

Introduction to statistics

(Day 2)

Agenda

- **Where:**
 - Mar 4: 
 - Mar 5: AULA DARWIN - MBC
 - Mar 6 : Online

- **When:**
 - 14-17
 - 1 coffee break

- **Who:**
 - Paola Dalmasso
paola.dalmasso@unito.it
 - Alessia Visconti
alessia.visconti@unito.it

- **How (to pass):**
 - Attend at least 2 lessons

How to ask questions/give feedback

- Interrupt me
- Take advantage of start/end/breaks
- Send emails alessia.visconti@unito.it
- Use the shared pad:
[\(or <https://t.ly/vRbvy>\)](https://etherpad.wikimedia.org/p/intro_stats_2024_specialita)



Recap



Recap

- The collection, organisation, summarisation, and analysis of data
→ *Descriptive* statistics
- The drawing of inferences about a body of data when only a part of the data is observed
→ *Inferential* statistics

Recap

- When can't study a population, we select a representative sample
- There are different sampling strategies
- There are different types of data
- Data are described with measures of centrality (mode, median, mean) and dispersion (range, IQR, standard deviation)
- Parameters (calculated on the population) *vs* statistics (calculated on the sample)

The Normal Distribution & the Central Limit Theorem

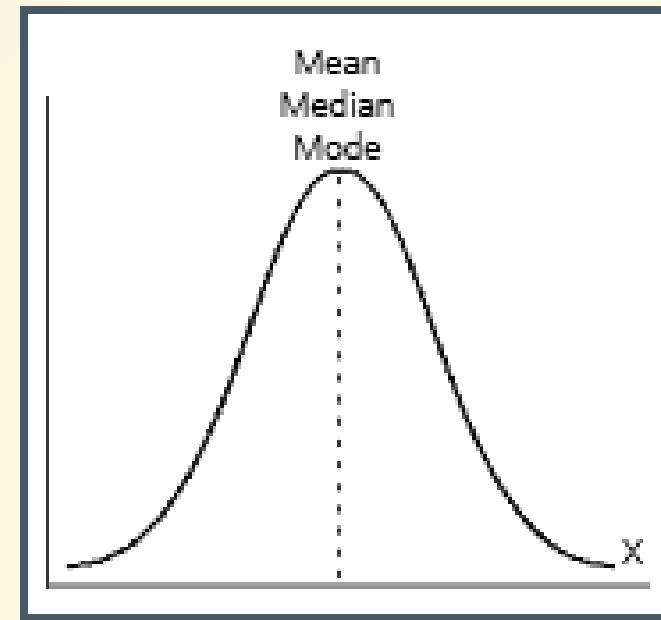


Learning objectives

- Know the characteristics of the Normal distribution
- Understand the Central Limit Theorem

The Normal distribution

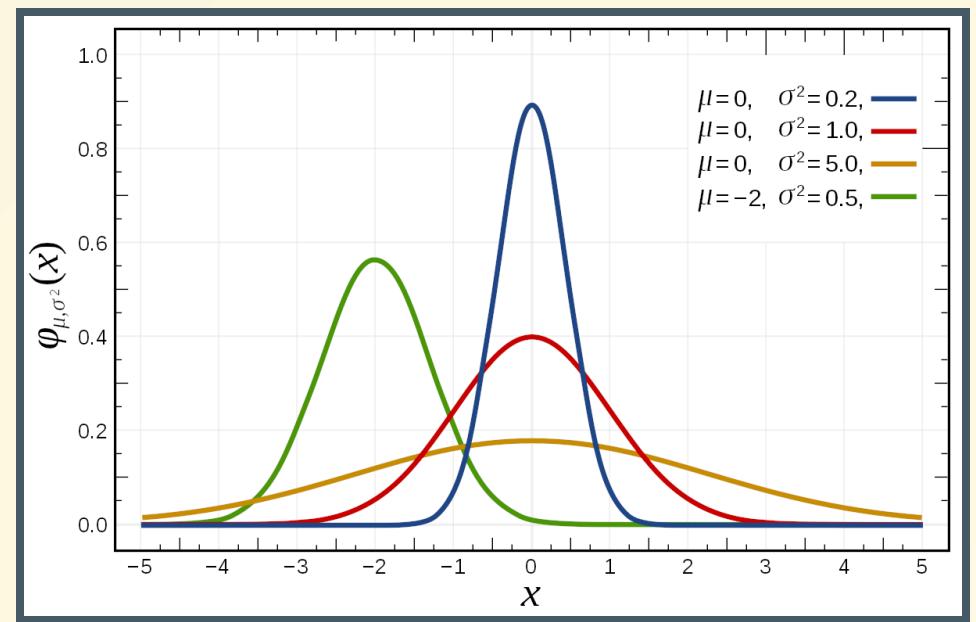
- Symmetrical
- $\mathcal{N} = (\mu, \sigma^2)$



$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The Normal distribution

- Symmetrical
- $\mathcal{N} = (\mu, \sigma^2)$

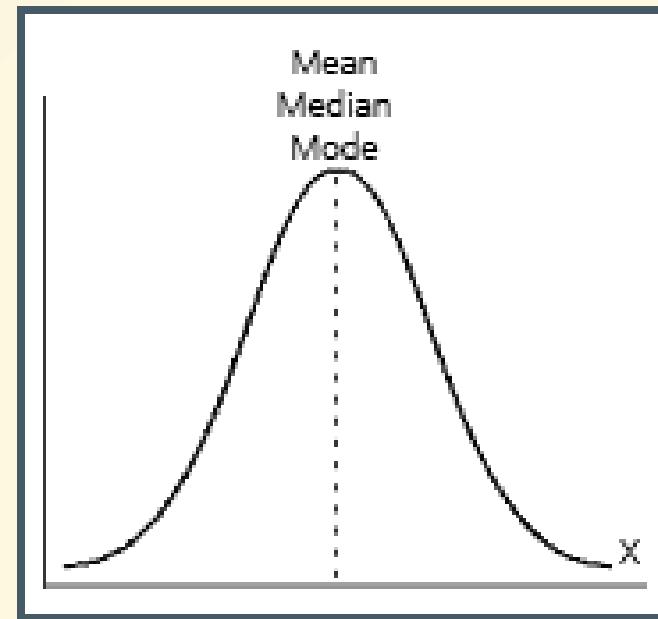


$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The Normal distribution

- Symmetrical
- $\mathcal{N} = (\mu, \sigma^2)$
- Area under the curve = 1

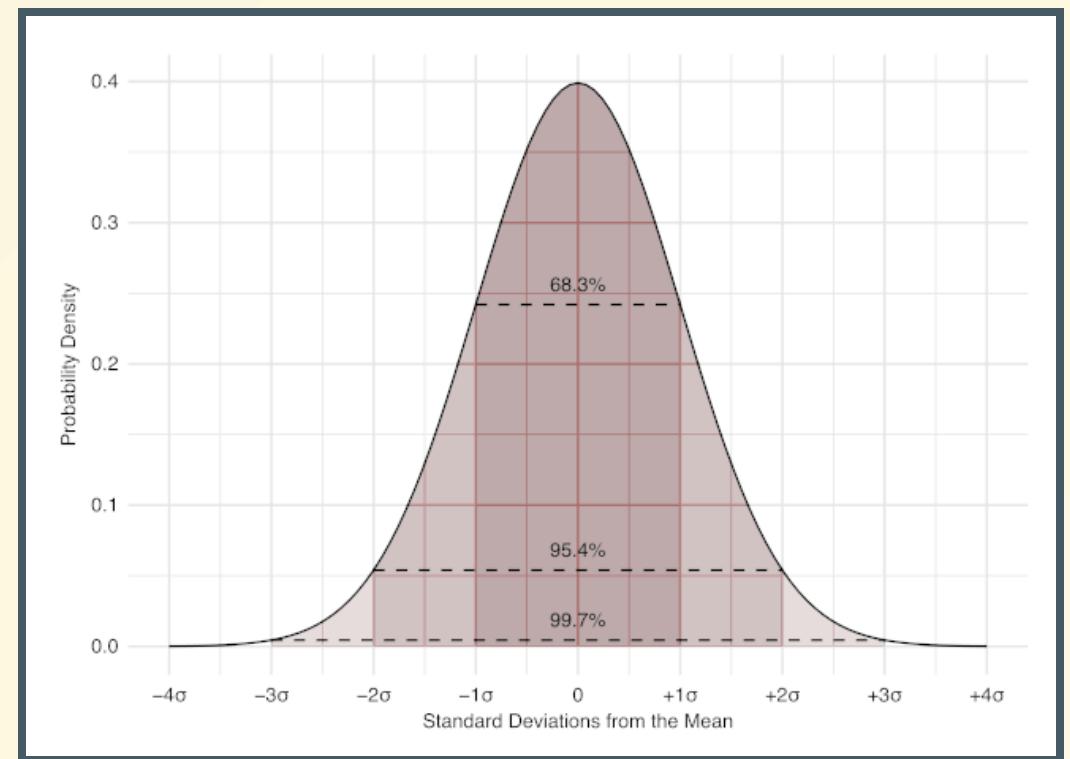
? What is the area left of the median? And right?



$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The Normal distribution

- Symmetrical
- $\mathcal{N} = (\mu, \sigma^2)$
- Area under the curve = 1
- *3-sigma* (or 68-95-99.7) rule

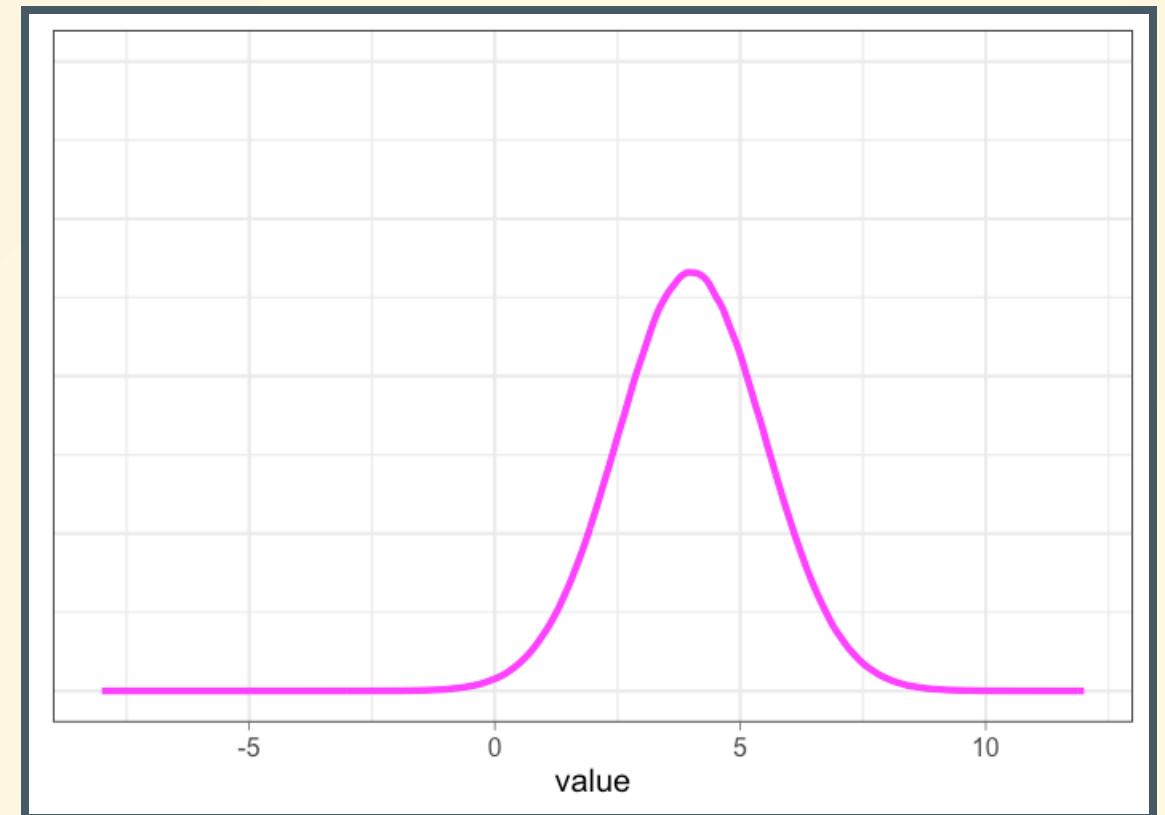


The Standard Normal distribution

- $\mathcal{N} = Z = (0, 1)$

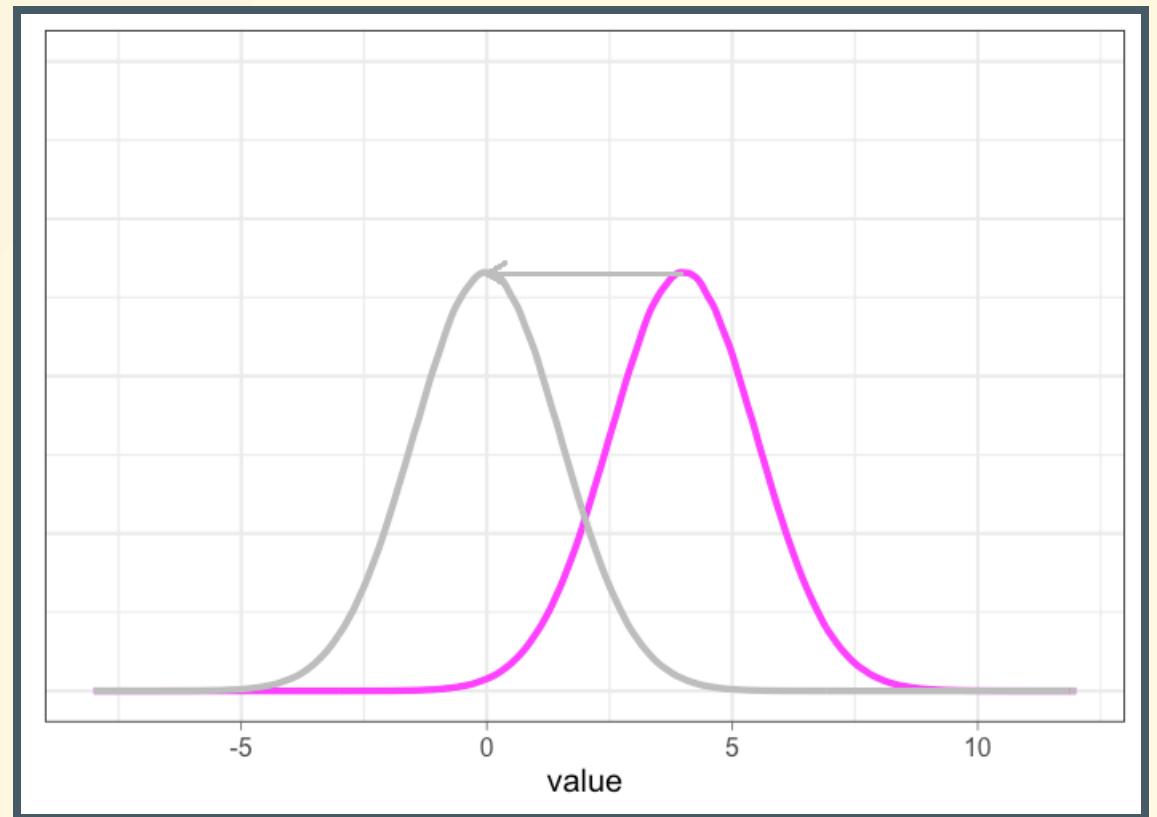
The Standard Normal distribution

- $\mathcal{N} = (\mu, \sigma^2) \rightarrow Z = (0, 1)$



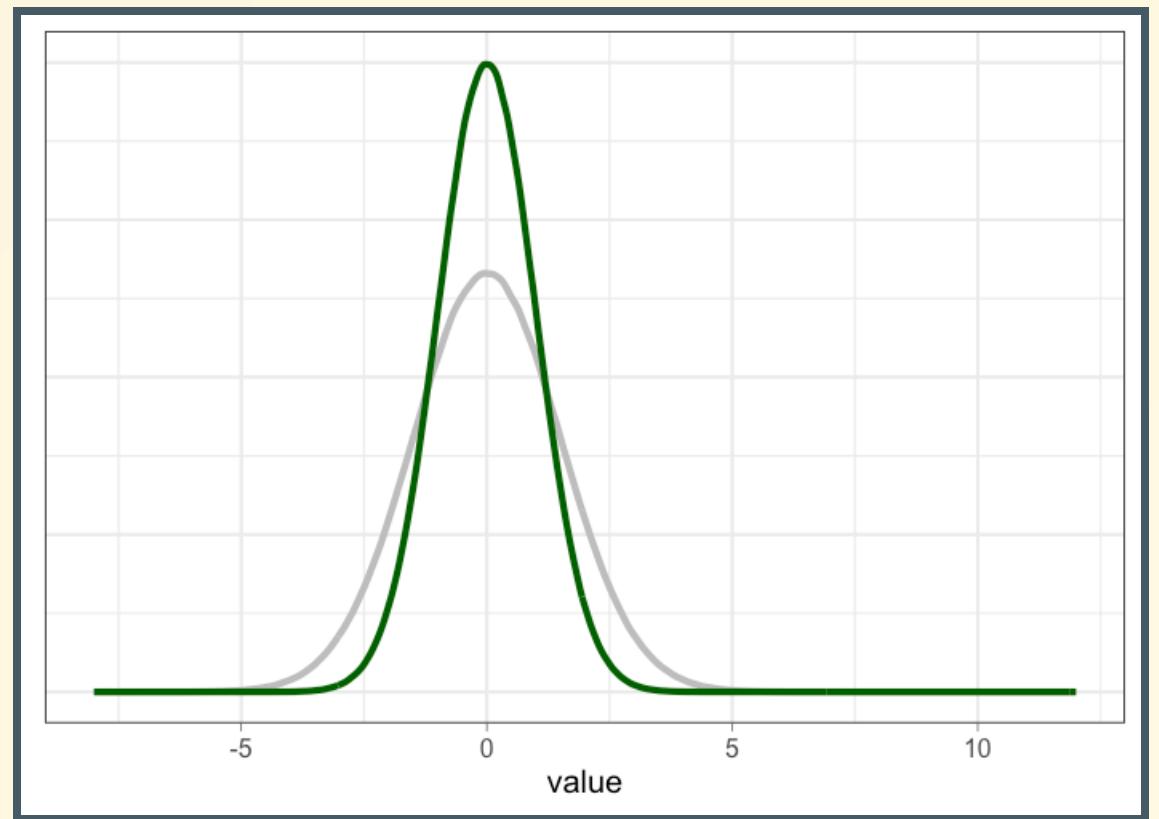
The Standard Normal distribution

- $\mathcal{N} = (\mu, \sigma^2) \rightarrow Z = (0, 1)$
- $z = \frac{x-\mu}{\sigma}$



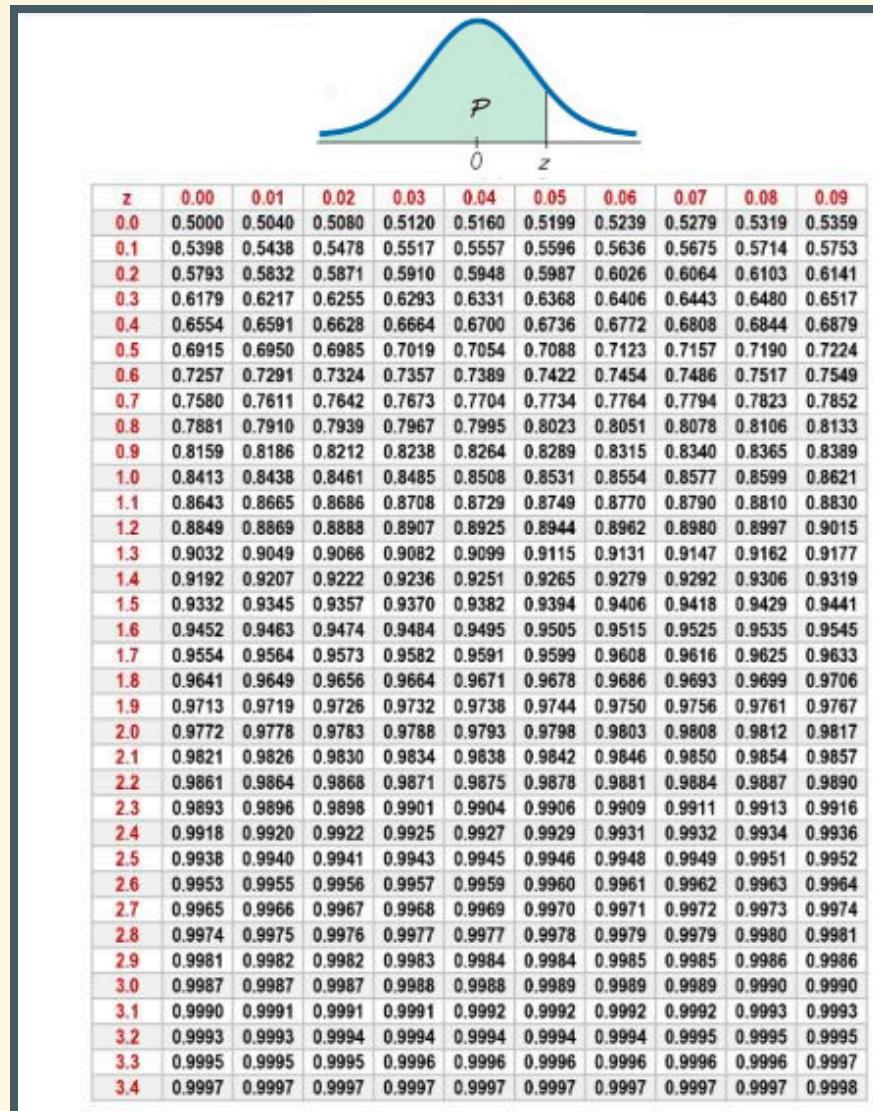
The Standard Normal distribution

- $\mathcal{N} = (\mu, \sigma^2) \rightarrow Z = (0, 1)$
- $z = \frac{x-\mu}{\sigma}$



The Standard Normal distribution

- $\mathcal{N} = (\mu, \sigma^2) \rightarrow Z = (0, 1)$
 - $z = \frac{x - \mu}{\sigma}$



The Standard Normal distribution in practice

📌 $\mu = 170 \text{ cm}$ $\sigma = 9.5 \text{ cm}$

$$\mathcal{P}(x \geq 176) = ?$$

The Standard Normal distribution in practice

📌 $\mu = 170 \text{ cm} \quad \sigma = 9.5 \text{ cm}$

$$z = \frac{x-\mu}{\sigma} = \frac{176-170}{9.5} = 0.63$$

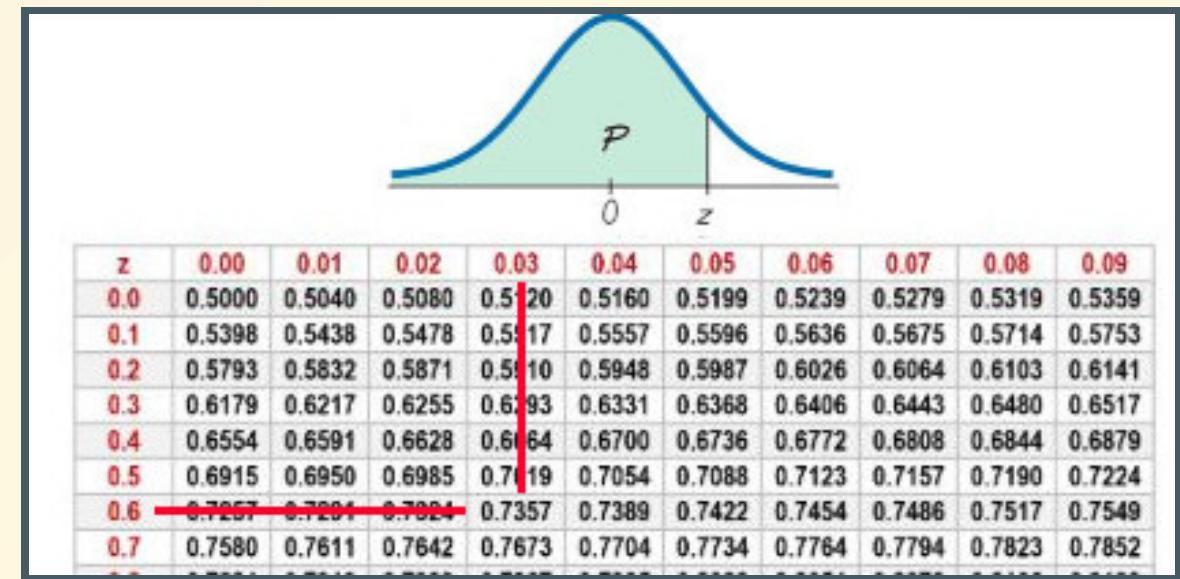
$$\mathcal{P}(x \geq 176) = ?$$

The Standard Normal distribution in practice



$$\mu = 170 \text{ cm} \quad \sigma = 9.5 \text{ cm}$$

$$z = \frac{x-\mu}{\sigma} = \frac{176-170}{9.5} = 0.63$$



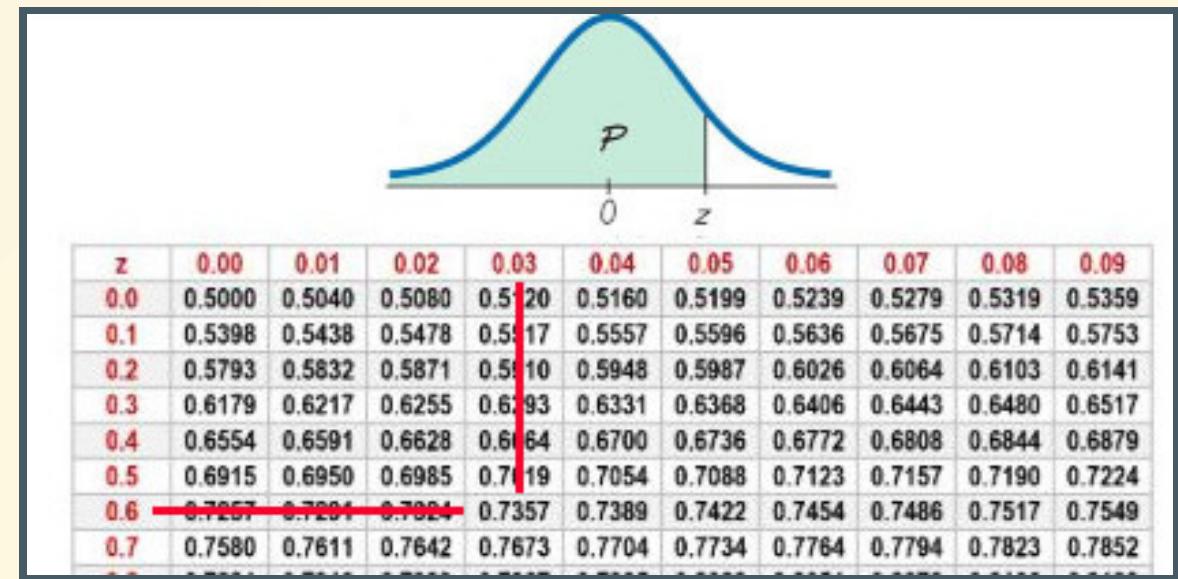
$$\mathcal{P}(x \geq 176) = ?$$

The Standard Normal distribution in practice



$$\mu = 170 \text{ cm} \quad \sigma = 9.5 \text{ cm}$$

$$z = \frac{x-\mu}{\sigma} = \frac{176-170}{9.5} = 0.63$$



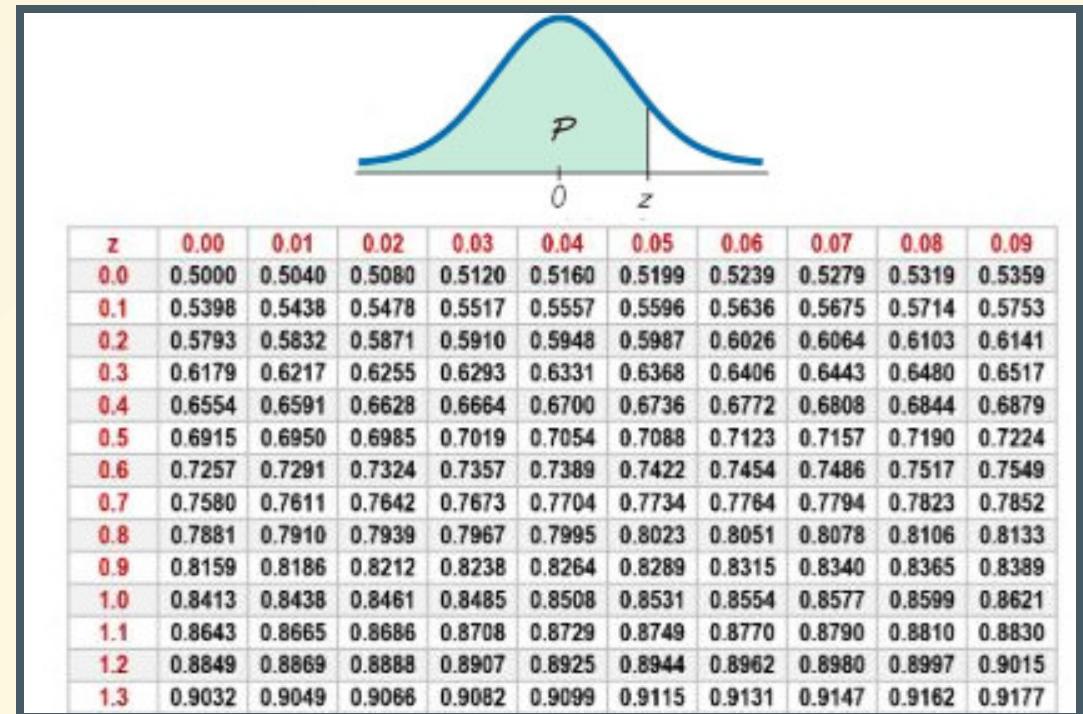
$$\mathcal{P}(x \geq 176) = 1 - 0.7357 = 0.264$$

Exercise #7



$$\mu = 170 \text{ cm} \quad \sigma = 9.5 \text{ cm}$$

$$\mathcal{P}(x \geq 179.5) = ?$$

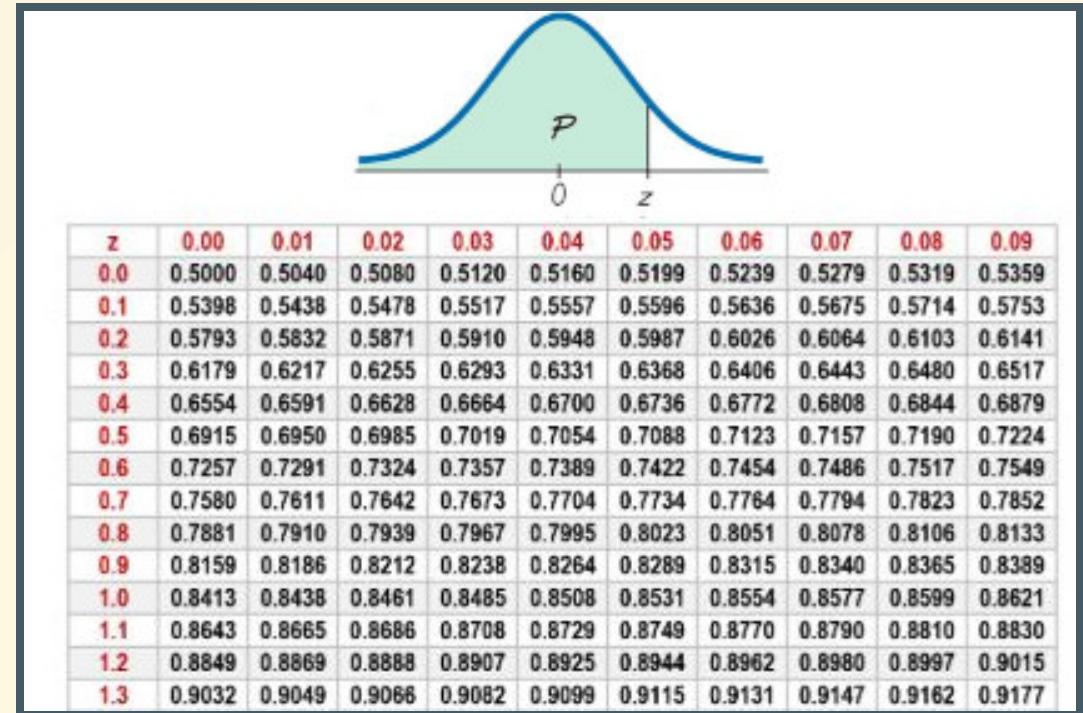


Exercise #7 -- Solution

📌 $\mu = 170 \text{ cm}$ $\sigma = 9.5 \text{ cm}$

$$z = \frac{x-\mu}{\sigma} = \frac{179.5-170}{9.5} = 1$$

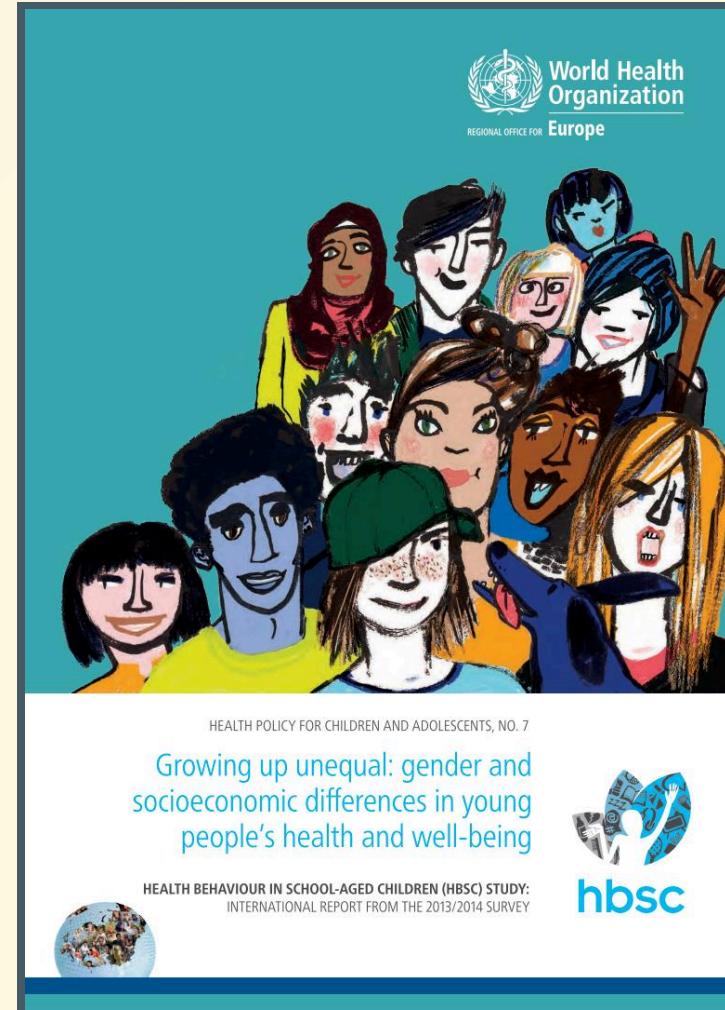
$$\begin{aligned}\mathcal{P}(x \geq 179.5) &= 1 - 0.8413 = \\ &= 0.16\end{aligned}$$



The Standard Normal distribution in practice

- 📌 $n = 6,705$ 15 y.o. males
 $\bar{x}_{\text{BMI}} = 21.5 \text{ kg/m}^2$
 $s_{\text{BMI}} = 3.1 \text{ kg/m}^2$

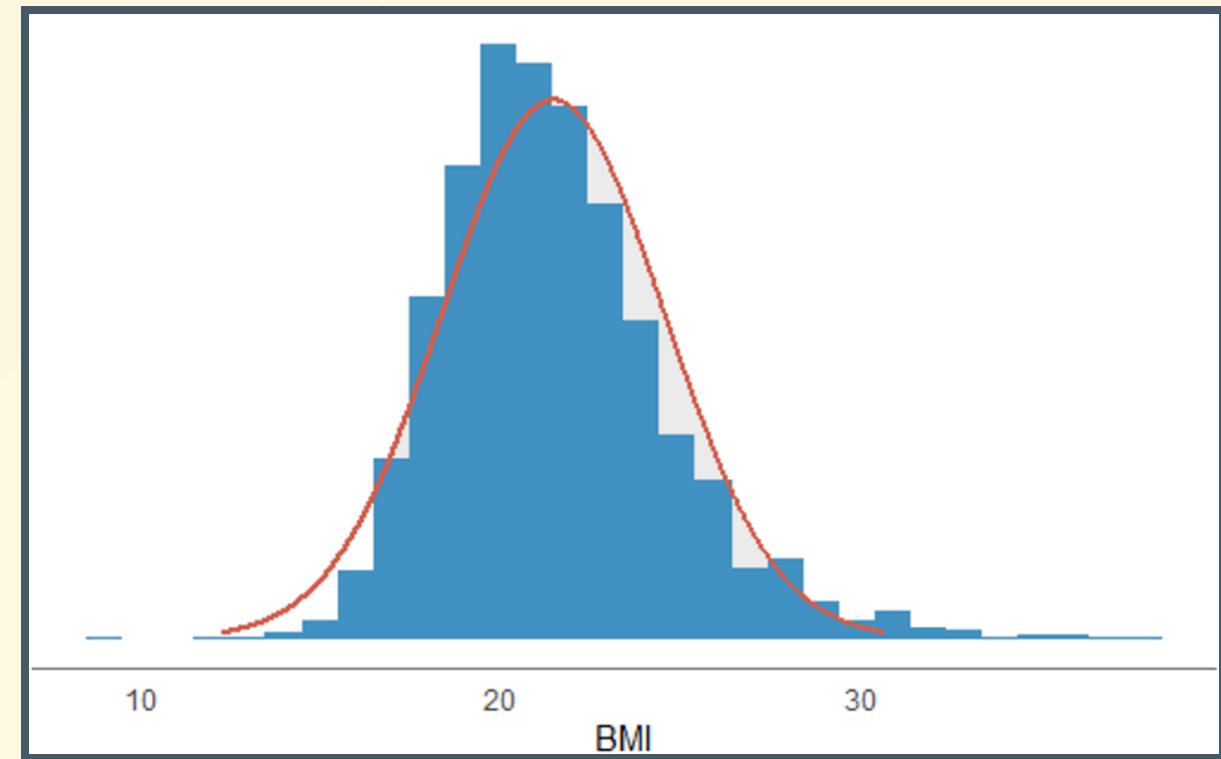
What percentage of 15 y.o. males have $\text{BMI} \geq 25$ in the population?



The Standard Normal distribution in practice

- 📌 $n = 6,705$ 15 y.o. males
 $\mu_{\text{BMI}} = 21.5 \text{ kg/m}^2$
 $\sigma_{\text{BMI}} = 3.1 \text{ kg/m}^2$

What percentage of 15 y.o. males have $\text{BMI} \geq 25$ in the population?



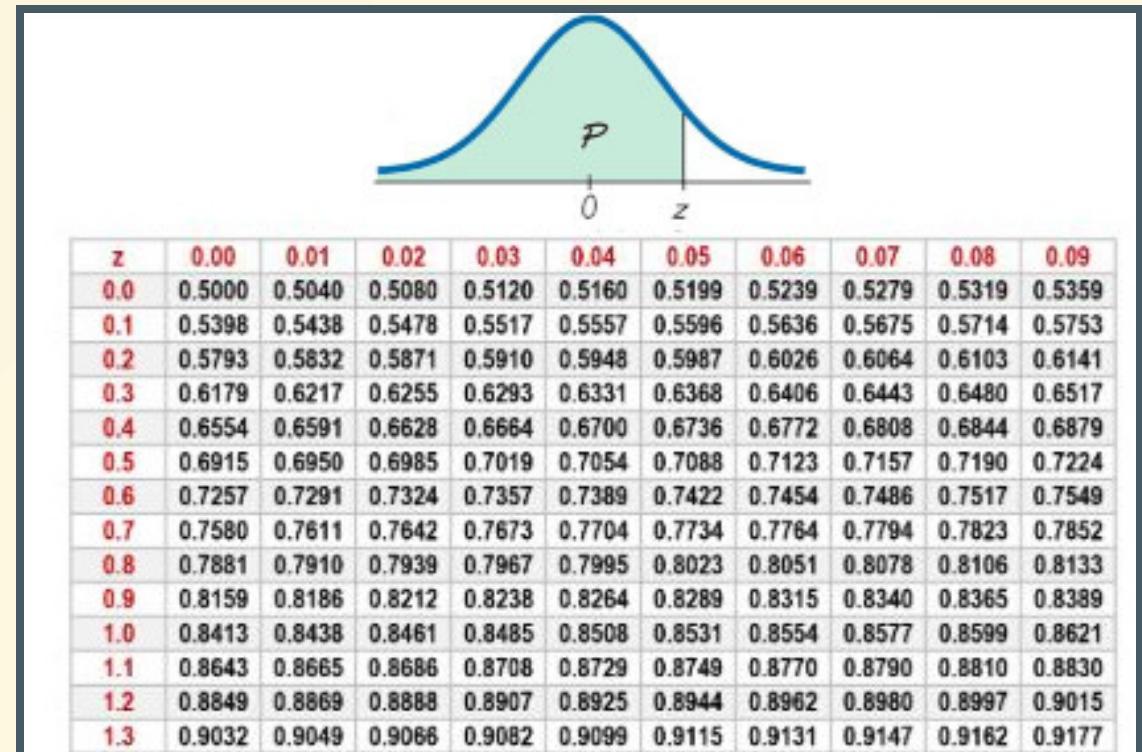
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The Standard Normal distribution in practice

- 📌 $n = 6,705$ 15 y.o. males
 $\mu_{\text{BMI}} = 21.5 \text{ kg/m}^2$
 $\sigma_{\text{BMI}} = 3.1 \text{ kg/m}^2$

What percentage of 15 y.o. males have $\text{BMI} \geq 25$ in the population?

$$z = \frac{x - \mu}{\sigma} = \frac{25 - 21.5}{3.1} = 1.12$$



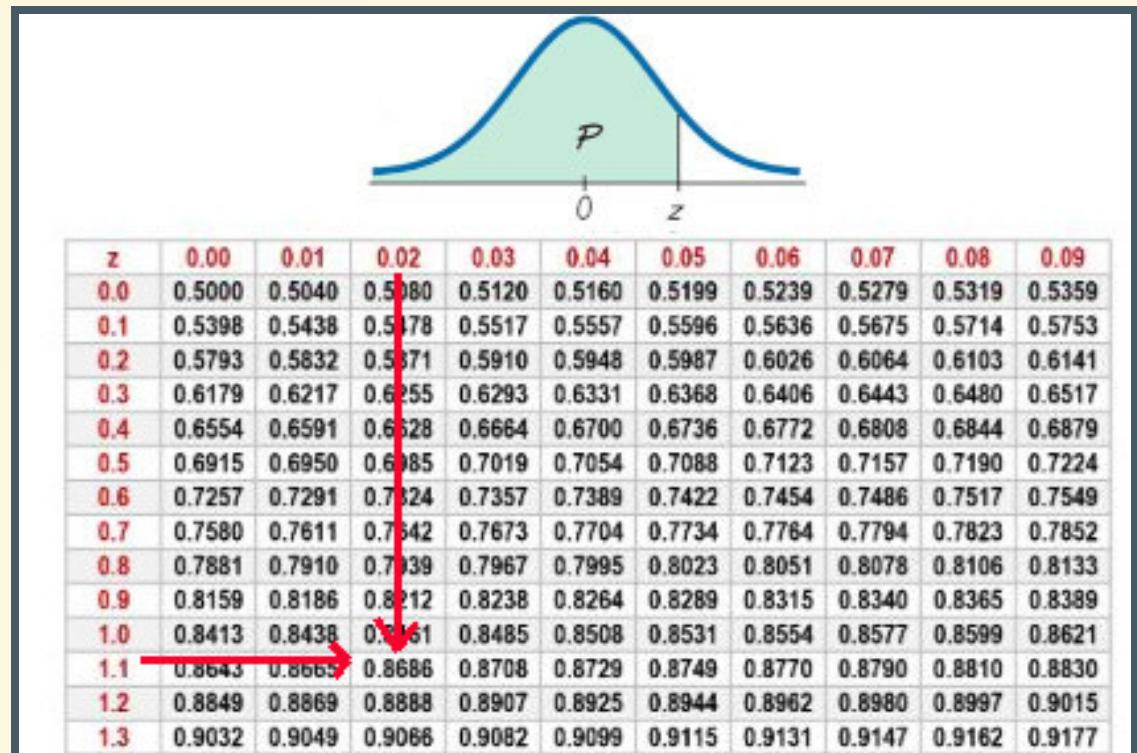
The Standard Normal distribution in practice

- 📌 $n = 6,705$ 15 y.o. males
 $\mu_{\text{BMI}} = 21.5 \text{ kg/m}^2$
 $\sigma_{\text{BMI}} = 3.1 \text{ kg/m}^2$

What percentage of 15 y.o. males have $\text{BMI} \geq 25$ in the population?

$$z = \frac{x - \mu}{\sigma} = \frac{25 - 21.5}{3.1} = 1.12$$

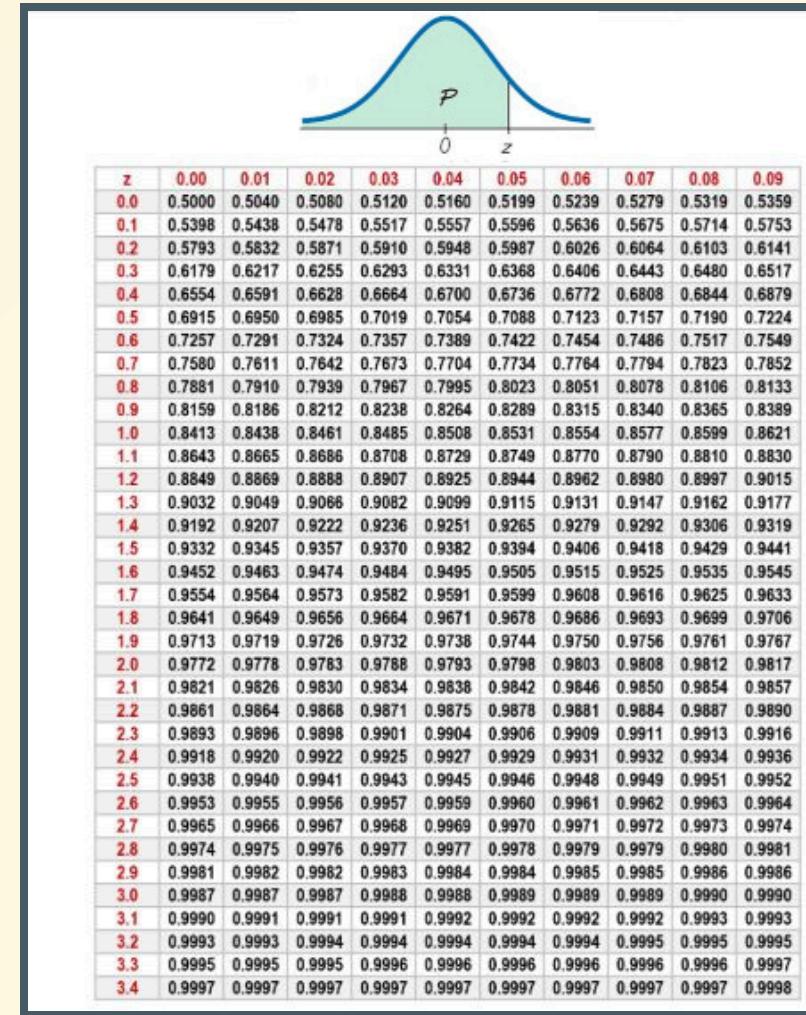
$$\mathcal{P}(\text{BMI} \geq 25) = 1 - 0.8686 = 0.1314 \rightarrow 13.1\%$$



Exercise #8

-  $n = 6,705$ 15 y.o. males
 $\mu_{\text{BMI}} = 21.5 \text{ kg/m}^2$
 $\sigma_{\text{BMI}} = 3.1 \text{ kg/m}^2$

What percentage of 15 y.o. males have $\text{BMI} \geq 30$ in the population?



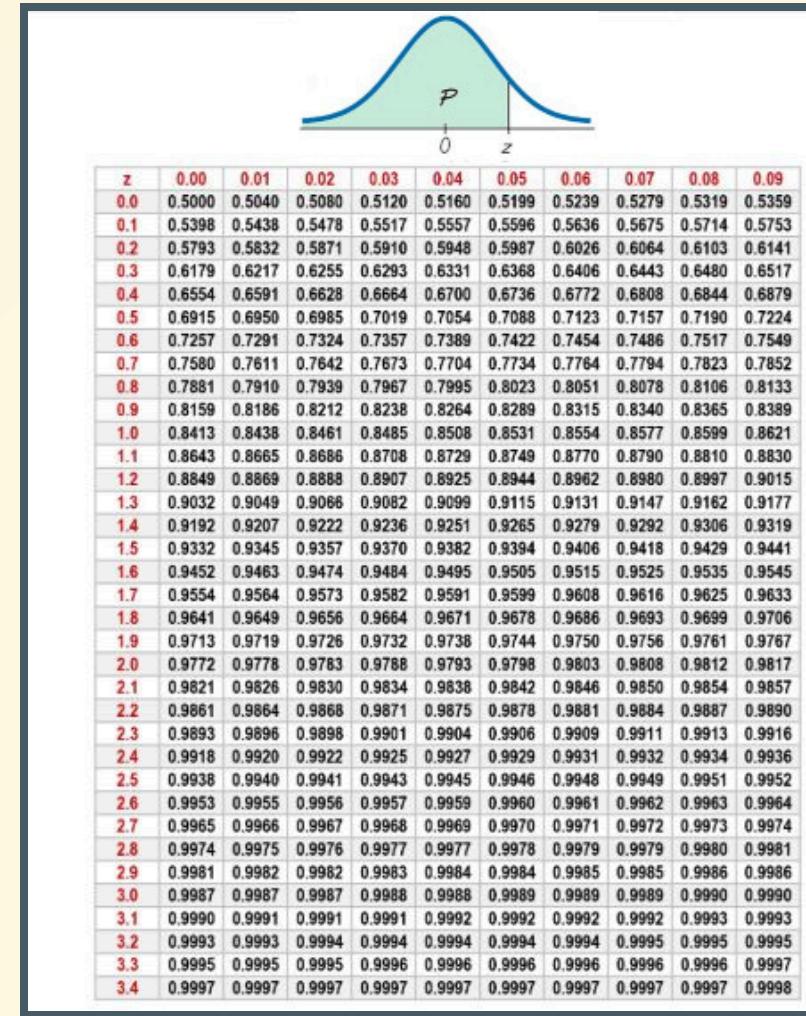
Exercise #8 -- Solution

-  $n = 6,705$ 15 y.o. males
 $\mu_{\text{BMI}} = 21.5 \text{ kg/m}^2$
 $\sigma_{\text{BMI}} = 3.1 \text{ kg/m}^2$

What percentage of 15 y.o. males have BMI ≥ 30 in the population?

$$z = \frac{x - \mu}{\sigma} = \frac{30 - 21.5}{3.1} = 2.74$$

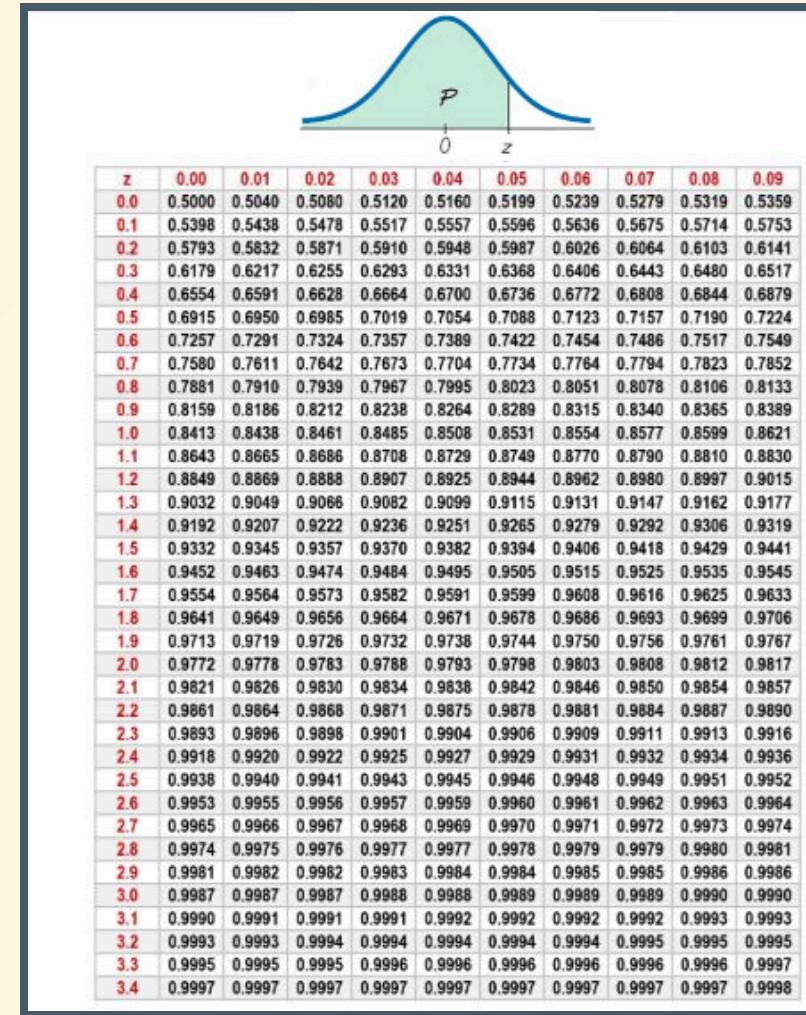
$$\mathcal{P}(\text{BMI} \geq 30) = 1 - 0.9969 = 0.0031 \rightarrow 0.31\%$$



Exercise #9

-  $n = 6,705$ 15 y.o. males
 $\mu_{\text{BMI}} = 21.5 \text{ kg/m}^2$
 $\sigma_{\text{BMI}} = 3.1 \text{ kg/m}^2$

What percentage of 15 y.o. males have $18.5 < \text{BMI} < 25$ in the population?



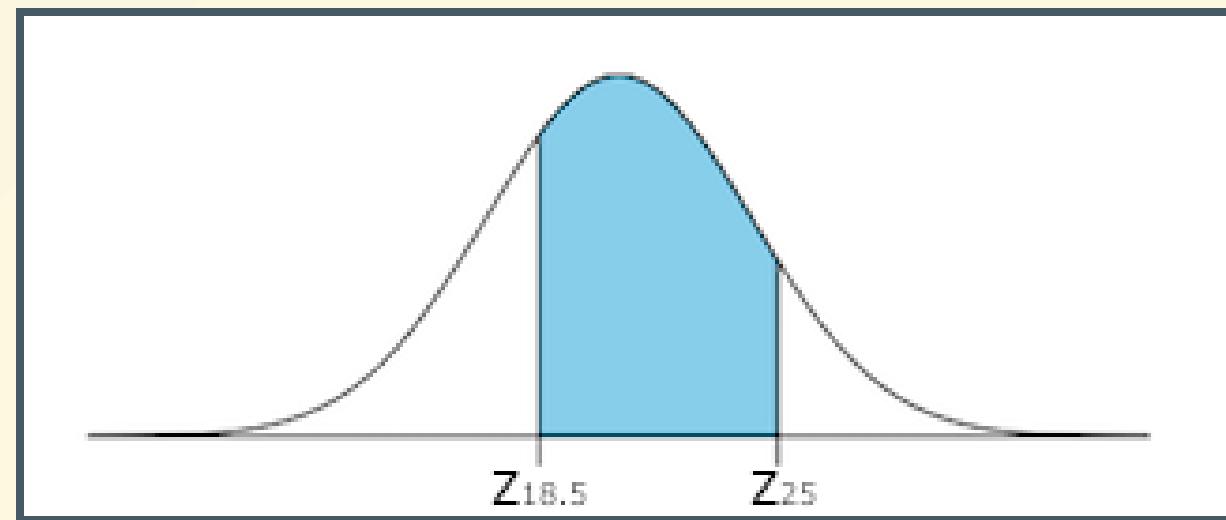
Exercise #9 -- Solution

- 📌 $n = 6,705$ 15 y.o. males

$$\mu_{\text{BMI}} = 21.5 \text{ kg/m}^2$$

$$\sigma_{\text{BMI}} = 3.1 \text{ kg/m}^2$$

What percentage of 15 y.o.
males have $18.5 < \text{BMI} < 25$
in the population?



Exercise #9 -- Solution

📌 $n = 6,705$ 15 y.o. males

$$\mu_{\text{BMI}} = 21.5 \text{ kg/m}^2$$

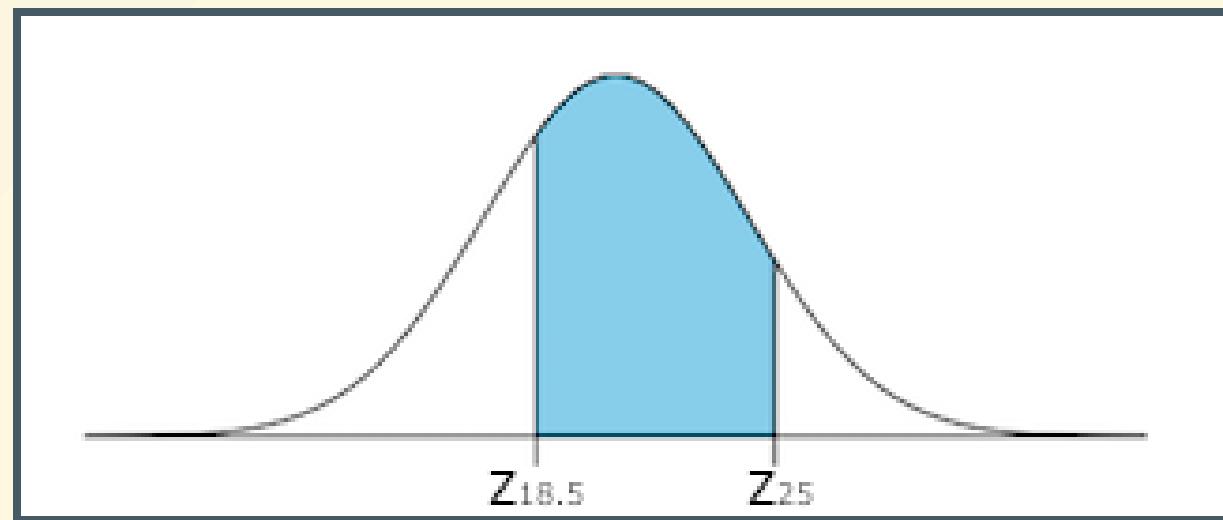
$$\sigma_{\text{BMI}} = 3.1 \text{ kg/m}^2$$

What percentage of 15 y.o.
males have $18.5 < \text{BMI} < 25$
in the population?

$$z_{25} = \frac{x - \mu}{\sigma} = \frac{25 - 21.5}{3.1} = 1.12$$

$$\mathcal{P}(\text{BMI} < 25) = 0.8686$$

$$\begin{aligned}\mathcal{P}(\text{BMI} \geq 25) &= 1 - 0.8686 = \\ &= 0.1314\end{aligned}$$



Exercise #9 -- Solution

📌 $n = 6,705$ 15 y.o. males

$$\mu_{\text{BMI}} = 21.5 \text{ kg/m}^2$$

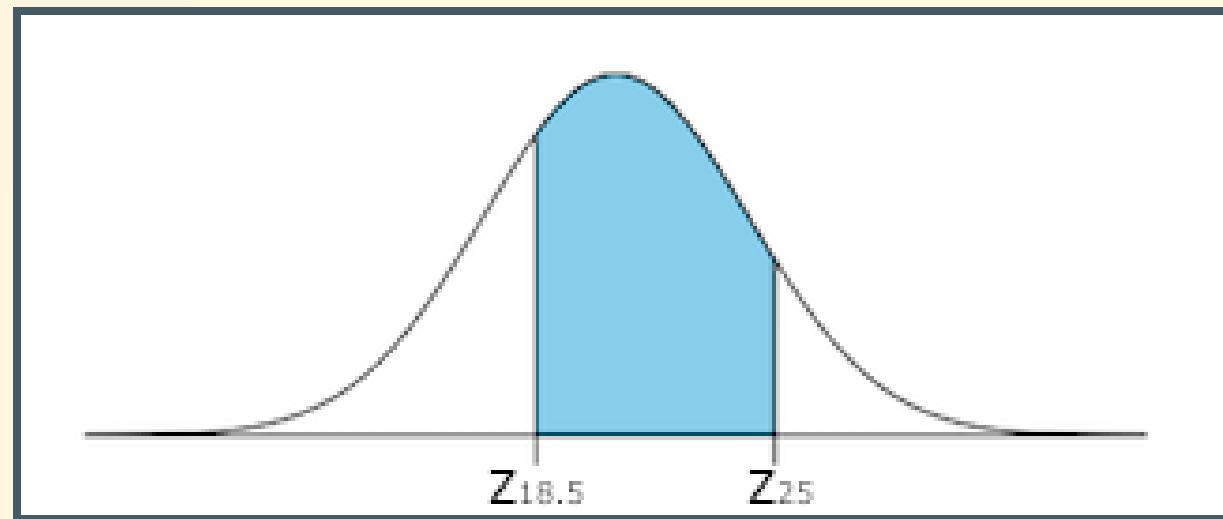
$$\sigma_{\text{BMI}} = 3.1 \text{ kg/m}^2$$

What percentage of 15 y.o.
males have $18.5 < \text{BMI} < 25$
in the population?

$$\mathcal{P}(\text{BMI} < 25) = 0.8686$$

$$\mathcal{P}(\text{BMI} \geq 25) = 0.1314$$

$$z_{18.5} = \frac{x - \mu}{\sigma} = \frac{18.5 - 21.5}{3.1} = -0.97$$



Exercise #9 -- Solution

📌 $n = 6,705$ 15 y.o. males

$$\mu_{\text{BMI}} = 21.5 \text{ kg/m}^2$$

$$\sigma_{\text{BMI}} = 3.1 \text{ kg/m}^2$$

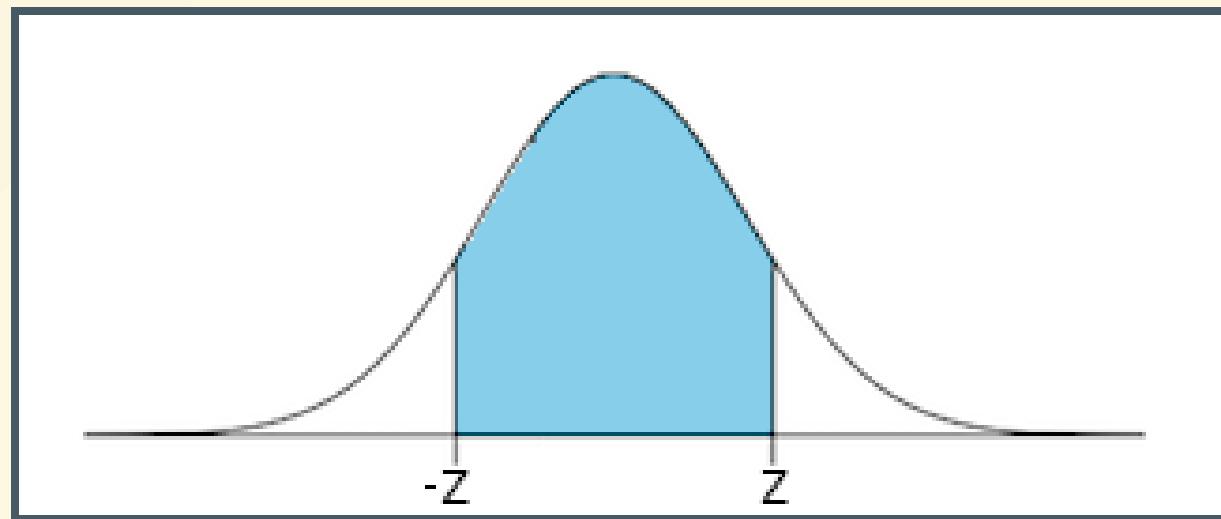
What percentage of 15 y.o.
males have $18.5 < \text{BMI} < 25$
in the population?

$$\mathcal{P}(\text{BMI} < 25) = 0.8686$$

$$\mathcal{P}(\text{BMI} \geq 25) = 0.1314$$

$$z_{18.5} = \frac{x - \mu}{\sigma} = \frac{18.5 - 21.5}{3.1} = -0.97$$

$$\mathcal{P}(\text{BMI} < 18.5) = \mathcal{P}(z < -0.97) = \mathcal{P}(z \geq 0.97)$$



Exercise #9 -- Solution

📌 $n = 6,705$ 15 y.o. males

$$\mu_{\text{BMI}} = 21.5 \text{ kg/m}^2$$

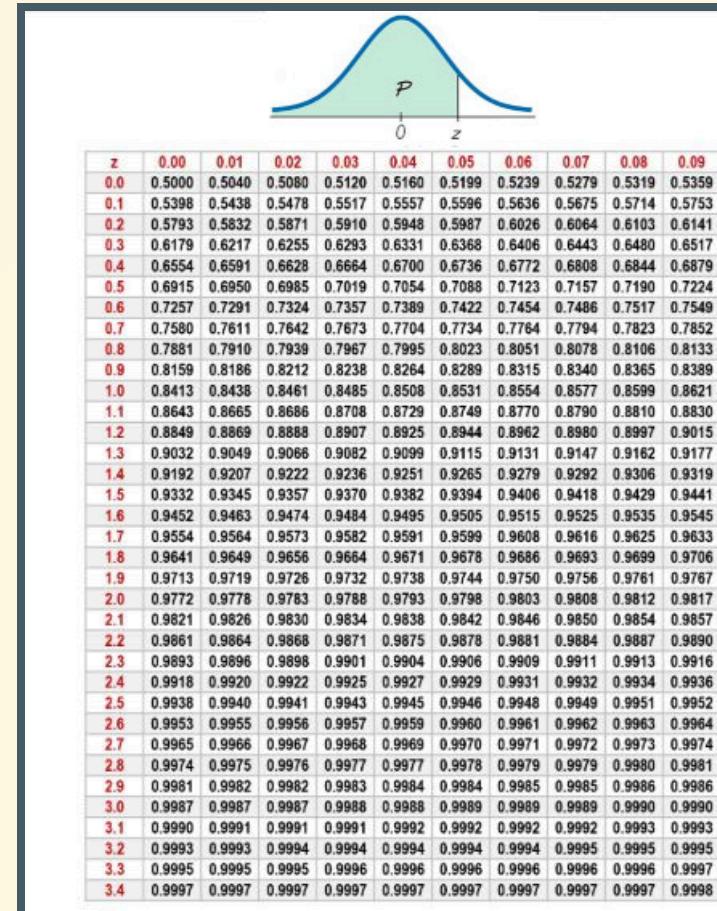
$$\sigma_{\text{BMI}} = 3.1 \text{ kg/m}^2$$

What percentage of 15 y.o.
males have $18.5 < \text{BMI} < 25$
in the population?

$$\mathcal{P}(\text{BMI} < 25) = 0.8686$$

$$\mathcal{P}(\text{BMI} \geq 25) = 0.1314$$

$$z_{18.5} = \frac{x - \mu}{\sigma} = \frac{18.5 - 21.5}{3.1} = -0.97$$



$$\mathcal{P}(\text{BMI} < 18.5) = \mathcal{P}(z < -0.97) = \mathcal{P}(z \geq 0.97) = 1 - 0.8340 = 0.166$$

Exercise #9 -- Solution

📌 $n = 6,705$ 15 y.o. males

$$\mu_{\text{BMI}} = 21.5 \text{ kg/m}^2$$

$$\sigma_{\text{BMI}} = 3.1 \text{ kg/m}^2$$

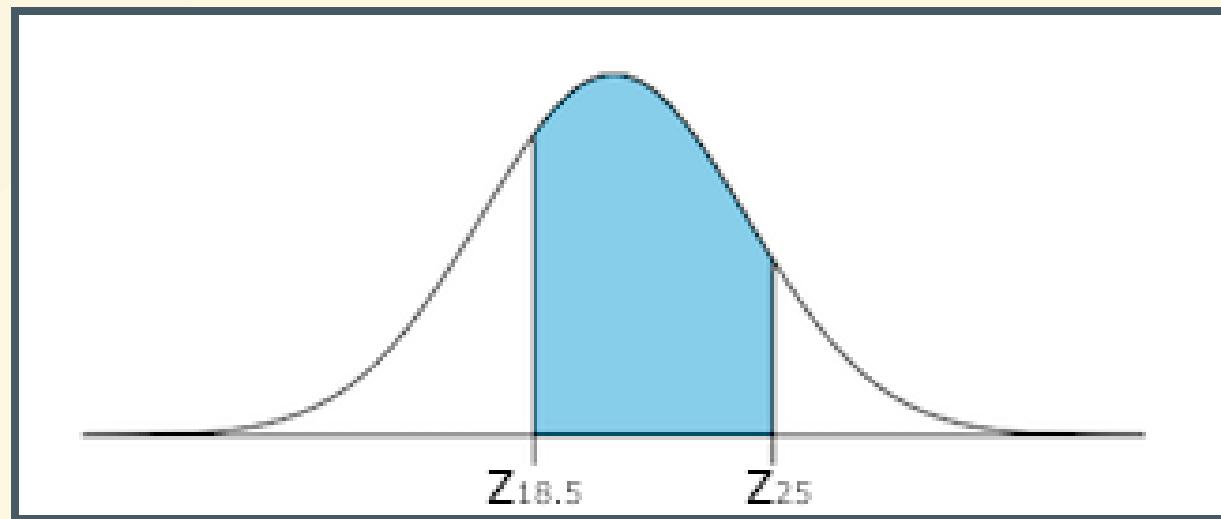
What percentage of 15 y.o.
males have $18.5 < \text{BMI} < 25$
in the population?

$$\mathcal{P}(\text{BMI} < 25) = 0.8686$$

$$\mathcal{P}(\text{BMI} \geq 25) = 0.1314$$

$$\mathcal{P}(\text{BMI} < 18.5) = 0.166$$

$$\mathcal{P}(18.5 < \text{BMI} < 25) = 0.8686 - 0.166 = 0.726 \rightarrow 70.3\%$$



Exercise #9 -- (Alternative) Solution

📌 $n = 6,705$ 15 y.o. males

$$\mu_{\text{BMI}} = 21.5 \text{ kg/m}^2$$

$$\sigma_{\text{BMI}} = 3.1 \text{ kg/m}^2$$

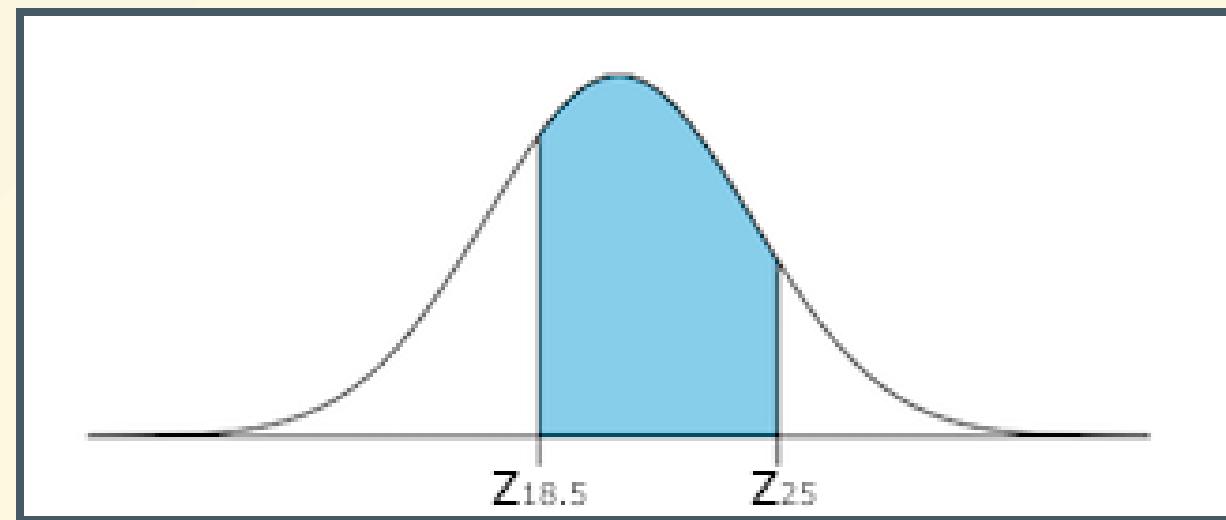
What percentage of 15 y.o.
males have $18.5 < \text{BMI} < 25$
in the population?

$$\mathcal{P}(\text{BMI} < 25) = 0.8686$$

$$\mathcal{P}(\text{BMI} \geq 25) = 0.1314$$

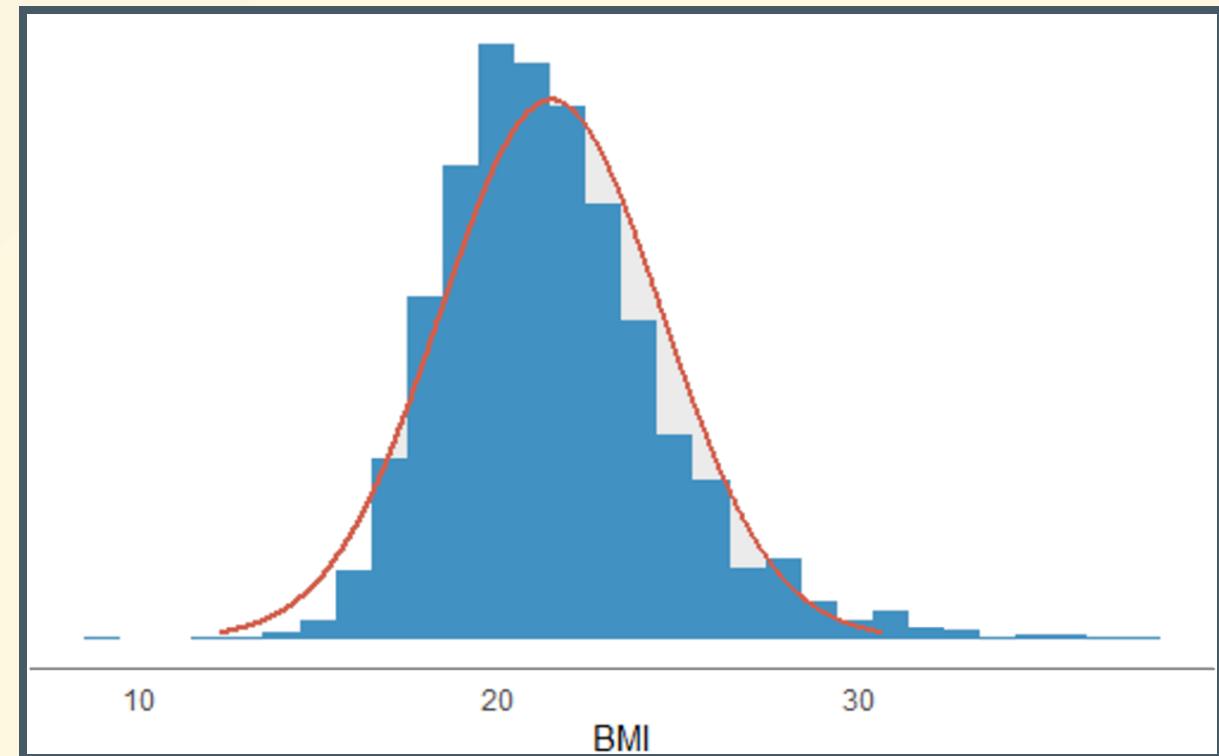
$$\mathcal{P}(\text{BMI} < 18.5) = 0.166$$

$$\mathcal{P}(18.5 < \text{BMI} < 25) = 1 - 0.1314 - 0.166 = 0.7026 \rightarrow 70.3\%$$



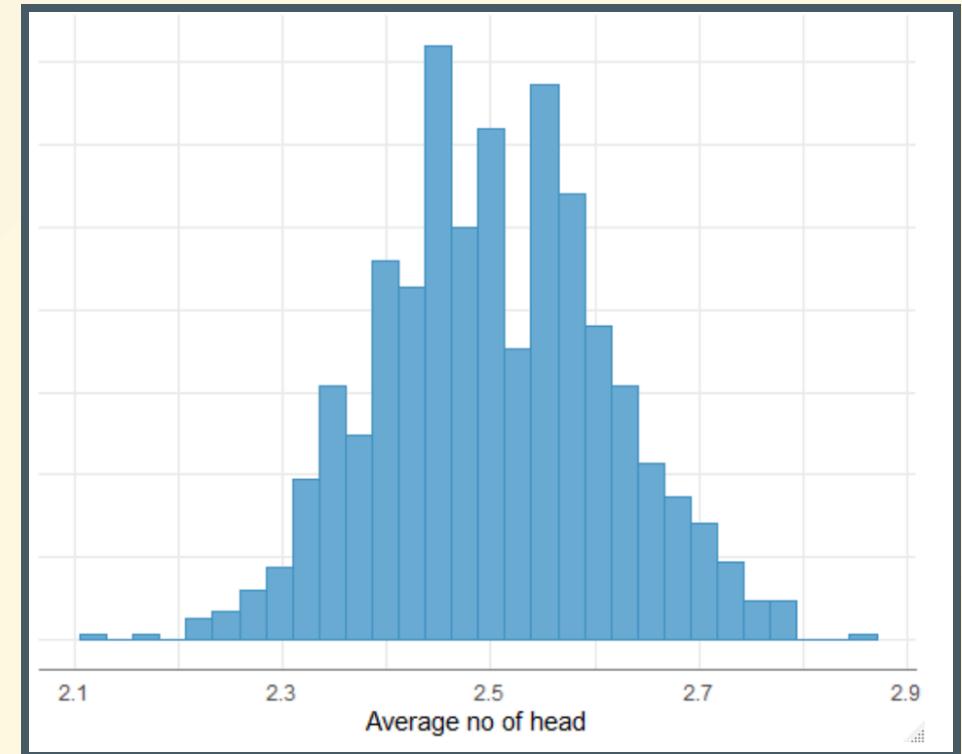
The Normal distribution in the wild

- Several natural phenomena approximate a Normal distribution

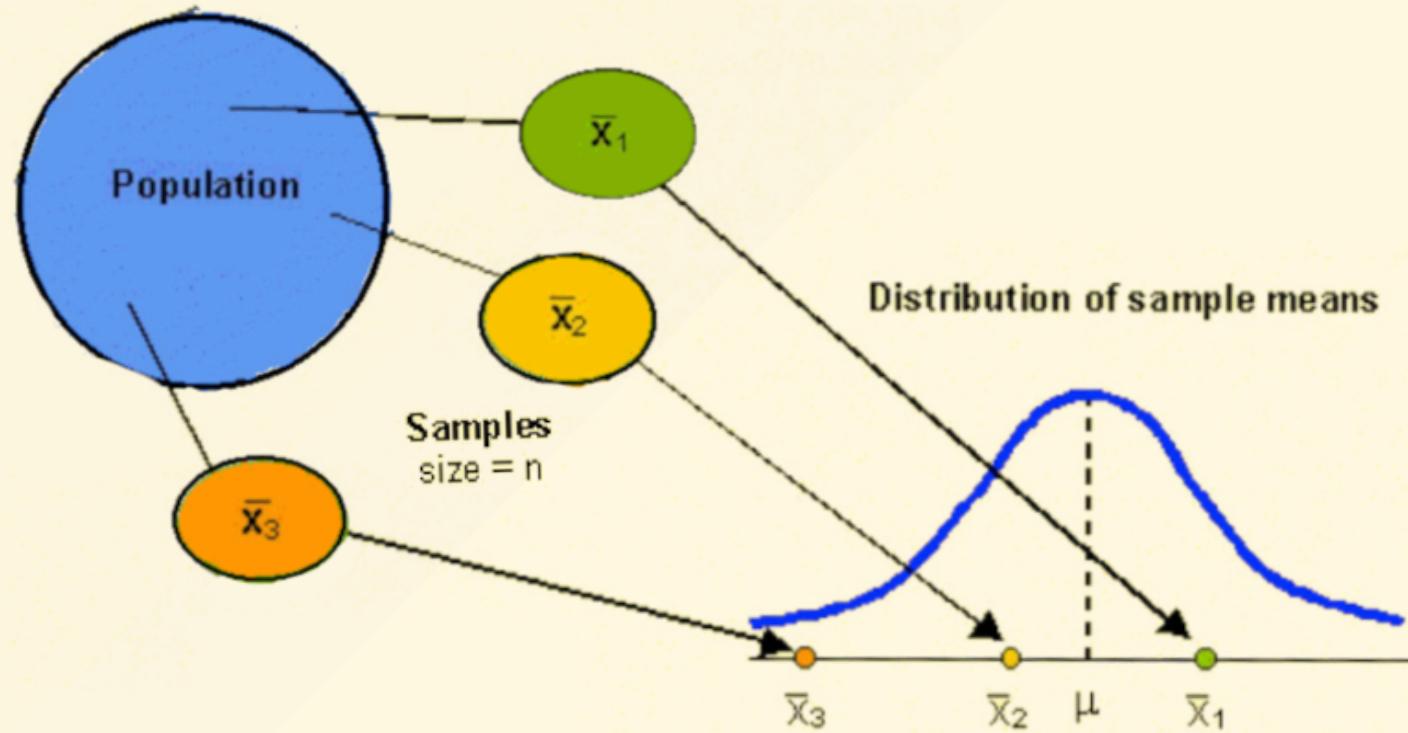


The Normal distribution in the wild

- Several natural phenomena approximate a Normal distribution
- Several probability distributions approximate a Normal distribution



Sampling distribution



The Central Limit Theorem (CLT)

- 🎯 As the sample size gets larger, the sampling distribution of the sample means approaches a normal distribution $\mathcal{N} = (\mu, \frac{\sigma^2}{n})$

$$\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} \rightarrow \text{standard error (SE)}$$

The Central Limit Theorem (CLT)

- 🎯 As the sample size gets larger, the sampling distribution of the sample means approaches a normal distribution $\mathcal{N} = (\mu, \frac{\sigma^2}{n})$

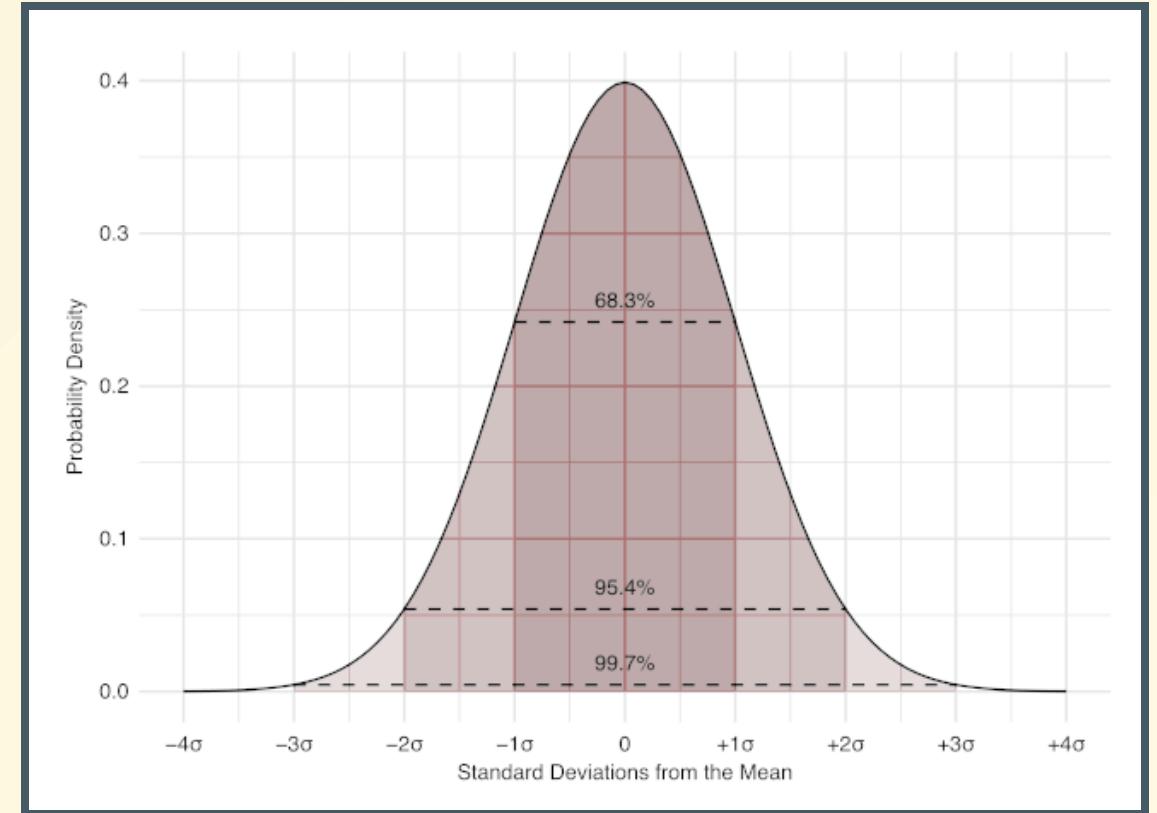
$$\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} \rightarrow \text{standard error (SE)}$$

- 🎯 standard error \neq standard deviation
standard deviation (SD) \rightarrow spread in the collected data
standard error (SE) \rightarrow error in the estimation

The Standard Error

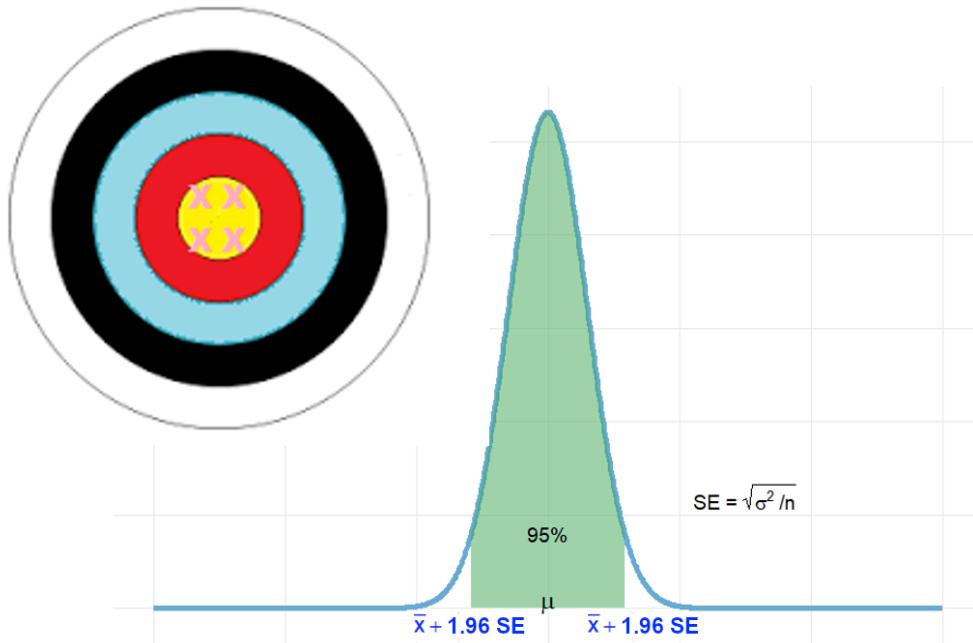


95% of the sample means
are within $\sim 2 \times \text{SE}$ from
population mean

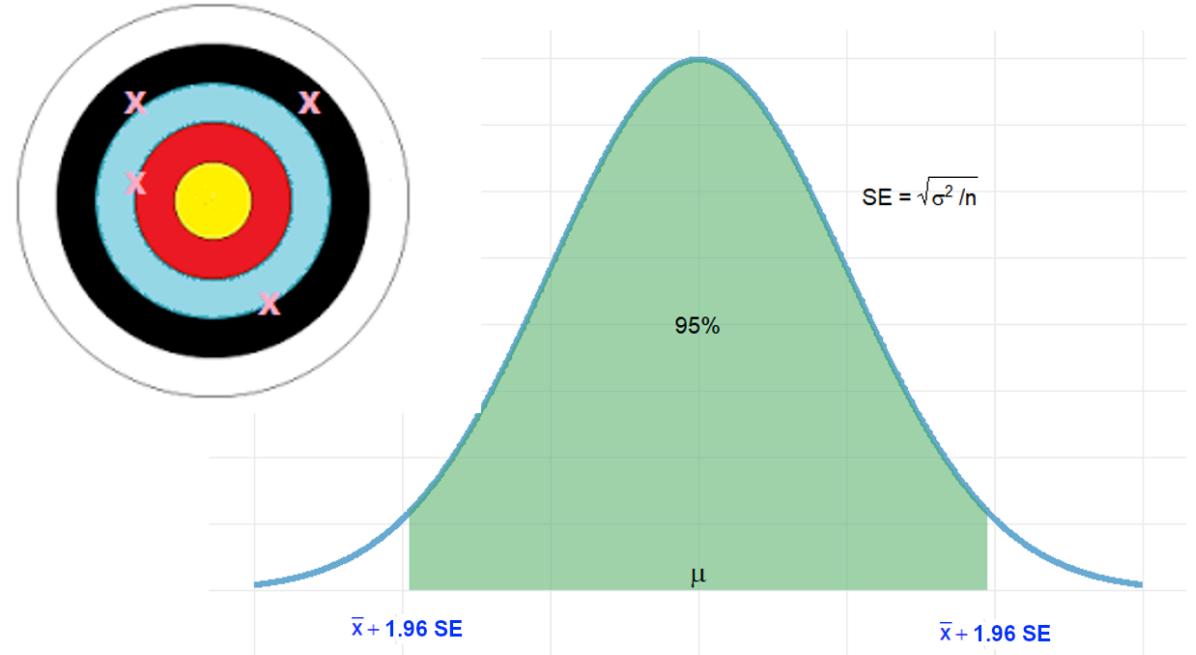


The Standard Error

IF STANDARD ERROR IS SMALL

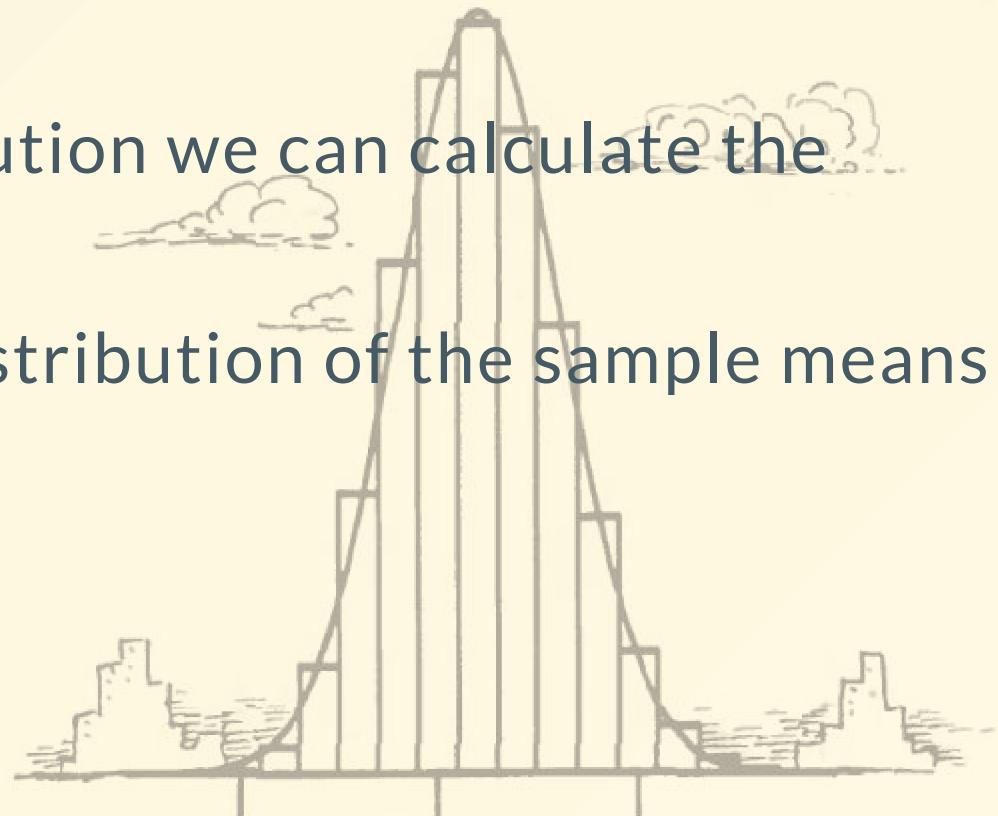


IF STANDARD ERROR IS LARGE



Summary

- Multiple phenomena and statistical distributions are normally distributed
- Using the standard normal distribution we can calculate the probability of an observation
- For large samples, the sampling distribution of the sample means will be normally distributed (CLT)



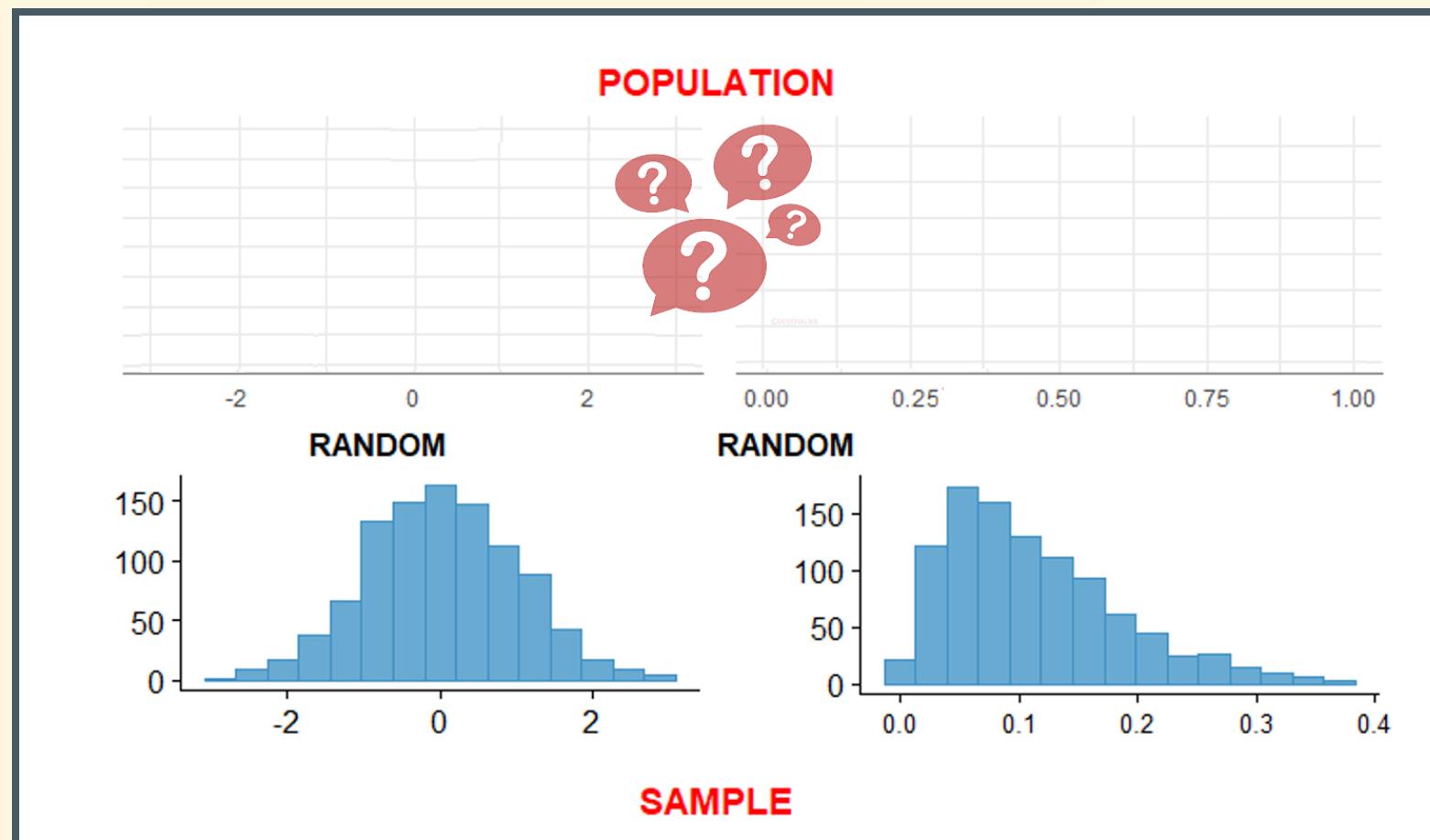
Inferential statistics



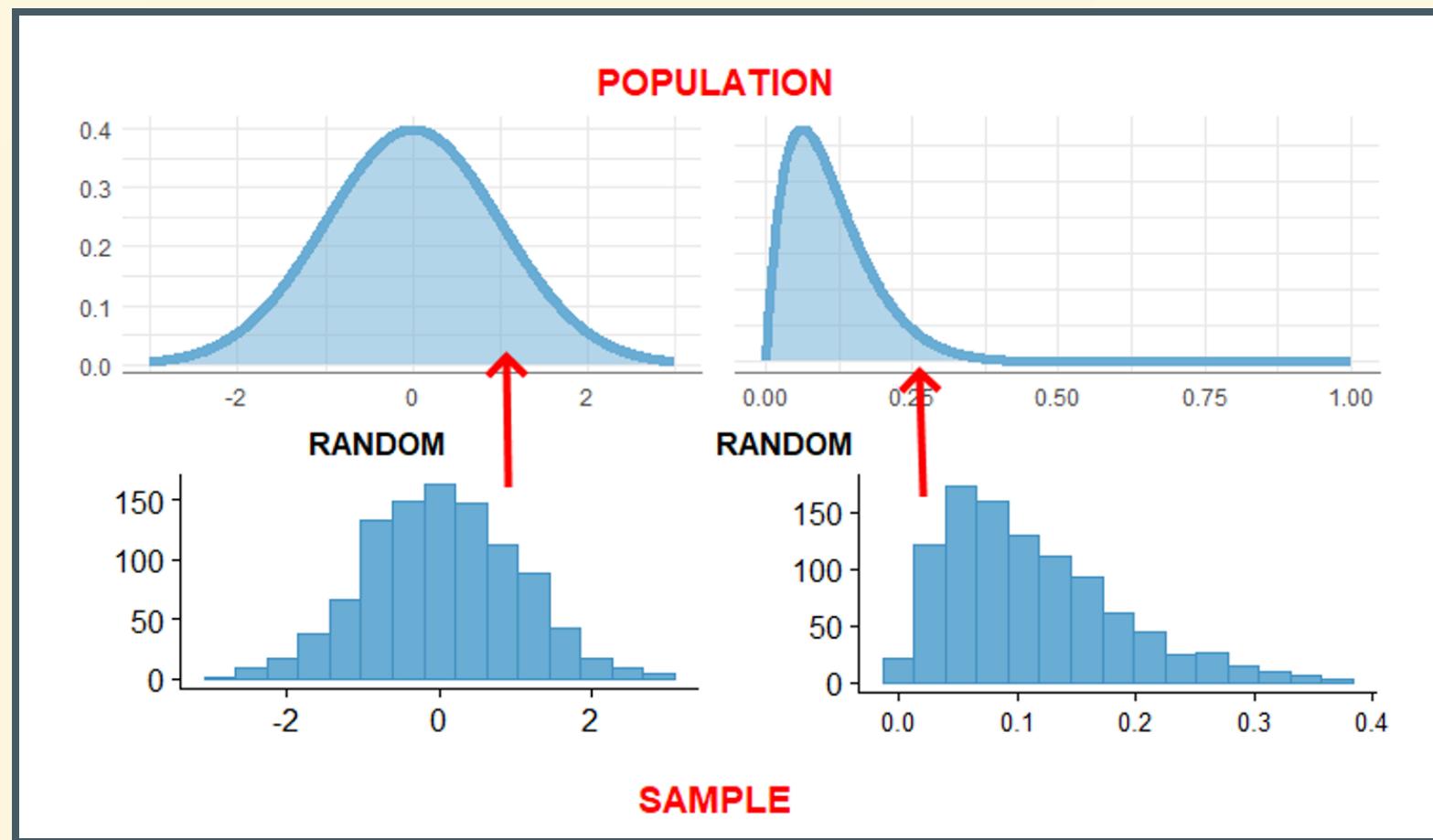
Learning objectives

- Understanding how to move from empirical to theoretical distributions
- Be able to calculate and interpret point and interval estimates (confidence intervals)

From sample to population



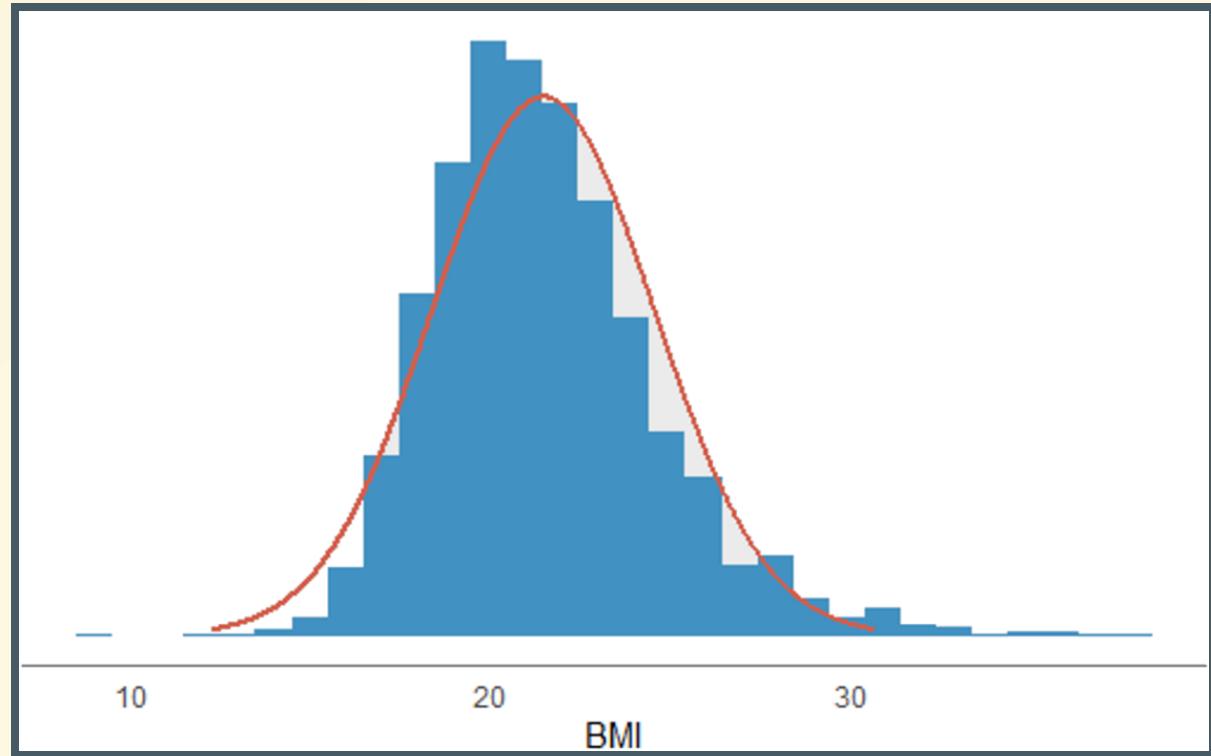
From sample to population



Point estimates

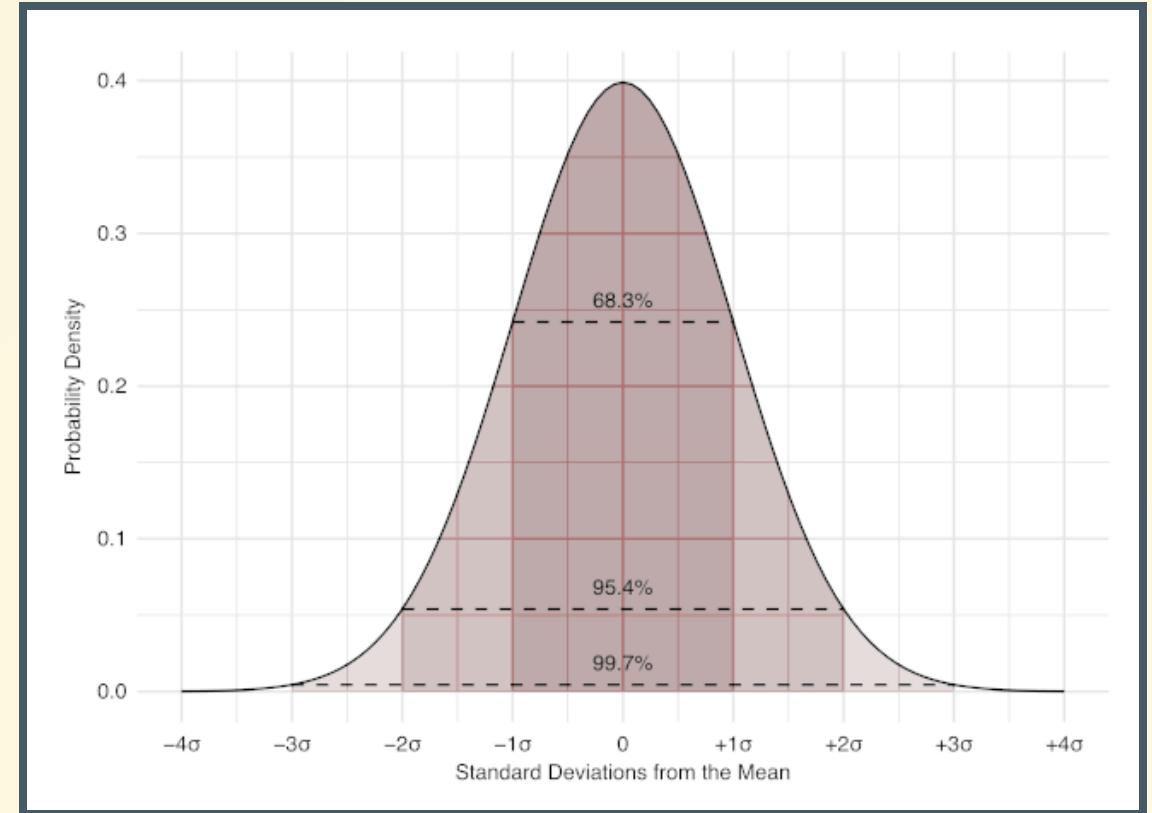
- 📌 $n = 6,705$ 15 y.o. males
 $\mu_{\text{BMI}} = 21.5 \text{ kg/m}^2$
 $\sigma_{\text{BMI}} = 3.1 \text{ kg/m}^2$

What percentage of 15 y.o. males have $\text{BMI} \geq 25$ in the population?



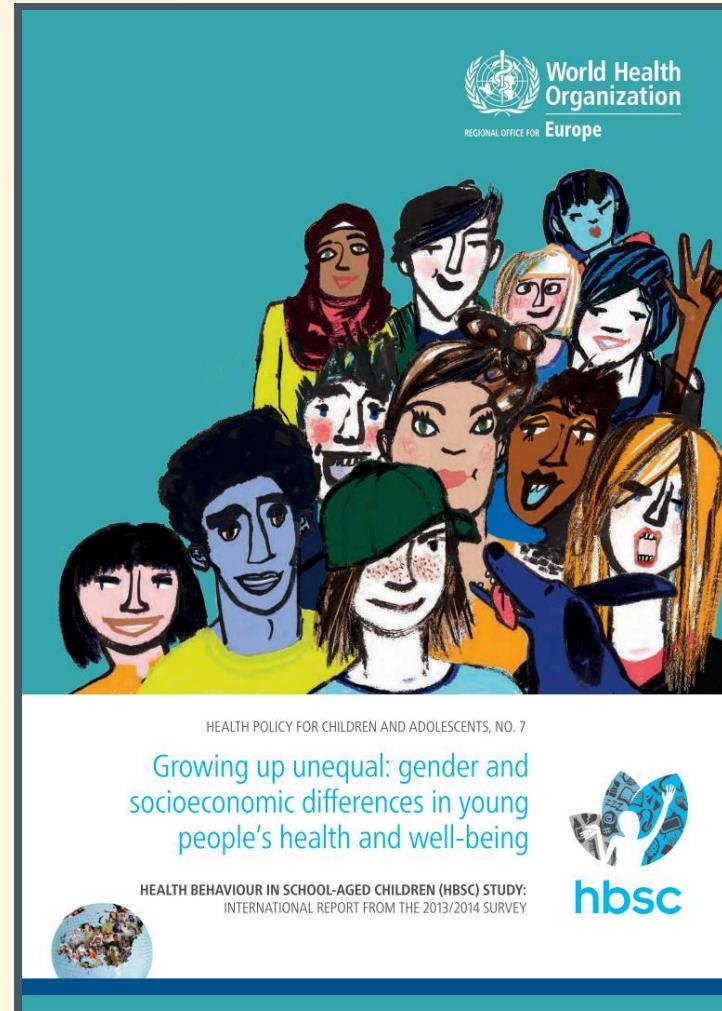
Confidence intervals

- 🎯 A confidence interval (CI) is a range of values which includes the estimated parameter with a given degree of confidence



Confidence intervals for means

- 📌 $n = 6,705$ 15 y.o. males
 $\bar{x}_{\text{BMI}} = 21.5 \text{ kg/m}^2$
 $s_{\text{BMI}} = 3.1 \text{ kg/m}^2$



<https://hbsc.org>

Confidence intervals for means

📌 $n = 6,705$ 15 y.o. males

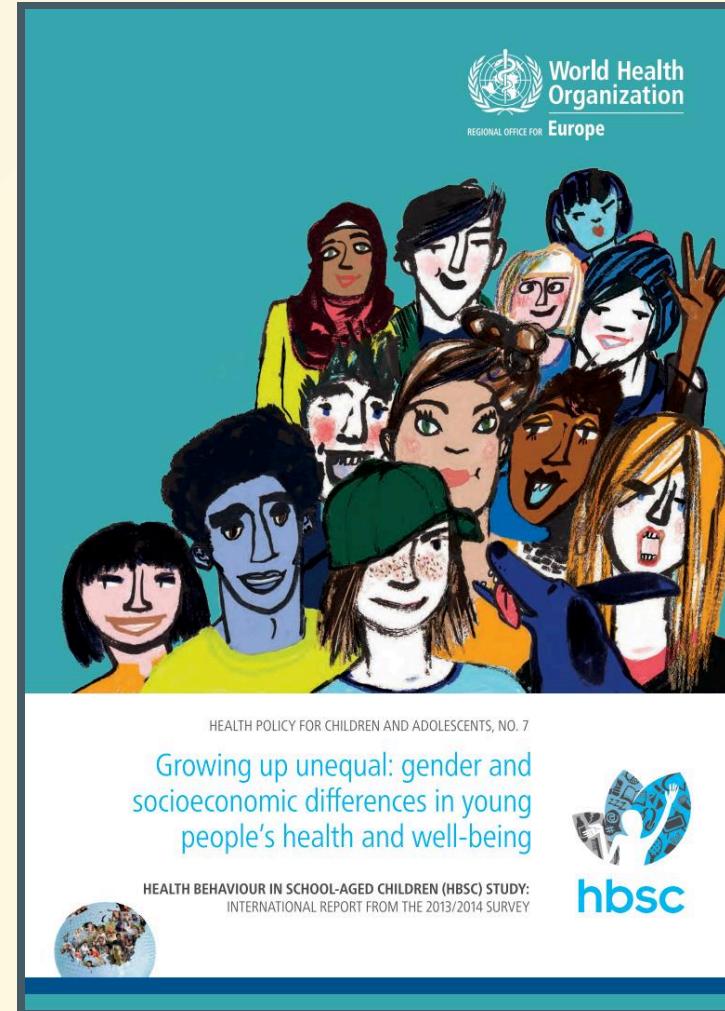
$$\bar{x}_{\text{BMI}} = 21.5 \text{ kg/m}^2$$

$$s_{\text{BMI}} = 3.1 \text{ kg/m}^2$$

$$\text{SE} = \sigma / \sqrt{n}, \text{ where } \sigma = ?$$

$$\hat{\text{SE}} = s / \sqrt{n} = \frac{3.1}{\sqrt{(6,705)}} = 0.038$$

<https://hbsc.org>



Confidence intervals for means

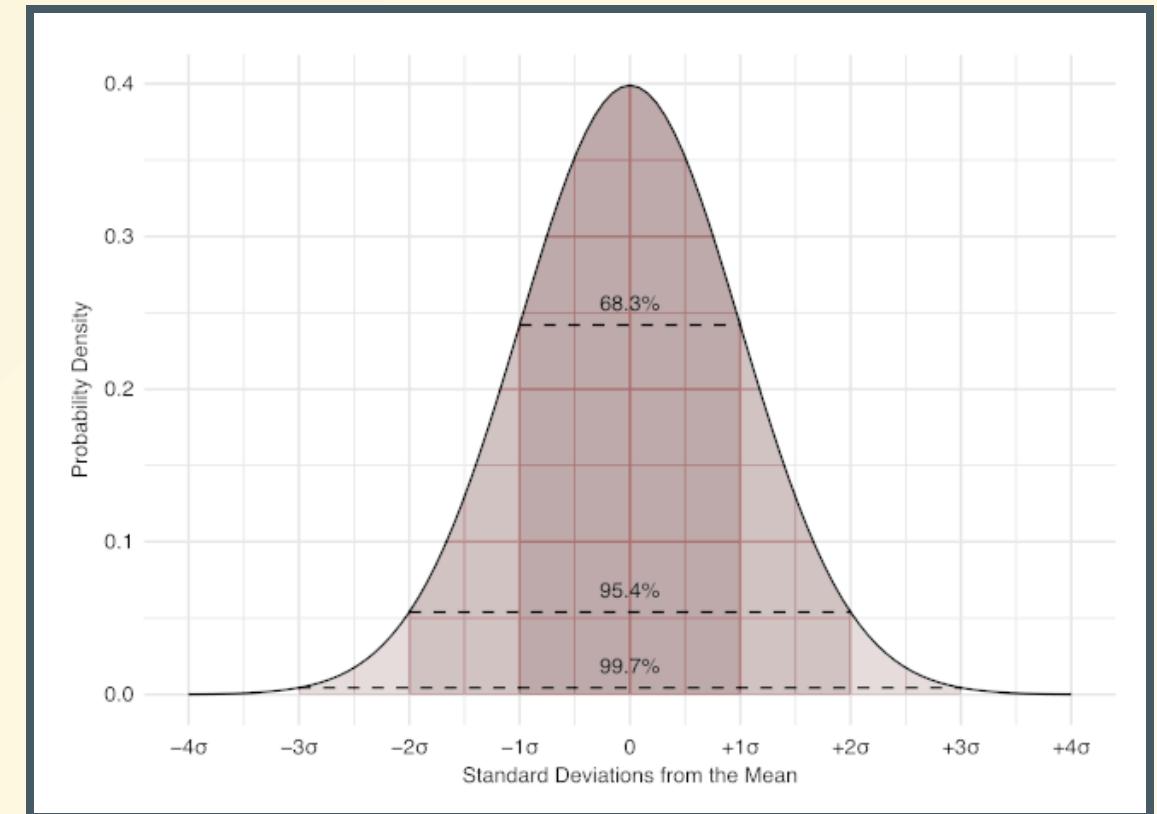
📌 $n = 6,705$ 15 y.o. males

$$\bar{x}_{\text{BMI}} = 21.5 \text{ kg/m}^2$$

$$s_{\text{BMI}} = 3.1 \text{ kg/m}^2$$

$$\text{SE} = \sigma / \sqrt{n}, \text{ where } \sigma = ?$$

$$\hat{\text{SE}} = s / \sqrt{n} = \frac{3.1}{\sqrt{(6,705)}} = 0.038$$



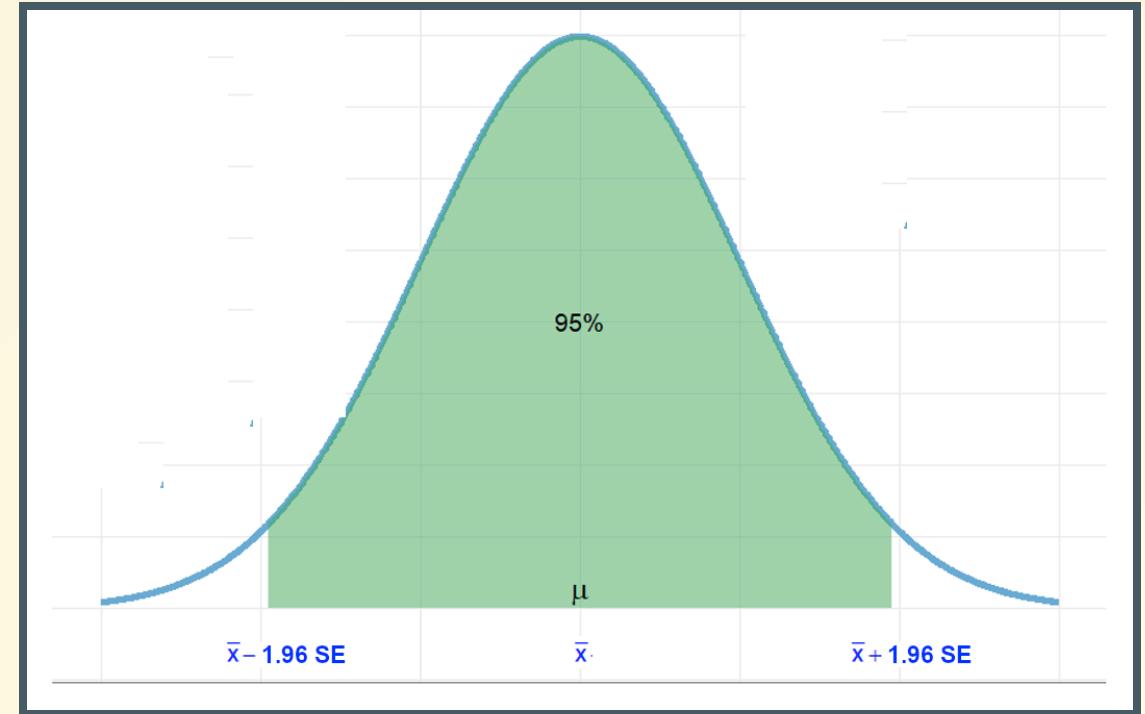
Confidence intervals for means

📌 $n = 6,705$ 15 y.o. males

$$\bar{x}_{\text{BMI}} = 21.5 \text{ kg/m}^2$$

$$s_{\text{BMI}} = 3.1 \text{ kg/m}^2$$

$$\hat{SE} = s/\sqrt{n} = \frac{3.1}{\sqrt{6,705}} = 0.038$$



$$\mathcal{P}(\bar{x} - 1.96 \times SE \leq \mu \leq \bar{x} + 1.96 \times \hat{SE}) = 95\%$$

Confidence intervals for means

📌 $n = 6,705$ 15 y.o. males

$$\bar{x}_{\text{BMI}} = 21.5 \text{ kg/m}^2$$

$$s_{\text{BMI}} = 3.1 \text{ kg/m}^2$$

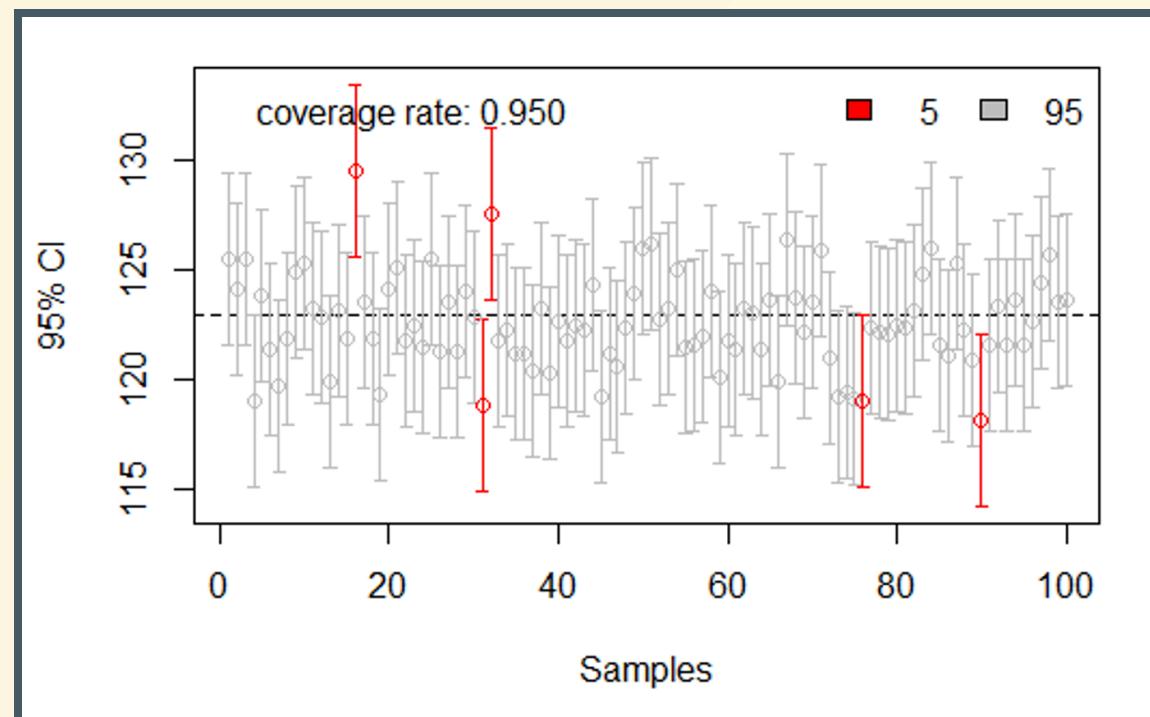
$$\hat{SE} = s/\sqrt{n} = \frac{3.1}{\sqrt{(6,705)}} = 0.038$$

$$\begin{aligned}\mathcal{P}(\bar{x} - 1.96 \times \hat{SE} \leq \mu \leq \bar{x} + 1.96 \times \hat{SE}) &= \\ &= \mathcal{P}(21.5 - 1.96 \times 0.038 \leq \mu \leq 21.5 + 1.96 \times 0.038) = \\ &= \mathcal{P}(21.42 \leq \mu \leq 21.58) = 95\%\end{aligned}$$

→ 95% Confidence Interval (CI) = (21.42; 21.58)

Interpreting confidence intervals

- Population: Italian women 25-74 years old
 $\mu = 123 \text{ mmHg}$



Exercise #10

- ? We take a random sample of 500 Italian women aged 25-74 with the following summary statistics

$$\bar{x} = 122.1 \text{ mmHg}$$

$$s = 19.8 \text{ mmHg}$$

What is the 95% CI for the true mean μ ?

Exercise #10 -- Solution

? $n = 500$

$$\bar{x} = 122.1 \text{ mmHg}$$

$$s = 19.8 \text{ mmHg}$$

What is the 95% CI for the true mean μ ?

$$\hat{SE} = s/\sqrt{n} = \frac{19.8}{\sqrt{(500)}} = 0.89$$

$$\begin{aligned} 95\% \text{ CI} &= (\bar{x} - 1.96 \times \hat{SE}; \bar{x} + 1.96 \times \hat{SE}) = \\ &= (122.1 - 1.96 \times 0.89; 122.1 + 1.96 \times 0.89) \\ &= (120.34; 123.86) \end{aligned}$$

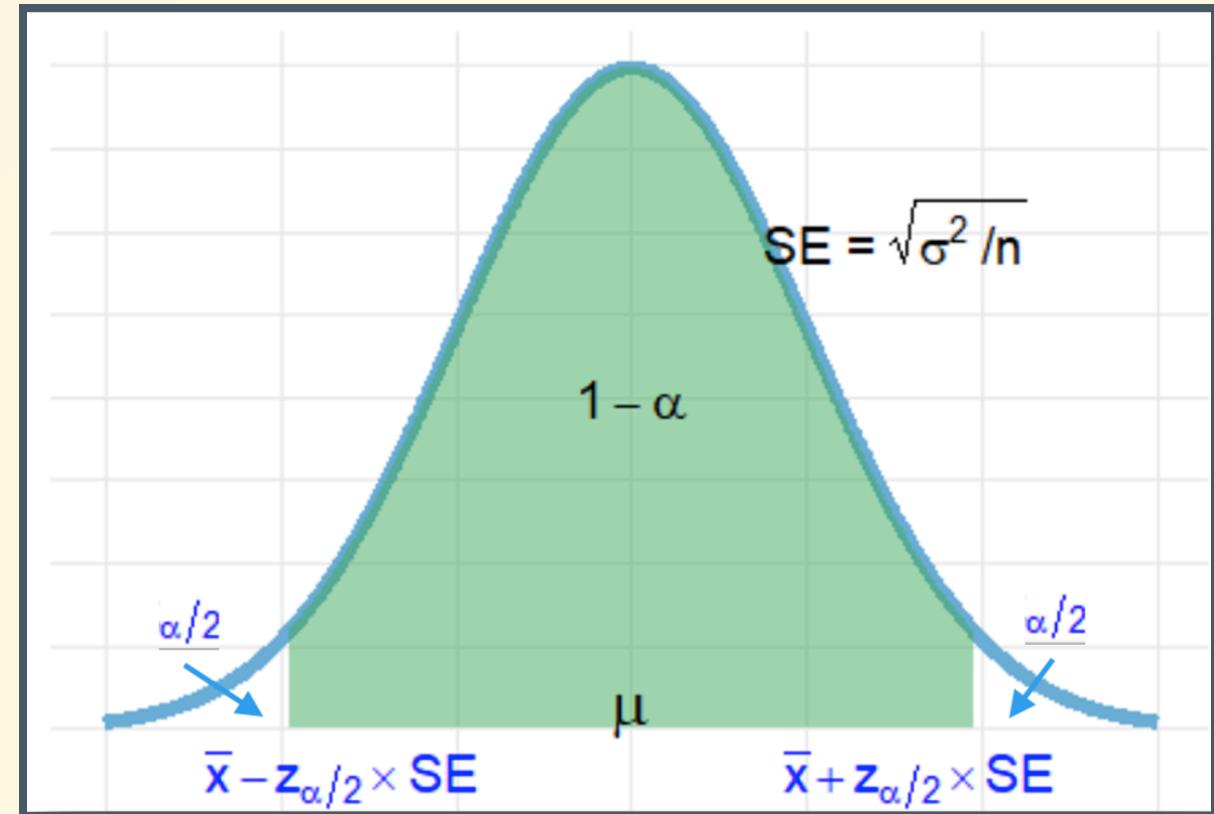
The α level



95% CI = $(\bar{x} - 1.96 \times \hat{SE}; \bar{x} + 1.96 \times \hat{SE})$

1.96 ?

Confidence Level	α	$\alpha/2$	$z_{\alpha/2}$
95%	5%	2.5%	



The α level

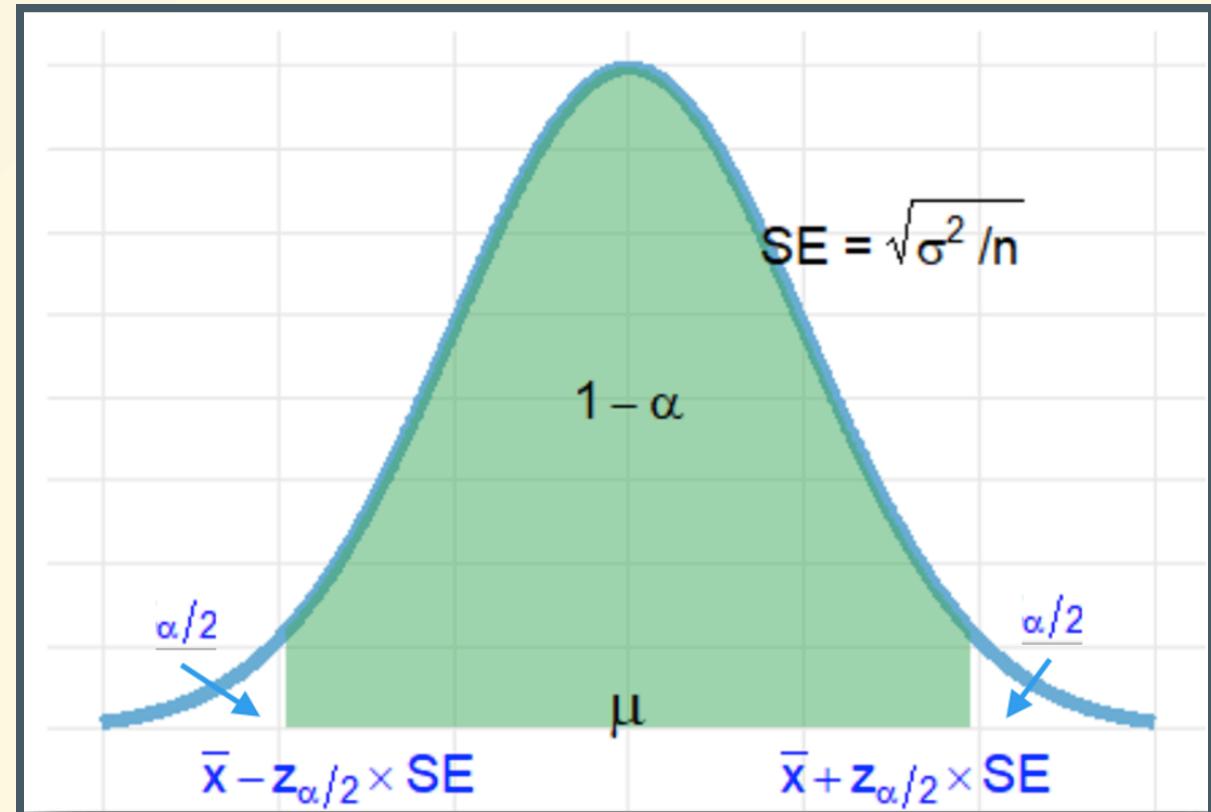


$$95\% \text{ CI} = (\bar{x} - 1.96 \times \hat{SE}; \bar{x} + 1.96 \times \hat{SE})$$

1.96 ?

Confidence Level	α	$\alpha/2$	$z_{\alpha/2}$
95%	5%	2.5%	

$$100\% - 2.5\% = 97.5\%$$



The α level



$$95\% \text{ CI} = (\bar{x} - 1.96 \times \hat{SE}; \bar{x} + 1.96 \times \hat{SE})$$

1.96 ?

Confidence Level	α	$\alpha/2$	$z_{\alpha/2}$
95%	5%	2.5%	1.96

$$100\% - 2.5\% = 97.5\% \rightarrow z = 1.96$$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9710	0.9710	0.9720	0.9732	0.9739	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817

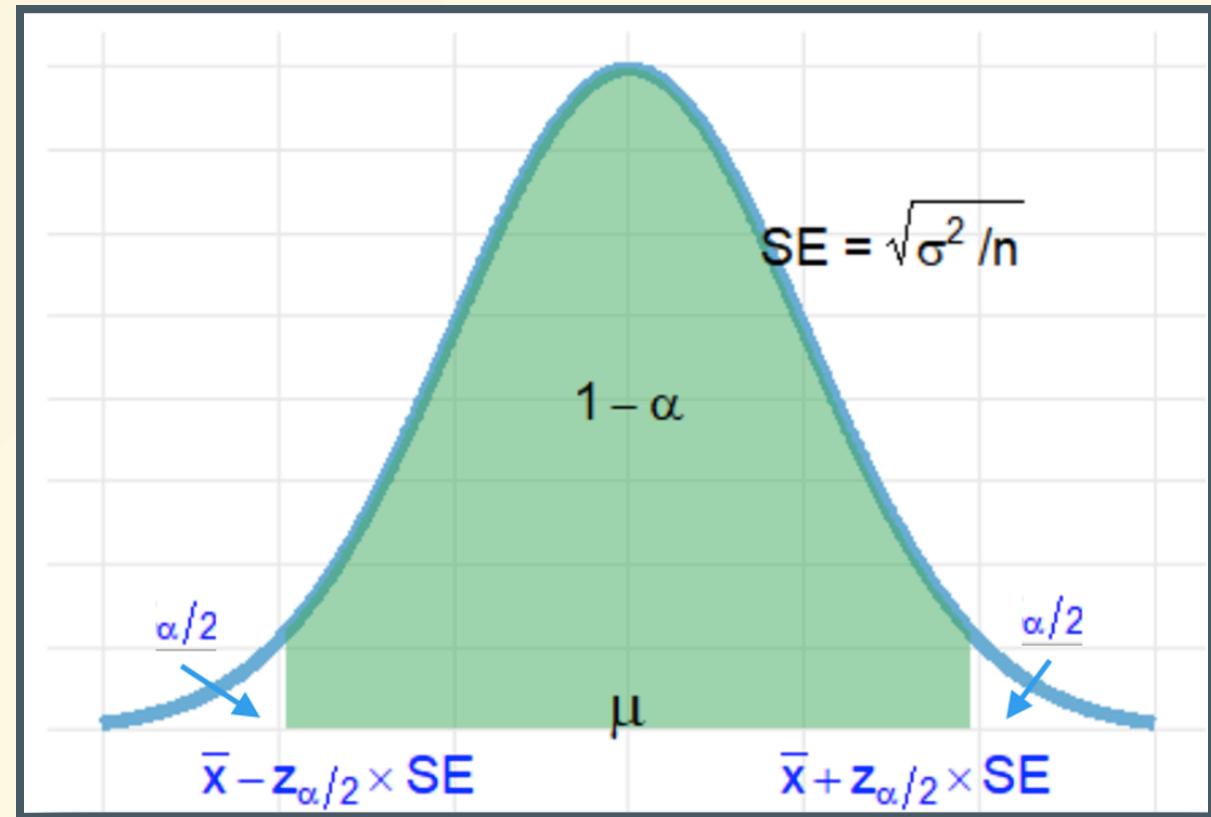
The α level

Confidence Level	α	$\alpha/2$	$z_{\alpha/2}$
95%	5%	2.5%	1.96
90%	10%	5.0%	1.65
99%	1%	0.5%	2.58

$$100\% - 2.5\% = 97.5\% \rightarrow z = 1.96$$

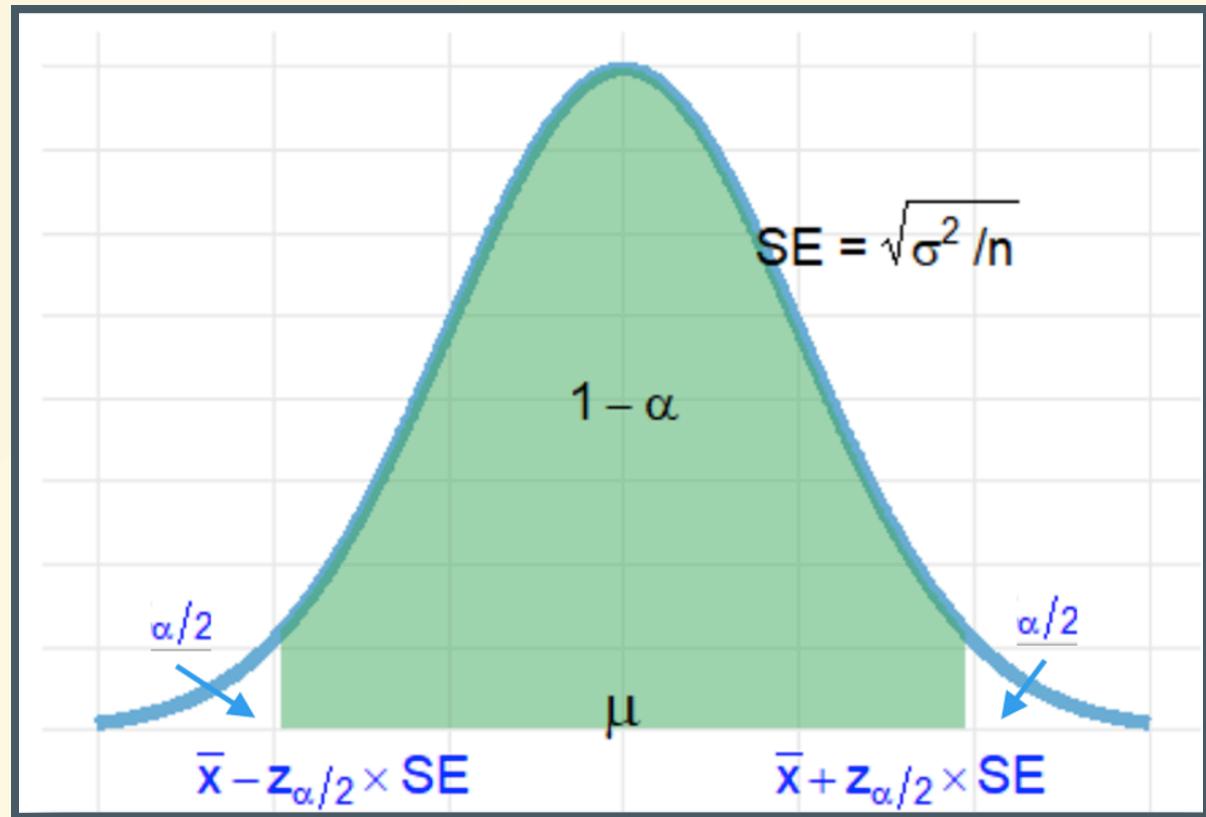
$$100\% - 5.0\% = 95.0\% \rightarrow z = 1.65$$

$$100\% - 0.5\% = 99.5\% \rightarrow z = 2.58$$



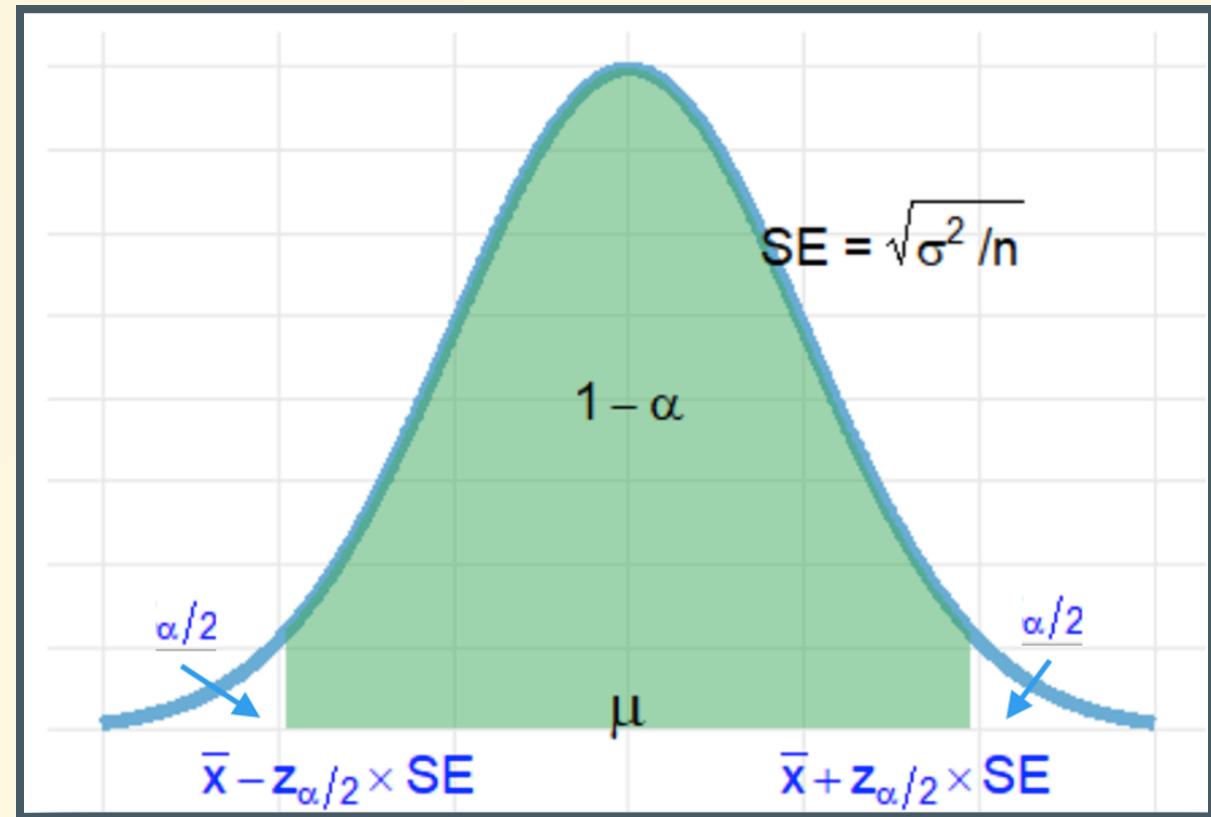
Exercise #11

- ? If the CI is large we are...
- a) more likely of including μ
 - b) less likely of including μ
 - c) there is no difference

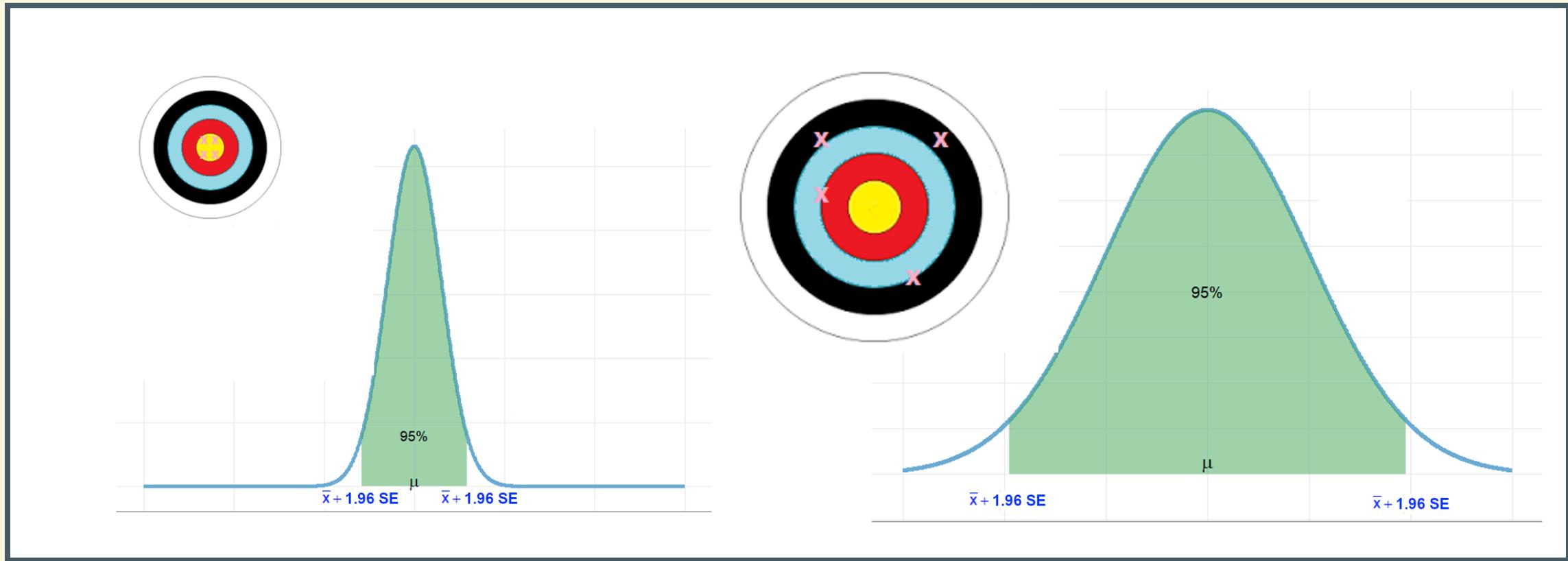


Exercise #11 -- Solution

- ? If the CI is large we are...
- a) more likely of including μ ✓
 - b) less likely of including μ
 - c) there is no difference

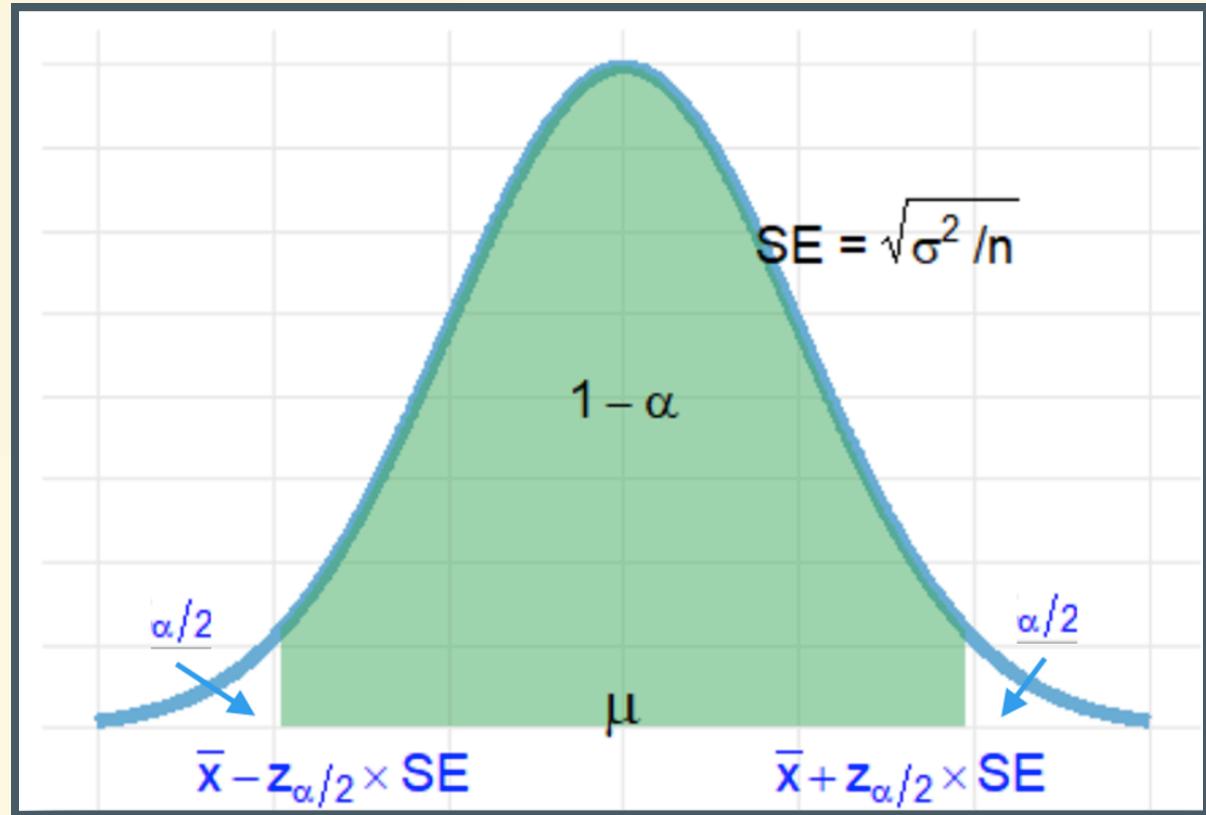


Exercise #11 -- Solution



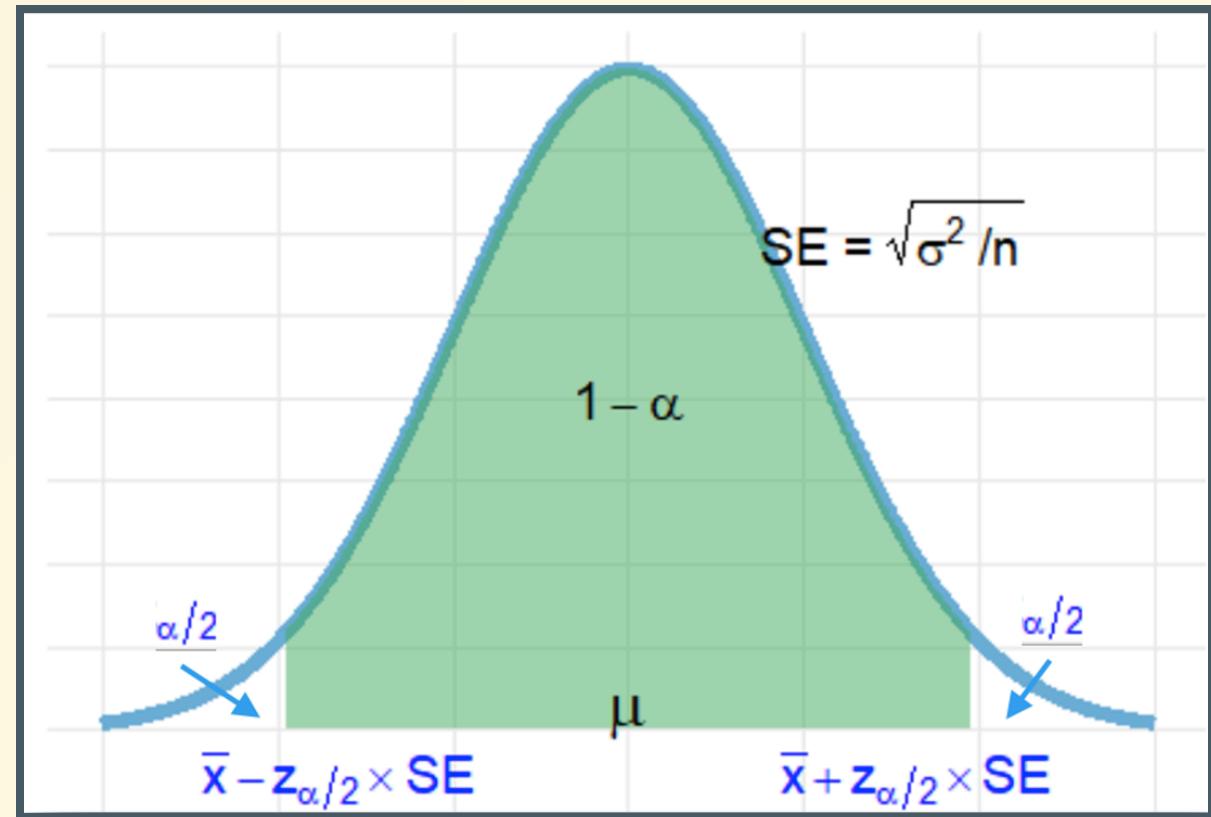
Exercise #12

- ? If the CI is large we are...
- a) more precise
 - b) less precise
 - c) there is no difference

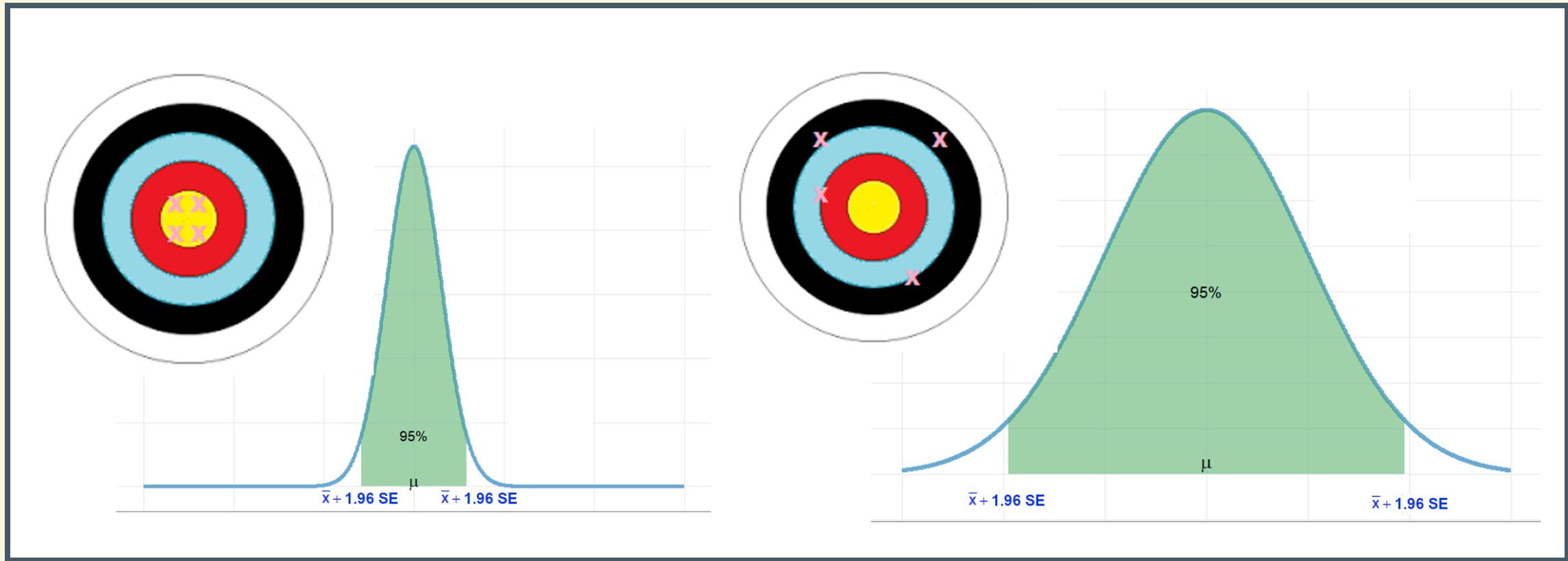


Exercise #12 -- Solution

- ? If the CI is large we are...
- a) more precise
 - b) less precise
 - c) there is no difference



Exercise #12 -- Solution



Confidence intervals for differences of means

- 📌 Are the intervention (i) and the control (c) group different?

Confidence intervals for differences of means

- 📌 Are the intervention (i) and the control (c) group different?

$$\mathcal{N} = \left(\mu_i - \mu_c, \sqrt{\frac{\sigma_i^2}{n_i} + \frac{\sigma_c^2}{n_c}} \right)$$

\rightarrow CLT

$$\hat{SE} = \sqrt{\frac{s_i^2}{n_i} + \frac{s_c^2}{n_c}}$$

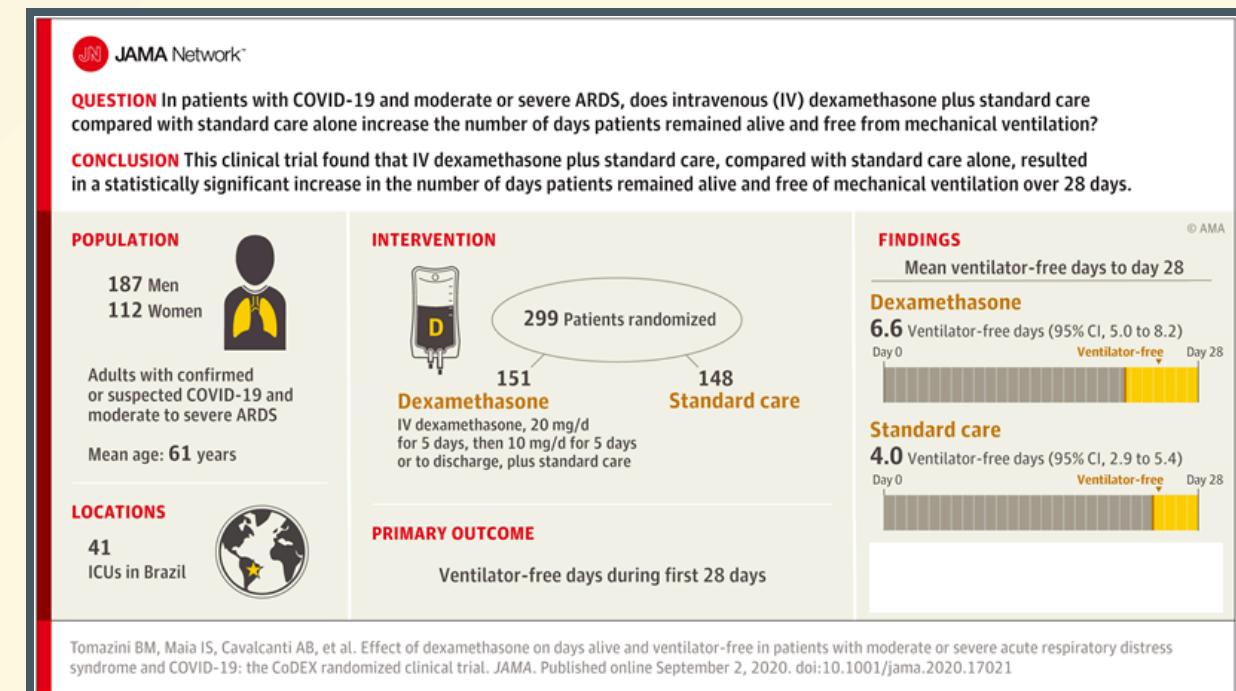
Confidence intervals for differences of means

- Are the intervention (i) and the control (c) group different?

$$n_i = 151, \bar{x}_i = 6.6, s_i = 10.0$$

$$n_c = 148, \bar{x}_c = 4.0, s_c = 8.7$$

$$95\% \text{ CI} = ?$$



Confidence intervals for differences of means

- 📌 $n_i = 151, \bar{x}_i = 6.6, s_i = 10.0$
 $n_c = 148, \bar{x}_c = 4.0, s_c = 8.7$

$$\bar{x}_i - \bar{x}_c = 6.6 - 4.0 = 2.6$$

$$\hat{SE} = \sqrt{\frac{s_i^2}{n_i} + \frac{s_c^2}{n_c}} = \sqrt{\frac{10.0^2}{151} + \frac{8.7^2}{148}} = 1.08$$

Confidence intervals for differences of means

- 📌 $n_i = 151, \bar{x}_i = 6.6, s_i = 10.0$
 $n_c = 148, \bar{x}_c = 4.0, s_c = 8.7$

$$\bar{x}_i - \bar{x}_c = 6.6 - 4.0 = 2.6$$

$$\hat{SE} = \sqrt{\frac{s_i^2}{n_i} + \frac{s_c^2}{n_c}} = \sqrt{\frac{10.0^2}{151} + \frac{8.7^2}{148}} = 1.08$$

$$\begin{aligned} 95\% \text{ CI} &= (\bar{x}_i - \bar{x}_c) - 1.96 \times \hat{SE}; (\bar{x}_i - \bar{x}_c) + 1.96 \times \hat{SE}) = \\ &= (2.6 - 1.96 \times 1.08; 2.6 + 1.96 \times 1.08) = \\ &= (0.48; 4.72) \end{aligned}$$

Exercise #13

? $n_i = 151, \bar{x}_i = 6.6, s_i = 10.0$
 $n_c = 148, \bar{x}_c = 4.0, s_c = 8.7$

$$\bar{x}_i - \bar{x}_c = 6.6 - 4.0 = 2.6$$

$$\hat{SE} = \sqrt{\frac{s_i^2}{n_i} + \frac{s_c^2}{n_i}} = \sqrt{\frac{10.0^2}{151} + \frac{8.7^2}{148}} = 1.08$$

$$90\% \text{ CI} = ? \quad (z_{\alpha/2} = z_{5/2} = 1.65)$$

Exercise #13 -- Solution

? $n_i = 151, \bar{x}_i = 6.6, s_i = 10.0$
 $n_c = 148, \bar{x}_c = 4.0, s_c = 8.7$

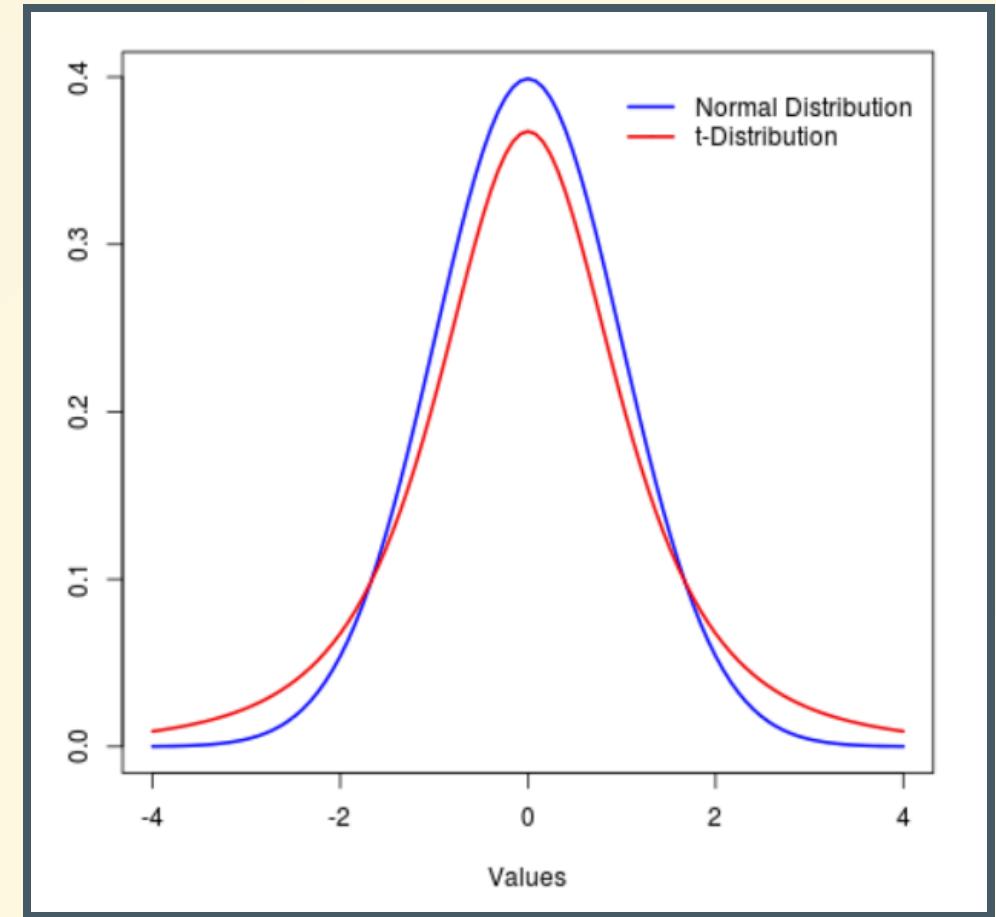
$$\bar{x}_i - \bar{x}_c = 6.6 - 4.0 = 2.6$$

$$\hat{SE} = \sqrt{\frac{s_i^2}{n_i} + \frac{s_c^2}{n_i}} = \sqrt{\frac{10.0^2}{151} + \frac{8.7^2}{148}} = 1.08$$

$$90\% \text{ CI} = (2.6 - 1.65 \times 1.08; 2.6 + 1.65 \times 1.08) = (0.82; 4.38)$$

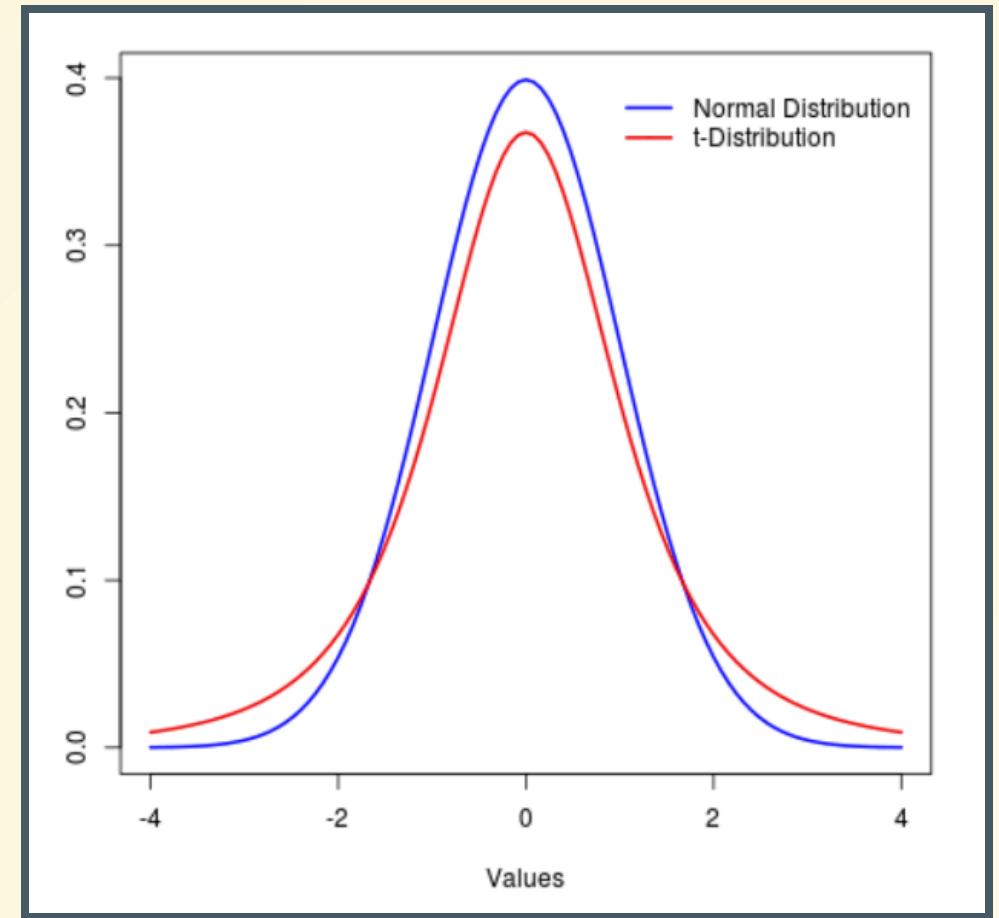
CTL and small sample size

- CTL works for large sample size
- t distribution



CTL and small sample size

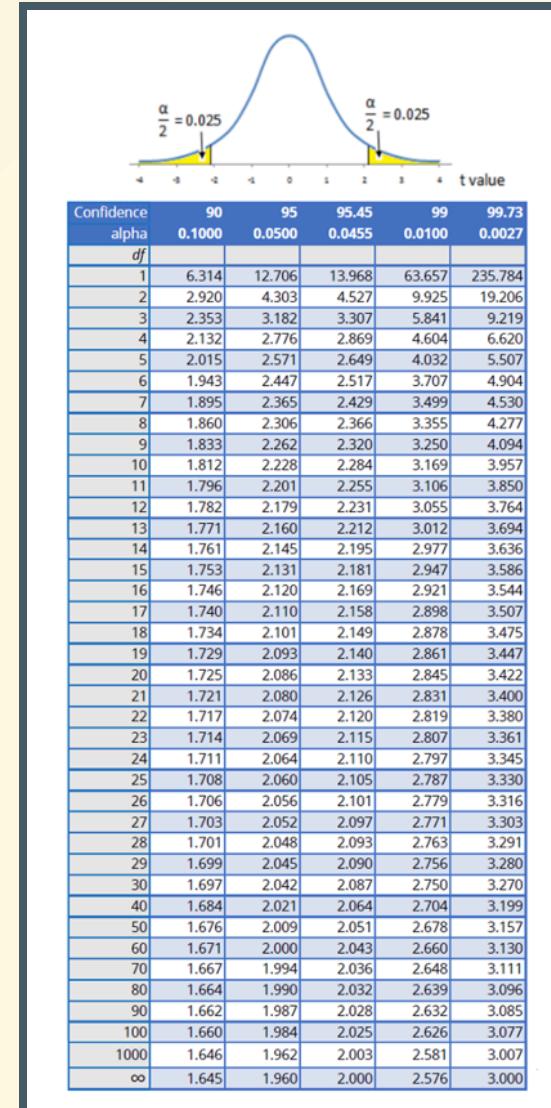
- CTL works for large sample size
- t distribution
 - keeps into account the degree of freedom (df)
 - one sample of size $n \rightarrow df = n - 1$
 - two samples of size
 $n_1 \wedge n_2 \rightarrow df = n_1 - 1 + n_2 - 1 =$
 $= n_1 + n_2 - 2$



CTL and small sample size

- CTL works for large sample size
- t distribution
 - keeps into account the degree of freedom (df)
 - one sample of size $n \rightarrow df = n - 1$
 - two samples of size
 $n_1 \wedge n_2 \rightarrow df = n_1 - 1 + n_2 - 1 = n_1 + n_2 - 2$

$$95\% \text{ CI} = (\bar{x} - t \times \hat{SE}; \bar{x} + t \times \hat{SE})$$



CTL and small sample size

📌 $n = 58$ patients with T2D

$$\bar{x}_{\text{BMI}} = 25.0 \text{ kg/m}^2$$

$$s_{\text{BMI}} = 2.7 \text{ kg/m}^2$$

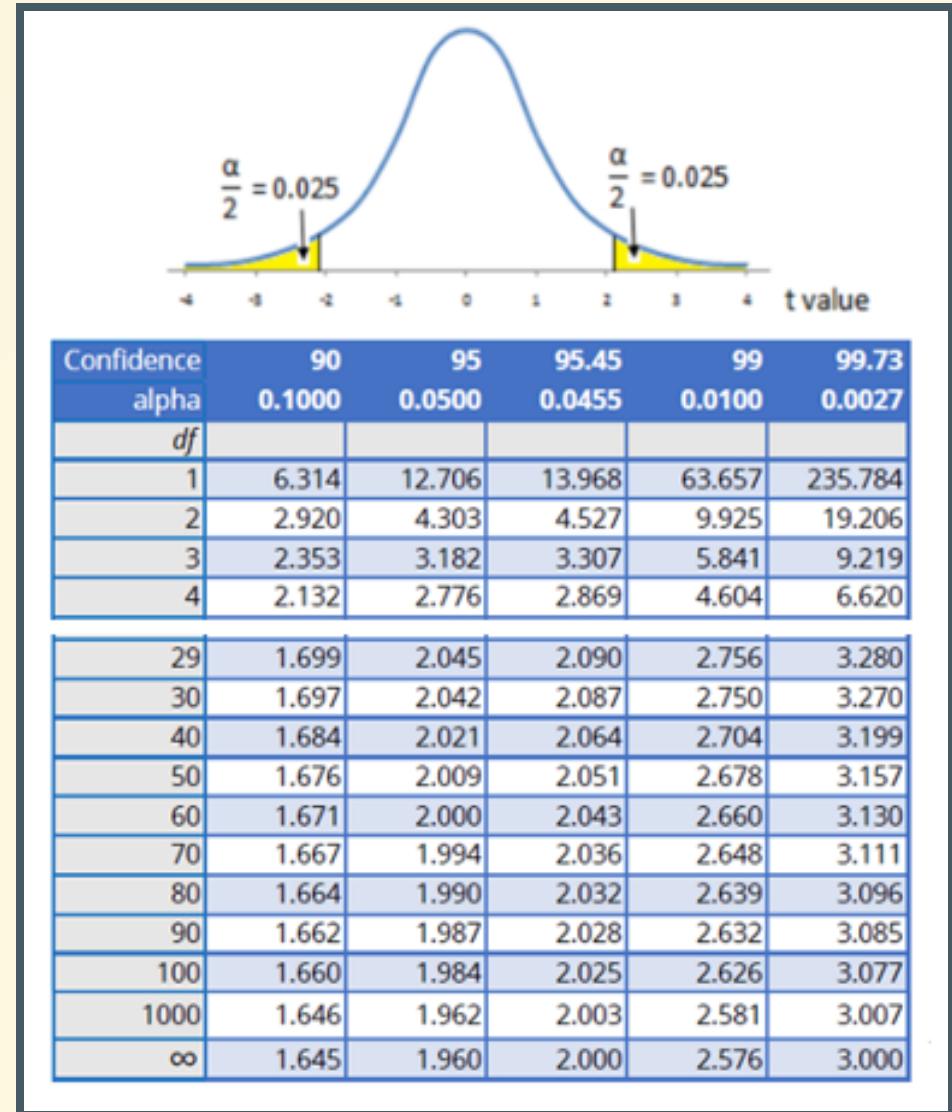
95% CI = ?

CTL and small sample size

- 📌 $n = 58$ patients with T2D
 $\bar{x}_{\text{BMI}} = 25.0 \text{ kg/m}^2$
 $s_{\text{BMI}} = 2.7 \text{ kg/m}^2$

$$\hat{\text{SE}} = s / \sqrt{n} = \frac{2.7}{\sqrt{58}} = 0.36$$

$$\text{df} = n - 1 = 58 - 1 = 57$$



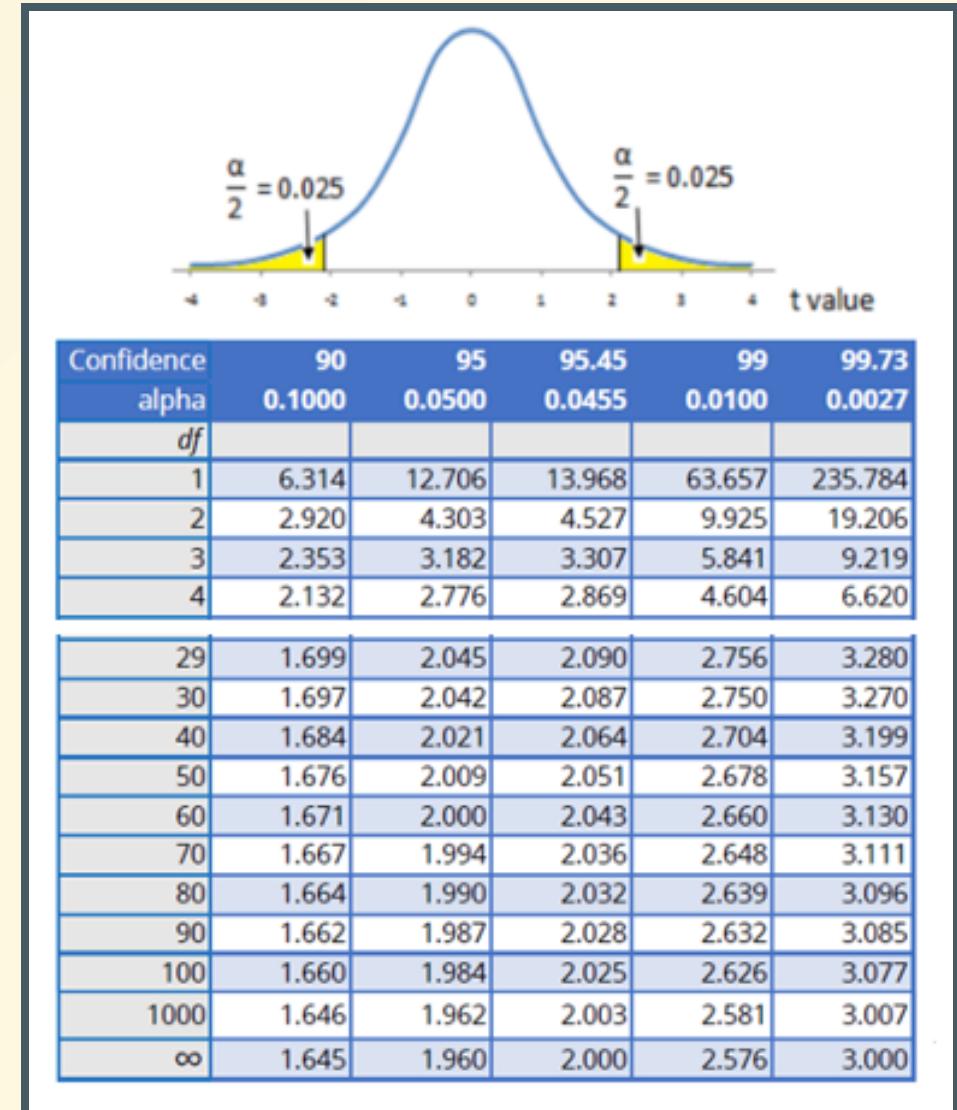
CTL and small sample size

- 📌 $n = 58$ patients with T2D
 $\bar{x}_{\text{BMI}} = 25.0 \text{ kg/m}^2$
 $s_{\text{BMI}} = 2.7 \text{ kg/m}^2$

$$\hat{SE} = s/\sqrt{n} = \frac{2.7}{\sqrt{58}} = 0.36$$

$$df = n - 1 = 58 - 1 = 57$$

$$\begin{aligned} 95\% \text{ CI} &= (\bar{x} - t \times \hat{SE}; \bar{x} + t \times \hat{SE}) = \\ &= (25.0 - t \times 0.36; 25.0 + t \times 0.36) = \\ &= (25.0 - 2 \times 0.36; 25.0 + 2 \times 0.36) = \\ &= (25.0 - 0.72; 25.0 + 0.72) = \\ &= (24.28; 25.72) \end{aligned}$$



Confidence intervals for proportions

- 📌 How many patients show gastrointestinal side effects from a new medication?

Confidence intervals for proportions

- 📌 How many patients show gastrointestinal side effects from a new medication?

$$\mathcal{N} = \left(\pi, \frac{\pi \times (1 - \pi)}{n} \right) \rightarrow \text{CLT}$$

$$\hat{SE} = \sqrt{\frac{\bar{p} \times (1 - \bar{p})}{n}}, \text{ where } \bar{p} = \frac{m}{n}$$

Confidence intervals for proportions

- 📌 How many patients show gastrointestinal side effects from a new medication?

$n = 100, m = 69$

95% CI = ?

Confidence intervals for proportions

📌 $n = 100, m = 69$

$$\bar{p} = \frac{m}{n} = \frac{69}{100} = 0.69$$

$$\hat{SE} = \sqrt{\frac{\bar{p} \times (1 - \bar{p})}{n}} = \sqrt{\frac{0.69 \times (1 - 0.69)}{100}} = 0.046$$

Confidence intervals for proportions

📌 $n = 100, m = 69$

$$\bar{p} = \frac{m}{n} = \frac{69}{100} = 0.69$$

$$\hat{SE} = \sqrt{\frac{\bar{p} \times (1 - \bar{p})}{n}} = \sqrt{\frac{0.69 \times (1 - 0.69)}{100}} = 0.046$$

$$\begin{aligned} 95\% \text{ CI} &= (\bar{p} - 1.96 \times \hat{SE}; \bar{p} + 1.96 \times \hat{SE}) = \\ &= (0.69 - 1.96 \times 0.046; 0.69 + 1.96 \times 0.046) = \\ &= (0.60; 0.78) \end{aligned}$$

Exercise #14

? $n = 100, m = 69$

$$\bar{p} = \frac{m}{n} = \frac{69}{100} = 0.69$$

$$\hat{SE} = \sqrt{\frac{\bar{p} \times (1 - \bar{p})}{n}} = \sqrt{\frac{0.69 \times (1 - 0.69)}{100}} = 0.046$$

99% CI = ? $(z_{\alpha/2} = z_{0.5/2} = 2.58)$

Exercise #14 -- Solution

? $n = 100, m = 69$

$$\bar{p} = \frac{m}{n} = \frac{69}{100} = 0.69$$

$$\hat{SE} = \sqrt{\frac{\bar{p} \times (1 - \bar{p})}{n}} = \sqrt{\frac{0.69 \times (1 - 0.69)}{100}} = 0.046$$

$$\begin{aligned}99\% \text{ CI} &= (\bar{p} - 2.58 \times \hat{SE}; \bar{p} + 2.58 \times \hat{SE}) = \\&= (0.69 - 2.58 \times 0.046; 0.69 + 2.58 \times 0.046) = (0.57; 0.81)\end{aligned}$$

Confidence intervals for differences of proportion

- 📌 Are the intervention (i) and the control (c) group different?

Confidence intervals for differences of proportion

- 📌 Are the intervention (i) and the control (c) group different?

$$\mathcal{N} = \left(\pi_i - \pi_c, \frac{\pi_i \times (1 - \pi_i)}{n_i} + \frac{\pi_c \times (1 - \pi_c)}{n_c} \right)$$

\rightarrow CLT

$$\hat{SE} = \sqrt{\frac{\bar{p}_i \times (1 - \bar{p}_i)}{n_i} + \frac{\bar{p}_c \times (1 - \bar{p}_c)}{n_c}}$$

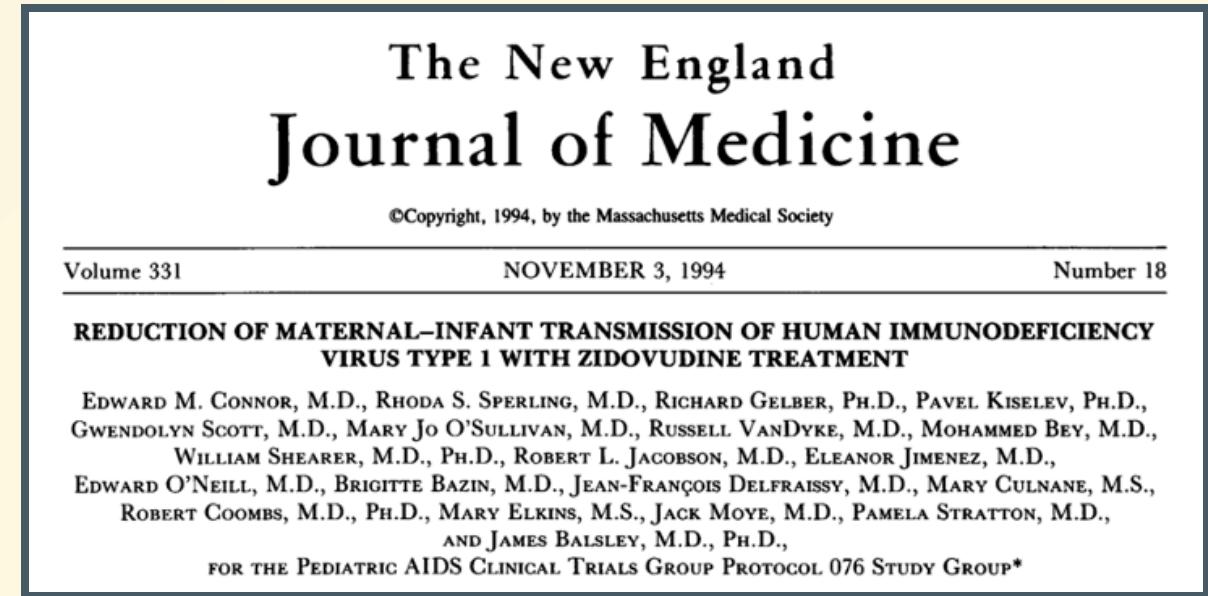
Confidence intervals for differences of proportion

- 📌 Are the intervention (i) and the control (c) group different?

$$n_i = 180, m_i = 13$$

$$n_c = 183, m_c = 40$$

$$95\% \text{ CI} = ?$$



Confidence intervals for differences of proportion

- 📌 Are the intervention (i) and the control (c) group different?

$$n_i = 180, m_i = 13$$

$$n_c = 183, m_c = 40$$

$$\bar{p}_i - \bar{p}_c = \frac{m_i}{n_i} - \frac{m_c}{n_c} = \frac{13}{180} - \frac{40}{183} = 0.07 - 0.22 = -0.15$$

$$\hat{SE} = \sqrt{\frac{\bar{p}_i \times (1 - \bar{p}_i)}{n_i} + \frac{\bar{p}_c \times (1 - \bar{p}_c)}{n_c}} = \sqrt{\frac{0.07 \times (1 - 0.07)}{180} + \frac{0.22 \times (1 - 0.22)}{183}} = 0.036$$

Confidence intervals for differences of proportion

- 📌 Are the intervention (i) and the control (c) group different?

$$n_i = 180, m_i = 13$$

$$n_c = 183, m_c = 40$$

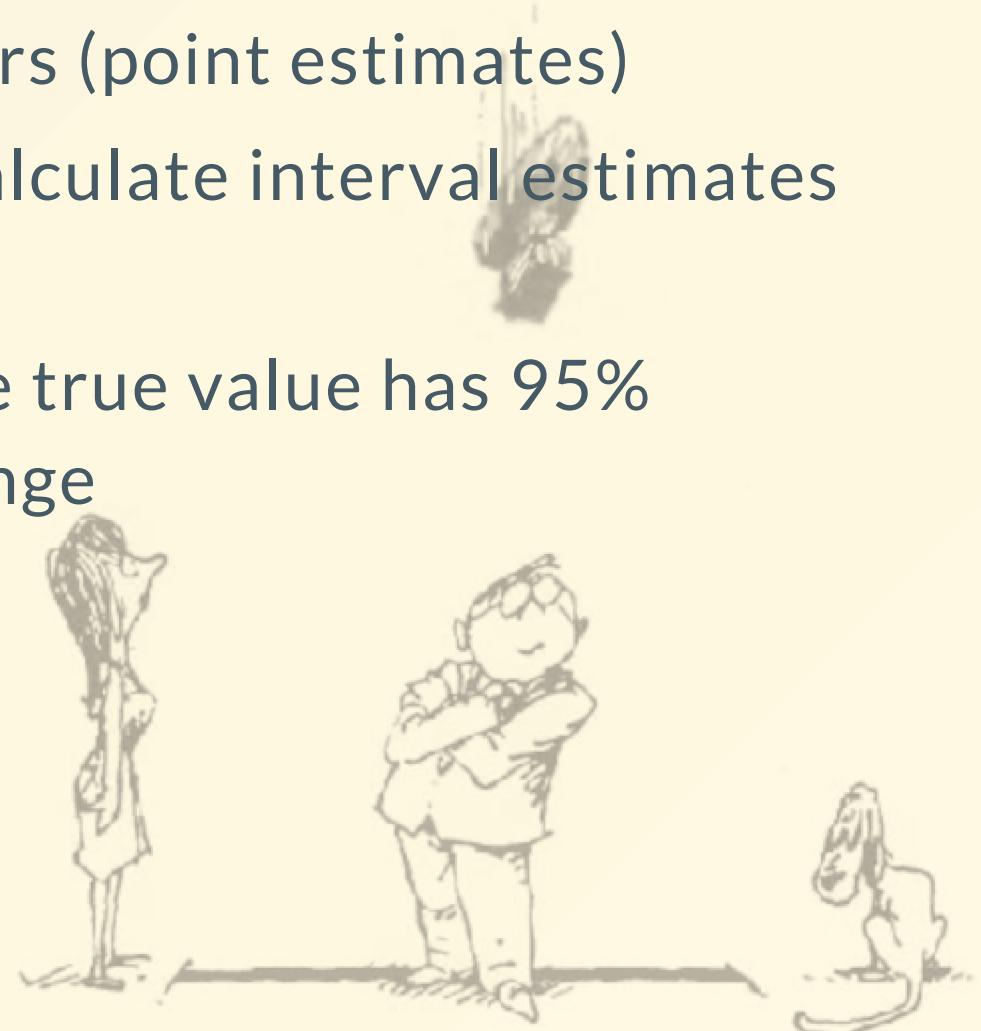
$$\bar{p}_i - \bar{p}_c = -0.15$$

$$\hat{SE} = 0.036$$

$$\begin{aligned} 95\% \text{ CI} &= (\bar{p}_i - \bar{p}_c - 1.96 \times \hat{SE}; \bar{p}_i - \bar{p}_c + 1.96 \times \hat{SE}) = \\ &= (-0.15 - 0.007; -0.015 + 0.07) = (-0.22; -0.08) \end{aligned}$$

Summary

- We use statistics to estimate parameters (point estimates)
- We can take advantage of the CLT to calculate interval estimates (CI)
- 95% confidence intervals tell us the true value has 95% probability of being inside the given range



See you tomorrow

