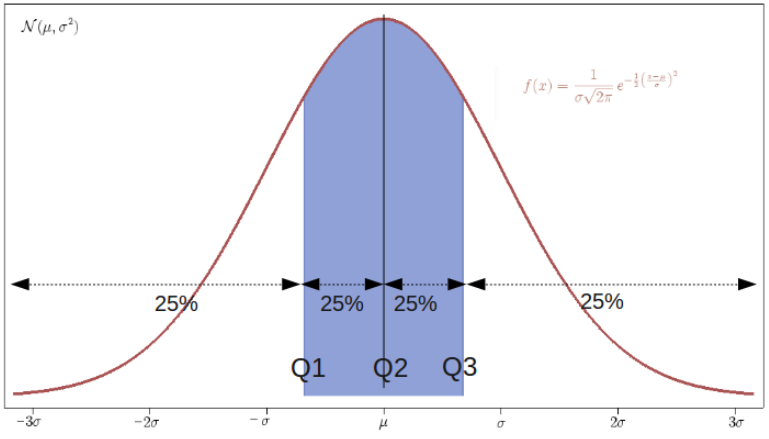
Quantiles, Quartiles and Percentiles

Quantiles

Cut points dividing the observations in a sample in intervals of equal dimension.

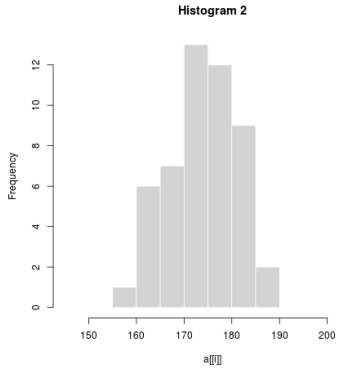
Special cases

- 2-quantile: median
- 4-quantile: quartiles
- 100-quantile: percentiles

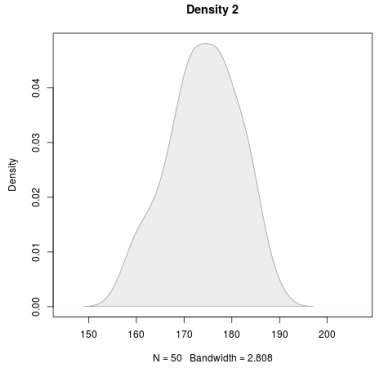


Representing Numerical variables

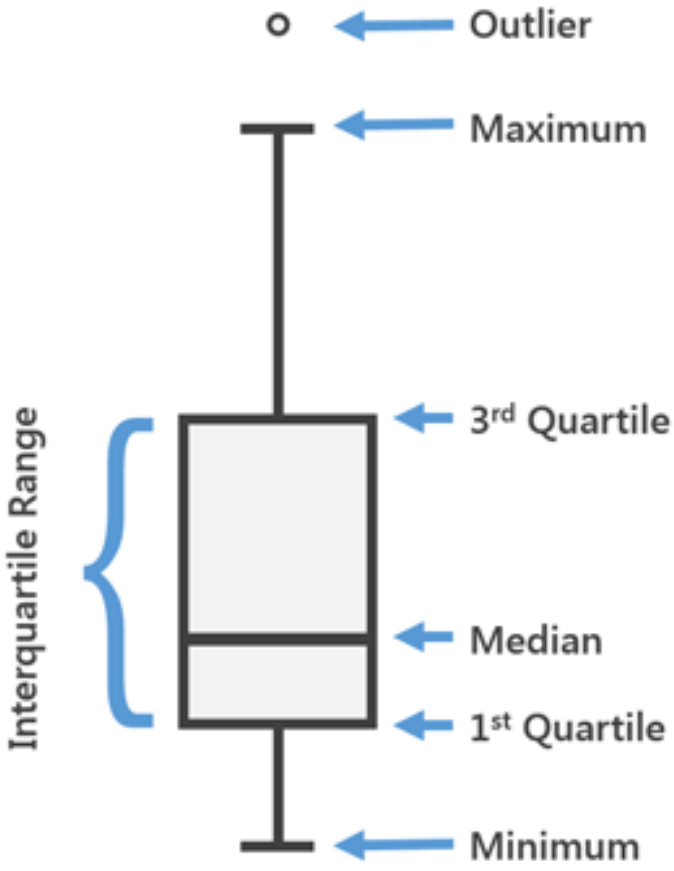
Histogram



Density



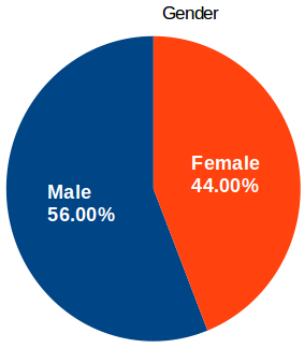
Boxplot



Representing cathegorical variables

Pie Chart

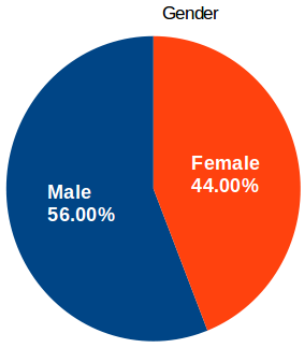
- One cathegorical variable



Representing cathegorical variables

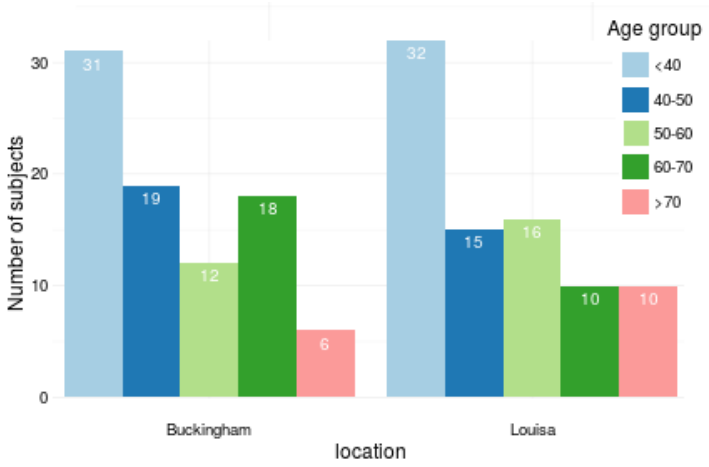
Pie Chart

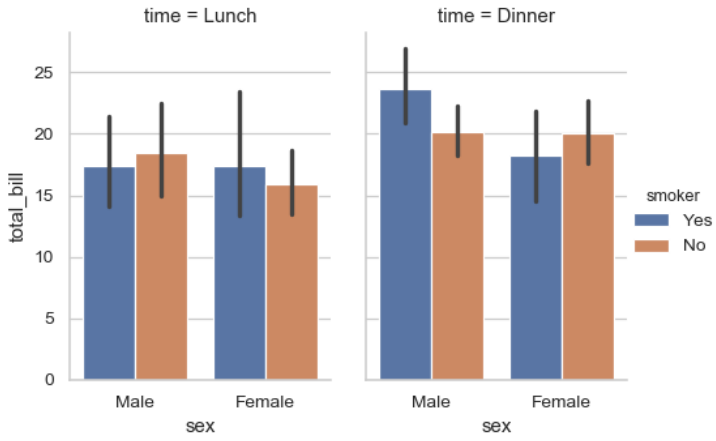
- One cathegorical variable



Bar plot

- Multiple cathegorical variables
- Relation cathegorical and numerical variables





Sampling distribution

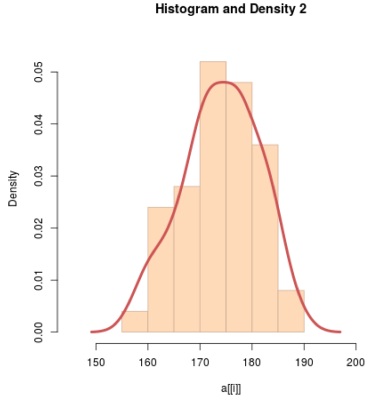
SAMPLE

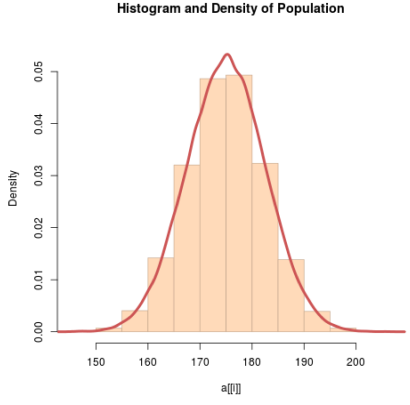
POPULATION

Sampling distribution

SAMPLE

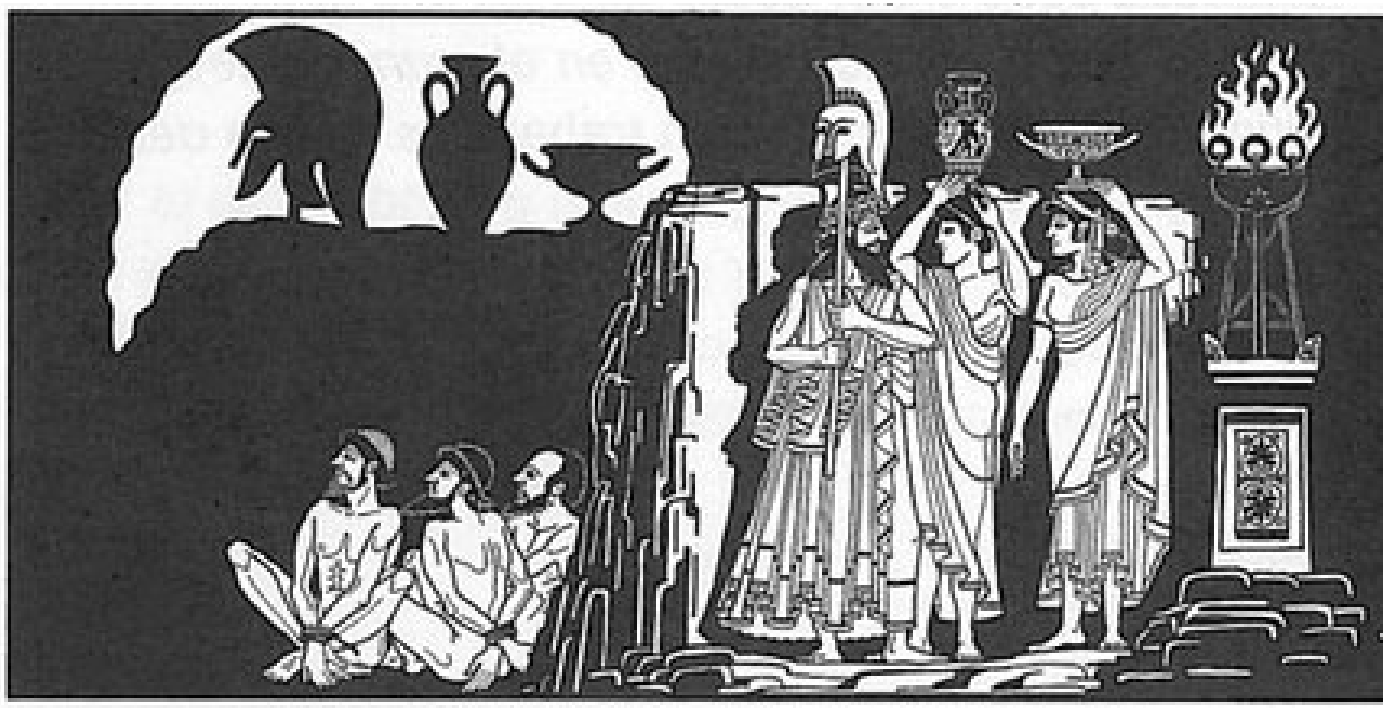
POPULATION





Sample / Population

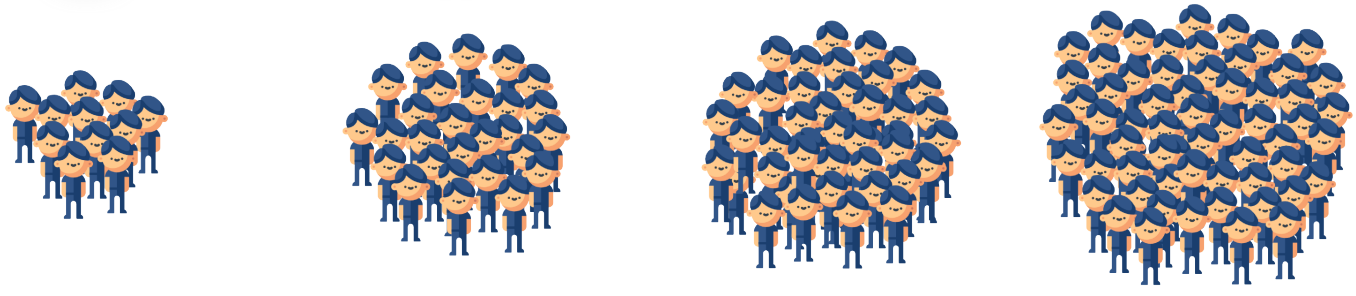
Plato's Cave Myth



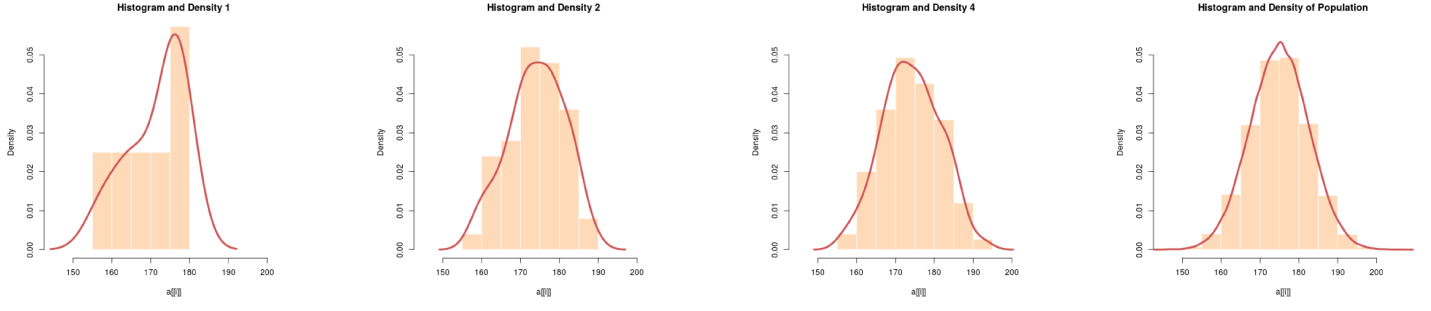
Sample Size

Shadow

N



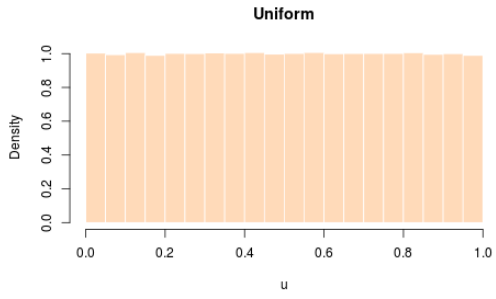
Histogram



Distributions

Uniform

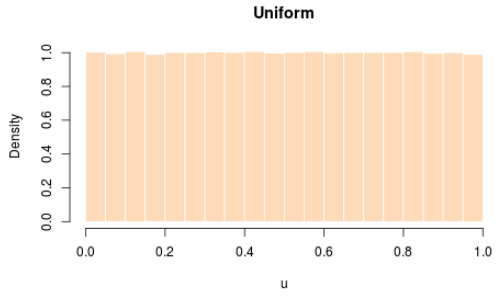
- All values equal probability
- Parameters: min, max



Distributions

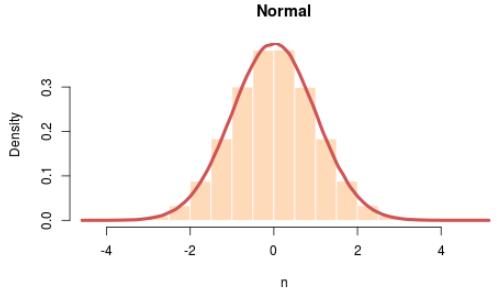
Uniform

- All values equal probability
- Parameters: min, max



Normal

- Gauss Bell
- Parameters: μ, σ



Distributions

Uniform

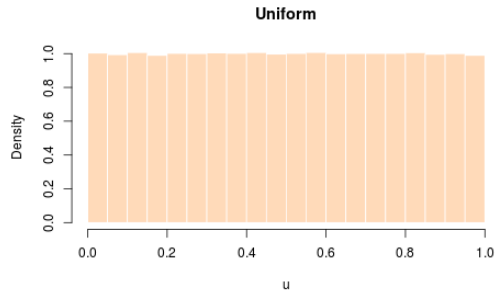
- All values equal probability
- Parameters: min, max

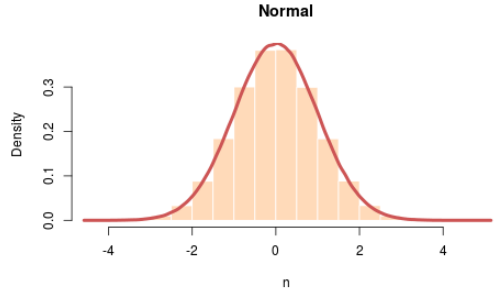
Normal

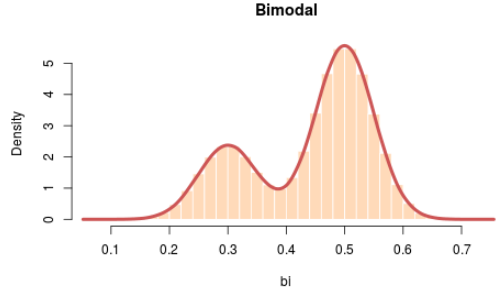
- Gauss Bell
- Parameters: μ, σ

Bimodal

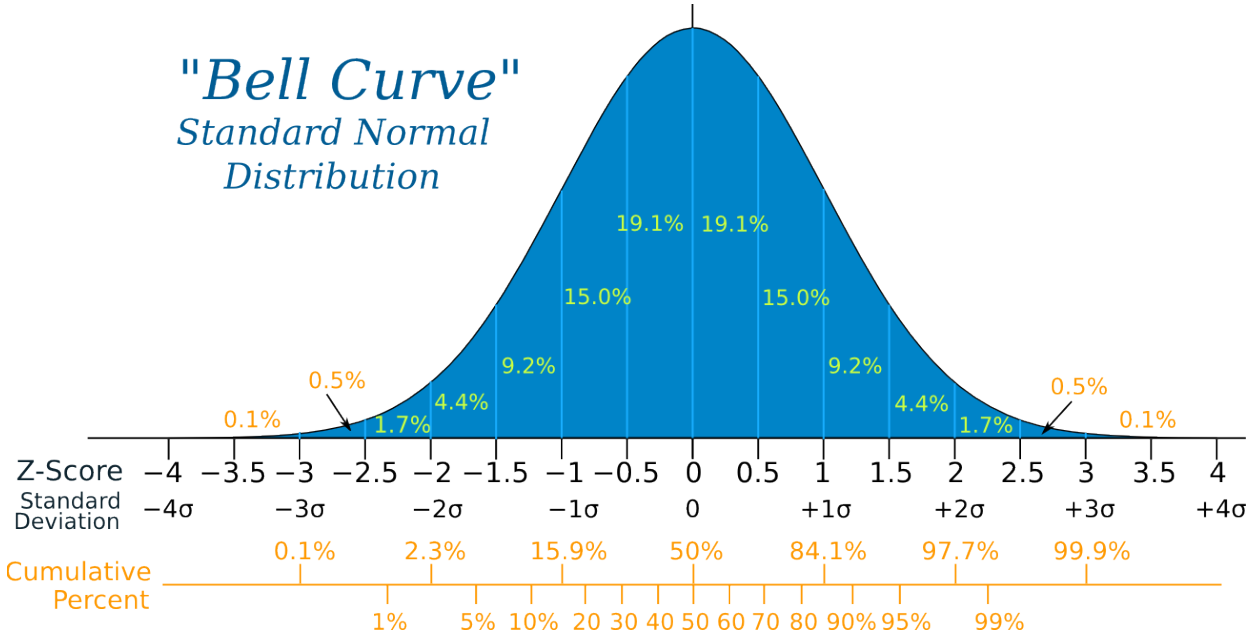
- Two different distributions combined
- Parameters: $(\mu_1, \sigma_1), (\mu_2, \sigma_2)$







Normal distribution



Inference

Inference

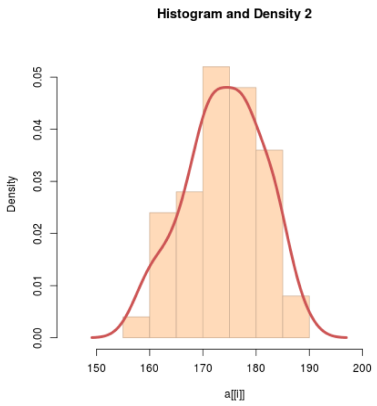
Conclude something about a parameter in the population from the data in the sample

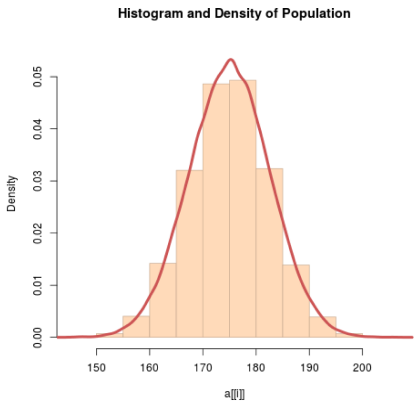
- Confidence intervals
- Hypothesis tests
- Specific tests

Confidence intervals

SAMPLE

POPULATION

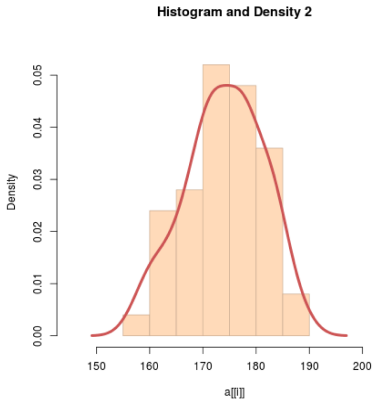




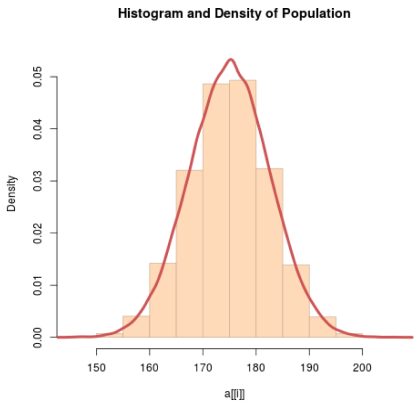
Confidence intervals

SAMPLE

POPULATION



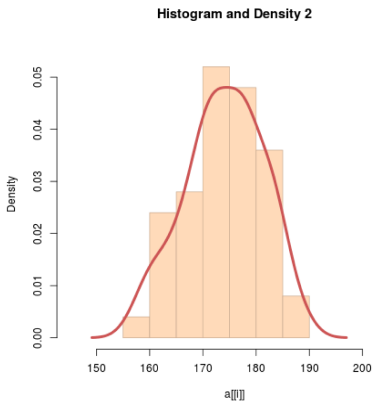
??

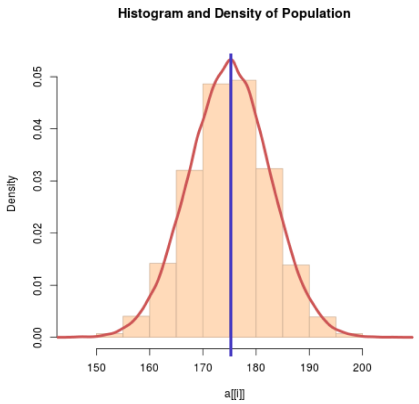


Confidence intervals

SAMPLE

POPULATION



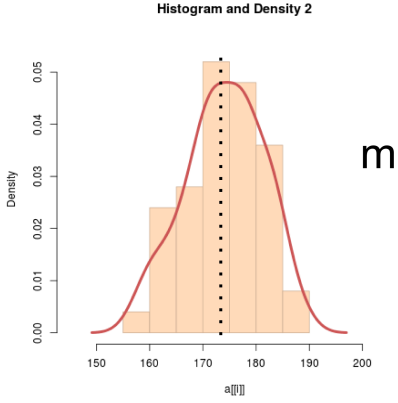


$\mu = 175$

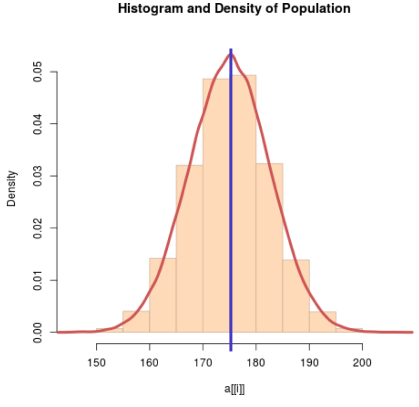
Confidence intervals

SAMPLE

POPULATION



mean = 173

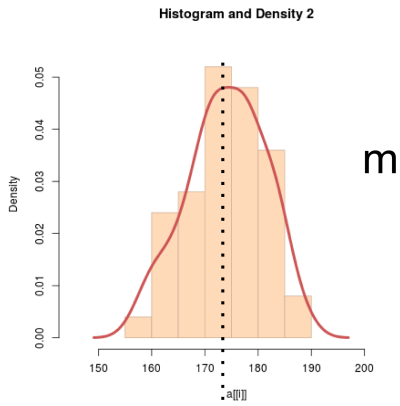


μ = 175

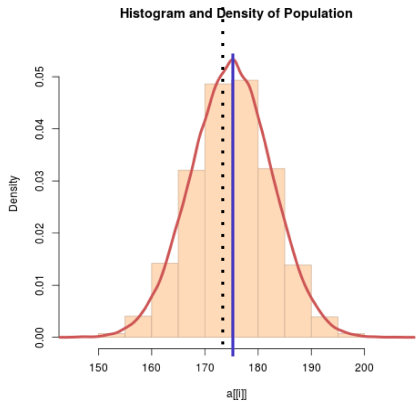
Confidence intervals

SAMPLE

POPULATION



mean = 173

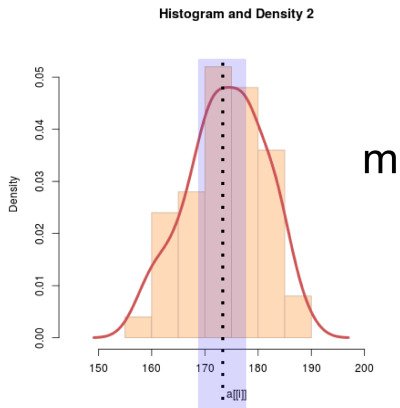


μ = 175

Confidence intervals

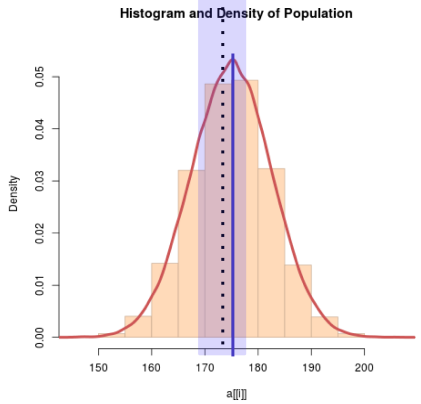
SAMPLE

POPULATION



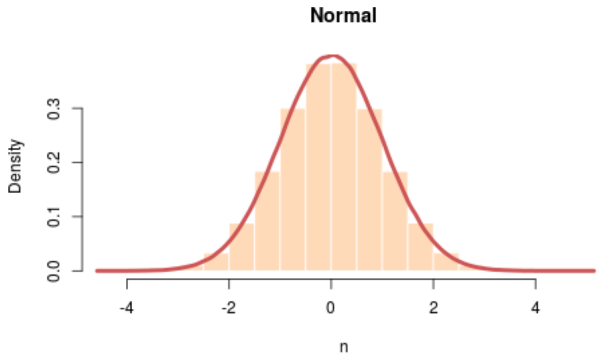
mean = 173

C.I. = [168, 178]

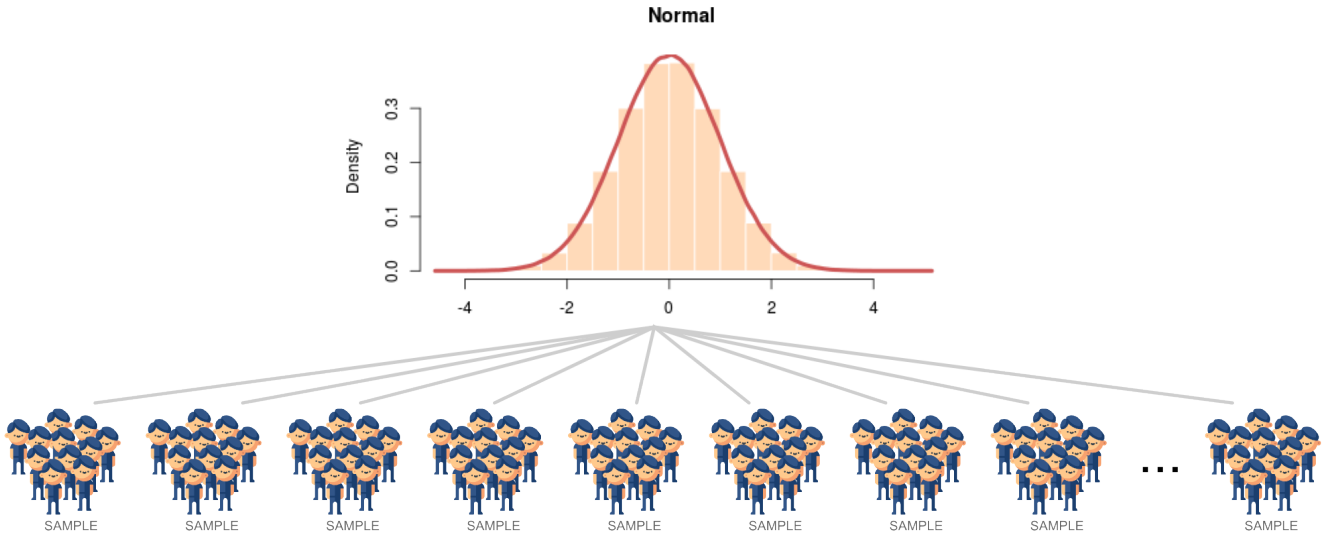


μ = 175

Confidence intervals

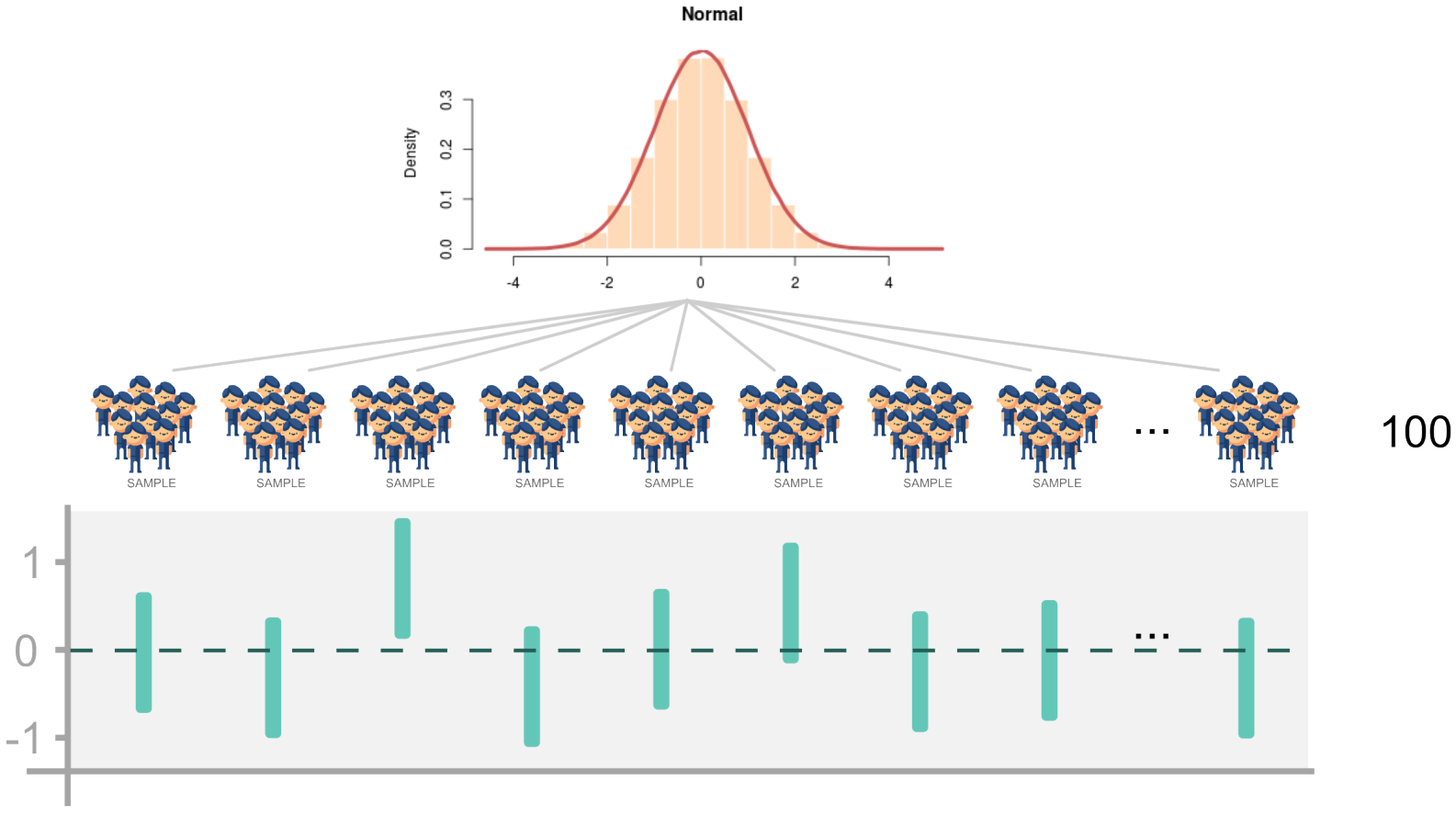


Confidence intervals

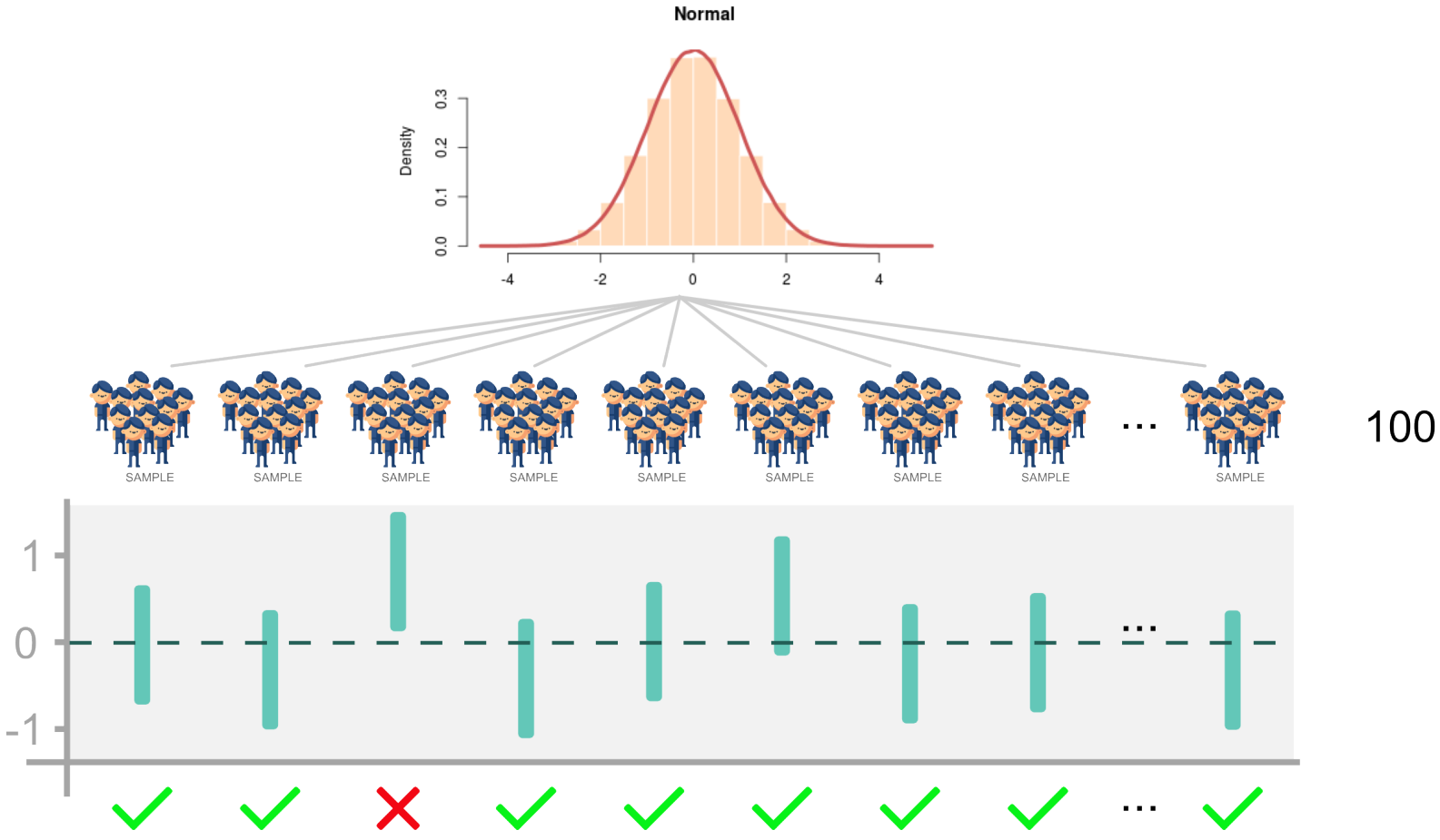


100

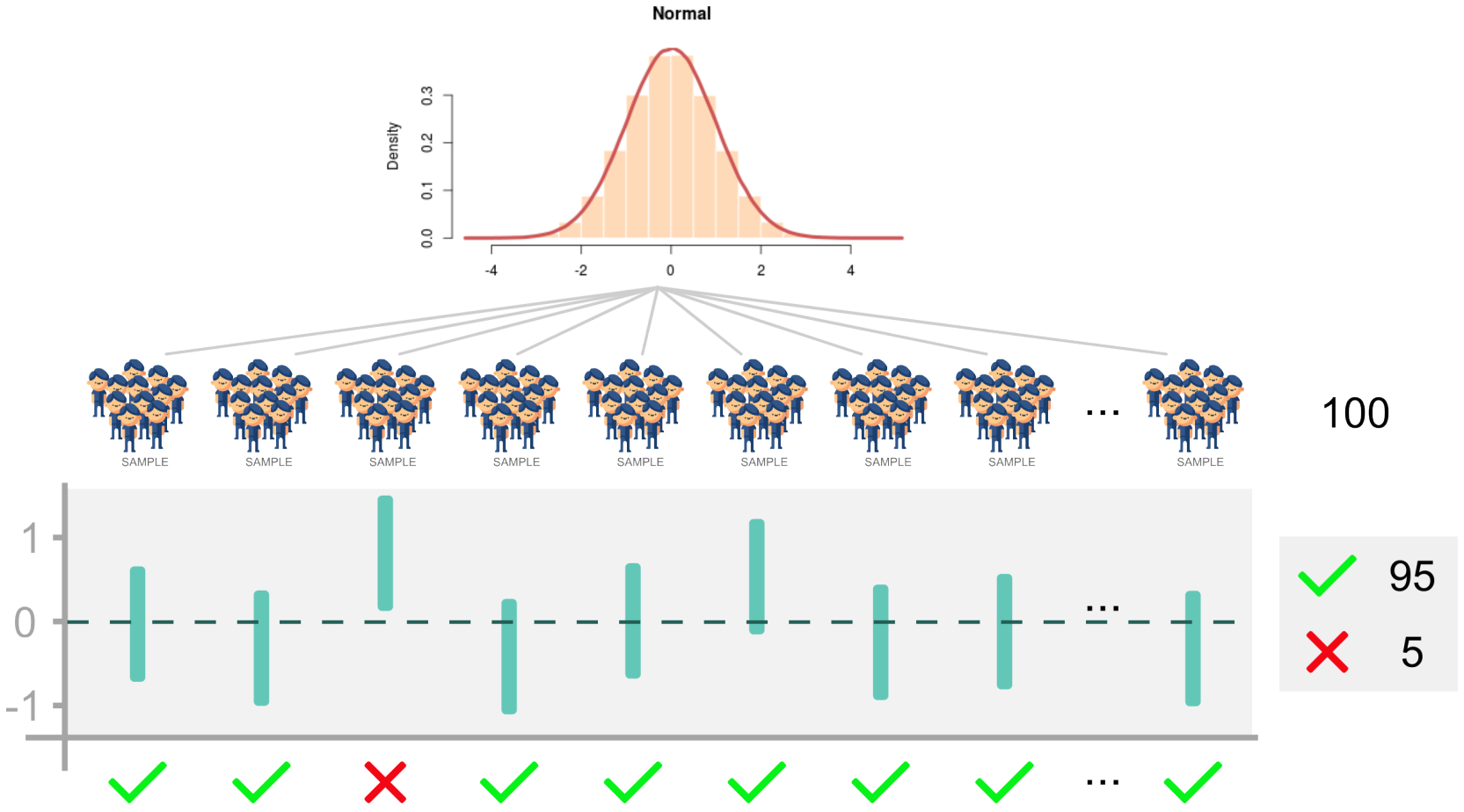
Confidence intervals



Confidence intervals



Confidence intervals



Hypothesis tests

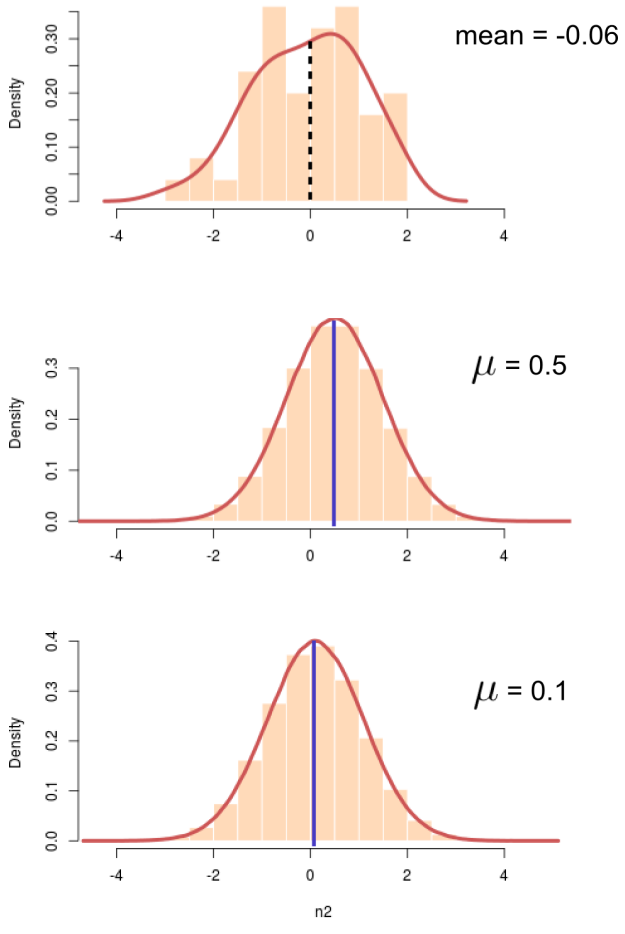
- Null hypothesis:

$$H_0 : \mu = 0.5$$

- Alternative hypothesis:

$$H_A : \mu \neq 0.5$$

- Compute confidence interval I.
- If $0.5 \notin I$, reject H_0.
- If $0.5 \in I$, H_0 is not rejected.



Hypothesis tests

t-test

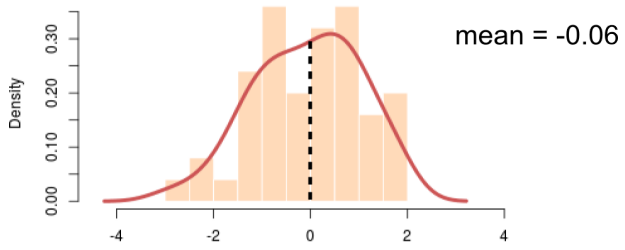
- Null hypothesis:

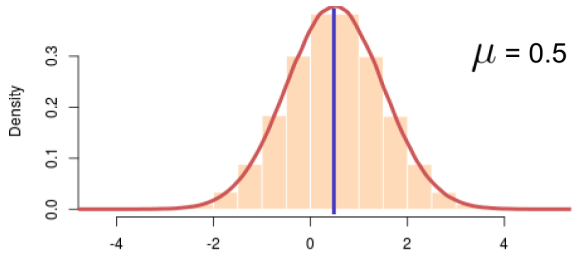
$$H_0 : \mu = 0.5$$

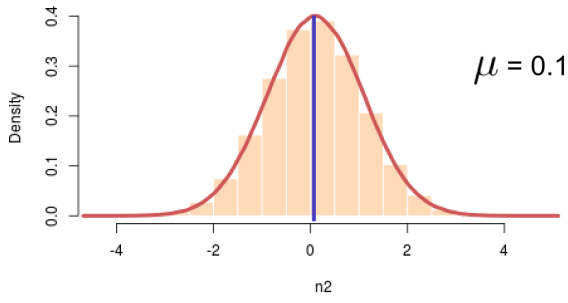
- Alternative hypothesis:

$$H_A : \mu \neq 0.5$$

- Compute confidence interval I.
- If $0.5 \notin I$, reject H_0.
- If $0.5 \in I$, H_0 is not rejected.







Hypothesis tests

t-test

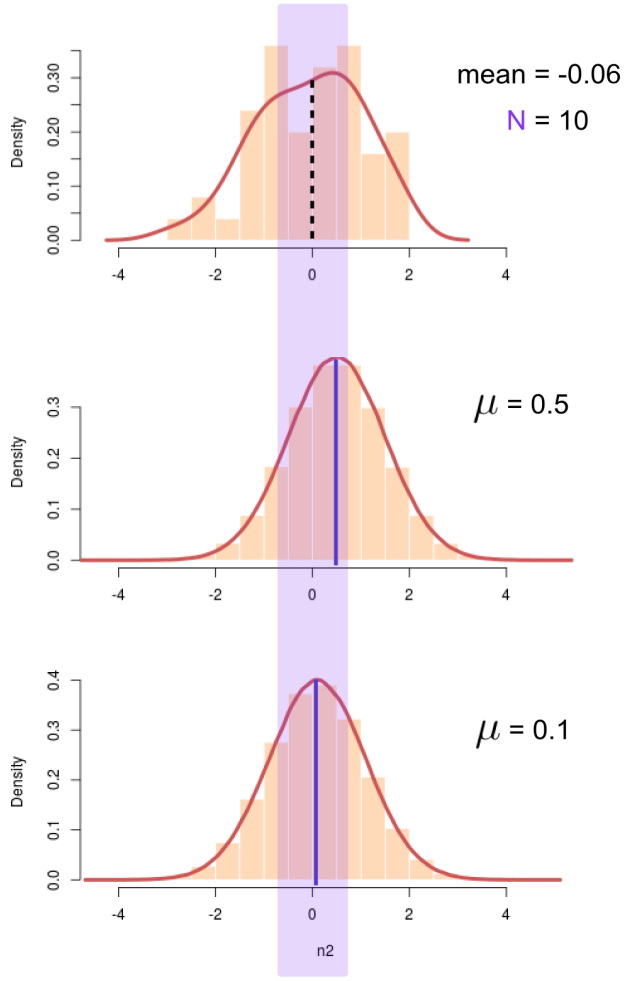
- Null hypothesis:

$$H_0 : \mu = 0.5$$

- Alternative hypothesis:

$$H_A : \mu \neq 0.5$$

- Compute confidence interval I.
- If $0.5 \notin I$, reject H_0.
- If $0.5 \in I$, H_0 is not rejected.



Hypothesis tests

t-test

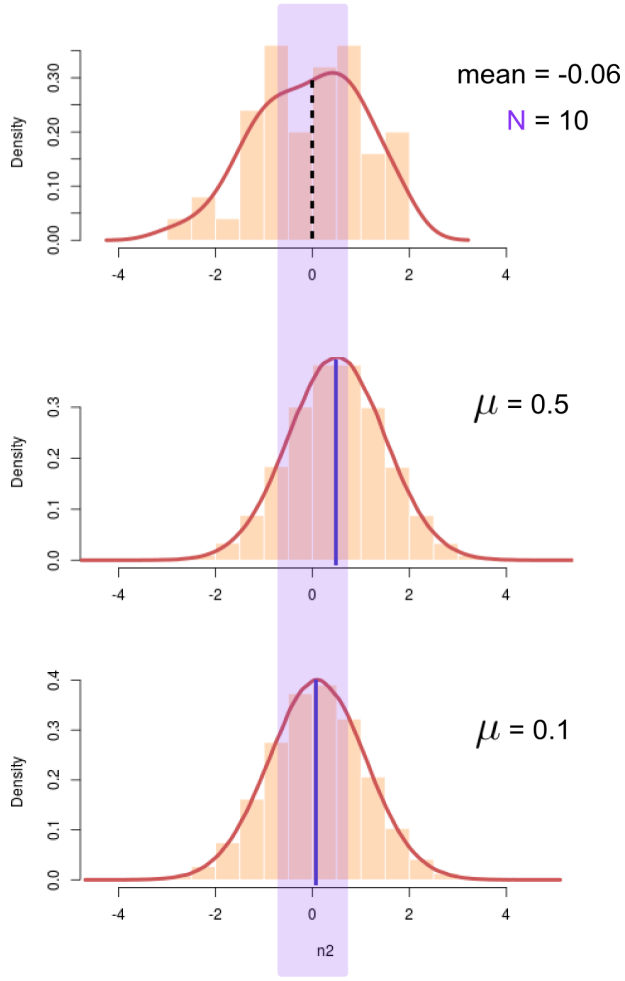
- Null hypothesis:

$$H_0 : \mu = 0.5$$

- Alternative hypothesis:

$$H_A : \mu \neq 0.5$$

- Compute confidence interval I.
- If $0.5 \notin I$, reject H_0.
- If $0.5 \in I$, H_0 is not rejected.



Hypothesis tests

t-test

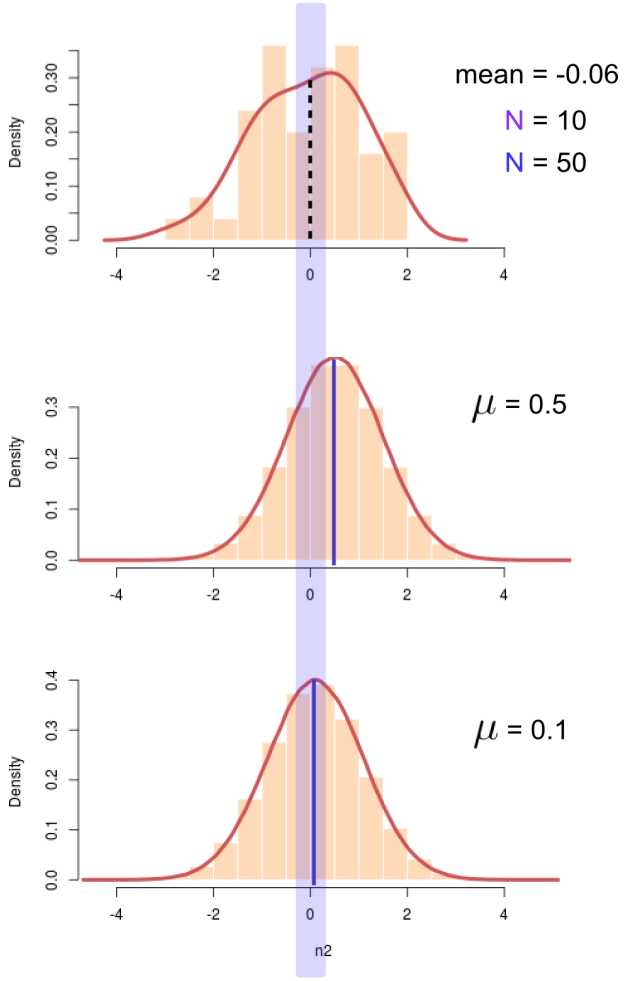
- Null hypothesis:

$$H_0 : \mu = 0.5$$

- Alternative hypothesis:

$$H_A : \mu \neq 0.5$$

- Compute confidence interval I.
- If $0.5 \notin I$, reject H_0.
- If $0.5 \in I$, H_0 is not rejected.



Mean comparison

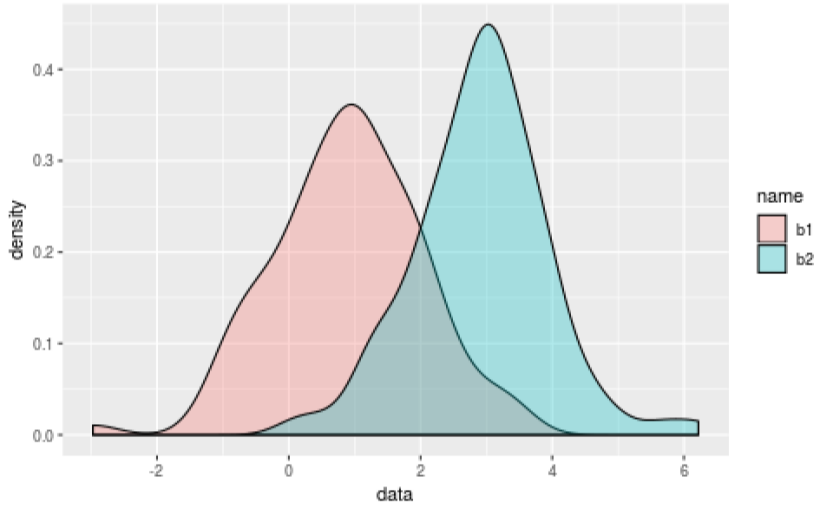
- Null hypothesis:

$$H_0 : \mu_1 = \mu_2$$

- Alternative hypothesis:

$$H_A : \mu_1 \neq \mu_2$$

- t-test, Wilcoxon
- Paired samples?



Mean comparison

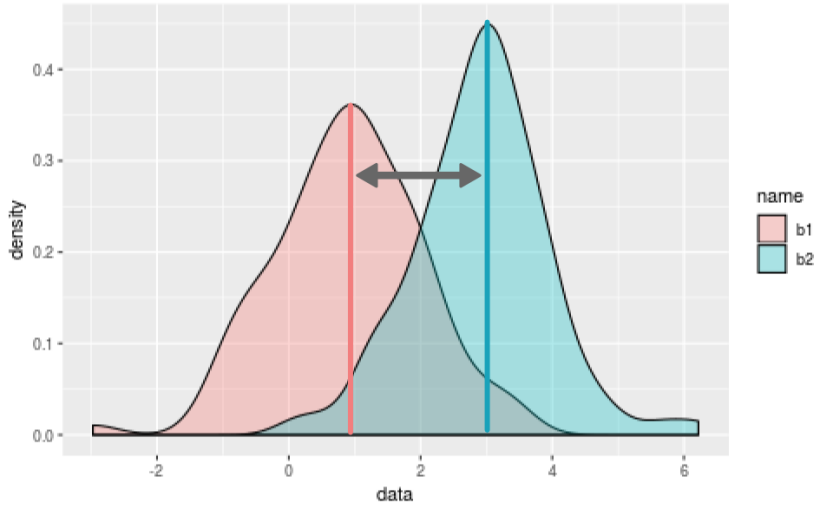
- Null hypothesis:

$$H_0 : \mu_1 = \mu_2$$

- Alternative hypothesis:

$$H_A : \mu_1 \neq \mu_2$$

- t-test, Wilcoxon
- Paired samples?



Mean comparison

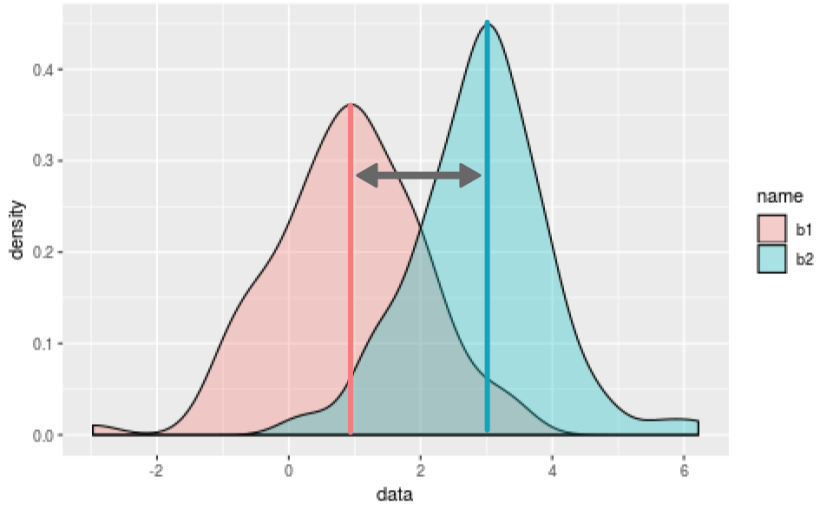
- Null hypothesis:

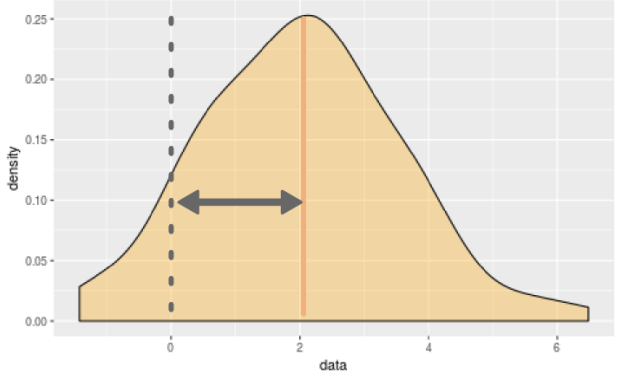
$$H_0 : \mu_1 - \mu_2 = 0$$

- Alternative hypothesis:

$$H_A : \mu_1 - \mu_2 \neq 0$$

- t-test, Wilcoxon
- Paired samples?





Mean comparison

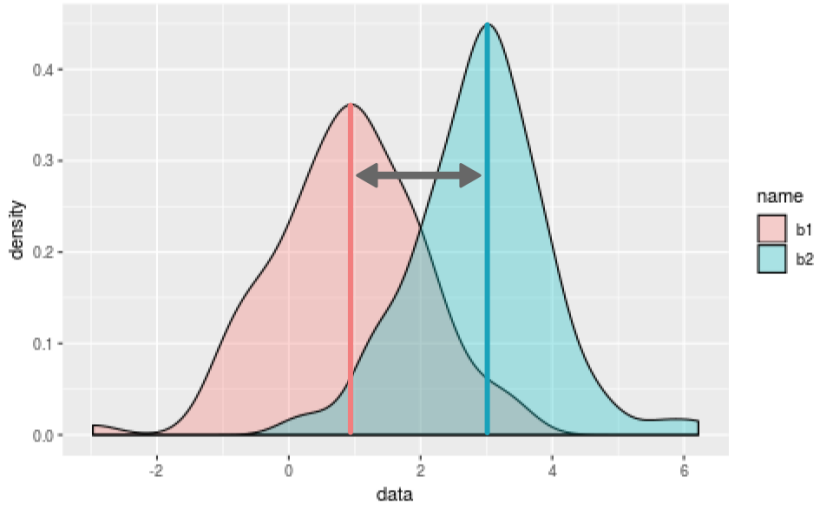
- Null hypothesis:

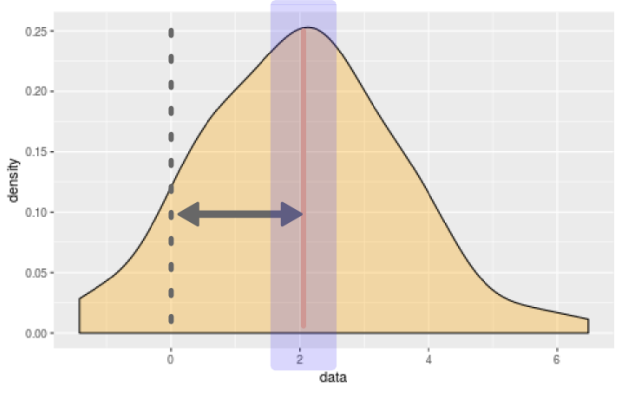
$$H_0 : \mu_1 - \mu_2 = 0$$

- Alternative hypothesis:

$$H_A : \mu_1 - \mu_2 \neq 0$$

- t-test, Wilcoxon
- Paired samples?





Mean comparison

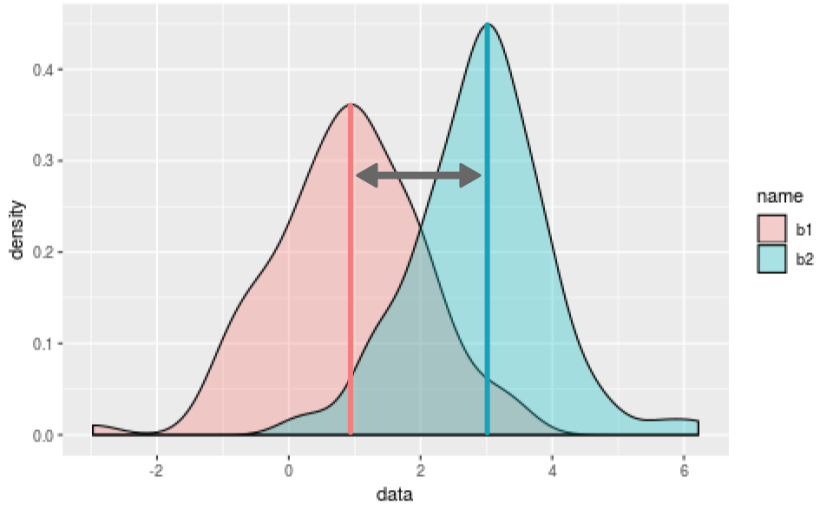
- Null hypothesis:

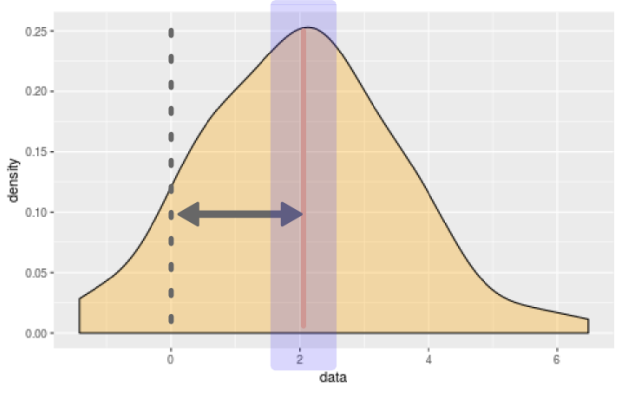
$$H_0 : \mu_1 - \mu_2 = 0$$

- Alternative hypothesis:

$$H_A : \mu_1 - \mu_2 \neq 0$$

- t-test, Wilcoxon
- Paired samples?





Fisher Test

Contingency table

Matrix that displays the frequency distribution of the variables

	Infected	Not infected	
Inoculated	3	276	279
Not inoculated	66	473	539
	69	749	818

Cholera Inoculation Study, 1894-96

- H_0 : Proportions are the same
- H_A: Proportions are different

Fisher Test

Contingency table

Matrix that displays the frequency distribution of the variables

	Infected	Not infected	
Inoculated	3 1%	276 99%	279
Not inoculated	66 12%	473 88%	539
	69	749	818

Cholera Inoculation Study, 1894-96

- H_0 : Proportions are the same
- H_A: Proportions are different

Fisher Test

Contingency table

Matrix that displays the frequency distribution of the variables

	Infected	Not infected	
Inoculated	3 1%	276 99%	279
Not inoculated	66 12%	473 88%	539
	69	749	818

Cholera Inoculation Study, 1894-96

- H_0 : Proportions are the same
- H_A: Proportions are different

The end

Enjoy the course!!