Introduction to statistics

(Day 2)

Recap

Recap

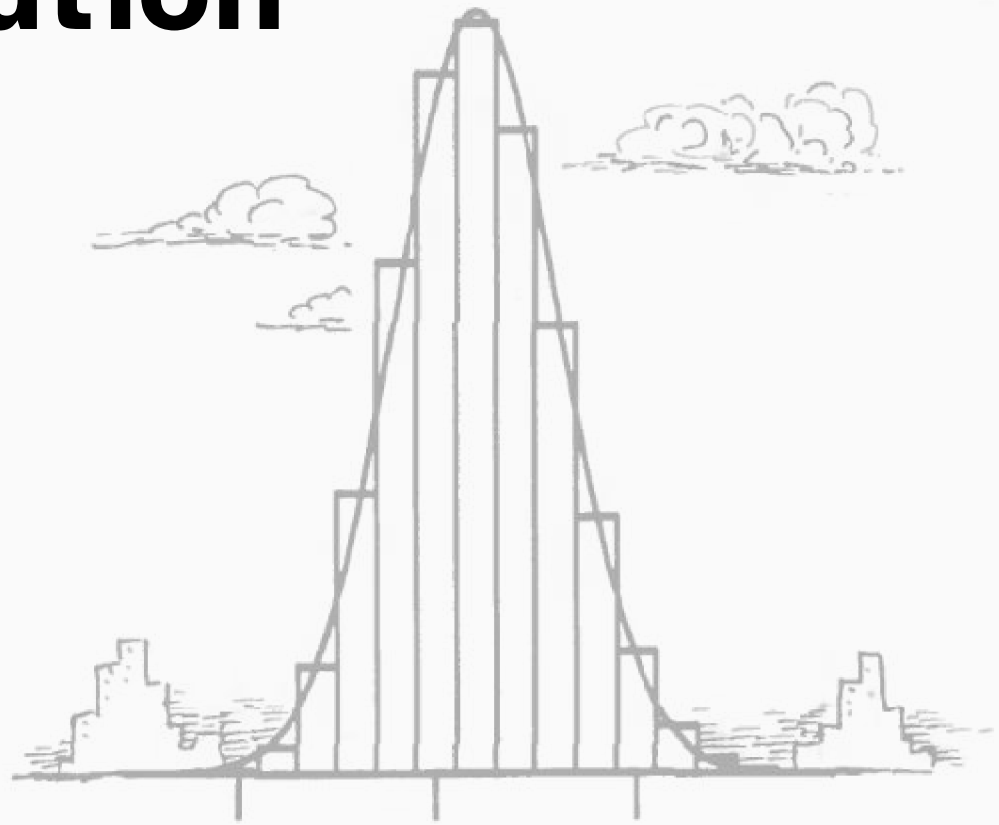
- The collection, organisation, summarisation, and analysis of data
  → *Descriptive* statistics
- The drawing of inferences about a body of data when only a part of the data is observed
  → *Inferential* statistics

- When can't study a population, we select a representative sample
- There are different sampling strategies

Recap

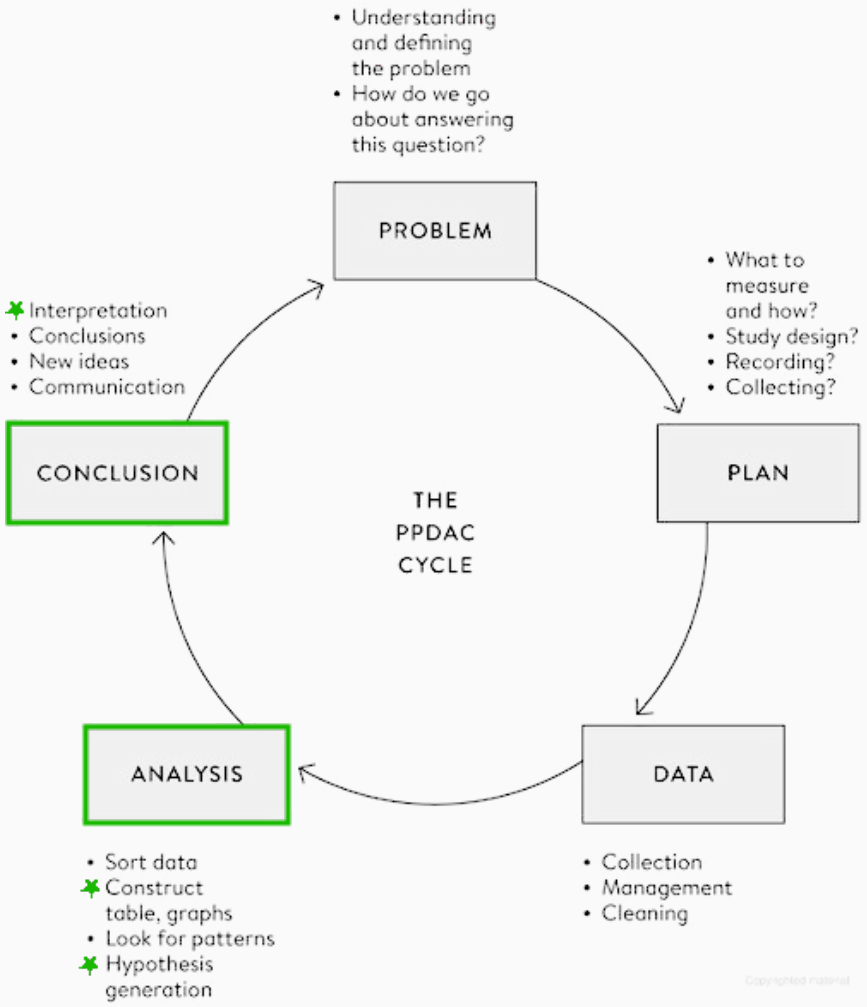
- There are different types of data
- Categorical variables are described with absolute and relative frequencies
- Numerical variables are described with measures of central tendency (mode, median, mean) and dispersion (range, IQR, standard deviation)
- Parameters (calculated on the population) *vs* statistics (calculated on the sample)

The Normal Distribution



Learning objectives

- Know the characteristics of the Normal distribution
- Know the characteristics of the Standard Normal distribution
- Know the characteristics of the Student's t distribution

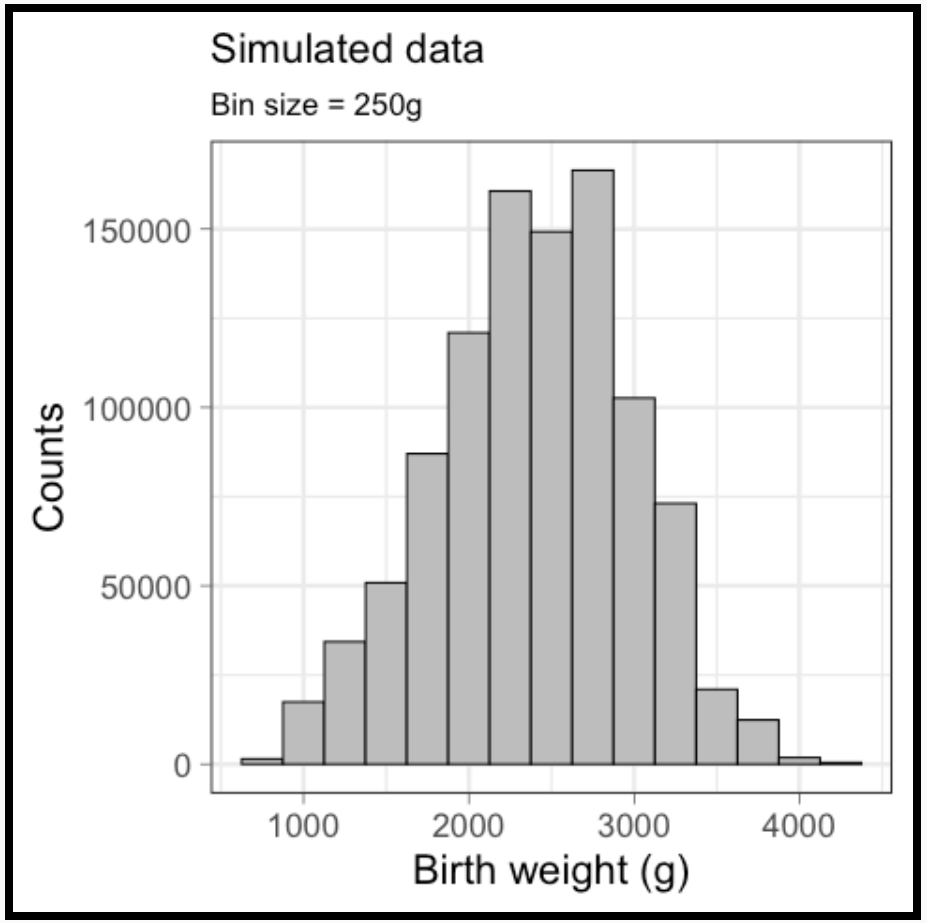


Spiegelhalter, D., *The Art of Statistics: Learning From Data*, Pelican, 2019

Let's go back to populations

Birth weight distribution for British twins

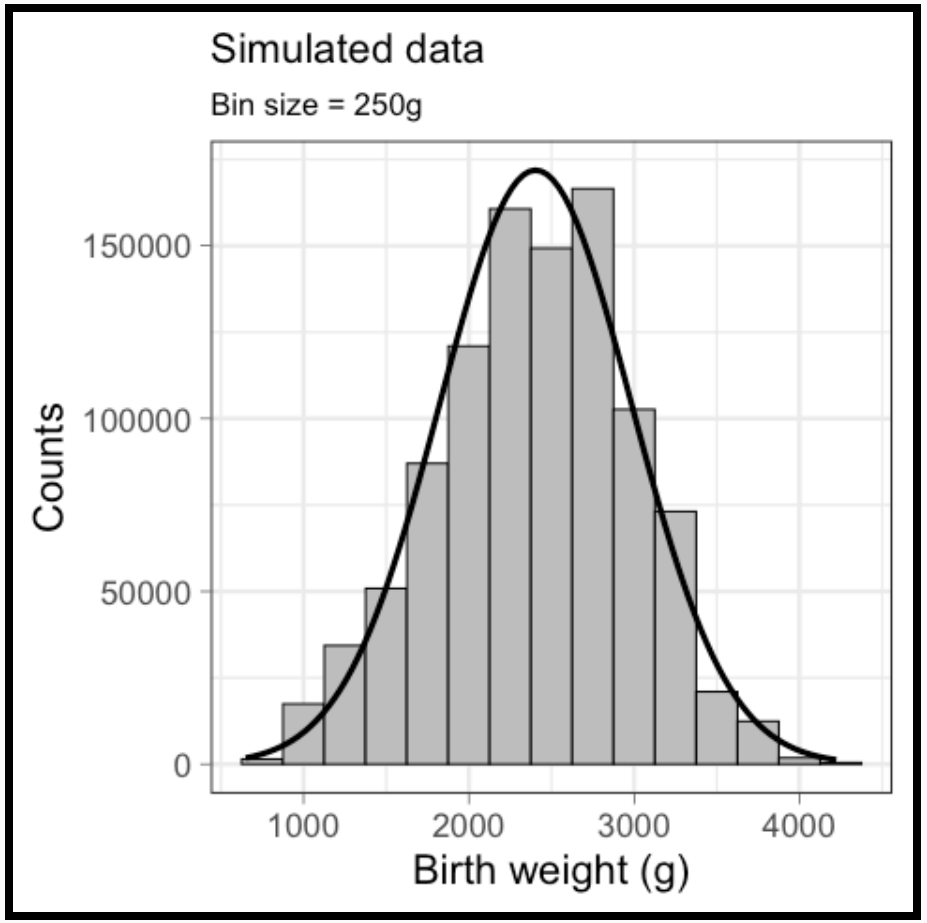
$$N = 1,000,000$$
$$\mu = 2404 \text{ g};\ \sigma = 580 \text{ g}$$
$$median = 2408 \text{ g}$$



Let's go back to populations

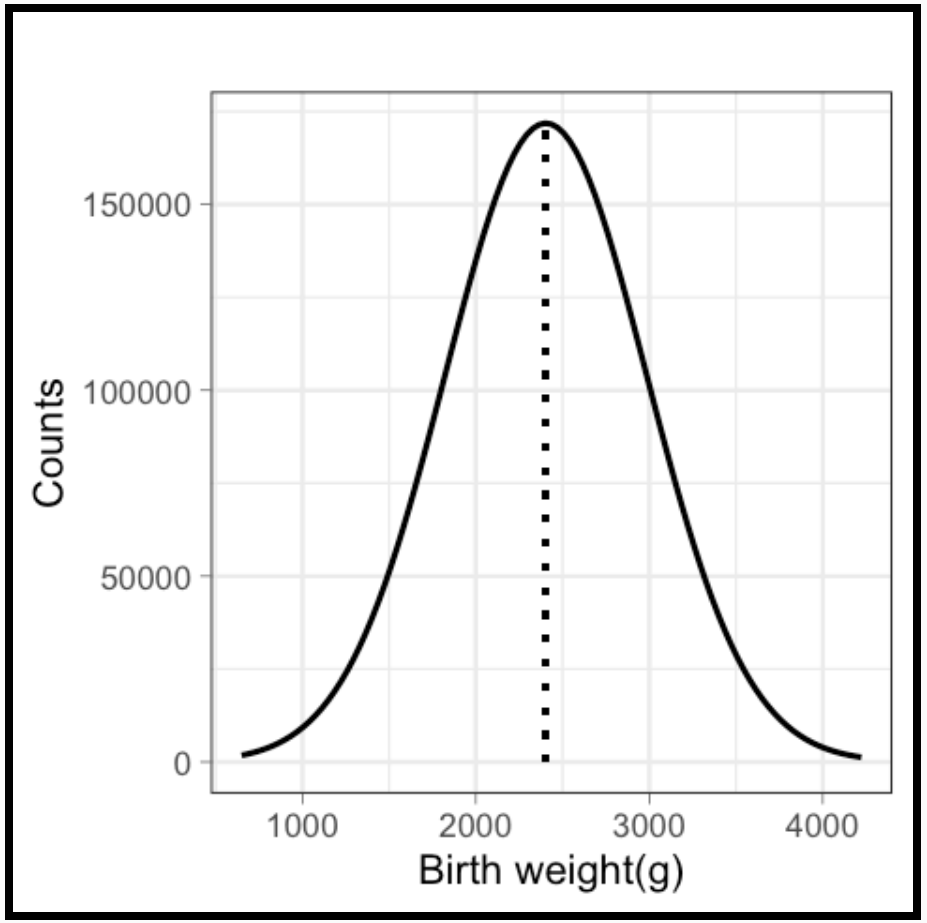
Birth weight distribution for British twins

$$N = 1,000,000$$
$$\mu = 2404 \text{ g};\ \sigma = 580 \text{ g}$$
$$median = 2408 \text{ g}$$



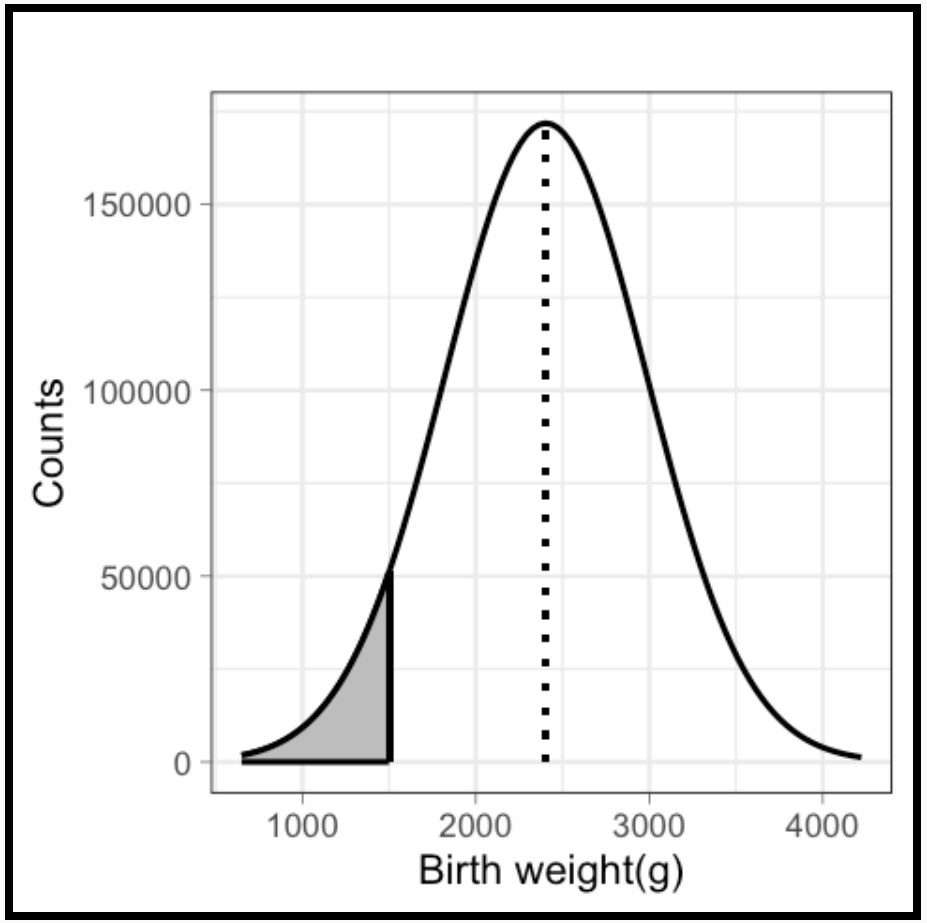
The Normal distribution

- $\mathcal{N} = (\mu, \sigma^2)$
- $\text{mode} \equiv \text{median} \equiv \text{mean}$
- Symmetrical



The Normal distribution

- Area under the curve $= 1$
- proportion $\equiv$ likelihood

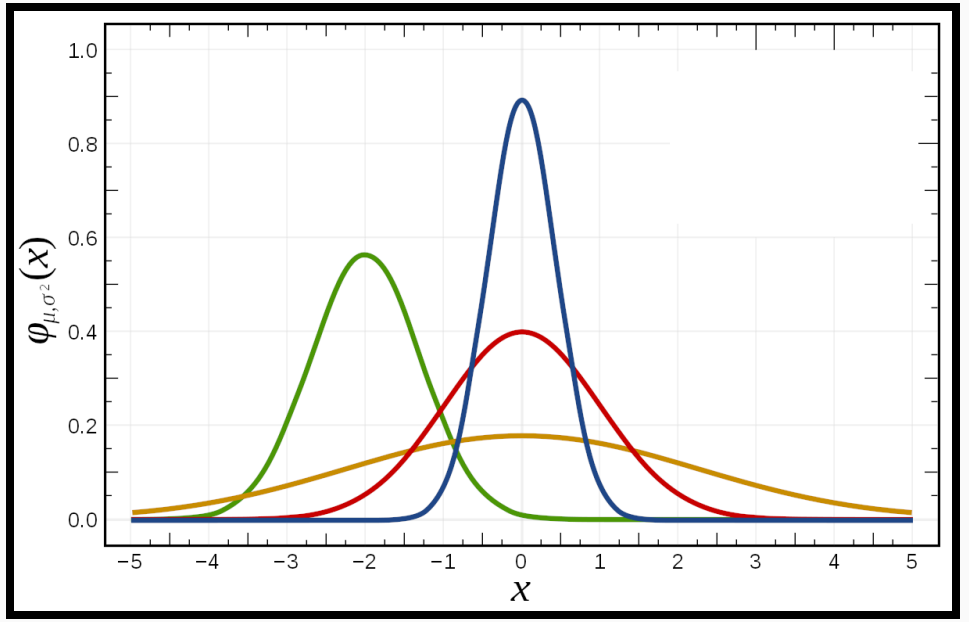


$$\text{very low birth weight (VLBW)} < 1500 \text{ g}$$
$$\text{Twins with VLBW} = 6\%$$
$$\mathcal{P}(\text{twins with VLBW}) = 0.06$$

Exercise #1

? Which distributions has the largest mean?

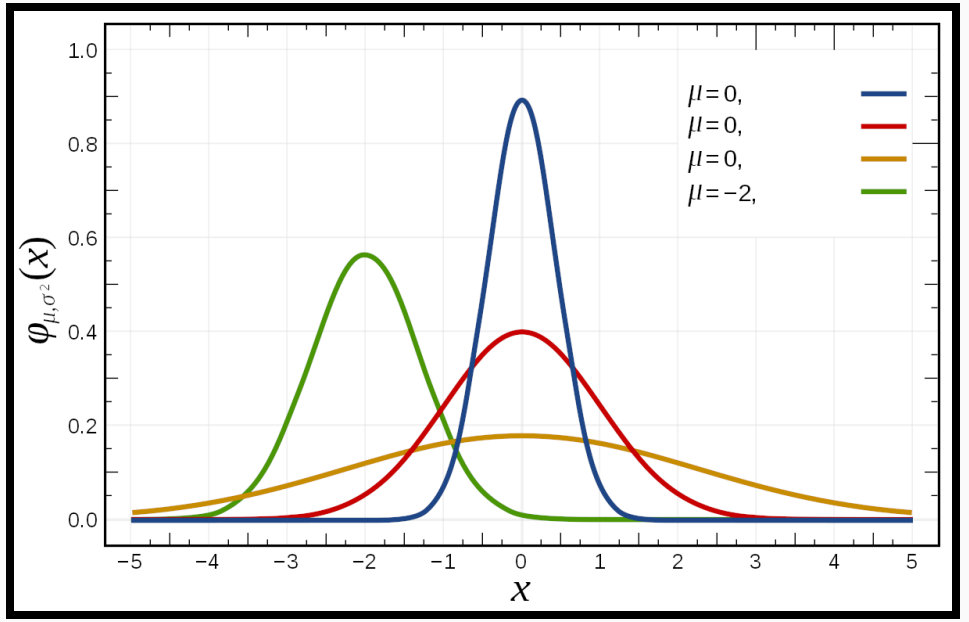
a) Green
b) Blue
c) Yellow
d) None of the above



Exercise #2

? Which distributions has the largest standard deviation?

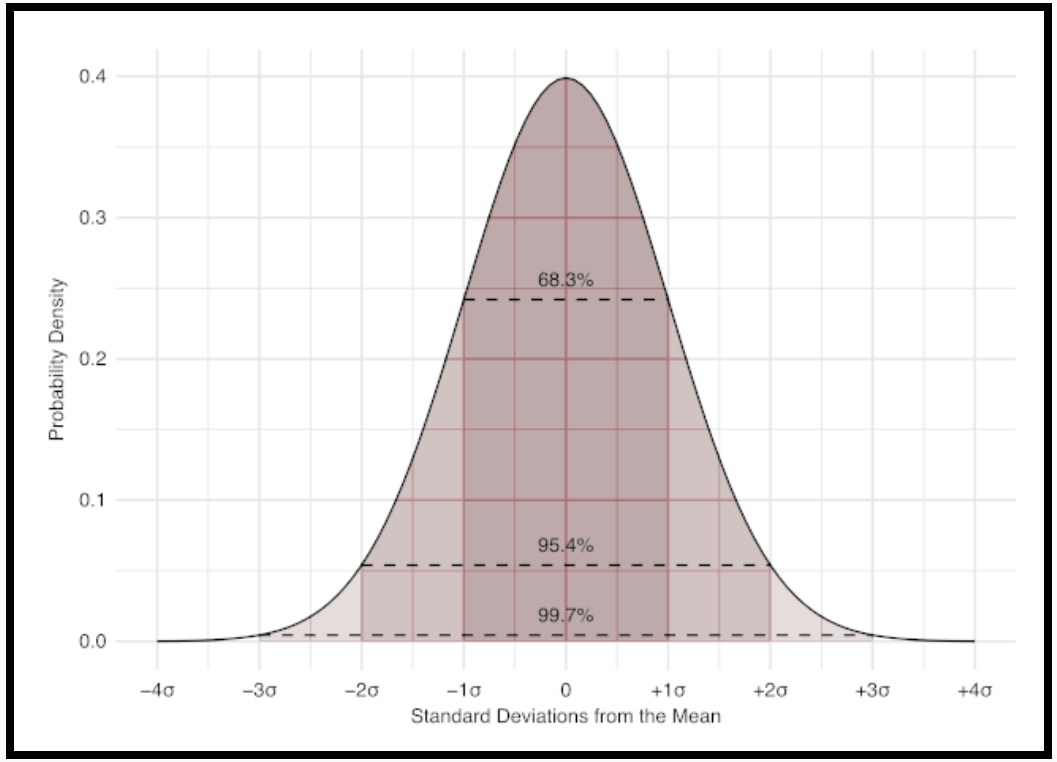
a) Green
b) Blue
c) Yellow
d) None of the above



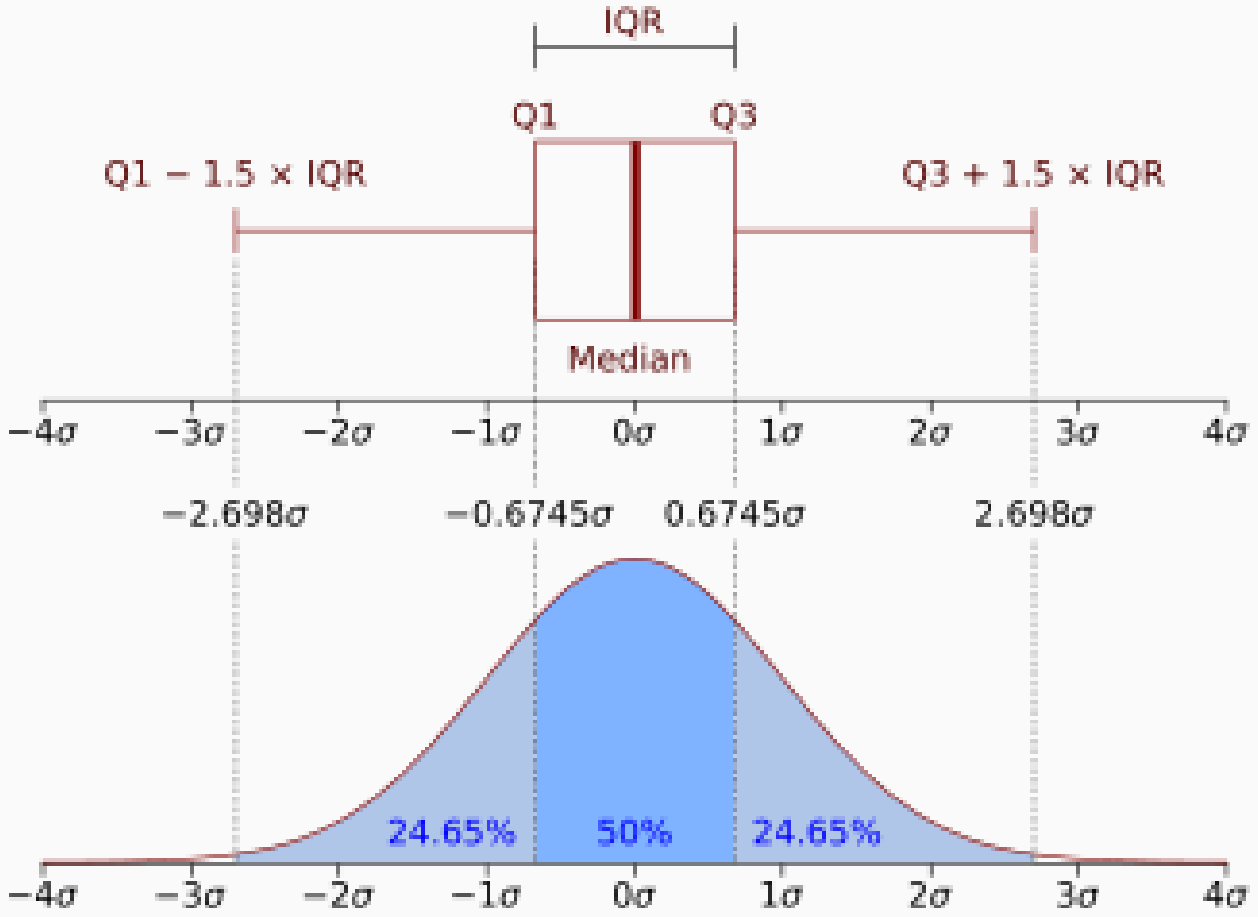
The Normal distribution

- 3 σ rule:
  - 68% of the observed values are at 1 σ from the mean
  - 95% at 2 σ
  - 99.7% at 3 σ

- Empirical rule:
  - values $< 2\sigma$ are "*common*"
  - values $> 2\sigma$ are "*unusual*"
  - values $> 3\sigma$ are "*outliers*"



Outliers


IQR
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%

Exercise #3

? The height of the Italian male population is distributed according to a Normal distribution with mean 170 cm and standard deviation 9.5 cm

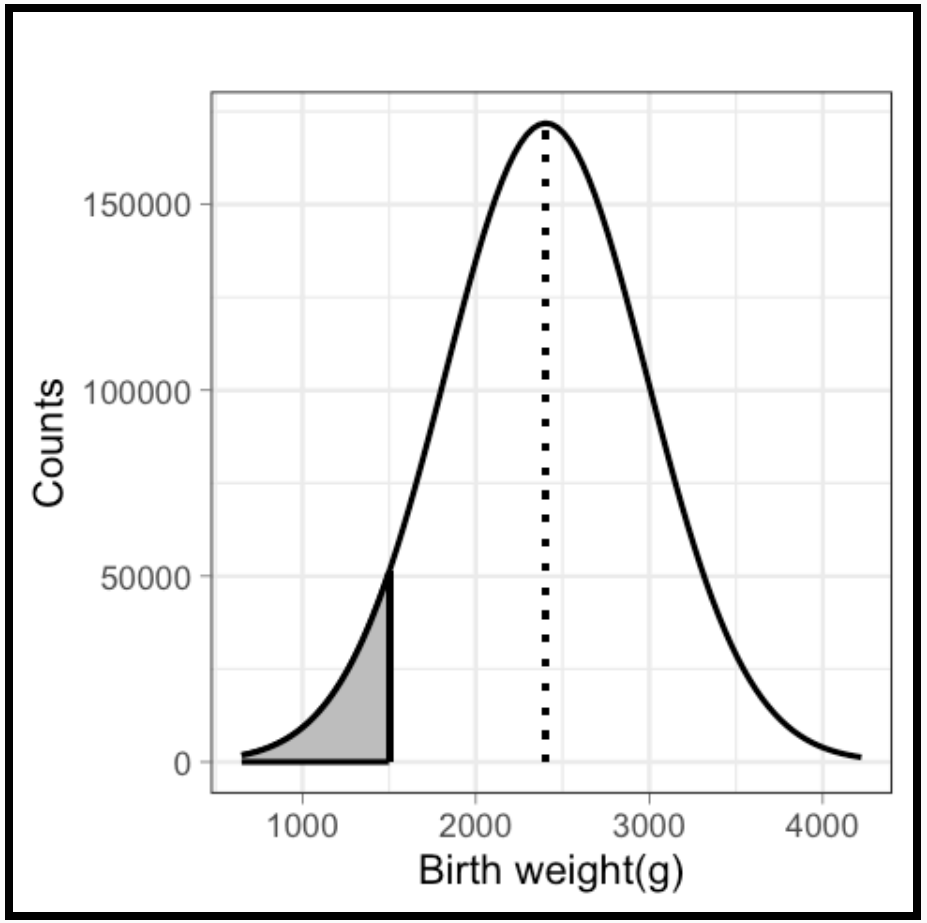
Calculate the following values, when possible

a) The median height
b) The proportion of Italian men taller than $> 170\,\mathrm{cm}$
c) The values considered unusual and/or outliers
d) The most common height
e) The range that includes 68% of the individuals
f) The tallest Italian man's height

Proportion ≡ likelihood

- 6% of the twins have a very low birth weight
- The probability of being very low birth weight is 0.06

How did we get these numbers?

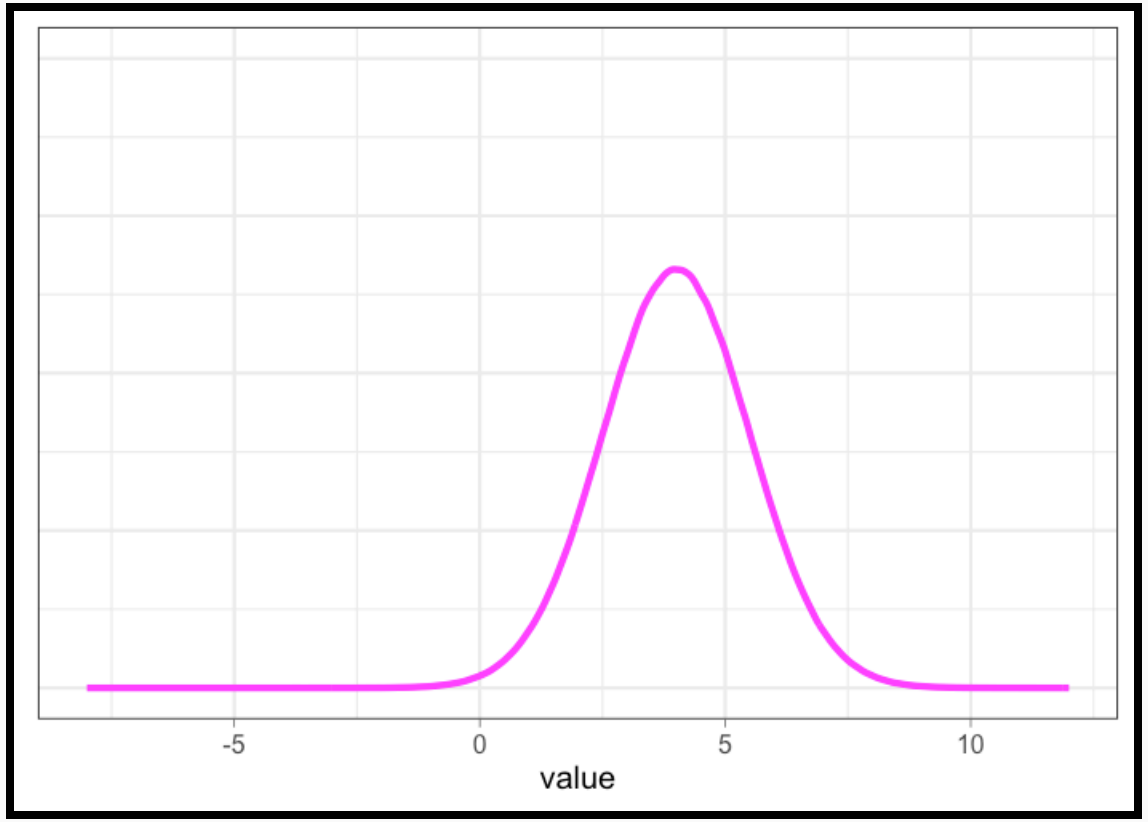


The Standard Normal distribution

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

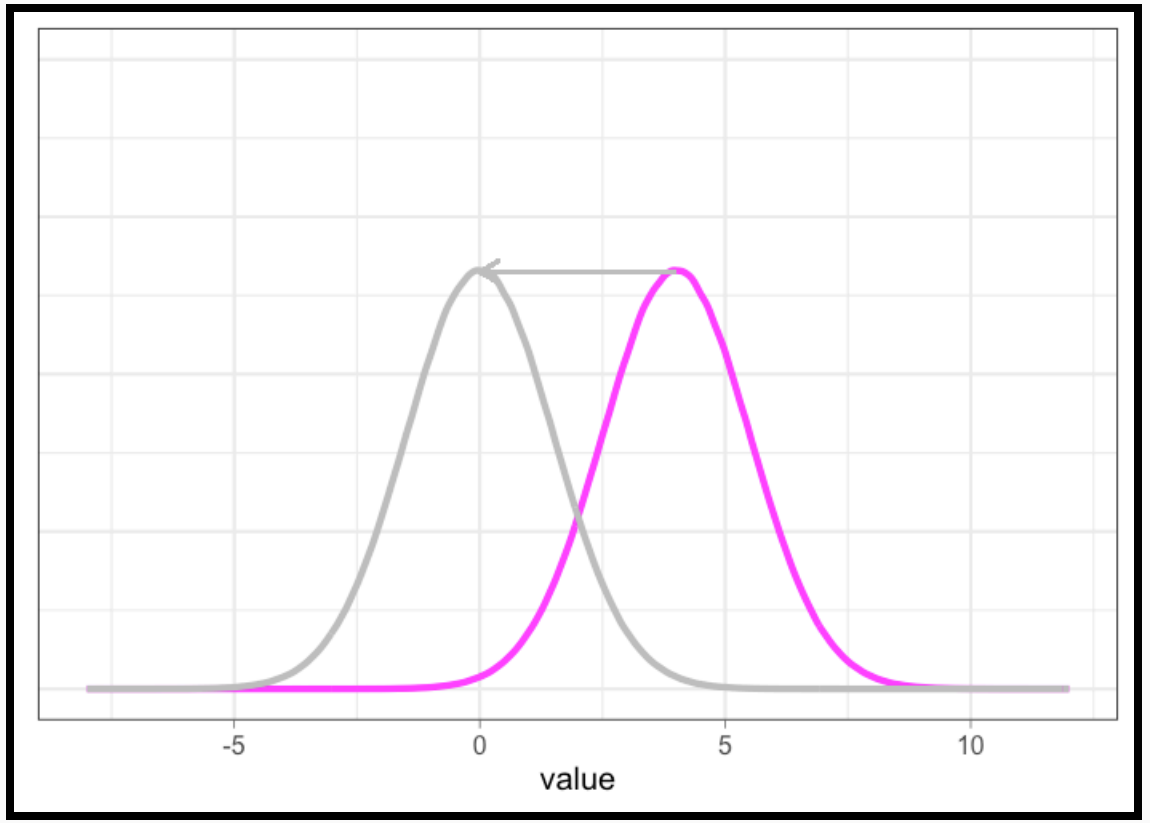
The Standard Normal distribution

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



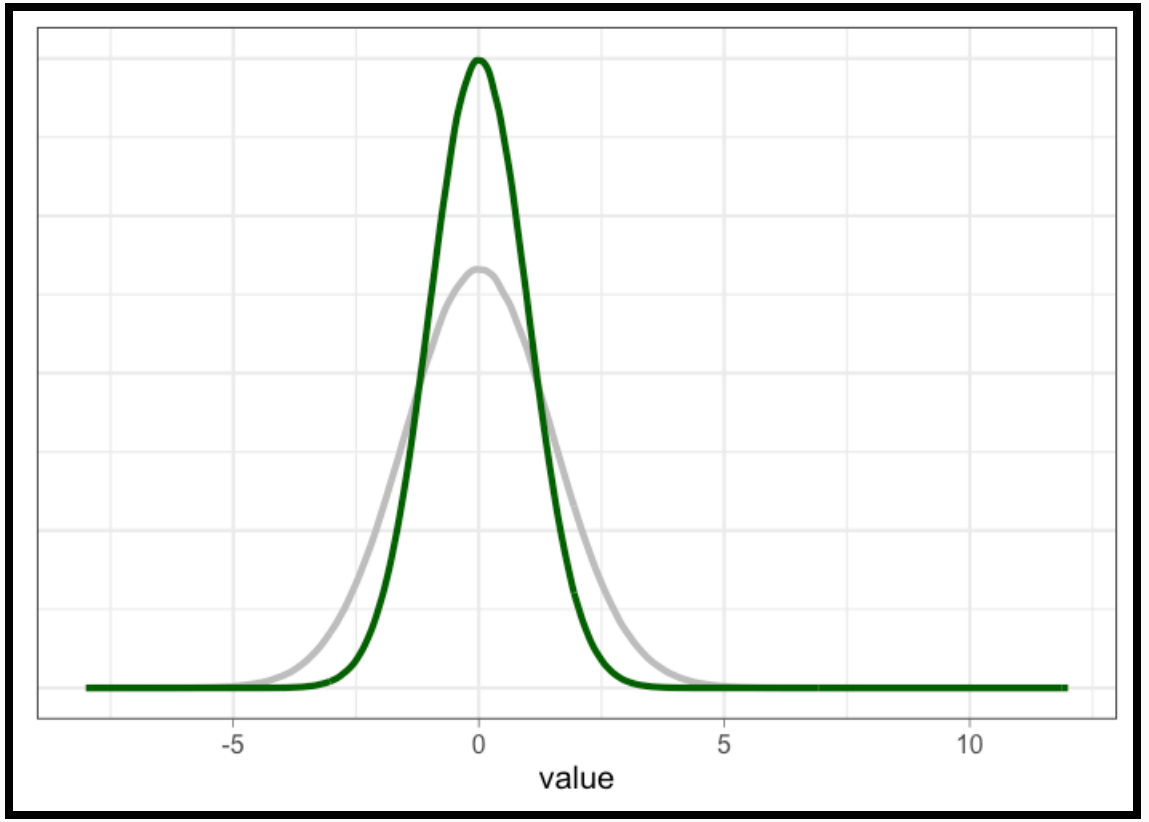
The Standard Normal distribution

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



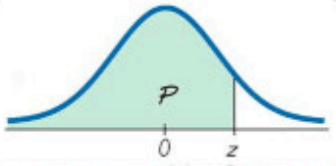
The Standard Normal distribution

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



The Standard Normal distribution

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



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	0.9713	0.9719	0.9726	0.9732	0.9738	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
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

The Standard Normal distribution in practice

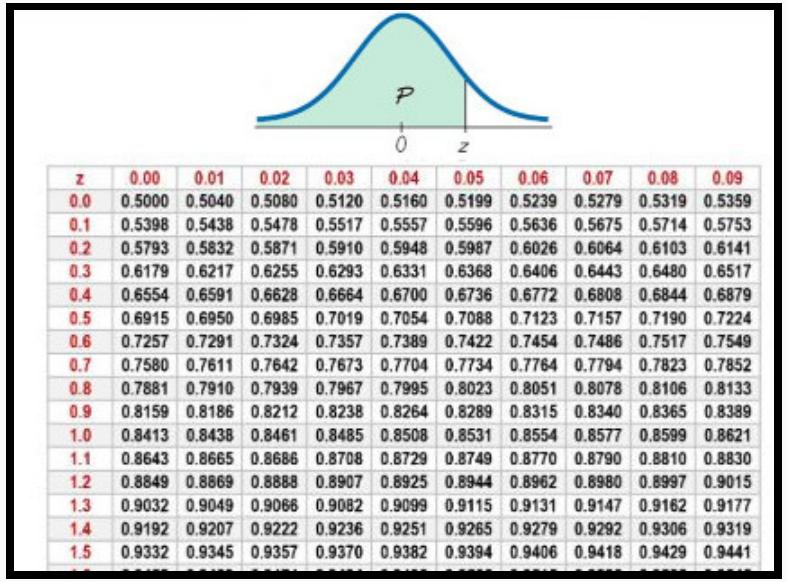
📌 $\mu = 2404\ \mathrm{g};\ \sigma = 580\ \mathrm{g}$

$\mathcal{P}(x < 1500\ \mathrm{g}) =\ ?$

The Standard Normal distribution in practice

📌 $\mu = 2404\ \text{g};\ \sigma = 580\ \text{g}$

$$z = \frac{x-\mu}{\sigma} = \frac{1500\ \text{g} - 2404\ \text{g}}{580\ \text{g}} = -1.56$$



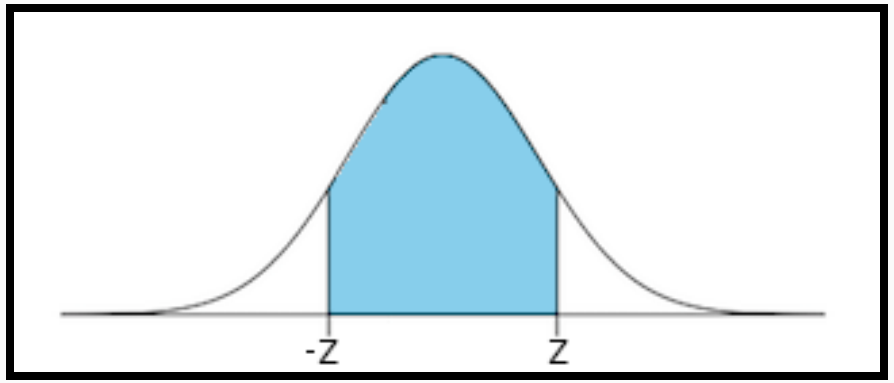
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	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

$$\mathcal{P}(x < 1500\ \text{g}) = \ ?$$

The Standard Normal distribution in practice

📌 $\mu = 2404\ \text{g};\ \sigma = 580\ \text{g}$

$$z = \frac{x-\mu}{\sigma} = \frac{1500\ \text{g} - 2404\ \text{g}}{580\ \text{g}} = -1.56$$

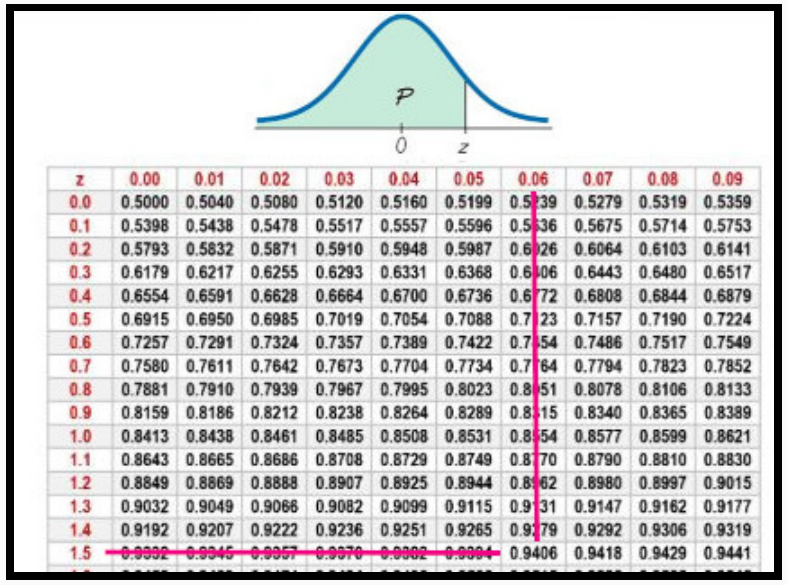


$$\mathcal{P}(x < 1500\ \text{g}) = \ ?$$

The Standard Normal distribution in practice

📌 $\mu = 2404\ \text{g};\ \sigma = 580\ \text{g}$

$$z = \frac{x-\mu}{\sigma} = \frac{1500\ \text{g} - 2404\ \text{g}}{580\ \text{g}} = -1.56$$



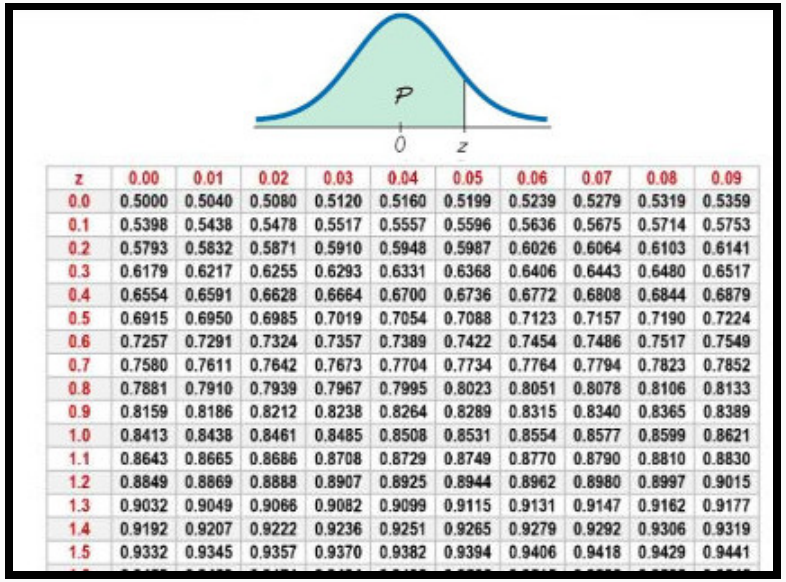
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	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	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.9406	0.9418	0.9429	0.9441

$$\mathcal{P}(x < 1500\ \text{g}) = 1 - 0.9406 = 0.0594 \rightarrow 5.94\%$$

Exercise #4

? Not knowing that the baby has a twin, the pediatrician tells to the mother that a birth weight lower than 2500 g is unusual. Should the mother be worried?

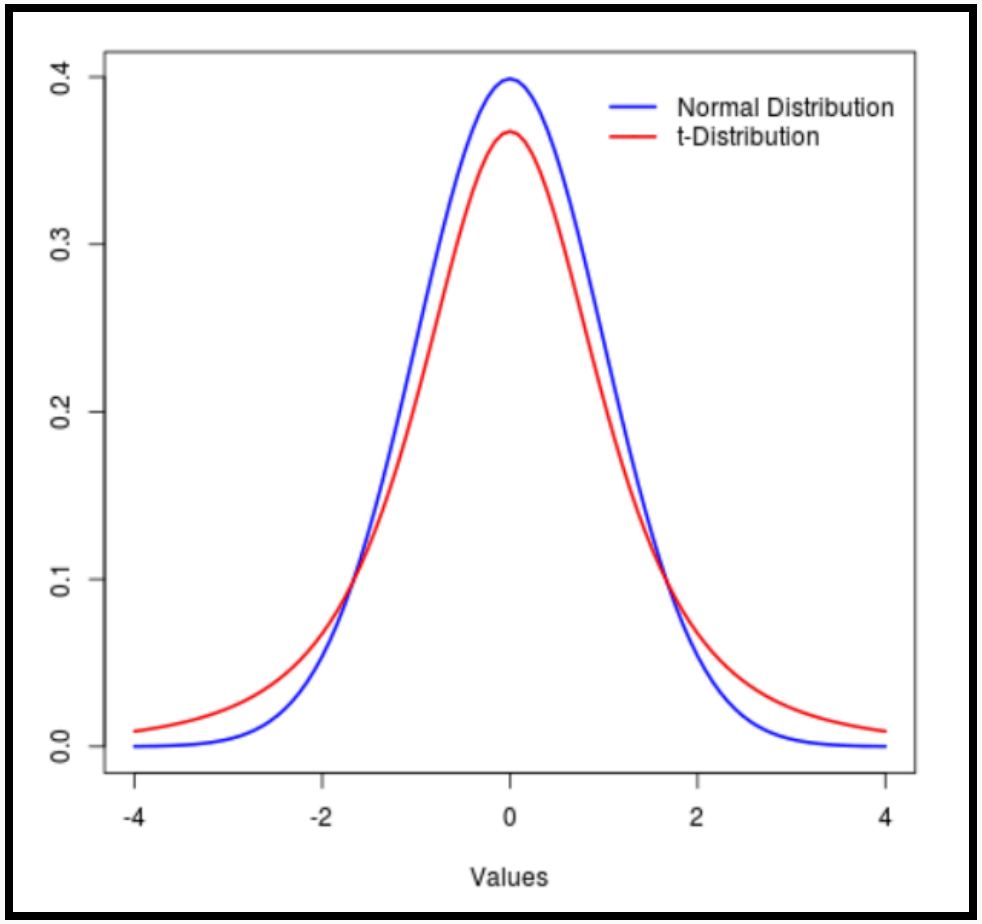
$\mu = 2404$ g; $\sigma = 580$ g



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

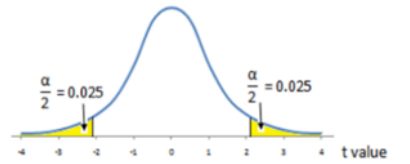
The Student's t distribution

- When observations are too little to be approximated to a Normal distribution



The Student's t distribution

- When observations are too little to be approximated to a Normal distribution
- Student's t distribution
  - keeps into account the degree of freedom (df)
  - one sample of size $n \rightarrow df = n - 1$

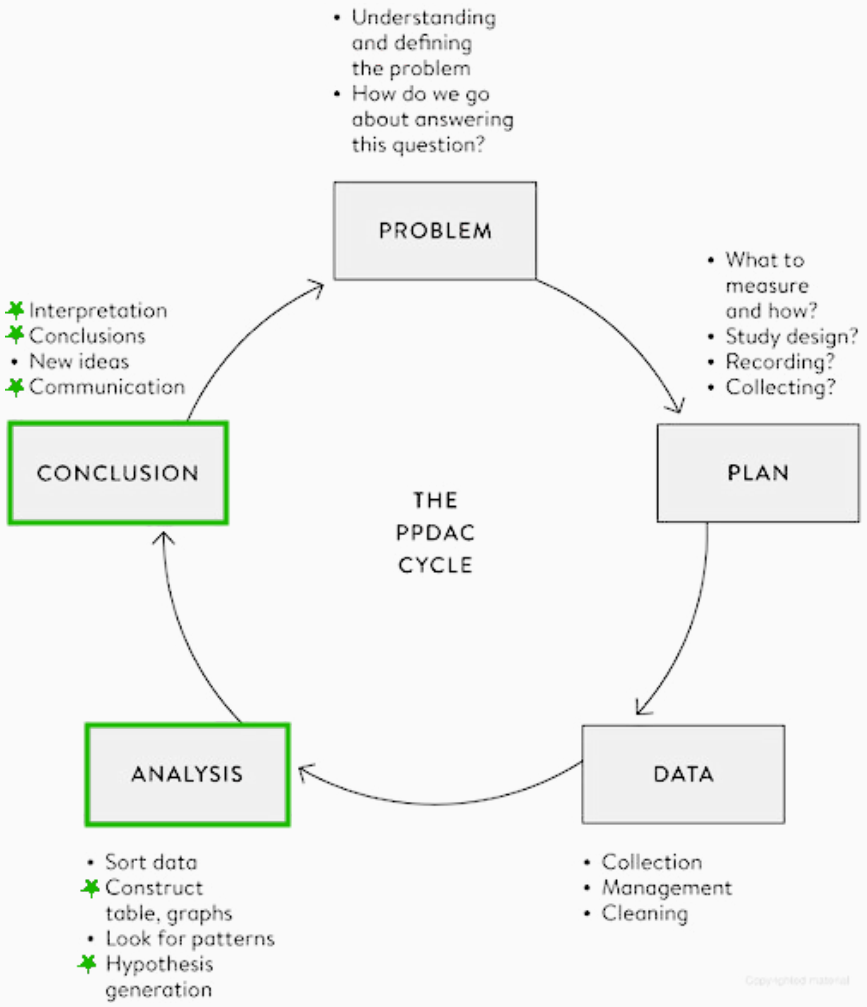


Confidence	90	95	95.45	99	99.73
alpha	0.1000	0.0500	0.0455	0.0100	0.0027
df					
1	6.314	12.706	13.968	63.657	235.784
2	2.920	4.303	4.527	9.925	19.206
3	2.353	3.182	3.307	5.841	9.219
4	2.132	2.776	2.869	4.604	6.620
5	2.015	2.571	2.649	4.032	5.507
6	1.943	2.447	2.517	3.707	4.904
7	1.895	2.365	2.429	3.499	4.530
8	1.860	2.306	2.366	3.355	4.277
9	1.833	2.262	2.320	3.250	4.094
10	1.812	2.228	2.284	3.169	3.957
11	1.796	2.201	2.255	3.106	3.850
12	1.782	2.179	2.231	3.055	3.764
13	1.771	2.160	2.212	3.012	3.694
14	1.761	2.145	2.195	2.977	3.636
15	1.753	2.131	2.181	2.947	3.586
16	1.746	2.120	2.169	2.921	3.544
17	1.740	2.110	2.158	2.898	3.507
18	1.734	2.101	2.149	2.878	3.475
19	1.729	2.093	2.140	2.861	3.447
20	1.725	2.086	2.133	2.845	3.422
21	1.721	2.080	2.126	2.831	3.400
22	1.717	2.074	2.120	2.819	3.380
23	1.714	2.069	2.115	2.807	3.361
24	1.711	2.064	2.110	2.797	3.345
25	1.708	2.060	2.105	2.787	3.330
26	1.706	2.056	2.101	2.779	3.316
27	1.703	2.052	2.097	2.771	3.303
28	1.701	2.048	2.093	2.763	3.291
29	1.699	2.045	2.090	2.756	3.280
30	1.697	2.042	2.087	2.750	3.270
40	1.684	2.021	2.064	2.704	3.199
50	1.676	2.009	2.051	2.678	3.157
60	1.671	2.000	2.043	2.660	3.130
70	1.667	1.994	2.036	2.648	3.111
80	1.664	1.990	2.032	2.639	3.096
90	1.662	1.987	2.028	2.632	3.085
100	1.660	1.984	2.025	2.626	3.077
1000	1.646	1.962	2.003	2.581	3.007
∞	1.645	1.960	2.000	2.576	3.000

Estimates and Confidence Intervals

Learning objectives

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



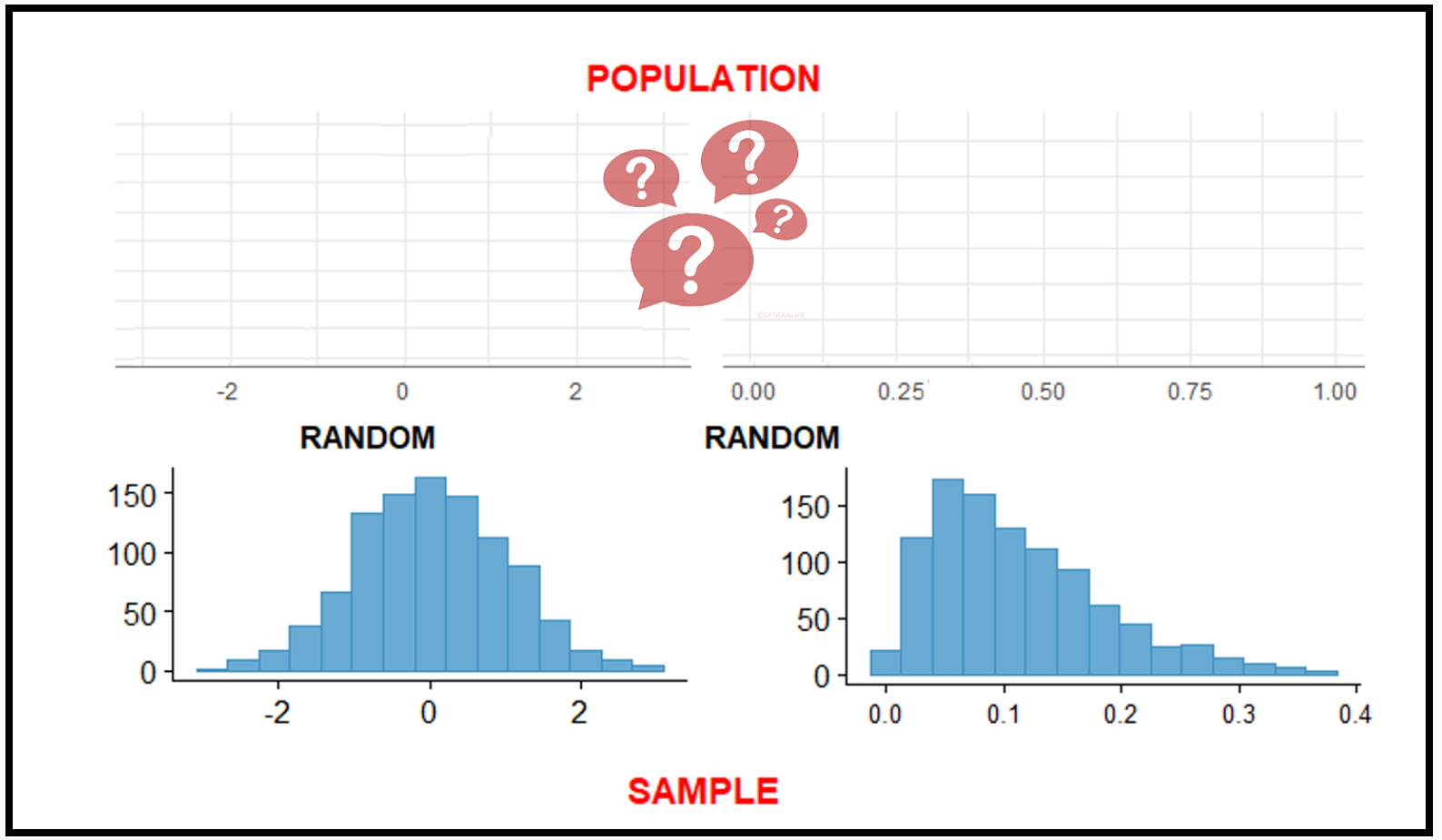
Spiegelhalter, D., *The Art of Statistics: Learning From Data*, Pelican, 2019

⚠️ Disclaimer ⚠️

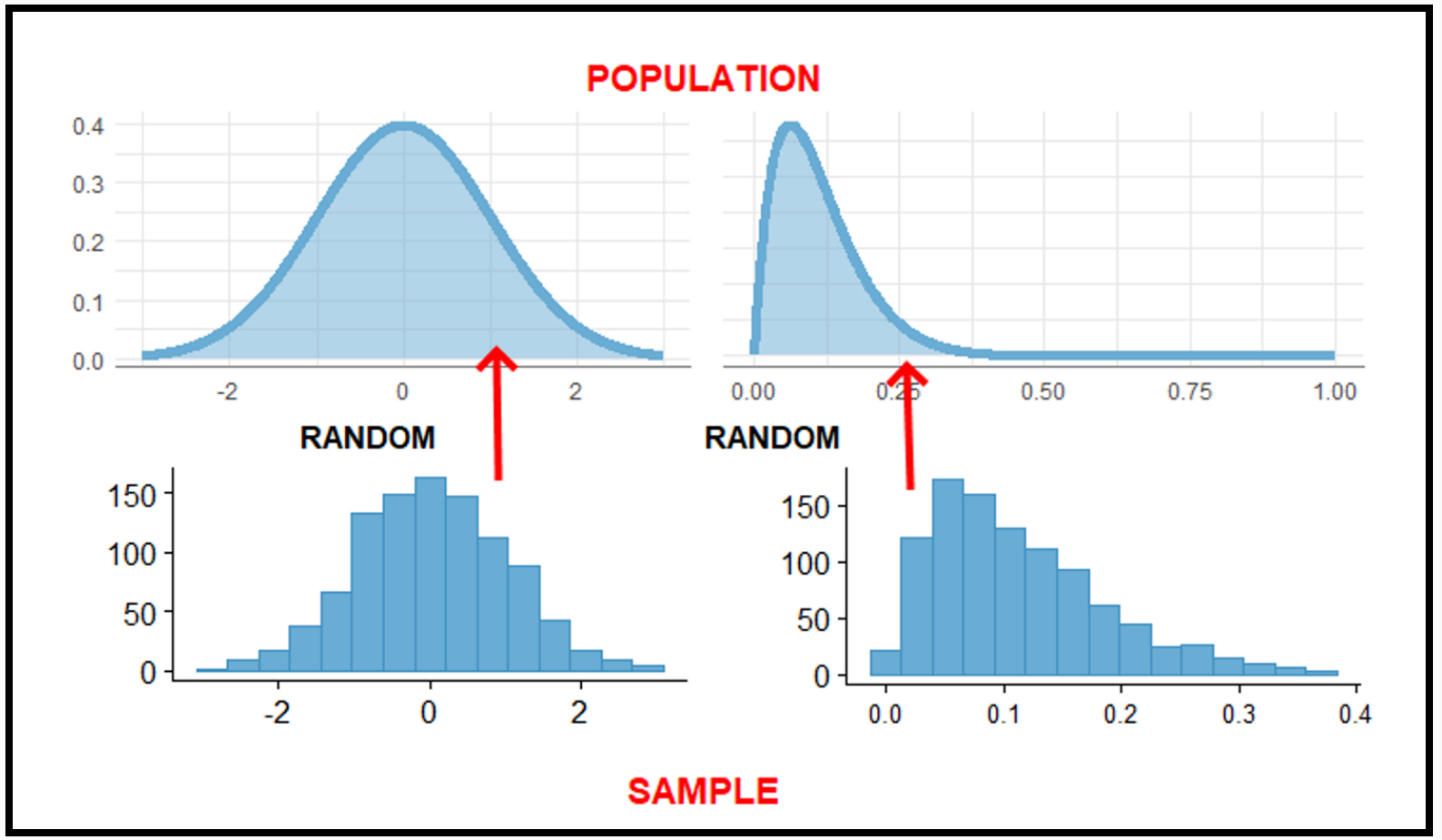
If this part seems difficult to you, it's because it's really difficult.

You may have to spend quite a bit of time before understanding it completely. Don't worry, we've all been there!

From sample to population

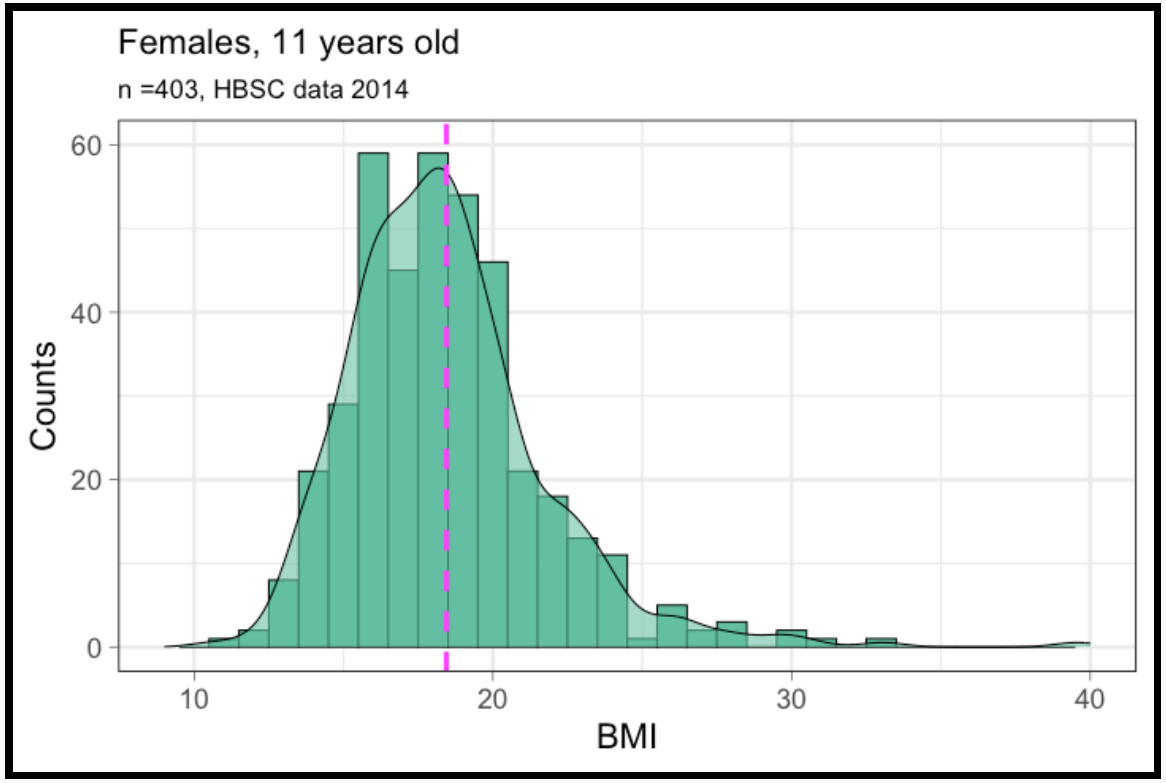


From sample to population



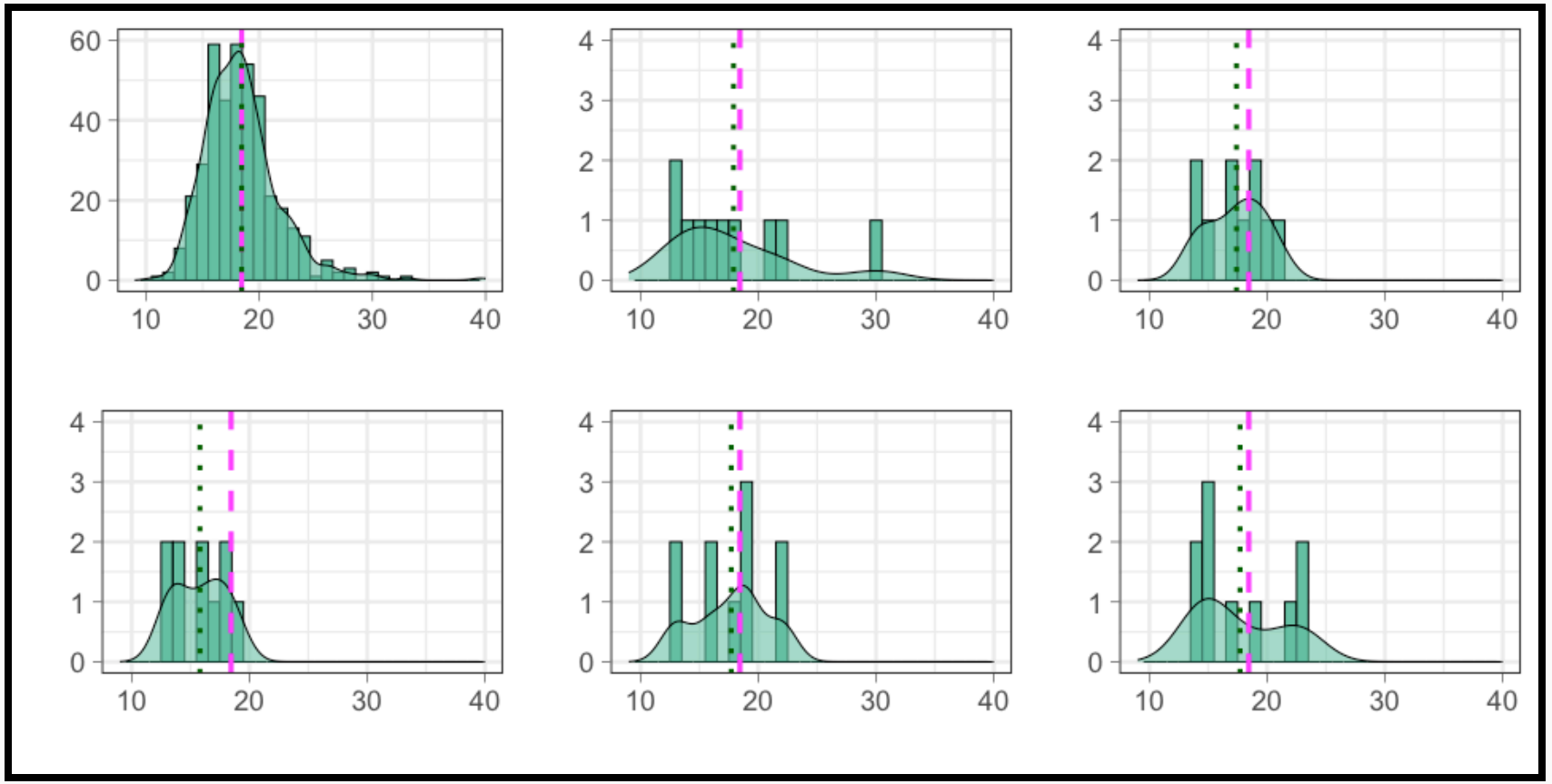
How accurate are we?

The mean BMI for Italian 11 years old girls is $18.4 \pm 3.3\ \mathrm{kg/m^2}$



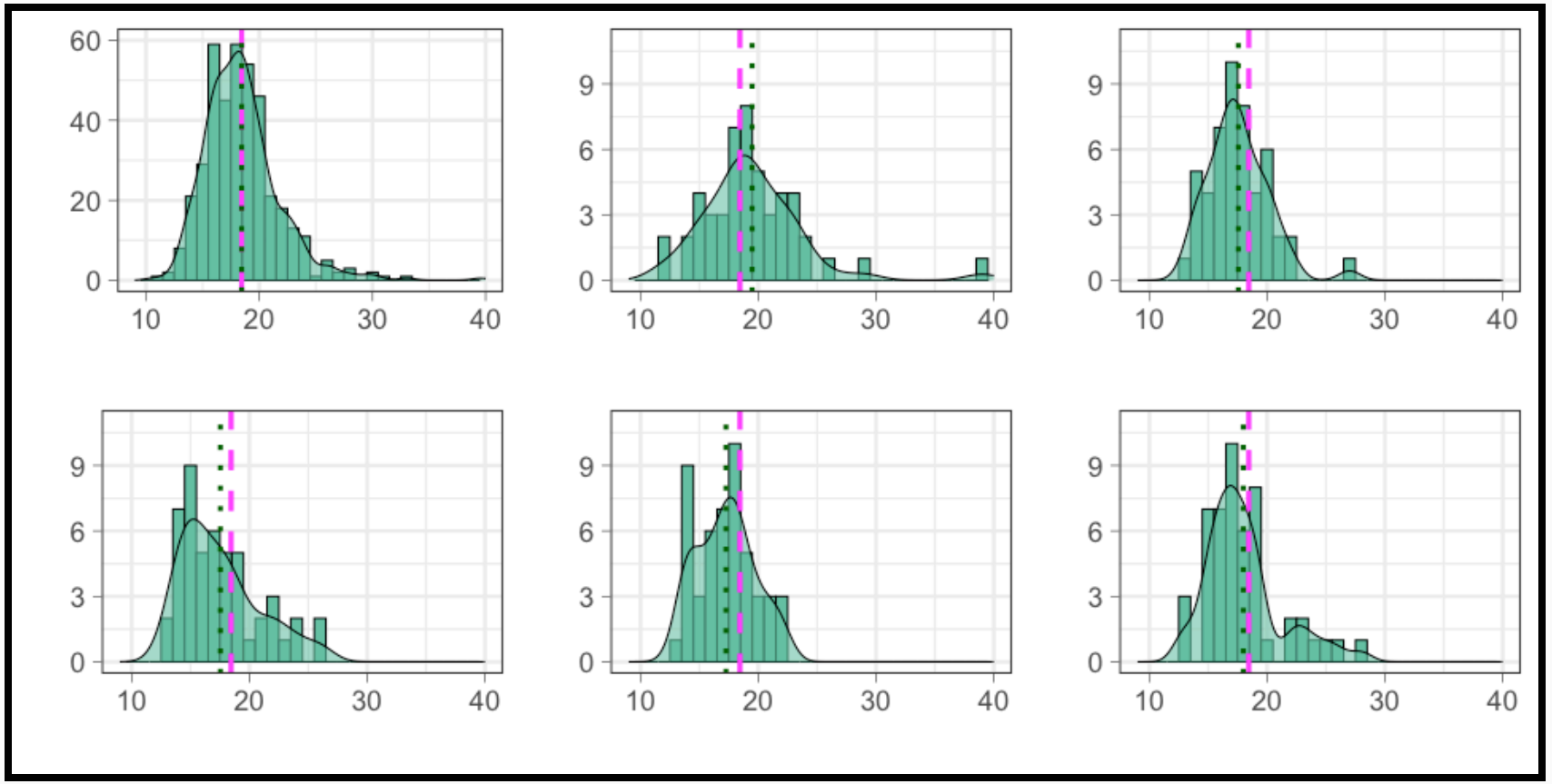
Sample size

$n = 10$



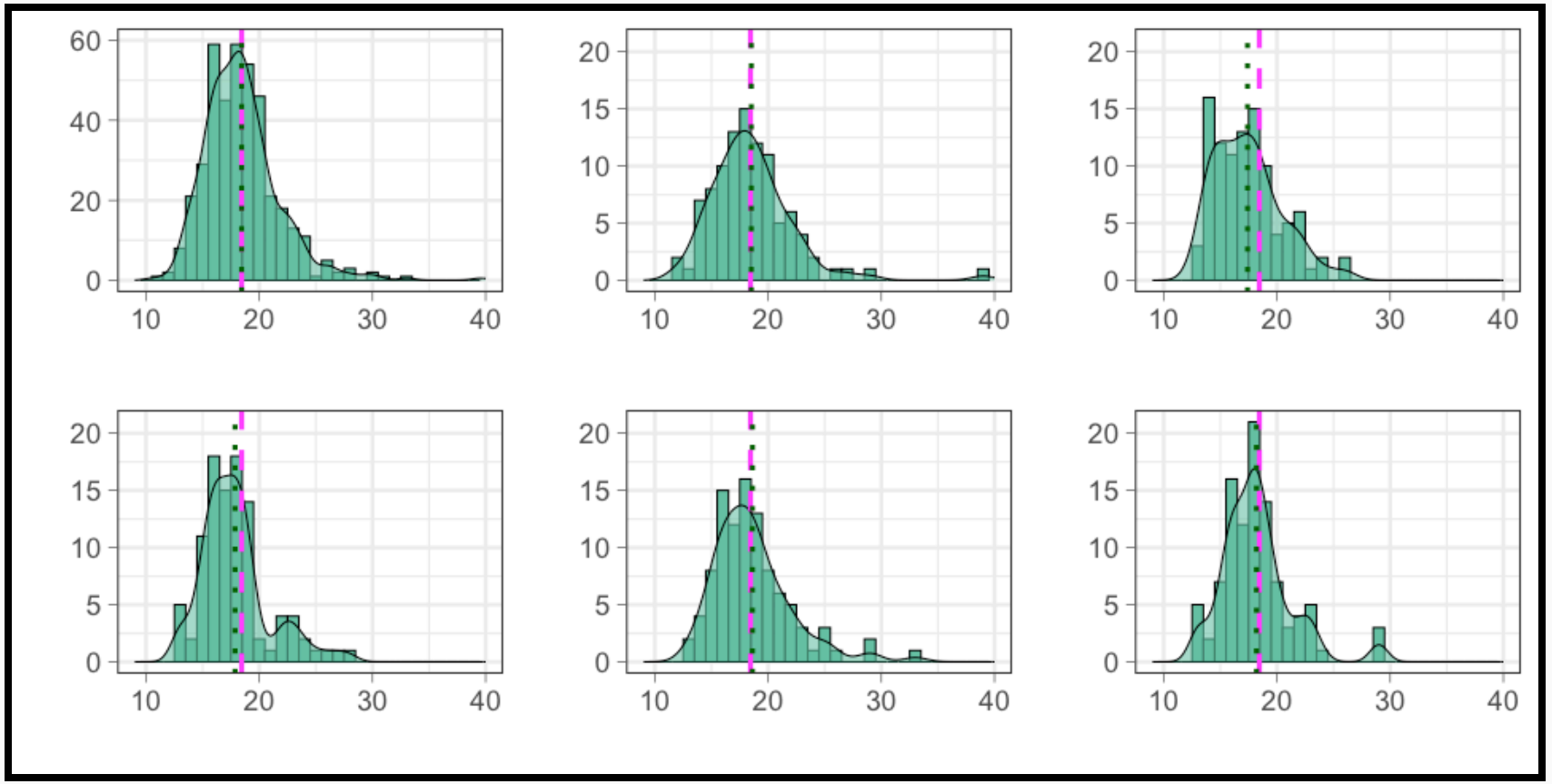
Sample size

$n = 50$



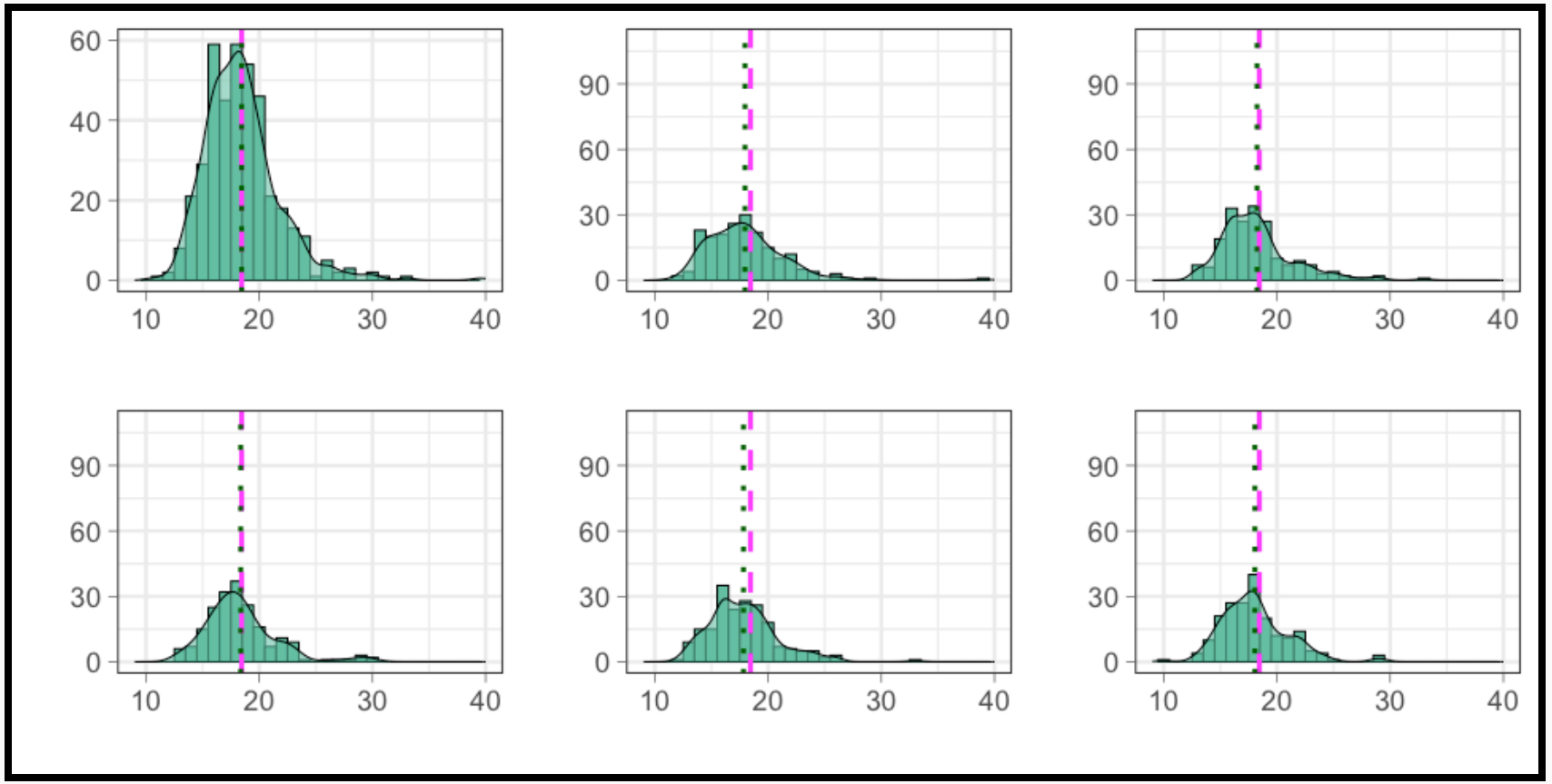
Sample size

$n = 100$



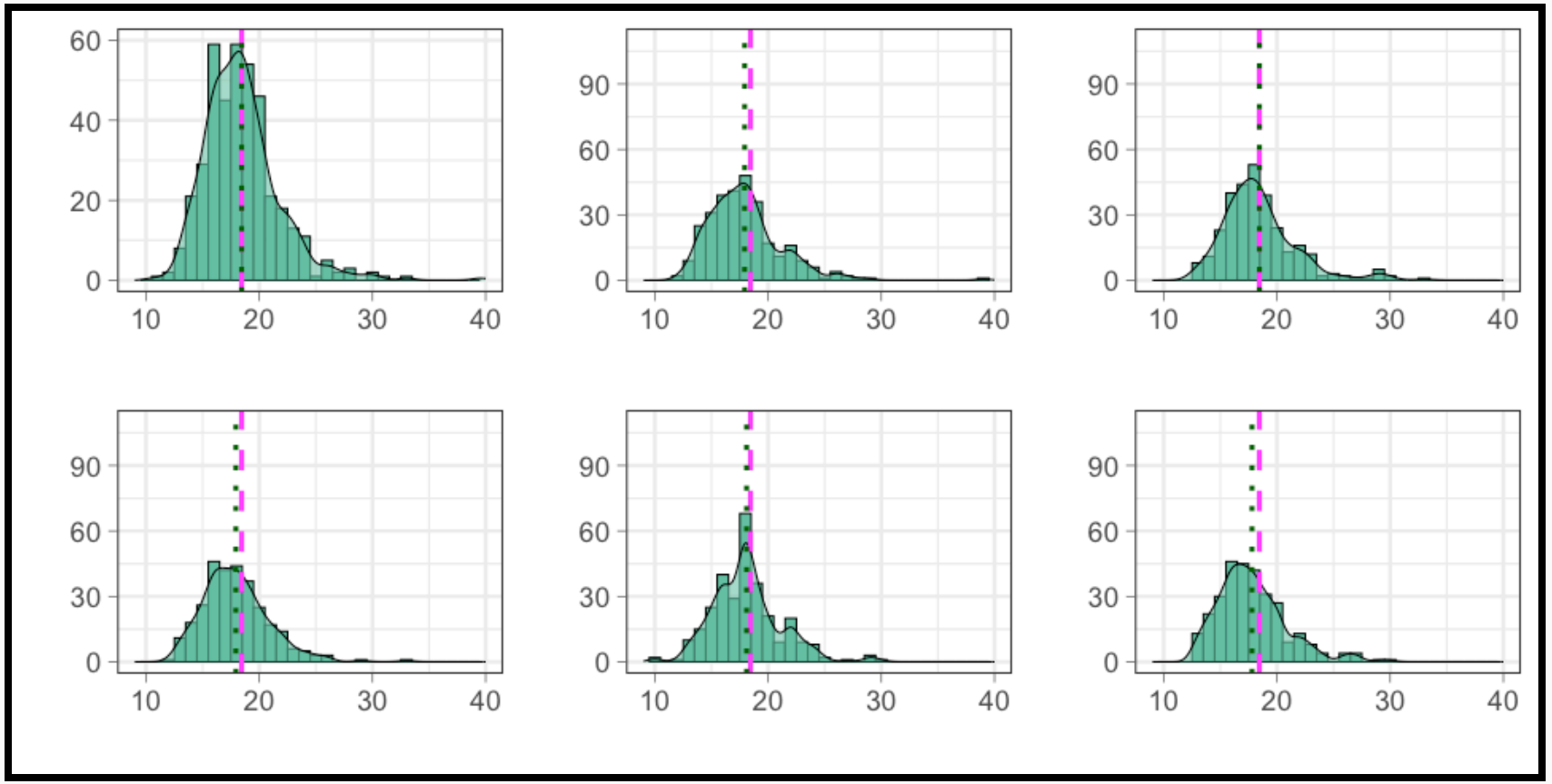
Sample size

$n = 200$



Sample size

$n = 300$



Exercise #5

? When the sample size increases, an estimate of a parameter

a) improves
b) worsens
c) is sensitive to single data-points
d) there is no difference

How accurate are we?

With this example, we introduced two ideas:

1. Larger samples improve population parameter estimation
2. If we keep extracting (sub)samples, we get a feeling for the variation (aka confidence interval) around the "plausible" value of the population parameter

and so, now?

Estimates & confidence intervals

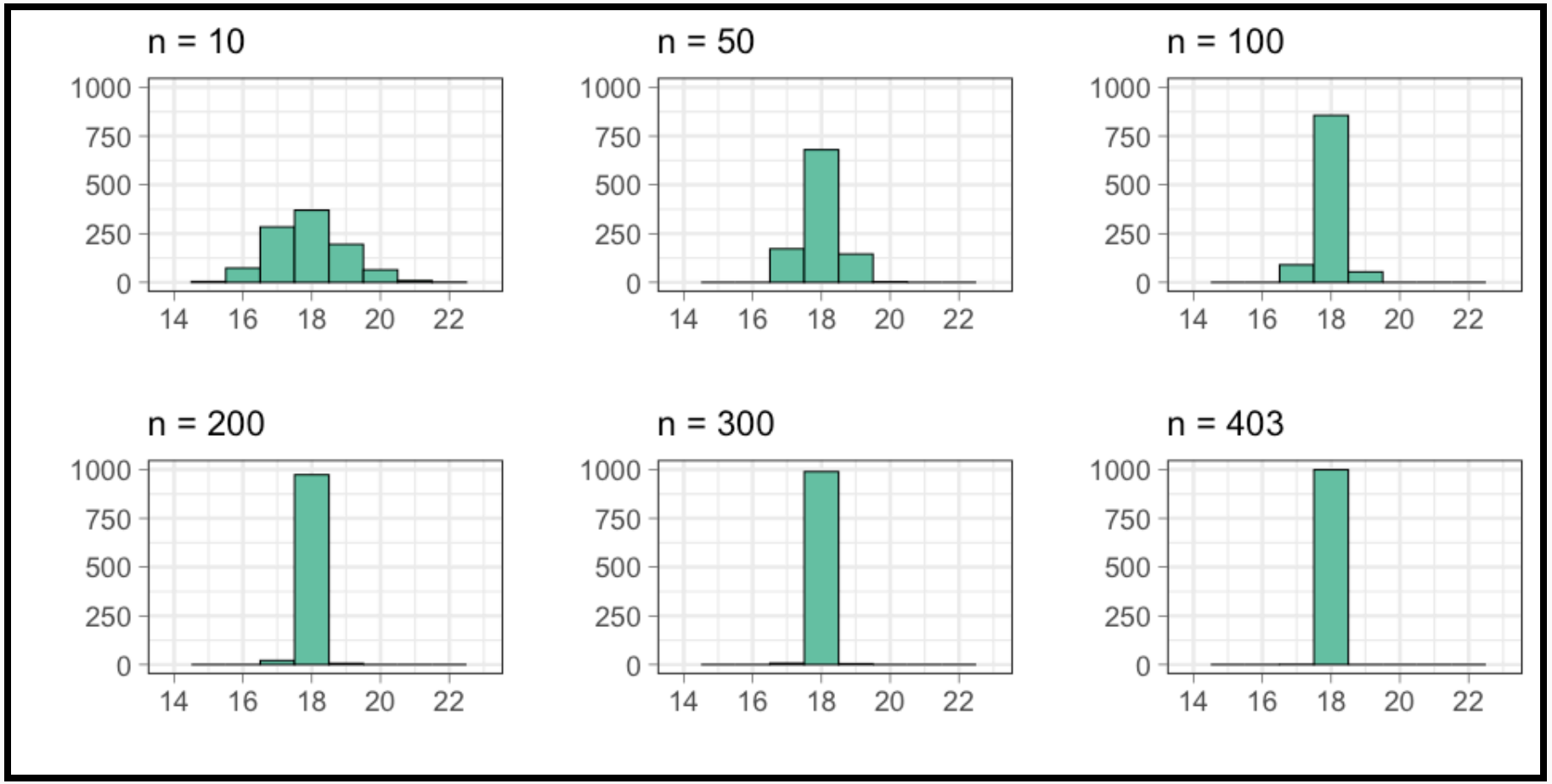
How does one estimate the variation around the true population parameter, if what they are looking for is the true population parameter?

Estimates & confidence intervals

1. Assuming that the population is similar to the sample
   $\rightarrow$ *via* bootstrapping
2. Making mathematical assumptions about the shape of the population distribution
   $\rightarrow$ *via* central limit theorem

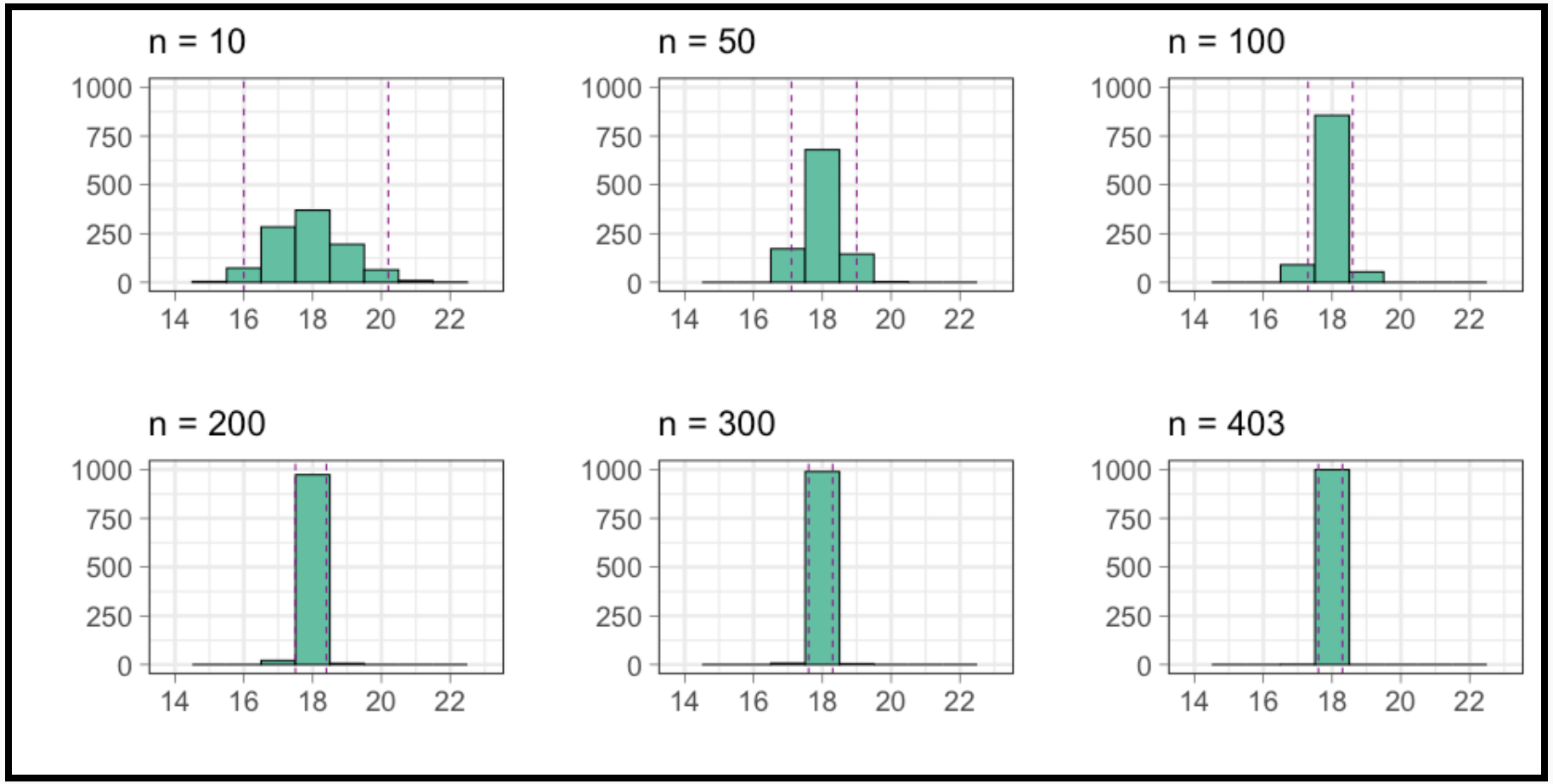
Estimates & confidence intervals

$$N_{\text{Bootstrapping}} = 1000$$



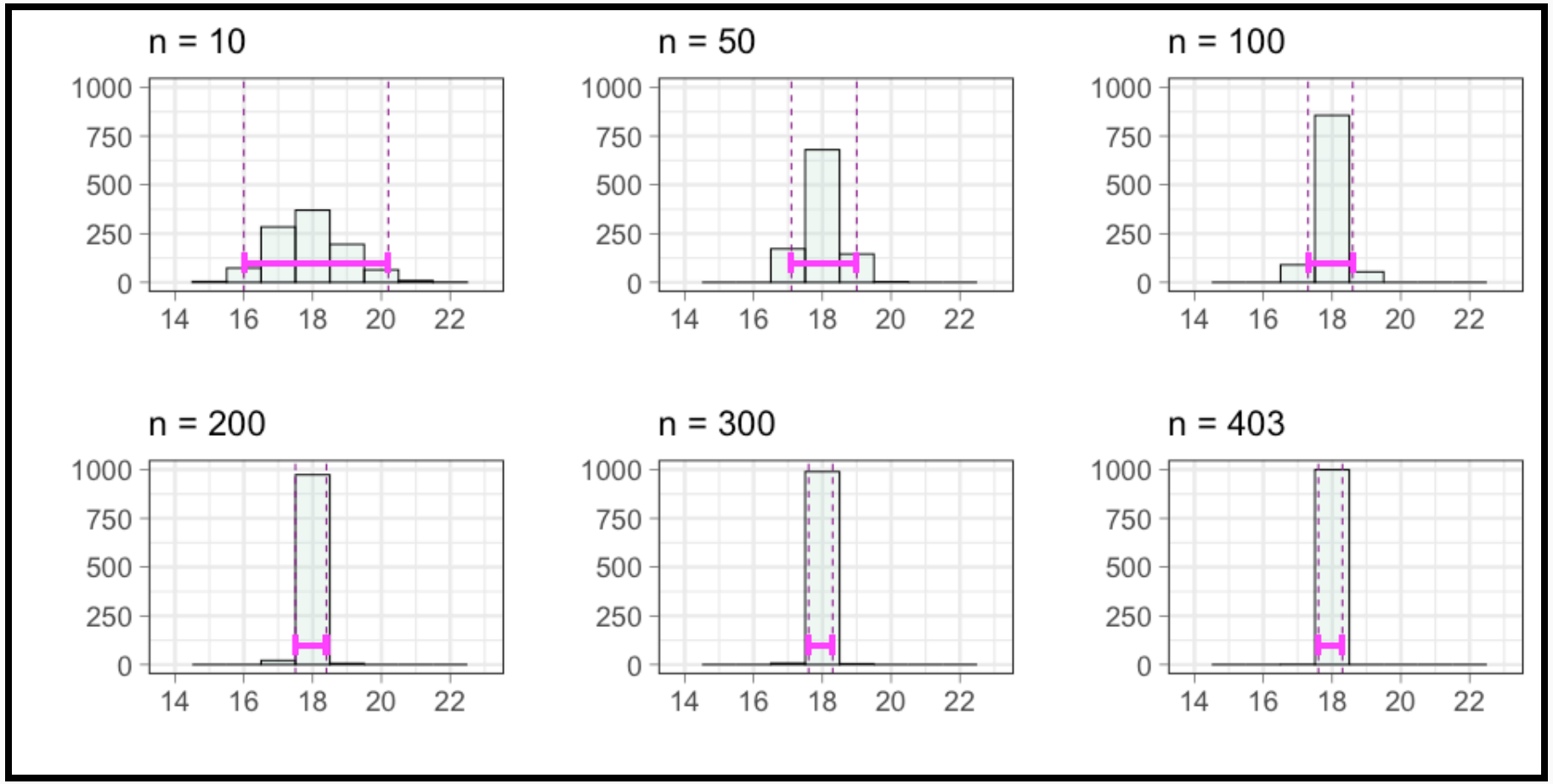
Estimates & confidence intervals

Interval including 95% of the means obtained *via* bootstrapping



Estimates & confidence intervals

Interval including 95% of the means obtained *via* bootstrapping



Exercise #6

? When the sample size increases, the accuracy of a parameter estimate...

a) improves
b) worsens
c) there is no difference

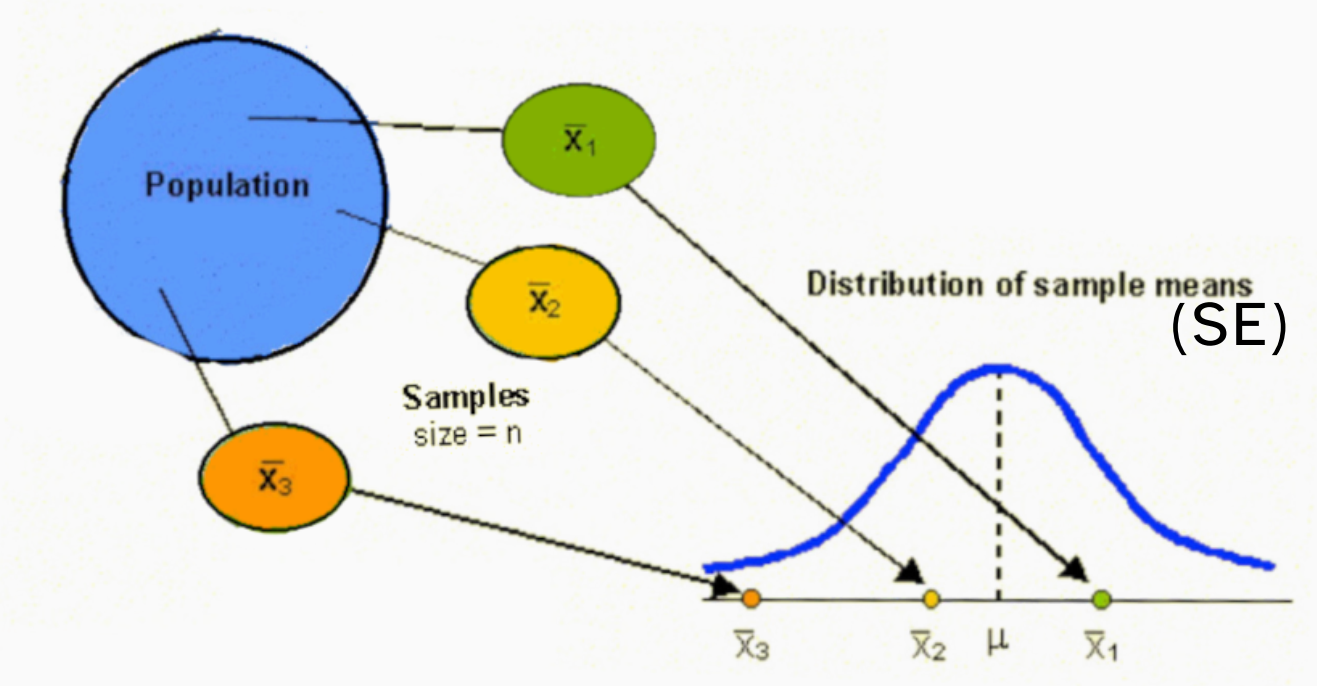
Let's stop for a minute

We introduced two difficult and important concepts:

1. there is a variability in the estimate of a parameter that depends on the sample
2. the shape of the sampling distribution doesn't depend on the shape of the empirical distribution, and can be approximate to a Normal distribution for large samples

We now have all the elements for facing the second approach that can be used to calculate parameters estimates and confidence interval

The sampling distribution & the central limit theorem



(SE)

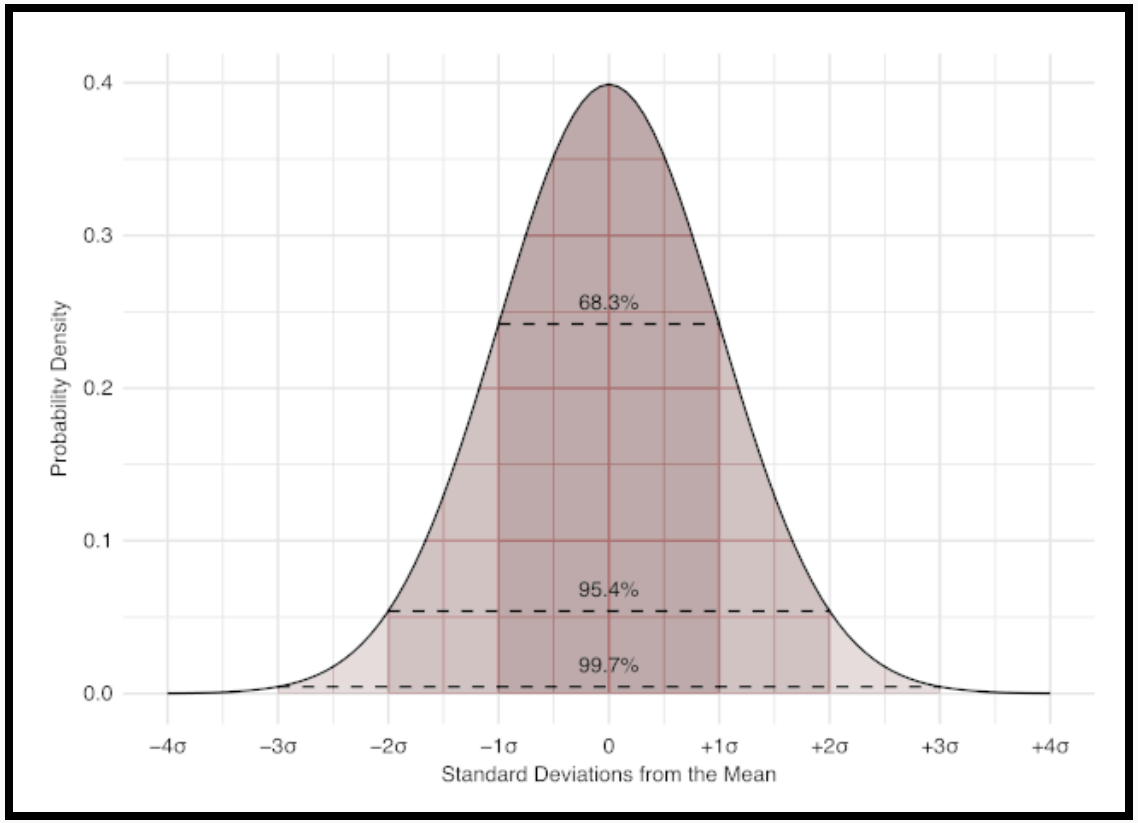
$\mathcal{N} = (\mu, \frac{\sigma^2}{n})$ with

$\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$

$\rightarrow$ standard error

Let's put the pieces together...

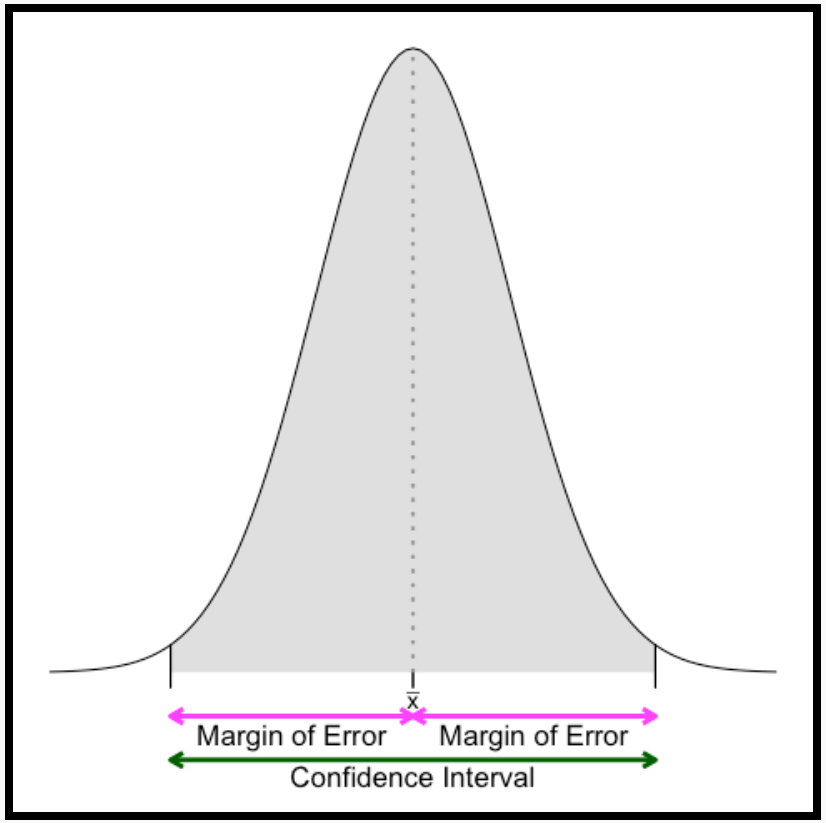
- the sampling distribution is a Normal distribution
- in a Normal distrubution, 95% of the data are at (about) $2 \times \mathrm{SD}$ from the mean
- a 95% confidence interval is at (about) $2 \times \mathrm{SE}$ from the sampling distribution mean



Calculating confidence intervals

🎯 A 95% CI is at (about) $2 \times \mathrm{SE}$ from the sampling distribution mean ($\bar{x}$)

1. Calculate SE
2. Calculate $2 \times \mathrm{SE}$, *i.e.*, 95% Margin of Error (ME)
3. Calculate the 95% CI as $(\bar{x} - \mathrm{ME}\,;\, \bar{x} + \mathrm{ME})$



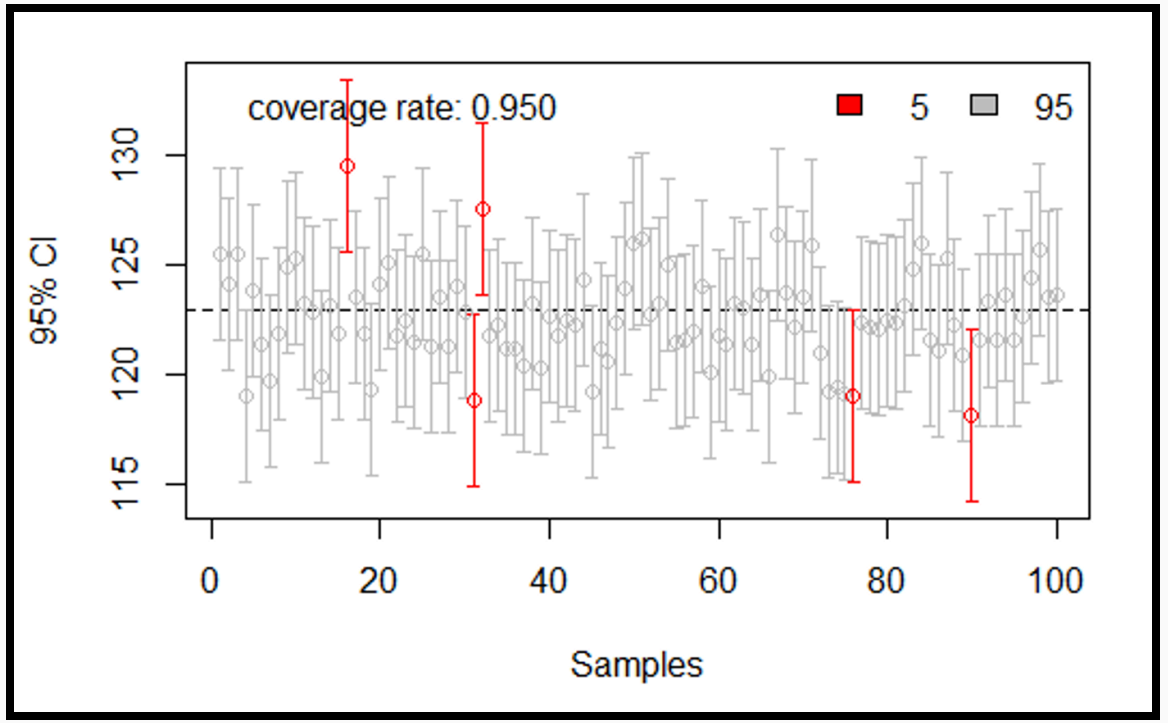
Interpreting confidence intervals

A confidence interval is a range of values which includes the estimated parameter with a given degree of confidence

Interpreting confidence intervals

If we could sample the population 100 times, 95 would estimate a confidence interval which includes the true population parameter

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



Exercise #7

? In the Results section, the authors reported the following

In Mexican Americans, the mean age at menarche was 12.09 years (95% CI = 11.81 to 12.37 years)

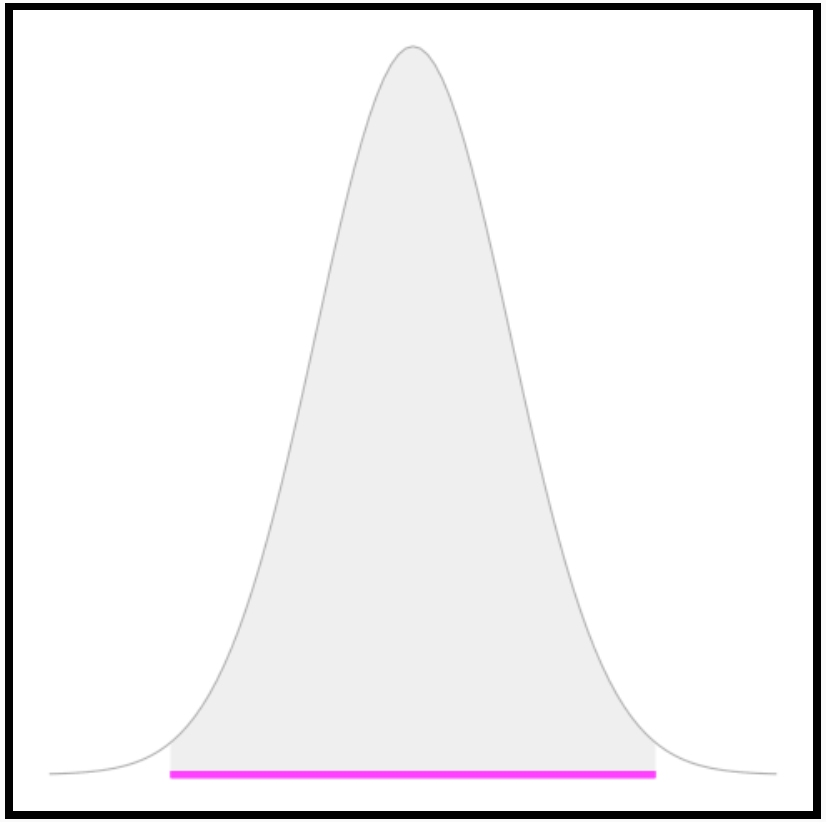
this means that...

a) the age at menarche for Mexican American girls is included between 11.81 and 12.37 years old
b) 95% of Mexican American girls experience menarche between 11.81 and 12.37 years old
c) the mean age at menarche for Mexican American girls has a 95% probability of being included between 11.81 and 12.37 years old
d) none of the above

Exercise #8

? If the CI is large we are...

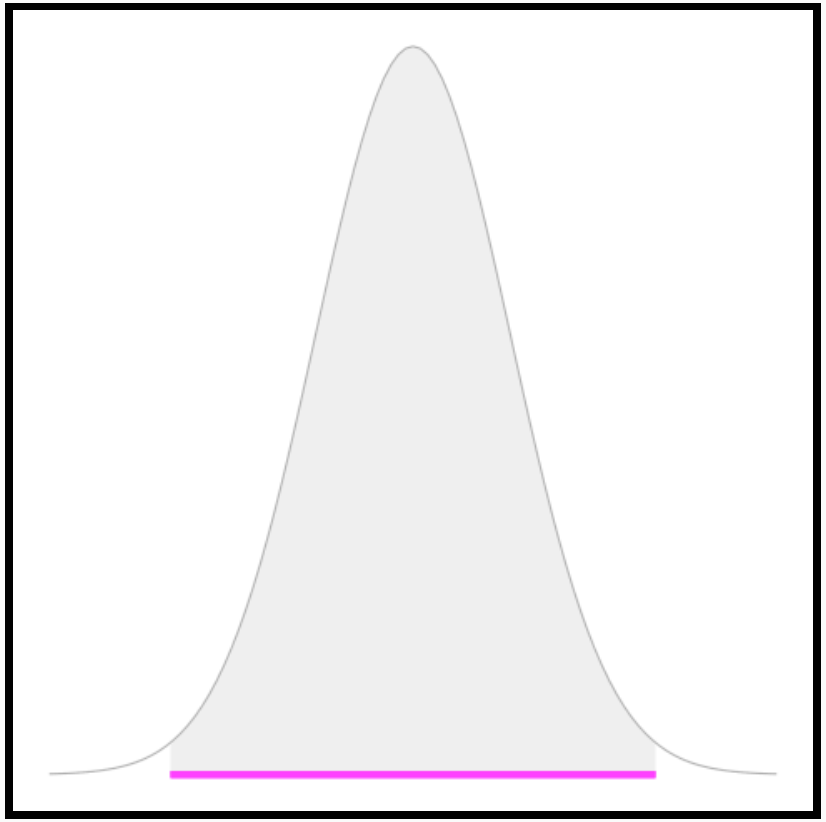
a) more likely of including μ
b) less likely of including μ
c) there is no difference



Exercise #9

? If the CI is large we are...

a) more precise
b) less precise
c) there is no difference



Exercise #10

The mean BMI for Italian 11 years old girls ($n = 403$) is $18.4 \pm 3.3\ \mathrm{kg/m^2}$

? Calculate the 95% CI for the true mean μ

$$\mathrm{SE} = \sigma/\sqrt{n} = \ ? \rightarrow \quad \hat{\mathrm{SE}} = s/\sqrt{n} = \ ?$$

Exercise #11

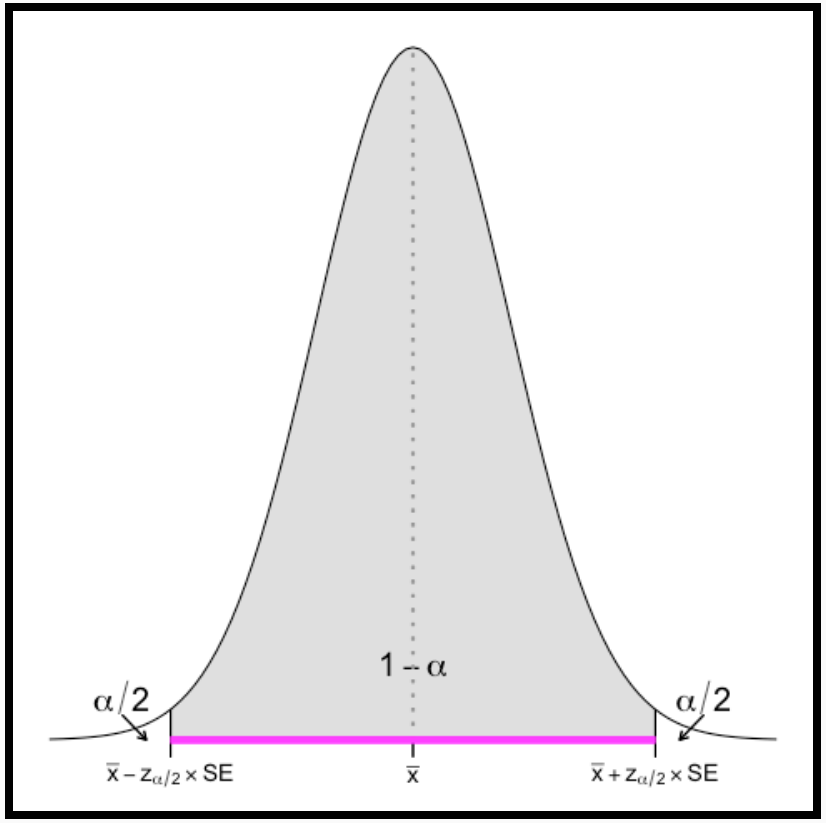
? Given that $\mathcal{N} = (\mu, \frac{\sigma^2}{n})$ with $\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} \rightarrow$ standard error (SE), how can one reduce the confidence interval?

a) increasing n
b) decreasing n
c) increasing σ
d) decreasing σ
e) none of the above

The α level

🎯 $$95\%\ \text{CI} = (\bar{x}- \approx 2 \times \hat{SE}\ ;\ \bar{x}+ \approx 2 \times \hat{SE})$$
$$\approx 2 \quad ?$$

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

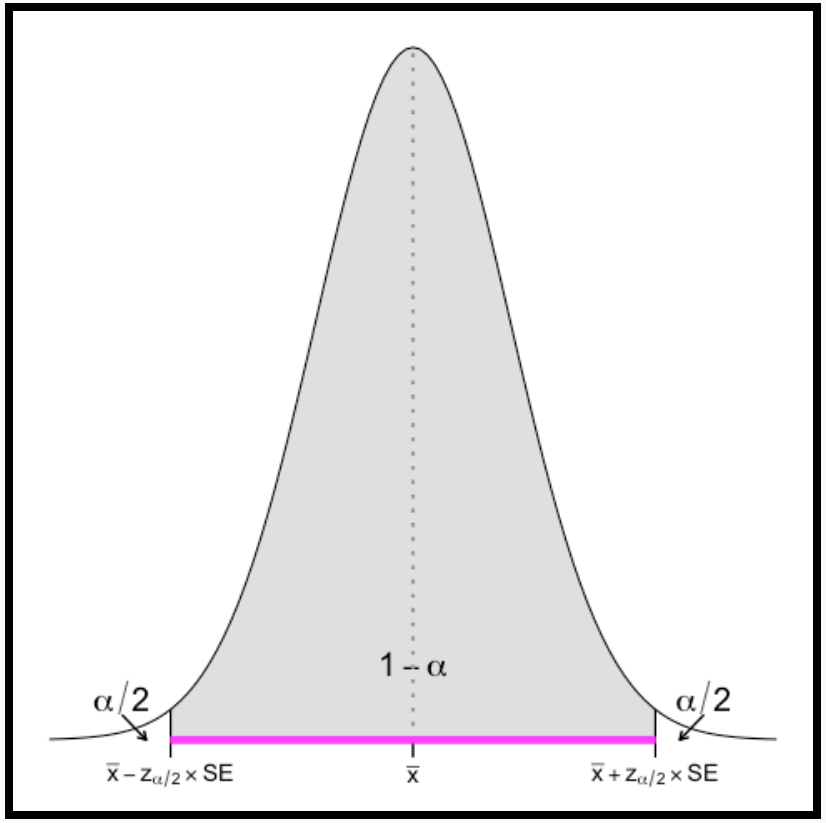


The α level

🎯 $$95\%\ \text{CI} = (\bar{x}- \approx 2 \times \hat{SE}\ ;\ \bar{x}+ \approx 2 \times \hat{SE})$$
$$\approx 2 \quad ?$$

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

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

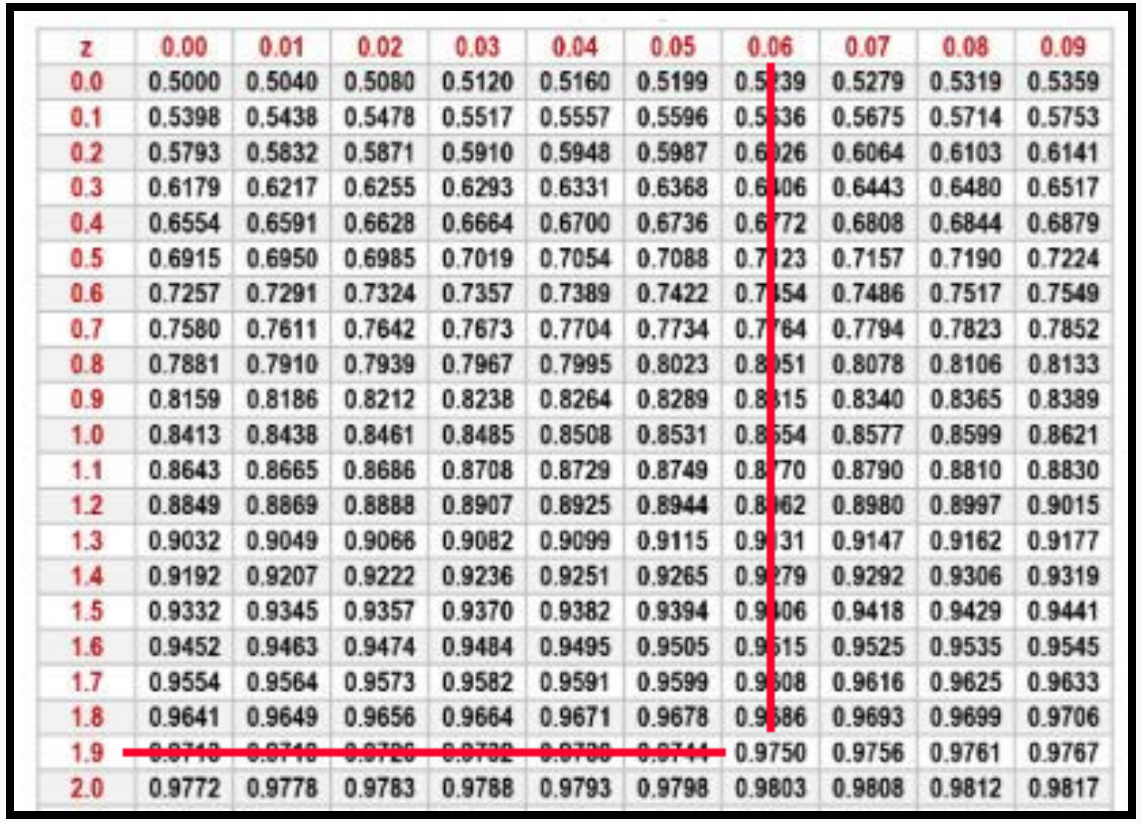


The α level

🎯 $$95\%\ \text{CI} = (\bar{x}- \approx 2 \times \hat{SE}\ ;\ \bar{x}+ \approx 2 \times \hat{SE})$$
$$\approx 2 \quad \rightarrow 1.96$$

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

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

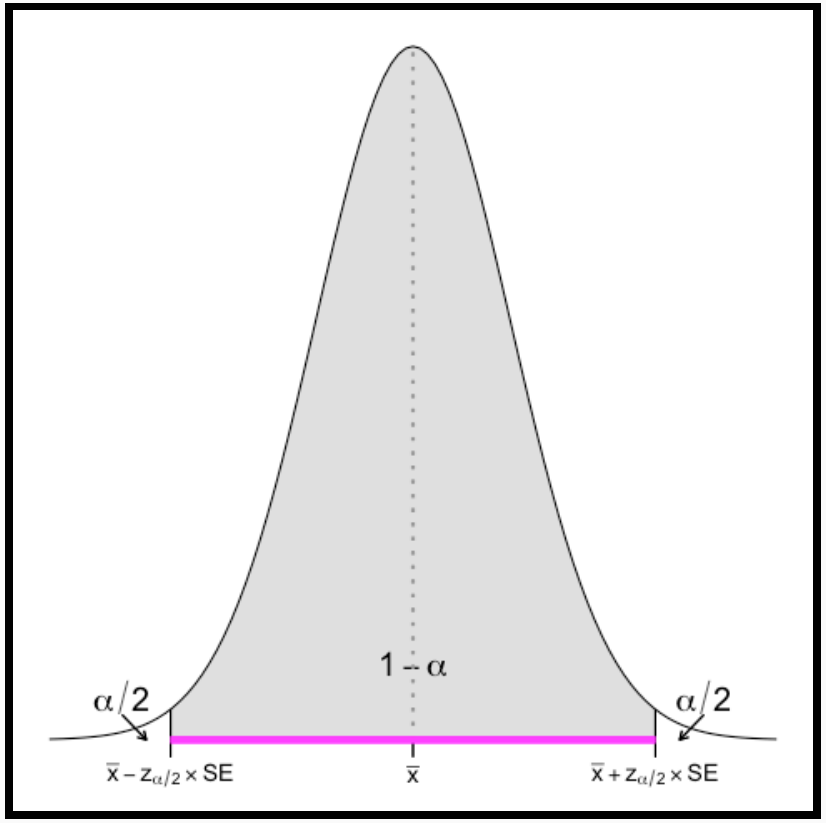


z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	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	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	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

$$100\% - 2.5\% = 97.5\% \rightarrow z = 1.96$$
$$100\% - 5.0\% = 95.0\% \rightarrow z = 1.65$$
$$100\% - 0.5\% = 99.5\% \rightarrow z = 2.58$$



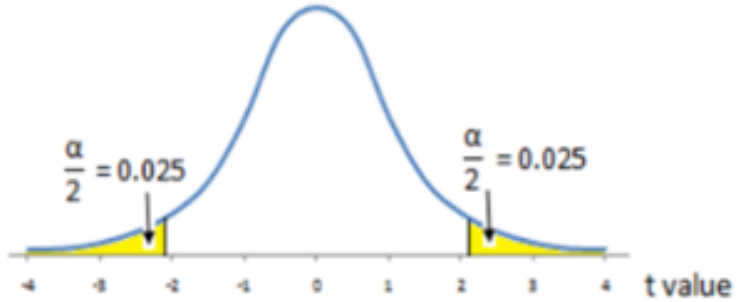
And for small samples?

📌 $n = 15$ patients with T2D

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

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

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



Confidence	90	95	95.45	99	99.73
alpha	0.1000	0.0500	0.0455	0.0100	0.0027
df					
1	6.314	12.706	13.968	63.657	235.784
2	2.920	4.303	4.527	9.925	19.206
3	2.353	3.182	3.307	5.841	9.219
4	2.132	2.776	2.869	4.604	6.620
5	2.015	2.571	2.649	4.032	5.507
6	1.943	2.447	2.517	3.707	4.904
7	1.895	2.365	2.429	3.499	4.530
8	1.860	2.306	2.366	3.355	4.277
9	1.833	2.262	2.320	3.250	4.094
10	1.812	2.228	2.284	3.169	3.957
11	1.796	2.201	2.255	3.106	3.850
12	1.782	2.179	2.231	3.055	3.764
13	1.771	2.160	2.212	3.012	3.694
14	1.761	2.145	2.195	2.977	3.636
15	1.753	2.131	2.181	2.947	3.586

Exercise 12

? $n = 15$ patients with T2D

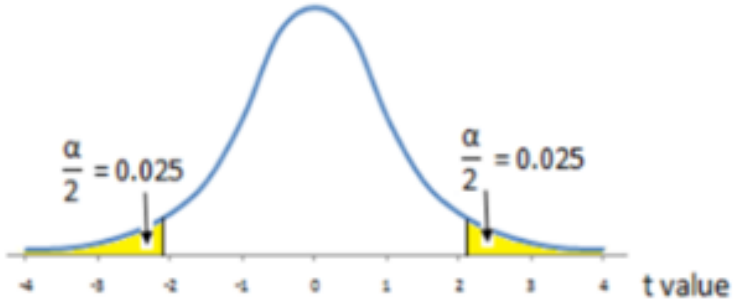
$\bar{x}_{\text{BMI}} = 25.0\ \text{kg/m}^2$

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

$\hat{\text{SE}} = s/\sqrt{n} = \ ?$

$\text{df} = n - 1 = \ ?$

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



Confidence	90	95	95.45	99	99.73
alpha	0.1000	0.0500	0.0455	0.0100	0.0027
df					
1	6.314	12.706	13.968	63.657	235.784
2	2.920	4.303	4.527	9.925	19.206
3	2.353	3.182	3.307	5.841	9.219
4	2.132	2.776	2.869	4.604	6.620
5	2.015	2.571	2.649	4.032	5.507
6	1.943	2.447	2.517	3.707	4.904
7	1.895	2.365	2.429	3.499	4.530
8	1.860	2.306	2.366	3.355	4.277
9	1.833	2.262	2.320	3.250	4.094
10	1.812	2.228	2.284	3.169	3.957
11	1.796	2.201	2.255	3.106	3.850
12	1.782	2.179	2.231	3.055	3.764
13	1.771	2.160	2.212	3.012	3.694
14	1.761	2.145	2.195	2.977	3.636
15	1.753	2.131	2.181	2.947	3.586

Confidence intervals for differences of means

📌 $$\mathcal{N} = (\mu_{\mathrm{i}} - \mu_{\mathrm{c}}, \sqrt{\frac{\sigma_{\mathrm{i}}^2}{n_{\mathrm{i}}} + \frac{\sigma_{\mathrm{c}}^2}{n_{\mathrm{c}}}})$$

$$\hat{\mathrm{SE}} = \sqrt{\frac{s_{\mathrm{i}}^2}{n_{\mathrm{i}}} + \frac{s_{\mathrm{c}}^2}{n_{\mathrm{c}}}}$$

Exercise #13

? Which is the *true* difference in mean between the two groups?

> **Interventions** Twenty mg of dexamethasone intravenously daily for 5 days, 10 mg of dexamethasone daily for 5 days or until ICU discharge, plus standard care (n=151) or standard care alone (n=148).
>
> **Results** A total of 299 patients (mean [SD] age, 61 [14] years; 37% women) were enrolled and all completed follow-up. Patients randomized to the dexamethasone group had a mean 6.6 ventilator-free days (95% CI, 5.0-8.2) during the first 28 days vs 4.0 ventilator-free days (95% CI, 2.9-5.4) in the standard care group

$$n_\mathrm{i} = \,?, \quad \bar{x}_\mathrm{i} = \,?, \quad s_\mathrm{i} = 10.0$$
$$n_\mathrm{c} = \,?, \quad \bar{x}_\mathrm{c} = \,?, \quad s_\mathrm{i} = 8.7$$

$$\hat{\mathrm{SE}} = \sqrt{\frac{s_\mathrm{i}^2}{n_\mathrm{i}} + \frac{s_\mathrm{c}^2}{n_\mathrm{c}}} = \,?$$

Tomazini B.M., *et al.*, *Effect of Dexamethasone on Days Alive and Ventilator-Free in Patients With Moderate or Severe Acute Respiratory Distress Syndrome and COVID-19*, JAMA, 2020

Confidence intervals for proportions

📌 $\mathcal{N} = (\pi, \frac{\pi \times (1-\pi)}{n})$

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

Exercise #14

? Which is the *true* proportion of women with endometriosis in the population?

> A total of 1291 women without a previous diagnosis of endometriosis were included in the study. On the basis of the symptoms, 108 women were referred to a gynecologist. After gynecological examination and transvaginal ultrasound, endometriosis was suspected in 51 women (47.2%). The diagnosis of endometriosis was confirmed by radiological investigations and/or surgery in 46 patients;

$$n = \ ?, \quad m = \ ?$$
$$\bar{p} = \frac{m}{n} = \ ?$$

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

Ferrero S.., *et al.*, *Prevalence of newly diagnosed endometriosis in women attending the general practitioner*, Int J Gynaecol Obstet, 2020

Confidence intervals for differences of proportion

📌 $$\mathcal{N} = (\pi_{\mathrm{i}} - \pi_{\mathrm{c}}, \frac{\pi_i \times (1-\pi_i)}{n_i} + \frac{\pi_c \times (1-\pi_c)}{n_c})$$

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

Exercise #15

? Which is the *true* difference in proportion between two groups?

> From April 1991 through December 20, 1993, the cutoff date for the first interim analysis of efficacy, 477 pregnant women were enrolled; during the study period, 409 gave birth to 415 live-born infants. HIV-infection status was known for 363 births (180 in the zidovudine group and 183 in the placebo group). Thirteen infants in the zidovudine group and 40 in the placebo group were HIV-infected.

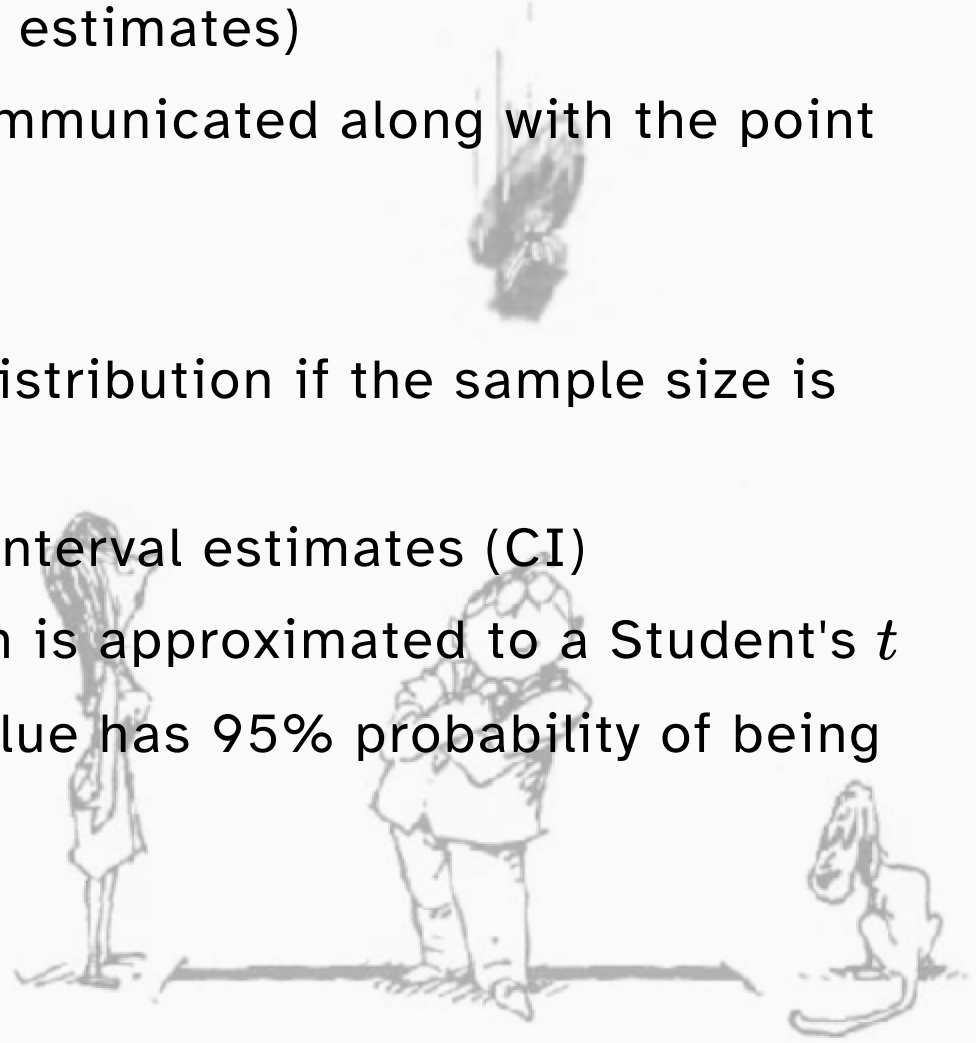
$$n_\mathrm{i} = ?, \quad m_\mathrm{i} = ?, \quad p_\mathrm{i} = \frac{m_\mathrm{i}}{n_\mathrm{i}} = ?$$
$$n_\mathrm{c} = ?, \quad m_\mathrm{c} = ?, \quad p_\mathrm{c} = \frac{m_\mathrm{c}}{n_\mathrm{c}} = ?$$

$$\hat{\mathrm{SE}} = \sqrt{\frac{\bar{p}_\mathrm{i} \times (1-\bar{p}_\mathrm{i})}{n_\mathrm{i}} + \frac{\bar{p}_\mathrm{c} \times (1-\bar{p}_\mathrm{c})}{n_\mathrm{c}}} = ?$$

Connor, E.M. *et al.*, *Reduction of Maternal-Infant Transmission of Human Immunodeficiency Virus Type 1 with Zidovudine Treatment*, NEJM, 1994

Summary

- We use statistics to estimate parameters (point estimates)
- Interval estimates (CI and/or ME) should be communicated along with the point estimates
- The sample size influences the size of the CI
- Sampling distributions tend to show a normal distribution if the sample size is large enough (CLT)
- We can take advantage of the CLT to calculate interval estimates (CI)
- For small sample size, the sampling distribution is approximated to a Student's t
- 95% confidence intervals tell us the the true value has 95% probability of being inside the given range



See you tomorrow