

D1EAD – Análise Estatística para Ciência de Dados

2021.1



Data Distributions (Part 3)

Prof. Ricardo Sovat

sovat@ifsp.edu.br

Prof. Samuel Martins (Samuka)

samuel.martins@ifsp.edu.br



Problem

Suppose the heights of the inhabitants of a city are **normally distributed** with **population standard deviation** of 20 cm.

We measure the heights of 40 randomly chosen people, and get a **mean height** of 1.75 m.

Construct a **confidence interval** for the **population mean** with a **significance level of 5%**.

Let's recap some concepts

fixed number
from population

change from
sample to sample

Population Parameter	Sample Statistic	Description
----------------------	------------------	-------------

N	n	Number of elements.
μ	\bar{x}	Mean
σ	s	Standard deviation
ρ	r	Correlation coefficient.



PARAMETER

A number that describes
the data from a population

MEAN

STANDARD
DEVIATION



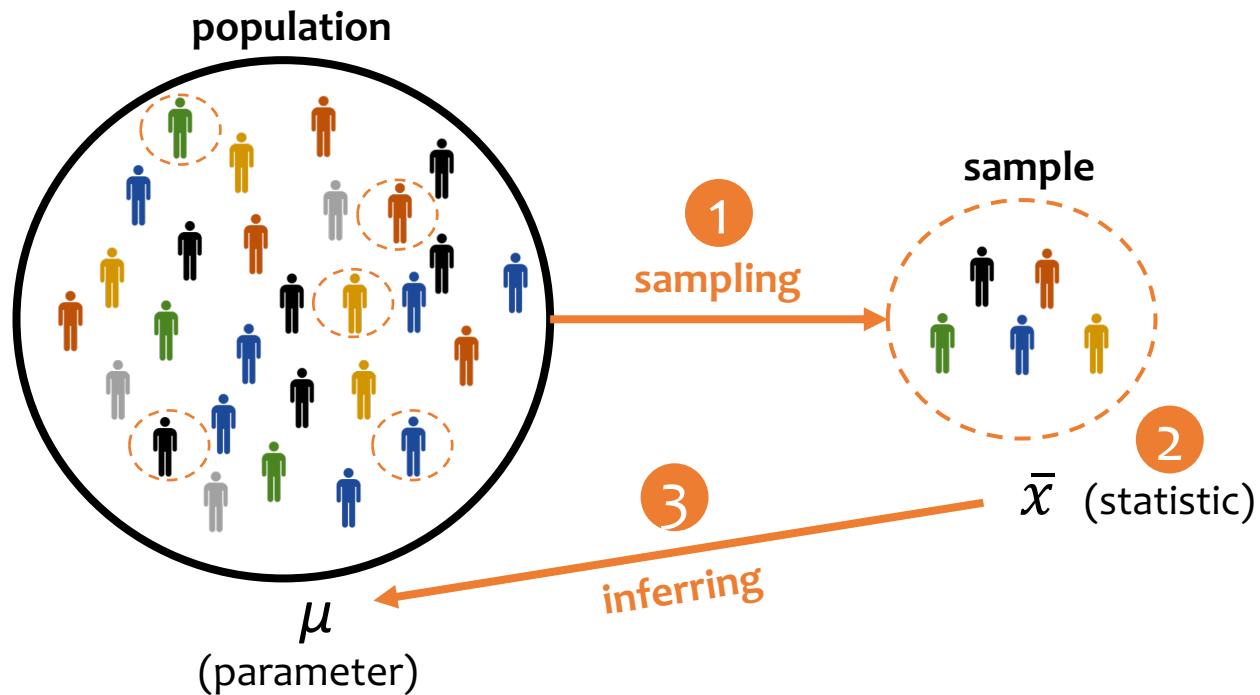
STATISTIC

A number that describes
the data from a sample

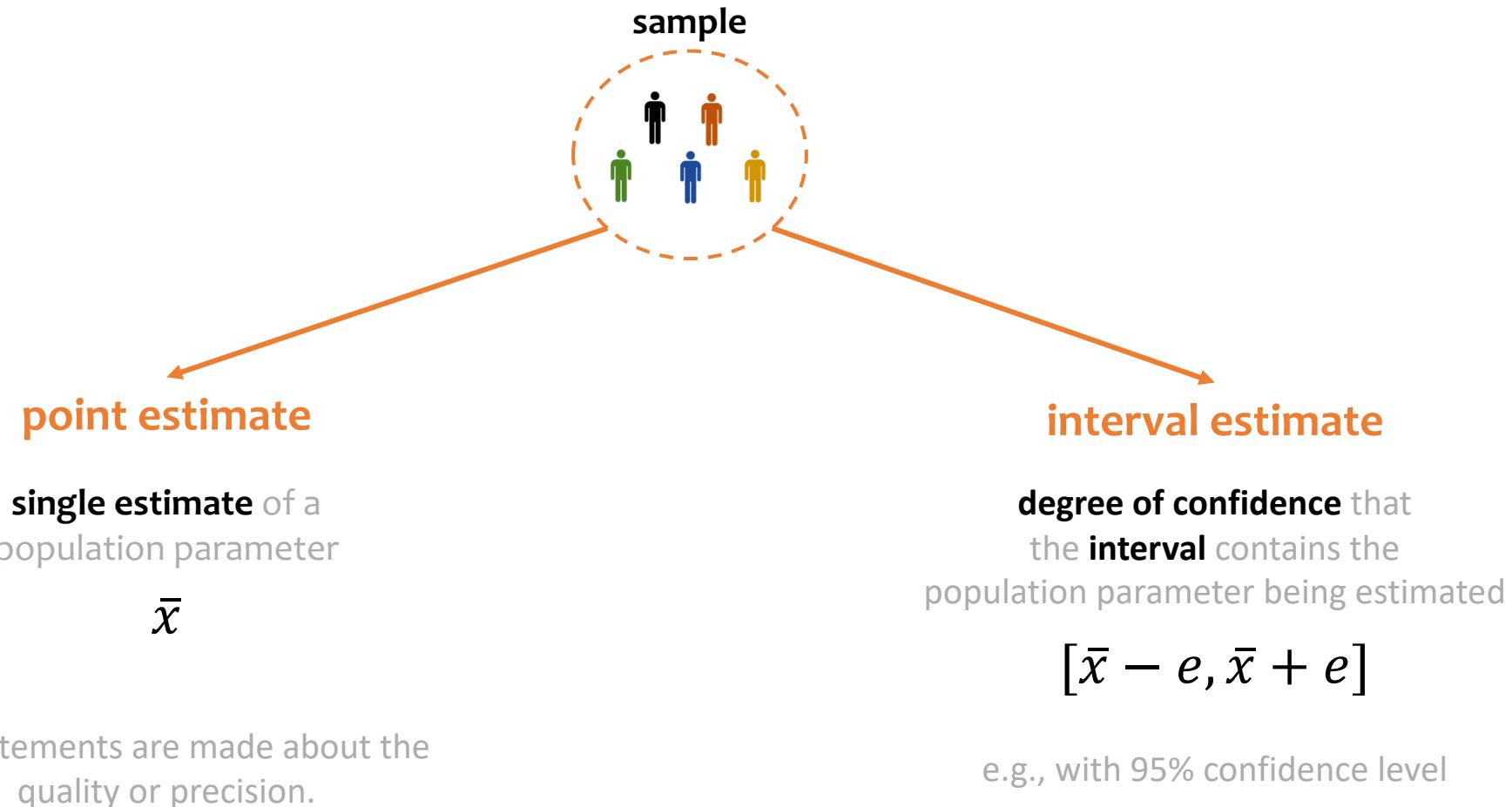
Estimation

Estimation

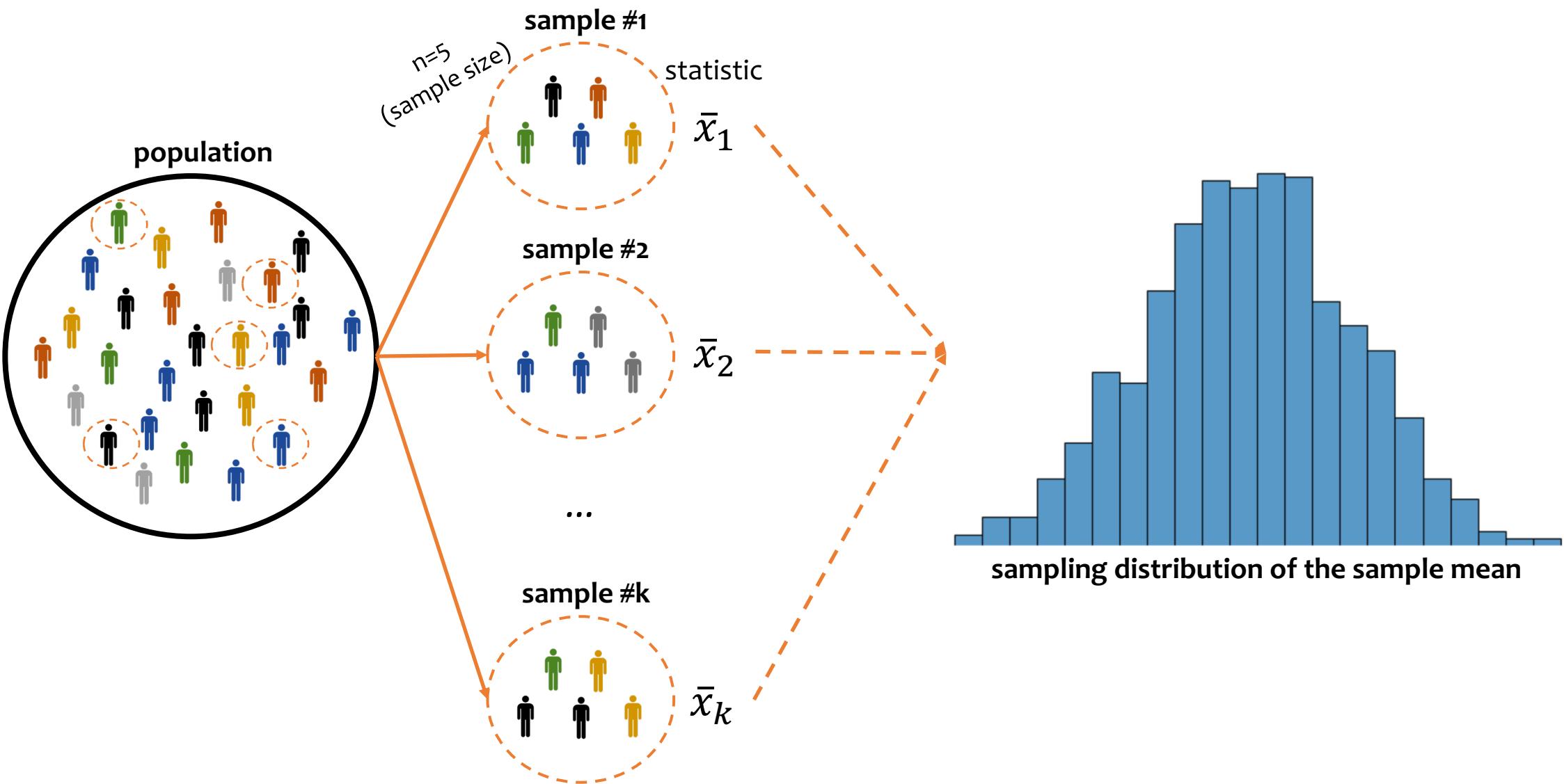
The process of using a **sample** to make **inferences** about a **population** is called **statistical inference**.



Types of Estimation

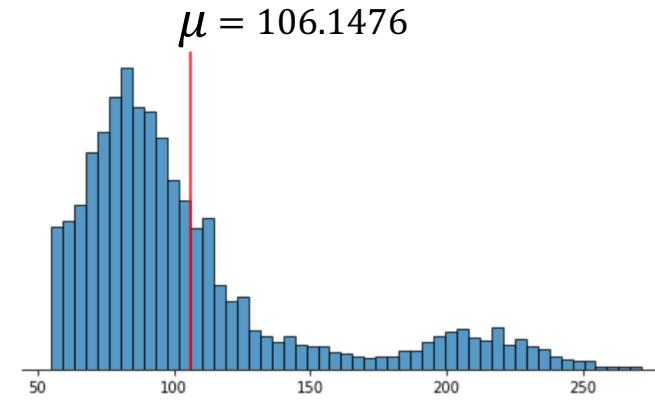
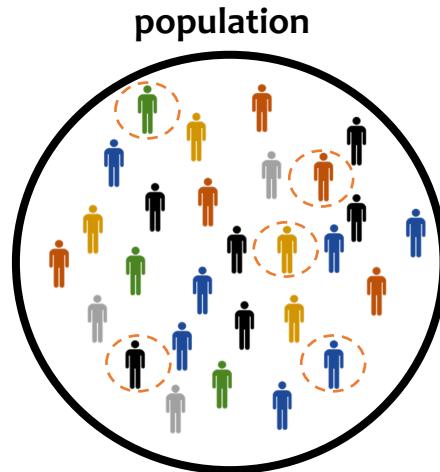


Sampling Distribution of a Statistic



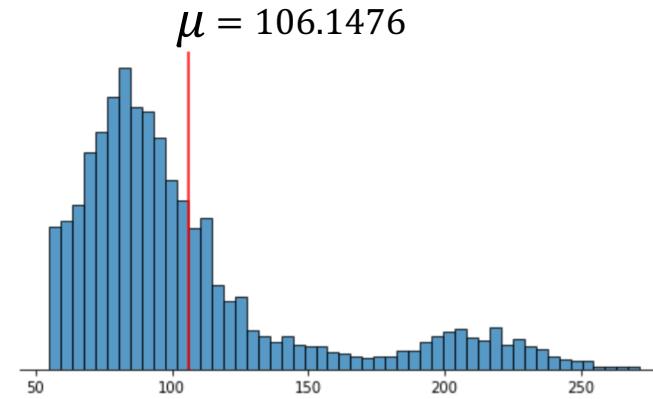
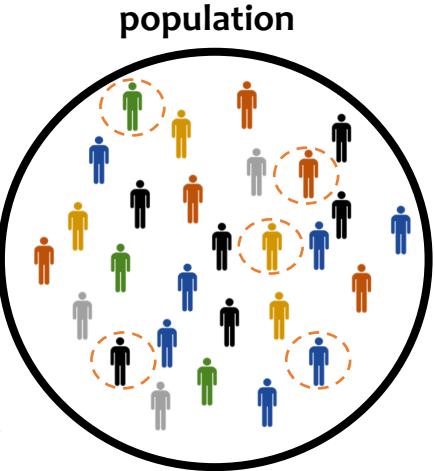
Central Limit Theorem

Central Limit Theorem



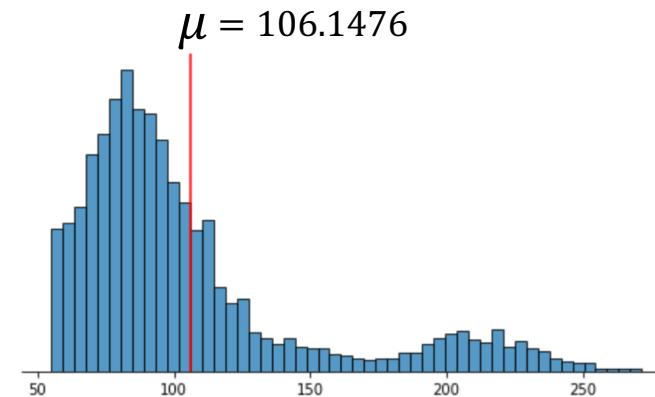
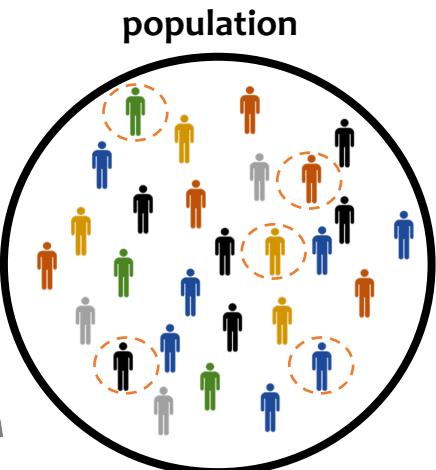
Central Limit Theorem

As the **sample size increases**,
the **sampling distribution** of
the mean approaches a
normal distribution
(no matter the population
distribution)

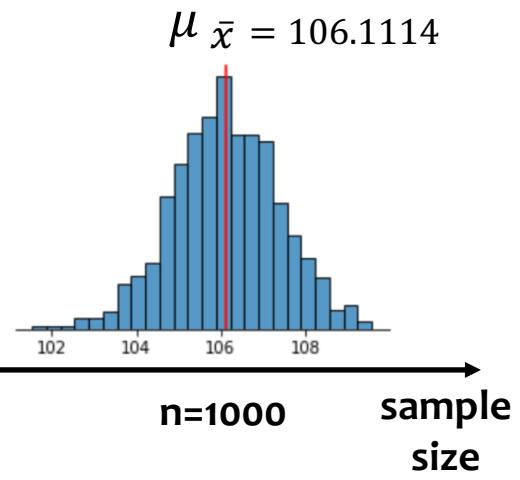
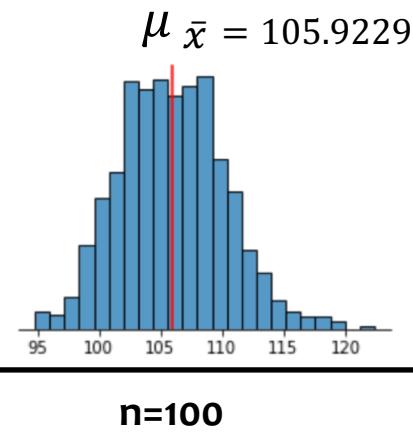
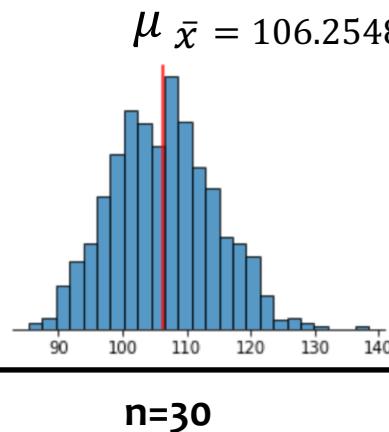
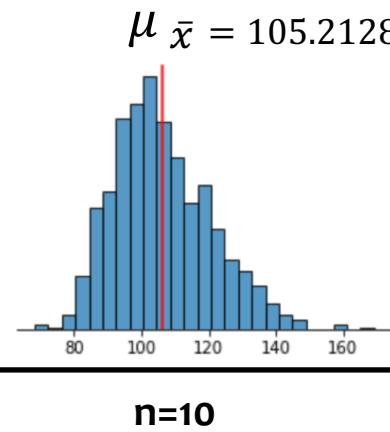
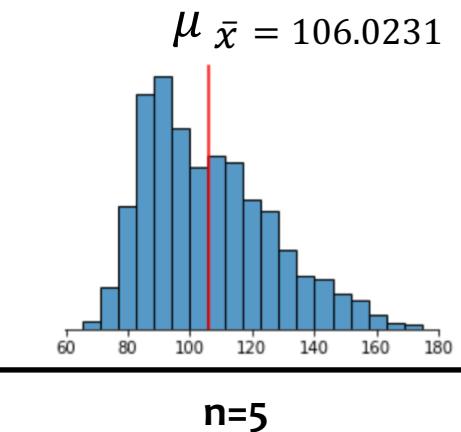


Central Limit Theorem

As the **sample size increases**,
the **sampling distribution** of
the mean approaches a
normal distribution
(no matter the population
distribution)

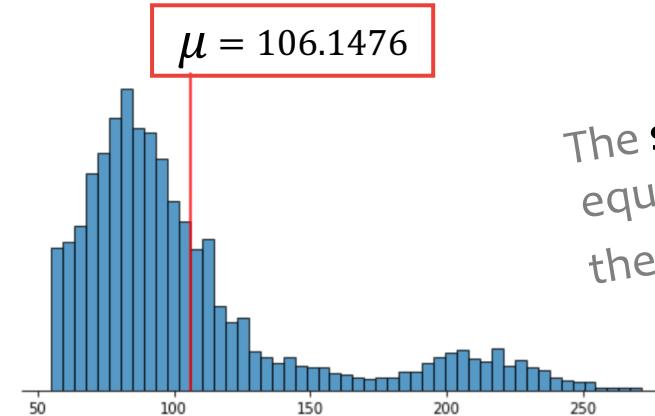
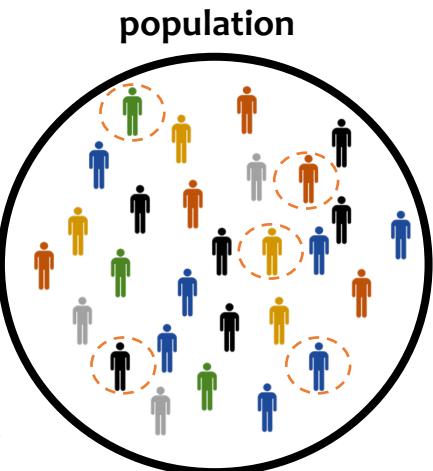


sampling distribution of the mean \bar{x}

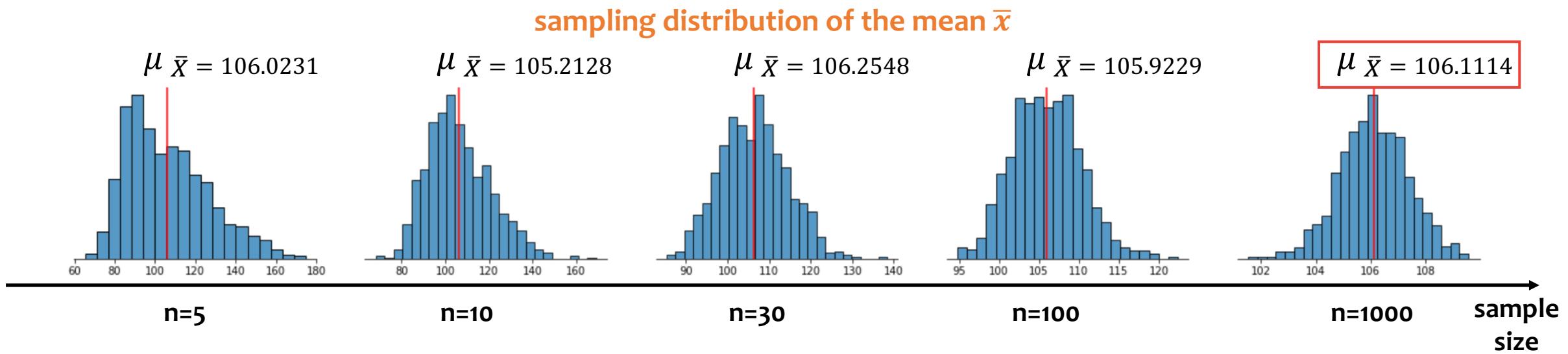


Central Limit Theorem

As the **sample size increases**,
the **sampling distribution** of
the mean approaches a
normal distribution
(no matter the population
distribution)

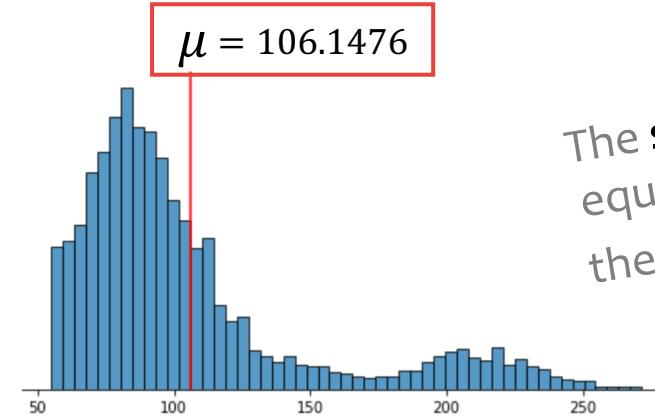
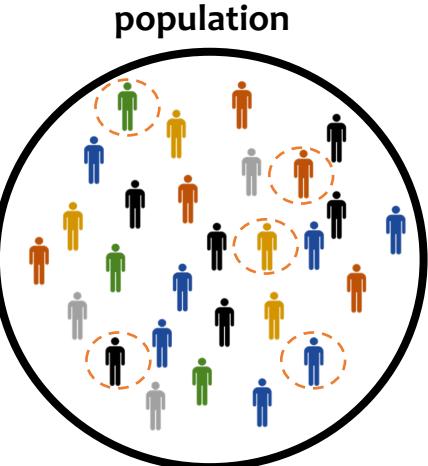


The **sampling distribution's mean**
equals the **population mean**, and
the **standard deviation** equals $\frac{\sigma}{\sqrt{n}}$.



Central Limit Theorem

As the **sample size increases**,
the **sampling distribution** of
the mean approaches a
normal distribution
(no matter the population
distribution)

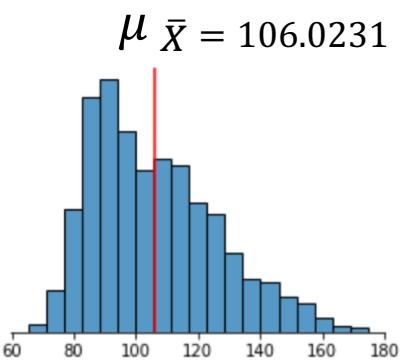


The **sampling distribution's mean**
equals the **population mean**, and
the **standard deviation** equals $\frac{\sigma}{\sqrt{n}}$

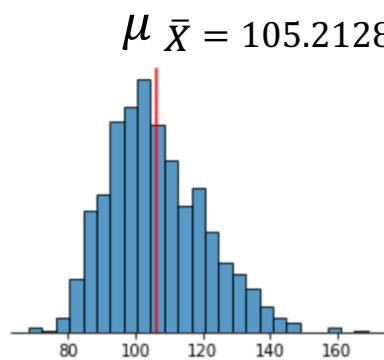
really useful to **estimate**
population parameters

sampling distribution of the mean \bar{x}

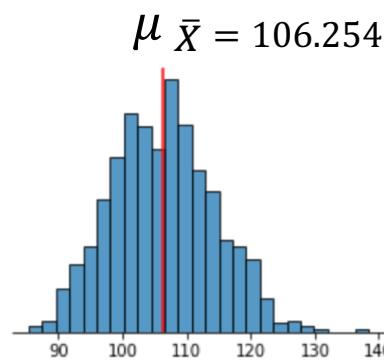
$$\mu \bar{X} = 106.0231$$



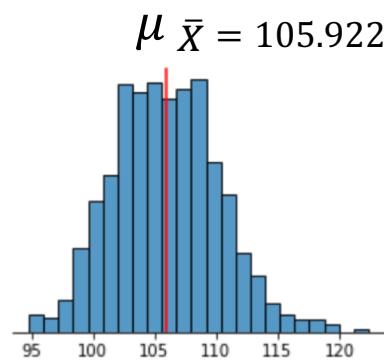
$$\mu \bar{X} = 105.2128$$



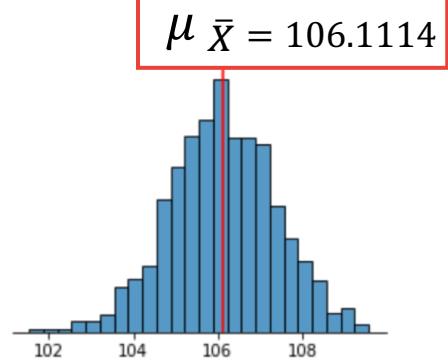
$$\mu \bar{X} = 106.2548$$



$$\mu \bar{X} = 105.9229$$



$$\mu \bar{X} = 106.1114$$



$n=5$

$n=10$

$n=30$

$n=100$

$n=1000$

sample
size

Properties

For a large **enough sample size n**:

- The sampling distribution for the mean tends to a **normal distribution** ($n \geq 30$);
- Sampling distribution's mean = Population mean

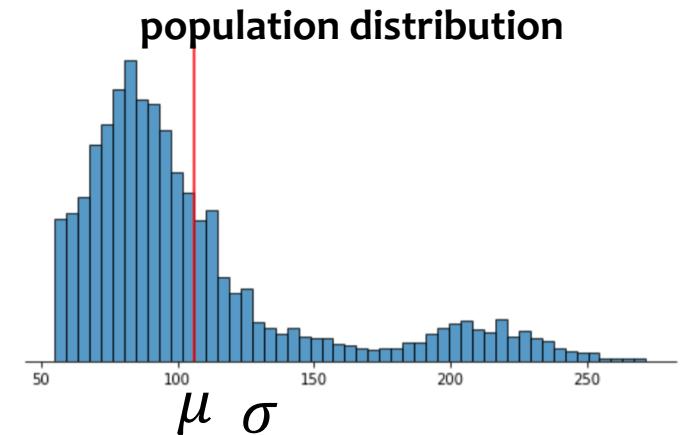
$$\bullet \mu_{\bar{X}} = \mu$$

- Sampling distribution's standard deviation (**standard error**):

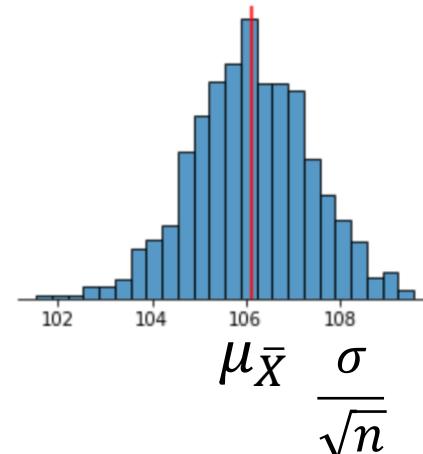
$$\bullet \sigma_{\bar{X}} = SE_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

how far the **sample mean \bar{x}** is
likely to be from the
population mean μ

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

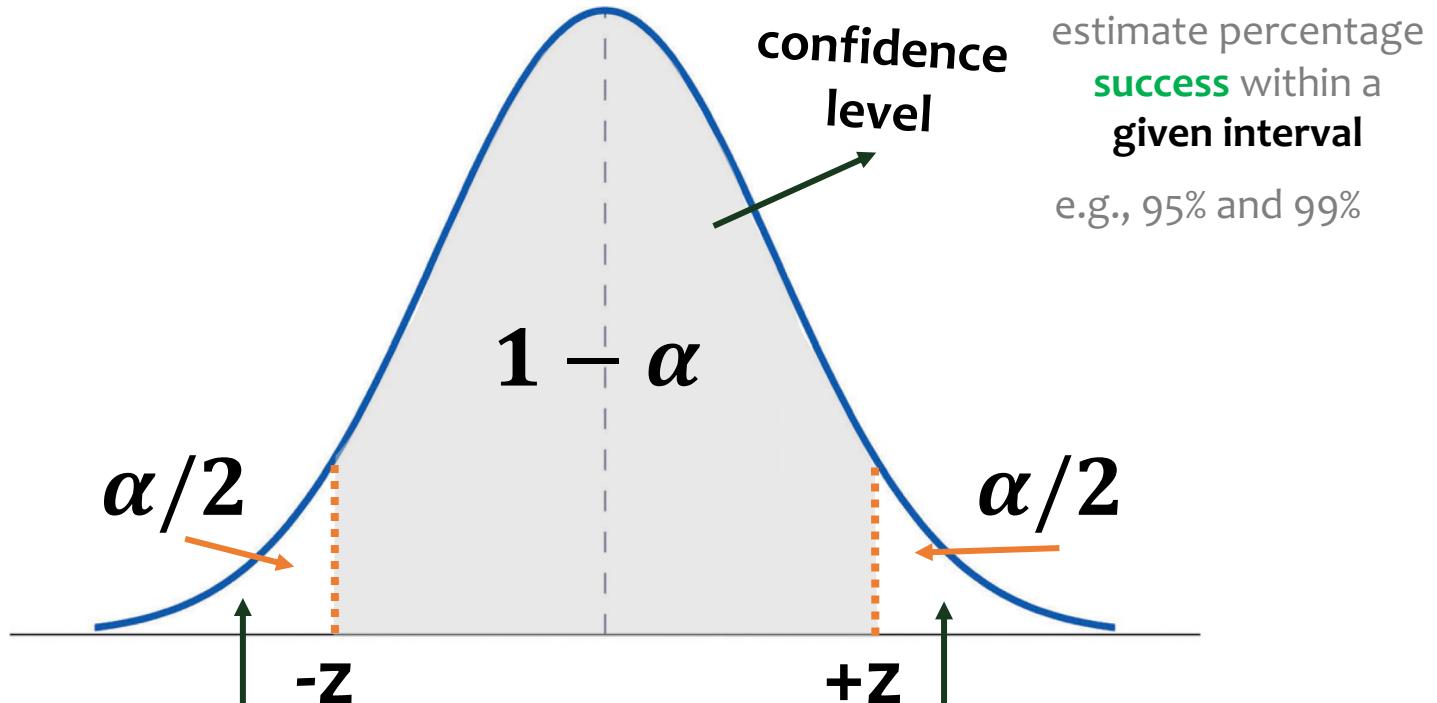


**sampling distribution
of the mean \bar{x}**



Confidence and Significance Levels

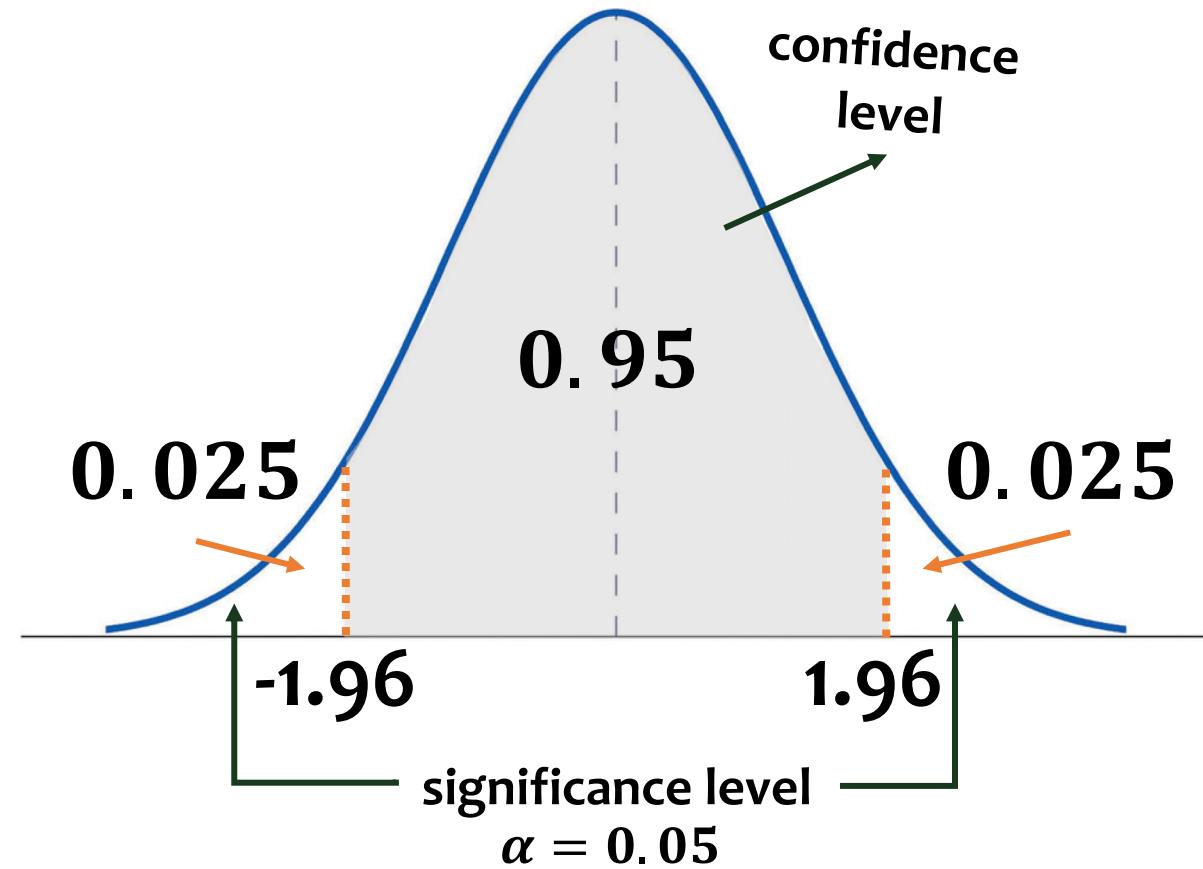
computed from a
(standard) normal
distribution

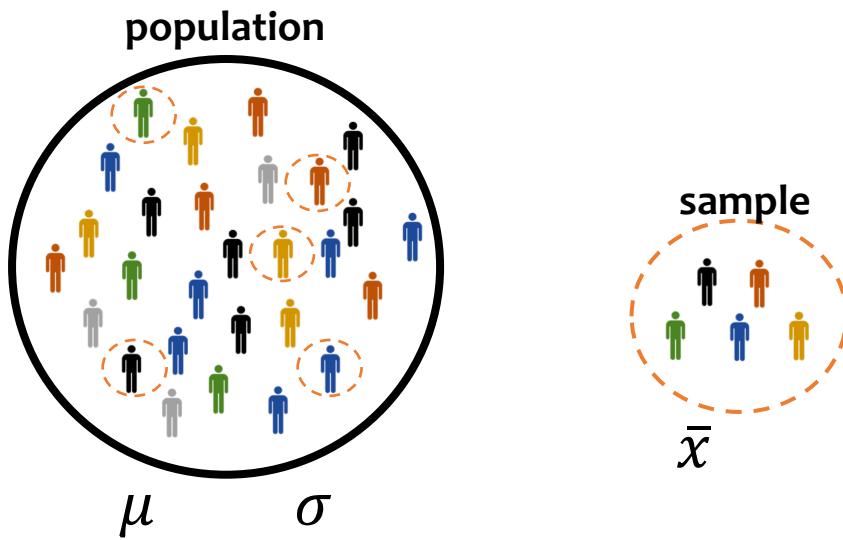


estimate percentage
success within a
given interval
e.g., 95% and 99%

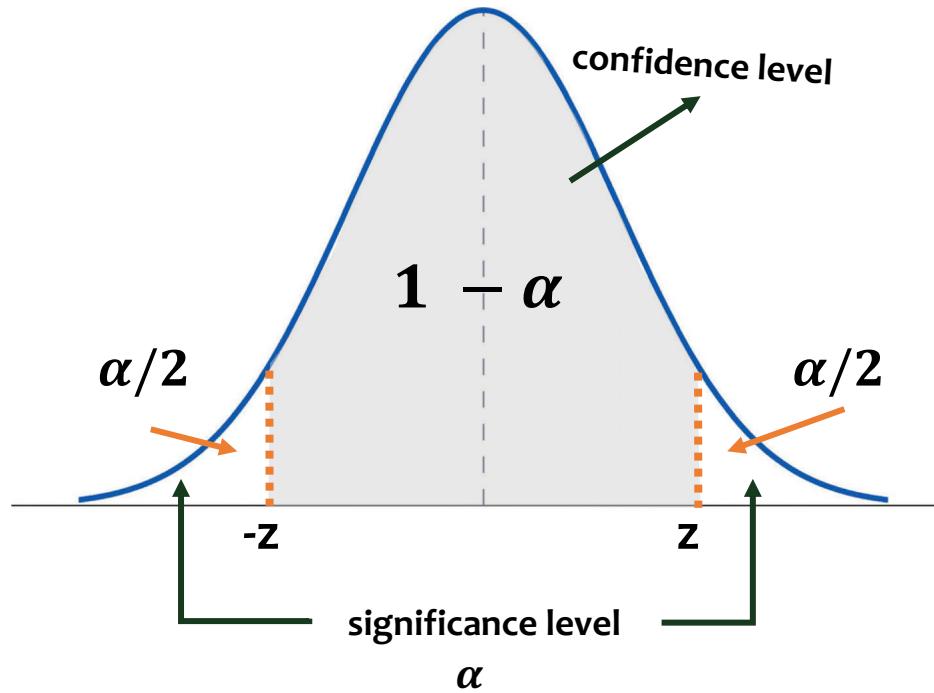
estimate percentage
failure within a
given interval

When we set a **95% confidence level** in a survey, for example, we assume that there is a **95% probability** that the survey results **represent reality**.





Provide an interval for the population mean μ .
with a significance level of α .



Margin of Error

$$e = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

population standard deviation

critical value

sample size

standard error

the z-score for which
an area $\alpha/2$ lies.

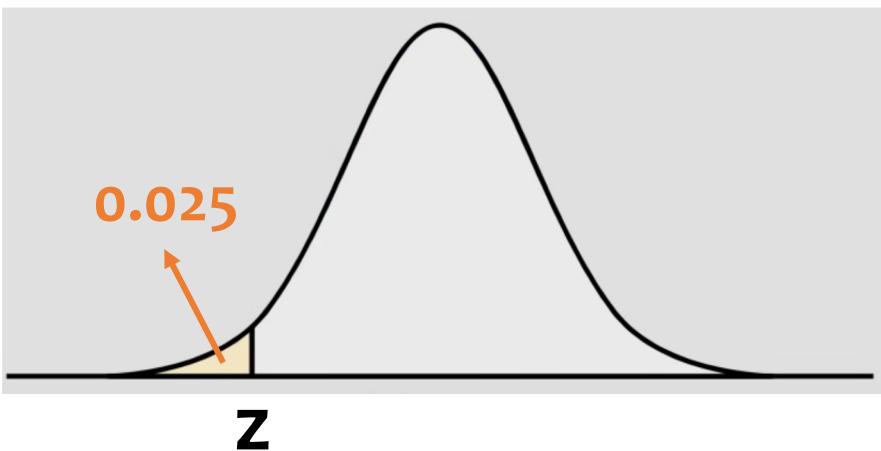
The diagram illustrates the components of the Margin of Error formula. The formula is $e = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$. Brackets group the terms into three parts: 'critical value' (containing $z_{\alpha/2}$), 'standard error' (containing $\frac{\sigma}{\sqrt{n}}$), and 'sample size' (containing \sqrt{n}). A blue arrow points from the 'critical value' bracket to the text 'the z-score for which an area $\alpha/2$ lies.'

Margin of Error

95% confidence level: $1 - \alpha = 0.95$

significance level: $\alpha = 0.05$

$z_{\alpha/2} = z_{0.05/2} = z_{0.025} = ???$



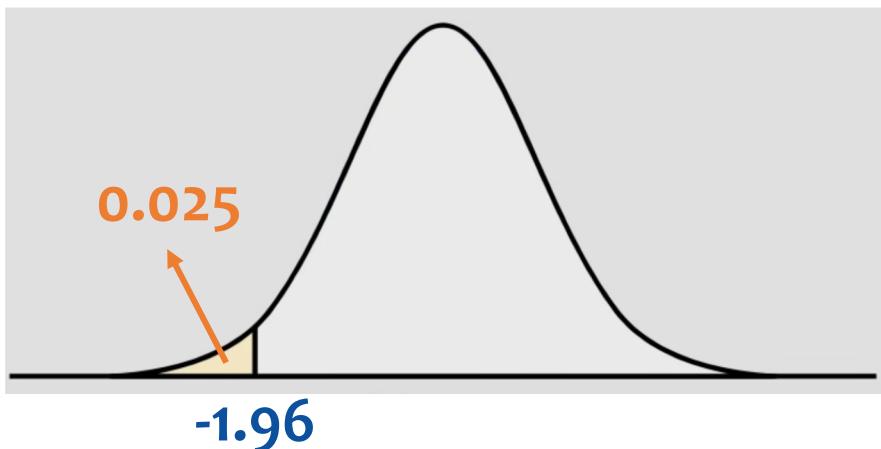
Z-Score Table

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0224
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582

Margin of Error

95% confidence level: $1 - \alpha = 0.95$
significance level: $\alpha = 0.05$

$$z_{\alpha/2} = z_{0.05/2} = z_{0.025} = -1.96$$



Z-Score Table

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0224
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582

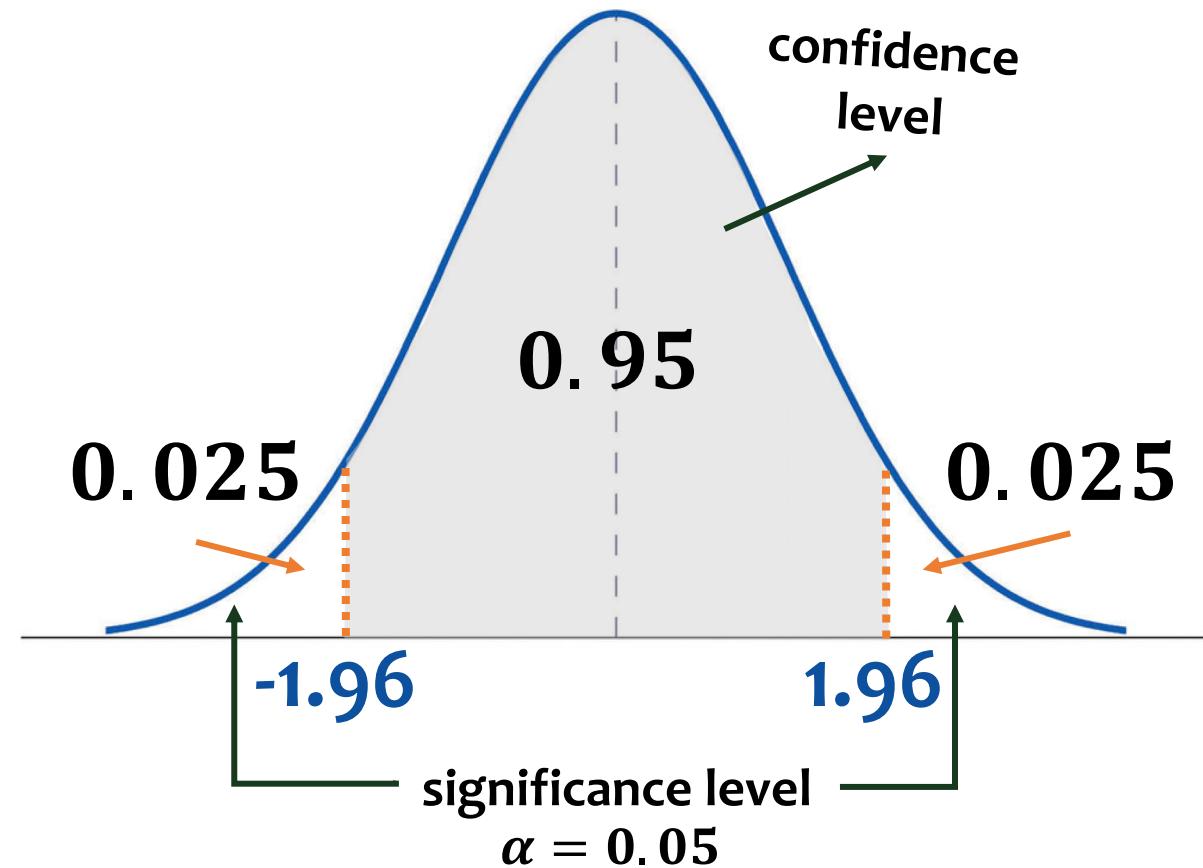
Margin of Error

95% confidence level: $1 - \alpha = 0.95$

significance level: $\alpha = 0.05$

$$z_{\alpha/2} = z_{0.05/2} = z_{0.025} = -1.96$$

$$e = 1.96 \frac{\sigma}{\sqrt{n}}$$



Margin of Error

- A **90% level of confidence** has $\alpha = 0.10$ and **critical value** of $z_{\alpha/2} = 1.64$.
- A **95% level of confidence** has $\alpha = 0.05$ and **critical value** of $z_{\alpha/2} = 1.96$.
- A **99% level of confidence** has $\alpha = 0.01$ and **critical value** of $z_{\alpha/2} = 2.58$.

Confidence Interval for the Mean

Provides an **interval estimate** that contains the **population mean** with a certain **confidence level**.

Known populational standard deviation

$$\mu = \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Unknown populational standard deviation

$$\mu \approx \bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

sample
standard deviation

Exercise 1

Suppose the heights of the inhabitants of a city are **normally distributed** with **population standard deviation** of 20 cm.

We measure the heights of 40 randomly chosen people, and get a **mean height** of 1.75 m.

Construct a **confidence interval** for the **population mean** with a **significance level of 5%**.

Exercise 2

Given a dataset from stroke patients, we want to study their **mean glucose level**.

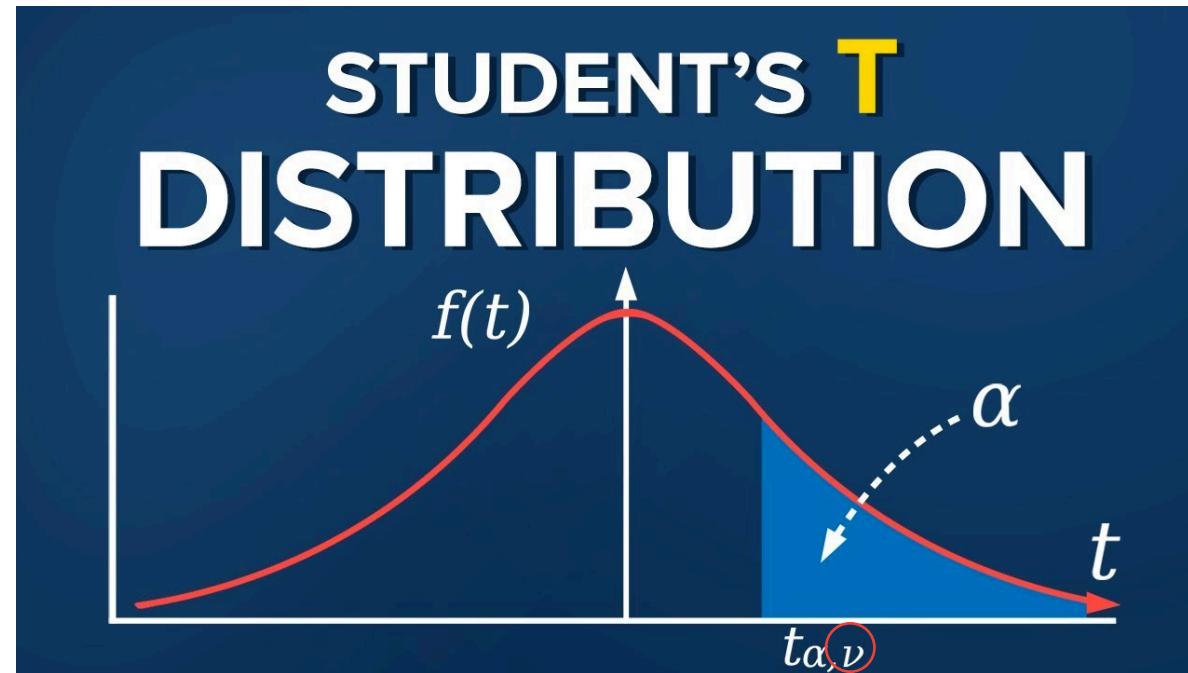
For two samples of 100 and 1000 observation, provide a **95% confidence intervals** for the following scenarios:

- (a) Known population standard deviation, and sample sizes of 100 and 1000
- (b) Unknown population standard deviation, and sample sizes of 100 and 1000

Confidence Interval for the Mean (Small Samples)

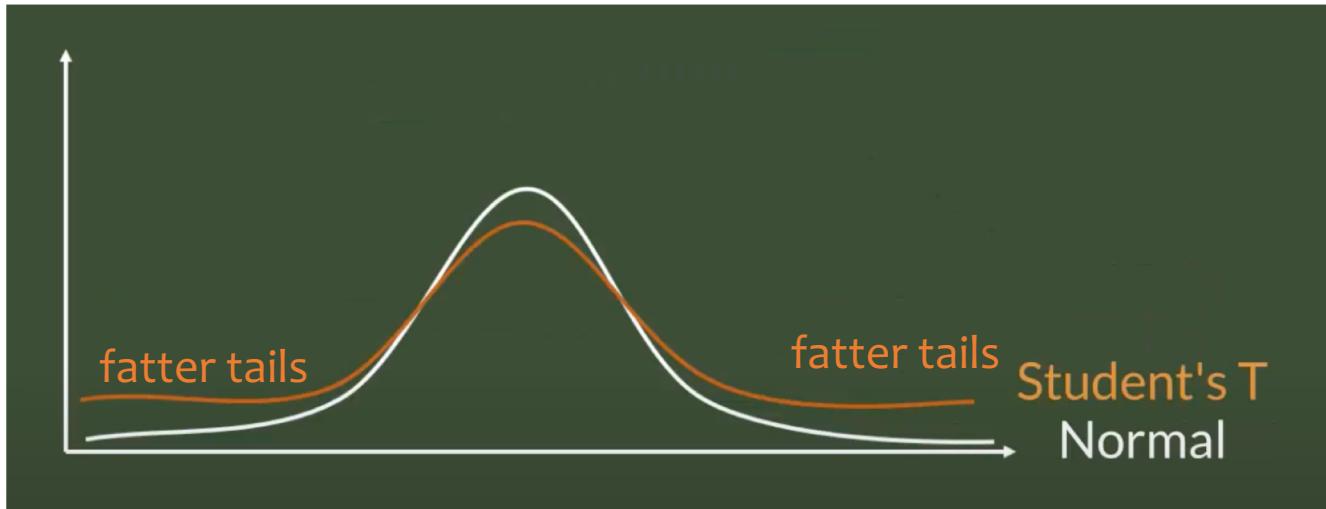
With **small sample sizes** ($n < 30$ observations), the distributions of sample means **is not always exactly normal**.

Then, we use a **t-distribution instead** of a **z-distribution (standard normal distribution)**.



$$\begin{aligned}\mu &= 0 \\ \sigma &= 1\end{aligned}$$

Degrees of freedom
(nº of independent
observations dataset)



$$\mu = 0, \sigma = 1$$

<https://www.youtube.com/watch?v=32CuxWdOlow>

	TABLE A-3 <i>t</i> Distribution: Critical <i>t</i> Values				
	0.005	0.01	Area in One Tail 0.025	0.05	0.10
Degrees of Freedom	0.01	0.02	Area in Two Tails 0.05	0.10	0.20
1	63.657	31.821	12.706	6.314	3.078
2	9.925	6.965	4.303	2.920	1.886
3	5.841	4.541	3.182	2.353	1.638
4	4.604	3.747	2.776	2.132	1.533
5	4.032	3.365	2.571	2.015	1.476
6	3.707	3.143	2.447	1.943	1.440
7	3.499	2.998	2.365	1.895	1.415
8	3.355	2.896	2.306	1.860	1.397
9	3.250	2.821	2.262	1.833	1.383
10	3.169	2.764	2.228	1.812	1.372
11	3.106	2.718	2.201	1.796	1.363
12	3.055	2.681	2.179	1.782	1.356
13	3.012	2.650	2.160	1.771	1.350
14	2.977	2.624	2.145	1.761	1.345
15	2.947	2.602	2.131	1.753	1.341
16	2.921	2.583	2.120	1.746	1.337
17	2.898	2.567	2.110	1.740	1.333
18	2.878	2.552	2.101	1.734	1.330
19	2.861	2.539	2.093	1.729	1.328
20	2.845	2.528	2.086	1.725	1.325
21	2.831	2.518	2.080	1.721	1.323
22	2.819	2.508	2.074	1.717	1.321
23	2.807	2.500	2.069	1.714	1.319

Confidence Interval for the Mean (Small Samples)

Provides an **interval estimate** that contains the **population mean** with a certain **confidence level**.

Known populational standard deviation

$$\mu = \bar{x} \pm t_{(\frac{\alpha}{2}, n-1)} \frac{\sigma}{\sqrt{n}}$$

Degrees of freedom

Unknown populational standard deviation

$$\mu = \bar{x} \pm t_{(\frac{\alpha}{2}, n-1)} \frac{s}{\sqrt{n}}$$

sample standard deviation

Calculating the Sample Size

margin of error
 $e = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

Infinite Population

Known populational standard deviation

$$n = (z_{\alpha/2} \frac{\sigma}{e})^2$$

Unknown populational standard deviation

$$n = (z_{\alpha/2} \frac{s}{e})^2$$

N : population size

z : z-score for area of $\alpha/2$

α : significance level

σ : population standard deviation

s : sample standard deviation

e : margin of error

Finite Population

Known populational standard deviation

$$n = \frac{z_{\alpha/2}^2 \sigma^2 N}{z_{\alpha/2}^2 \sigma^2 + e^2(N - 1)}$$

Unknown populational standard deviation

$$n = \frac{z_{\alpha/2}^2 s^2 N}{z_{\alpha/2}^2 s^2 + e^2(N - 1)}$$

Exercise 1

Suppose we want to investigate the **mean income** from the Brazilian population. According to the last census, the **population standard deviation** is **R\$ 4000,00**. Assuming a **95% confidence level** and a **margin of error** (with respect to the sample mean) of **R\$ 100,00**, what should be the **sample size**?

Exercise 2

In a batch of **10,000** cans of soda, a simple **random sample** of **200 cans** was collected to validation the amount of soda. The found **sample standard deviation** was **16 ml**. Each can costs **R\$ 4,00** and is thrown away after the experiment.

The manufacturer stipulates a **maximum error on the population average** of only **4 ml**. To ensure a **95% confidence level** what sample size should be selected for this study? How much will the company save in the next experiment/validation?

D1EAD – Análise Estatística para Ciência de Dados

2021.1



Data Distributions (Part 3)

Prof. Ricardo Sovat

sovat@ifsp.edu.br

Prof. Samuel Martins (Samuka)

samuel.martins@ifsp.edu.br

