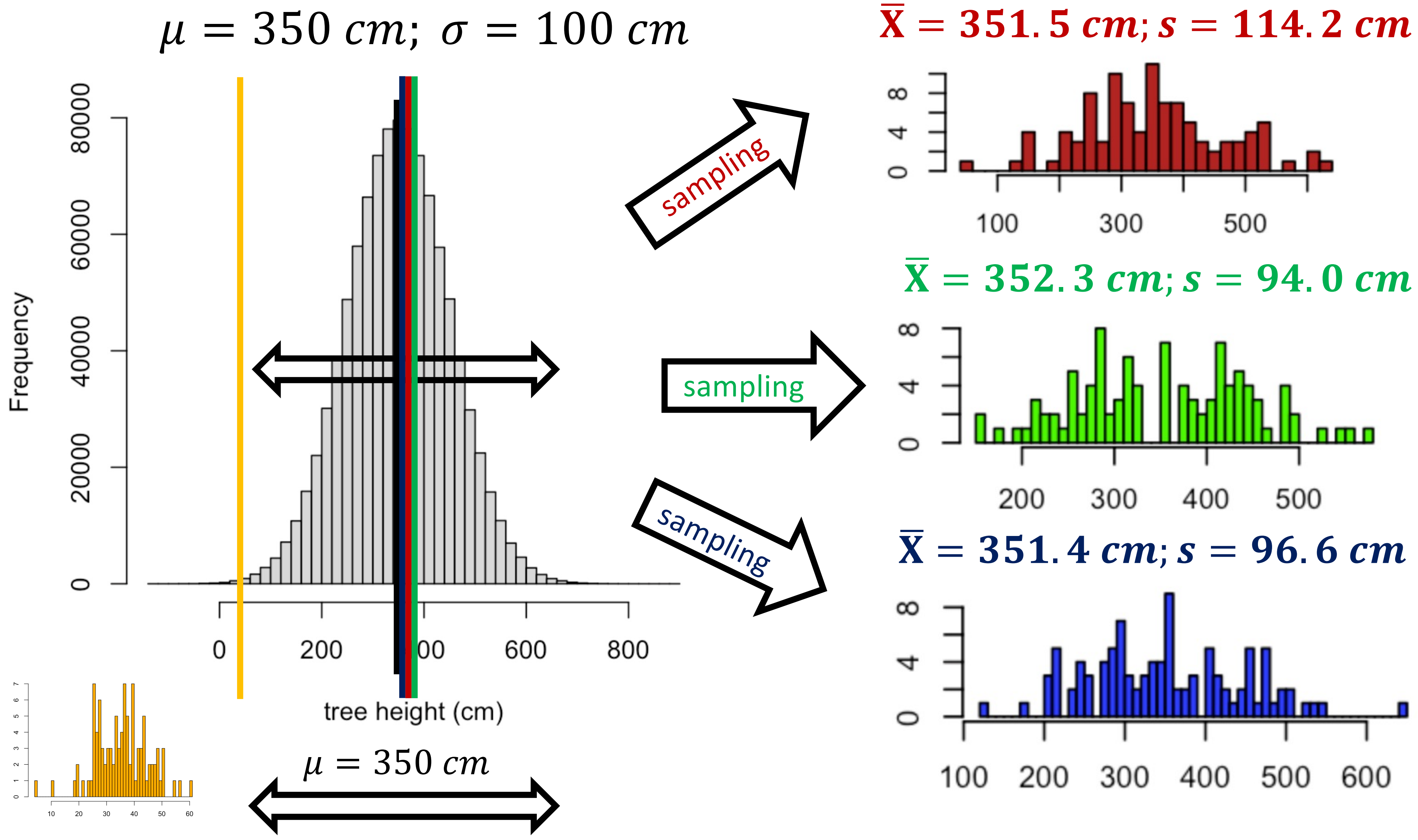
Lecture 10: Estimating with uncertainty, but with a degree of certainty (i.e., with some confidence), part 2

Key statistical concepts for understanding confidence intervals, statistical procedures, and statistical reasoning

Sampling variation generates uncertainty, i.e., sampling error



$\mu = 350\ cm;\ \sigma = 100\ cm$

$\bar{\mathbf{X}} = \mathbf{351.5}\ \boldsymbol{cm}; \boldsymbol{s} = \mathbf{114.2}\ \boldsymbol{cm}$

$\bar{\mathbf{X}} = \mathbf{352.3}\ \boldsymbol{cm}; \boldsymbol{s} = \mathbf{94.0}\ \boldsymbol{cm}$

$\bar{\mathbf{X}} = \mathbf{351.4}\ \boldsymbol{cm}; \boldsymbol{s} = \mathbf{96.6}\ \boldsymbol{cm}$

sampling

sampling

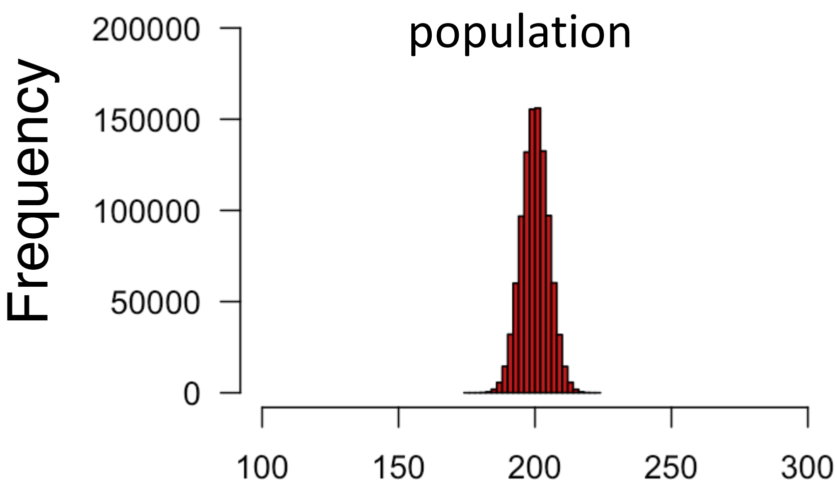
sampling

$\mu = 350\ cm$

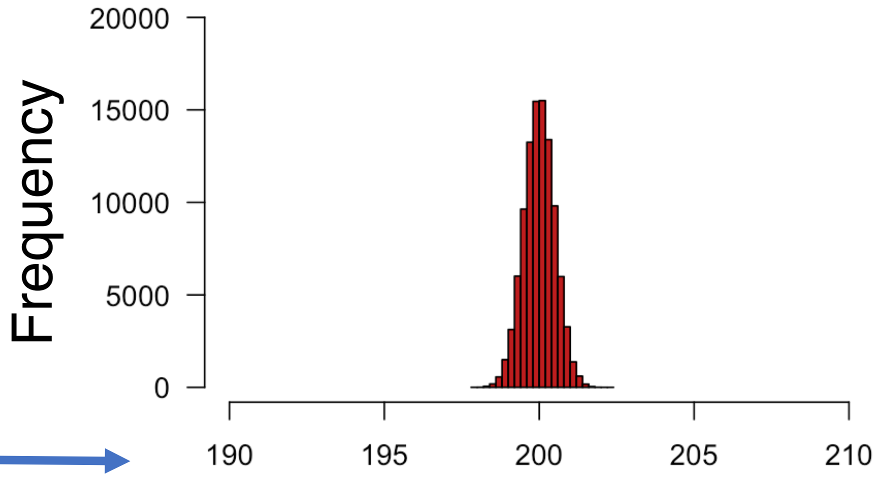
Uncertainty (samples means varying around the true population mean)

The variation within a sample, summarized by the sample standard deviation, provides information about how much sample means are expected to fluctuate around the true population mean—allowing us to estimate the typical uncertainty in our estimate.

$\mu = 200\ cm, \sigma$=5 cm



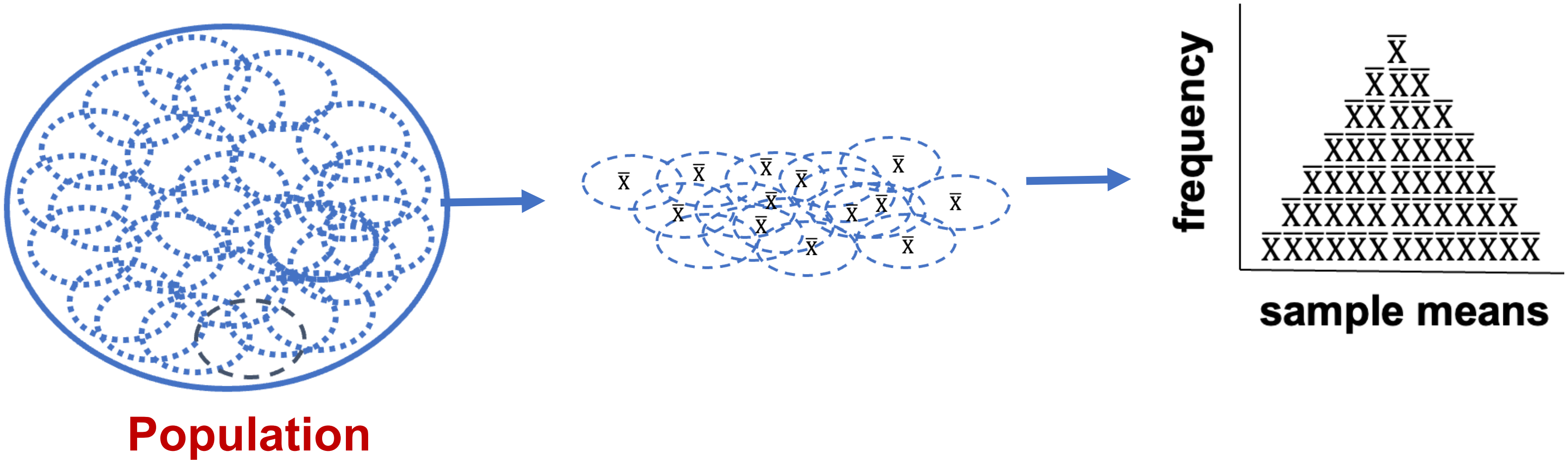
Sampling distribution of means



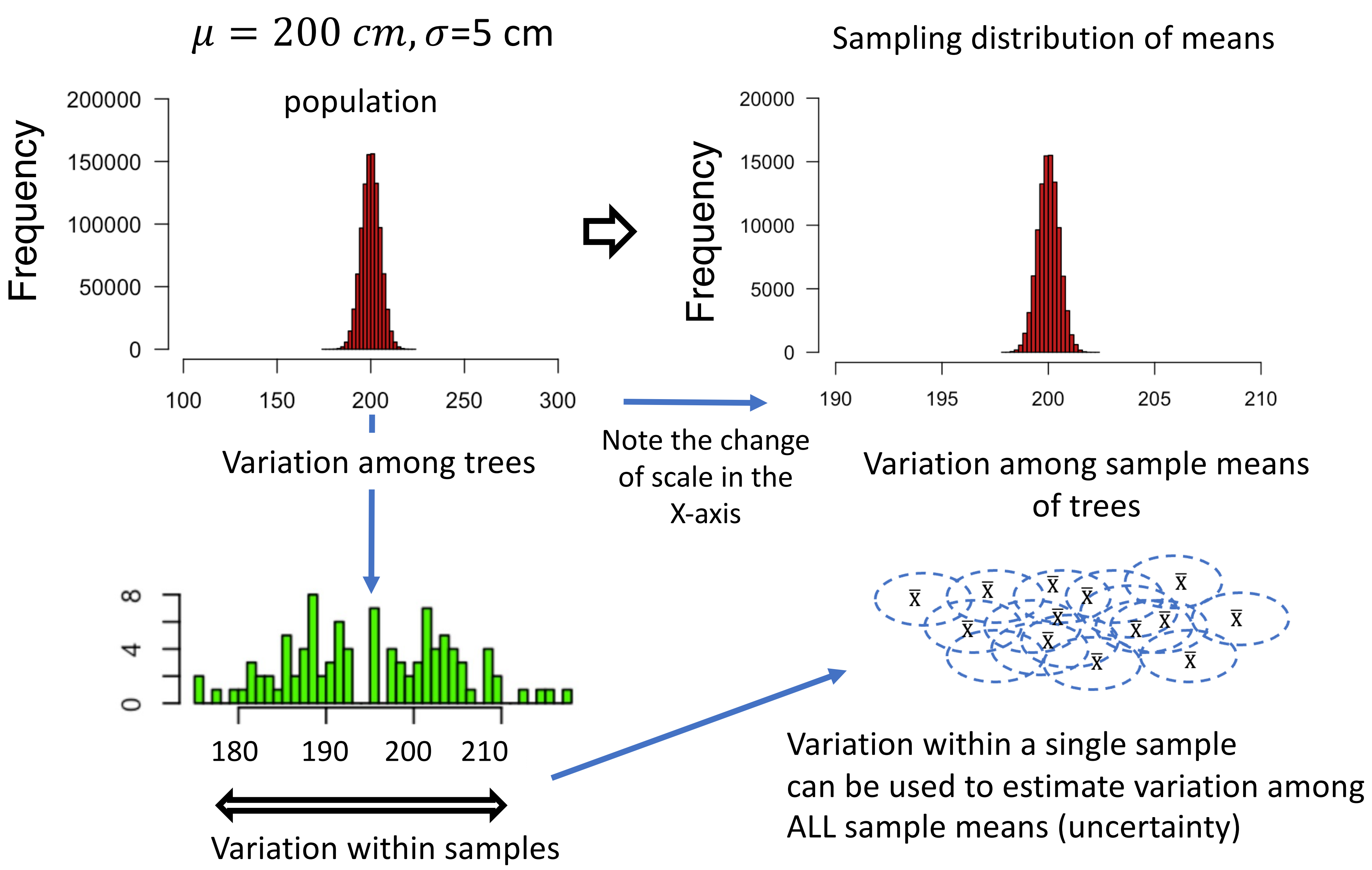
Variation among trees (small trees)

Note the change of scale in the X-axis

Variation among sample means of trees



From variation among observations to sampling variation: the variation within a sample, summarized by the sample standard deviation, provides information about how much sample means are expected to fluctuate around the true population mean, allowing us to estimate the typical uncertainty in our estimate.



$\mu = 200\ cm, \sigma$=5 cm

Sampling distribution of means

Variation among trees

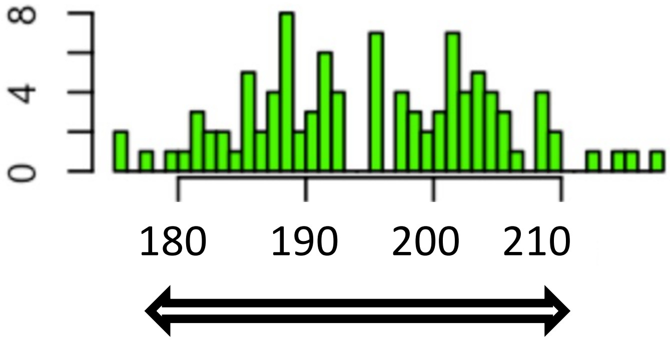
Note the change of scale in the X-axis

Variation among sample means of trees

Variation within samples

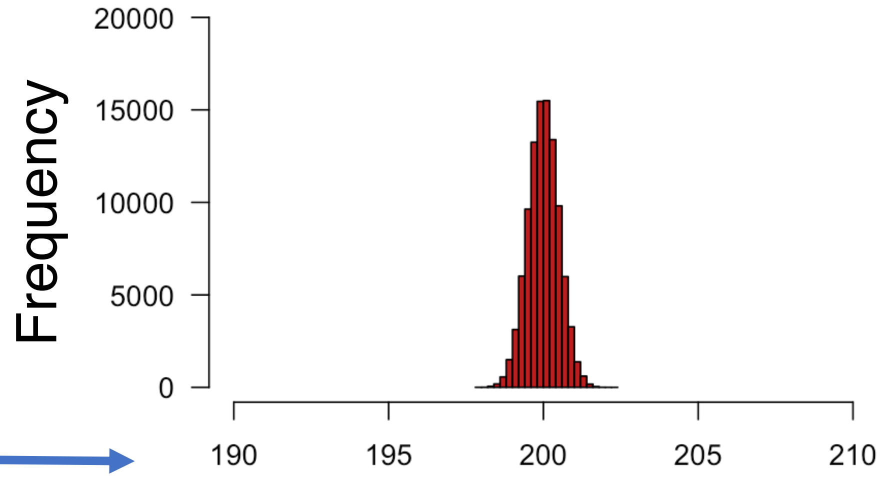
Variation within a single sample can be used to estimate variation among ALL sample means (uncertainty)

From variation among observations to sampling variation: the variation within a sample, summarized by the sample standard deviation, provides information about how much sample means are expected to fluctuate around the true population mean, allowing us to estimate the typical uncertainty in our estimate.



Variation within a sample

Sampling distribution of means



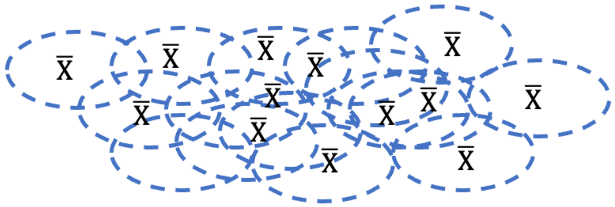
Note the change of scale

Variation among sample means of trees

s Standard deviation of the sample

n Sample size

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$



$s_{\bar{X}}$ Standard error of the mean = the standard deviation of the sampling distribution of the mean (uncertainty).

Variation within a single sample can be used to estimate variation among ALL sample means (uncertainty)

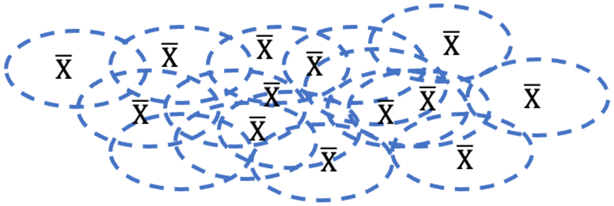
$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

THE STANDARD ERROR OF THE MEAN

Standard deviation (*s*): If you measure 25 students and the SD is 10 cm, it means a typical student's height differs from the sample average by about 10 cm.

If you repeatedly (randomly) sampled different groups of 25 students, their averages would vary slightly around the true population mean. And the same for any number of students that build a sample.

$$s_{\bar{X}} = \frac{10}{\sqrt{25}} = 2\text{cm}$$

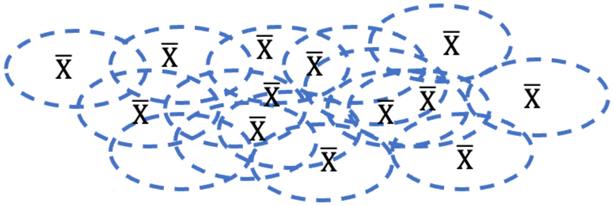


The 2 cm is the standard error and is a measure of uncertainty. It tells you: if you repeatedly took random samples of 25 students and calculated their average height, those averages would typically differ from the true population mean by about 2 cm.

UNCERTAINTY: $s_{\bar{X}}$ (the standard error) estimates how wrong our sample mean is expected to be, on average, from the true population mean—purely because we relied on a sample.

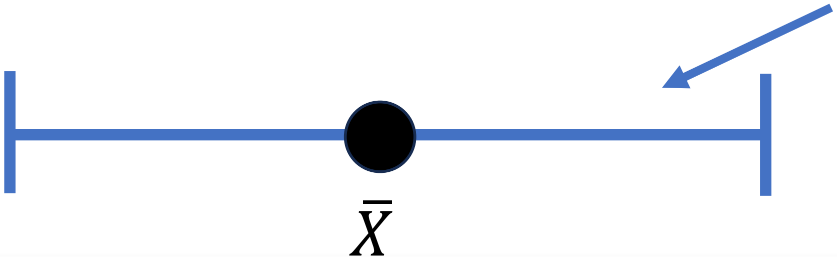
CONFIDENCE INTERVAL FOR THE MEAN

$$s_{\bar{X}} = \frac{10}{\sqrt{25}} = 2\text{cm}$$



UNCERTAINTY: $s_{\bar{X}}$ (the standard error) estimates how wrong our sample mean is expected to be, on average, from the true population mean—purely because we relied on a sample.

$$\mathit{Confidence\ interval} = \bar{X} \pm \mathit{quantity} * s_{\bar{X}}$$



Very plausible (high confidence) that the population parameter μ is somewhere within the 95% confidence interval.

The quantity (i.e., which establishes the critical value or margin of error) varies with the confidence level we choose (e.g., 95% or 99%).

Although we have not yet shown how to estimate a confidence interval, we already know that it depends on the standard error of the mean, a measure of how much sample means vary due to sampling.

The standard deviation of the sampling distribution $\sigma_{\bar{Y}}$ is called standard error (SE) and is exactly:

$$\sigma_{\bar{Y}} = \sqrt{\lim_{N\to\infty} \sum_{i=1}^{N} \frac{(\bar{X}_i - \mu)^2}{N}} = \frac{\sigma}{\sqrt{n}}$$

N is all infinite samples and *n* is sample size

The left side is a definition using infinite repetition and the right side is a mathematical consequence of the probability model. Statistics closes these gaps.

$$\sqrt{\lim_{N\to\infty} \sum_{i=1}^{N} \frac{(\bar{X}_i - \mu)^2}{N}}$$

→ Not observable: infinitely many samples

$$\frac{\sigma}{\sqrt{n}}$$

→ closed-form result from probability theory.
variability of sample means depends only on:
the population standard deviation σ,
and the sample size n.

[REVISED]

Although we have not yet shown how to estimate a confidence interval, we already know that it depends on the standard error of the mean, a measure of how much sample means vary due to sampling.

Given that we almost never know the population standard deviation, we estimate it with the sample value basec on the sample standard error:

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

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

$\sigma_{\bar{X}} =$ the standard deviation of the sampling distribution of means (standard error) ; $\sigma =$ the standard deviation of the population.

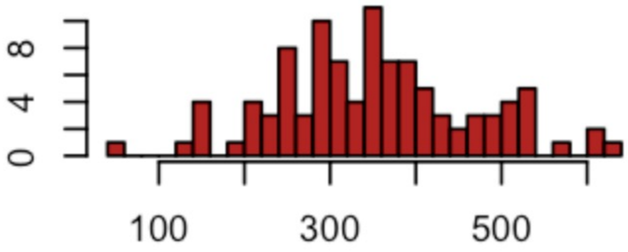
The standard error of the mean, SE_X , estimates the standard deviation of the sampling distribution of the mean; that is, how much sample means are expected to vary around the true population mean across repeated samples. Remarkably, this quantity can be estimated from a single sample (a principle we will examine in detail in Tutorial 5).

[REVISED]

How to calculate a confidence interval in practice:

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{114.2}{\sqrt{100}} = 11.42\ cm$$

$$\mathbf{\bar{X} = 351.5\ \mathit{cm}; \mathit{s} = 114.2\ \mathit{cm}}$$



$$Confidence\ interval = 351.5\ cm \pm \mathbf{\mathit{quantity}} * 11.42\ cm$$

S_X estimates the standard deviation of the sampling distribution of the mean (uncertainty).

Based on this sample, we estimate that sample means typically differ from the true population value by about 11.42 cm due to sampling variability.

Obviously, a different sample will generate a different estimate of this error.

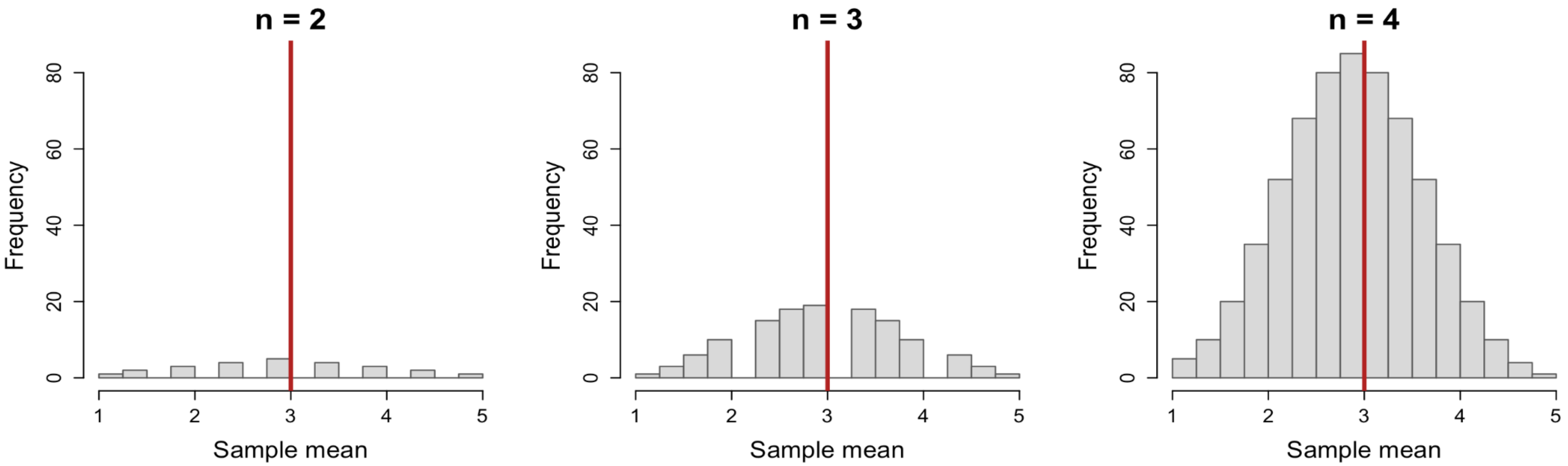
[REVISED]

Sampling distribution of the **mean, variance, standard deviation, and standard error** obtained from the population (1, 2, 3, 4, 5), considering all possible samples drawn with replacement for sample sizes n = 2, 3, 4.

Obs 1	Obs 2	Sample mean	Mean − 3	(Mean − 3)^2	Sample Var (n−1)	Sample SD (n−1)	Sample SD / √2
1	1	1.0	−2.0	4.00	0.000	0.000	0.000
1	2	1.5	−1.5	2.25	0.500	0.707	0.500
1	3	2.0	−1.0	1.00	2.000	1.414	1.000
1	4	2.5	−0.5	0.25	4.500	2.121	1.500
1	5	3.0	0.0	0.00	8.000	2.828	2.000
2	1	1.5	−1.5	2.25	0.500	0.707	0.500
2	2	2.0	−1.0	1.00	0.000	0.000	0.000
2	3	2.5	−0.5	0.25	0.500	0.707	0.500
2	4	3.0	0.0	0.00	2.000	1.414	1.000
2	5	3.5	0.5	0.25	4.500	2.121	1.500
3	1	2.0	−1.0	1.00	2.000	1.414	1.000
3	2	2.5	−0.5	0.25	0.500	0.707	0.500
3	3	3.0	0.0	0.00	0.000	0.000	0.000
3	4	3.5	0.5	0.25	0.500	0.707	0.500
3	5	4.0	1.0	1.00	2.000	1.414	1.000
4	1	2.5	−0.5	0.25	4.500	2.121	1.500
4	2	3.0	0.0	0.00	2.000	1.414	1.000
4	3	3.5	0.5	0.25	0.500	0.707	0.500
4	4	4.0	1.0	1.00	0.000	0.000	0.000
4	5	4.5	1.5	2.25	0.500	0.707	0.500
5	1	3.0	0.0	0.00	8.000	2.828	2.000
5	2	3.5	0.5	0.25	4.500	2.121	1.500
5	3	4.0	1.0	1.00	2.000	1.414	1.000
5	4	4.5	1.5	2.25	0.500	0.707	0.500
5	5	5.0	2.0	4.00	0.000	0.000	0.000

Sampling distribution of the **mean** obtained from the population (1, 2, 3, 4, 5), considering all possible samples drawn with replacement for sample sizes n = 2, 3, 4.

▬ True population mean = mean of sample means



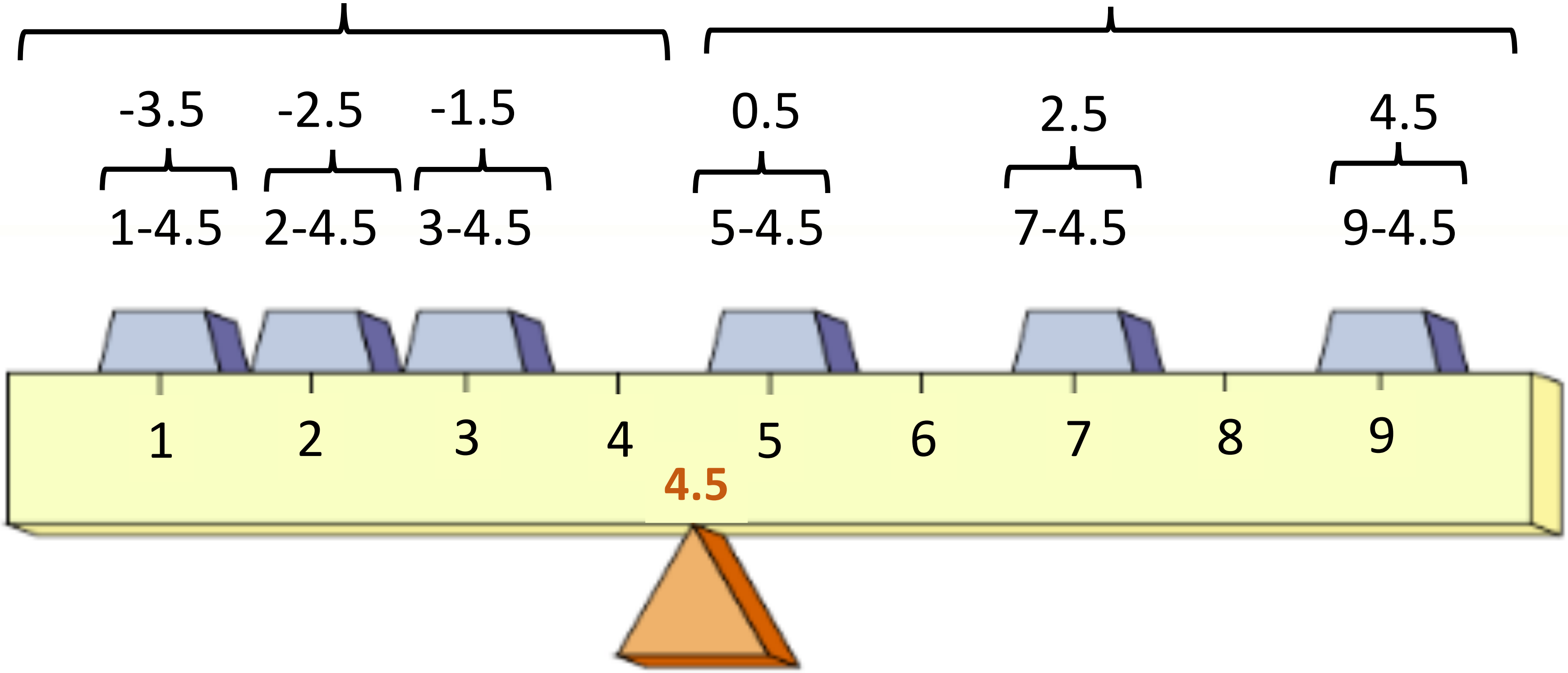
One critical observation is that **the mean of all possible sample means is exactly the population mean**. This is important because it tells us that, on average, the sampling process is **unbiased**: repeated sampling does not systematically overestimate or underestimate the true population value.

Even though individual samples can yield means that are far from the population mean, especially when sample size is small, these deviations balance out across all possible samples.

Recall from lecture 6: The mean can be understood as the center of gravity of a distribution: the point at which the values on either side balance each other

Sum (left) = -7.5

Sum (right) = 7.5

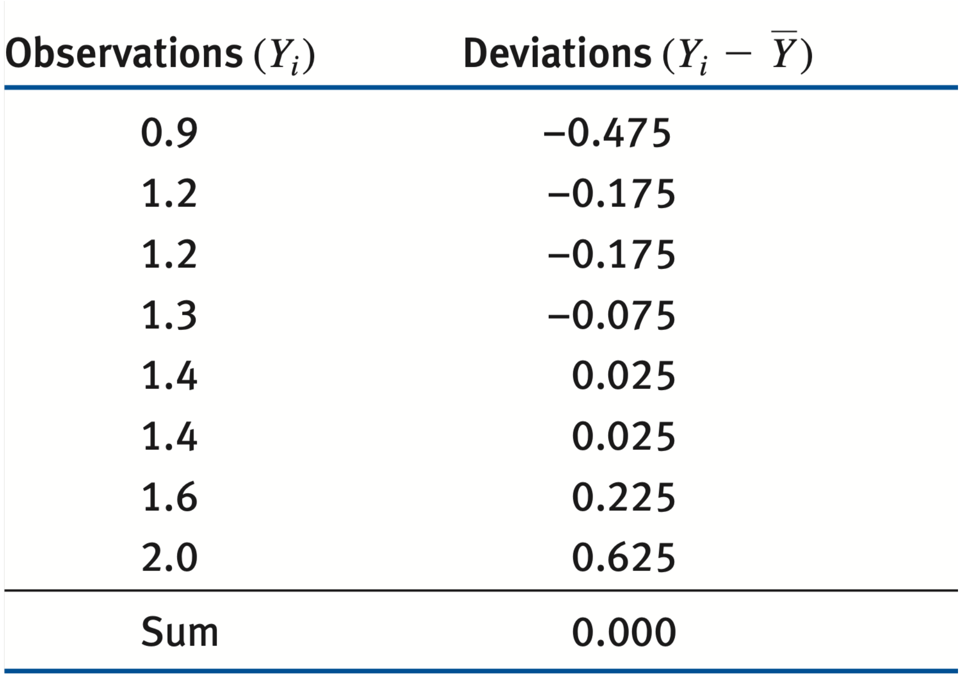


$$\underbrace{1-4.5}_{-3.5} \quad \underbrace{2-4.5}_{-2.5} \quad \underbrace{3-4.5}_{-1.5} \quad \underbrace{5-4.5}_{0.5} \quad \underbrace{7-4.5}_{2.5} \quad \underbrace{9-4.5}_{4.5}$$

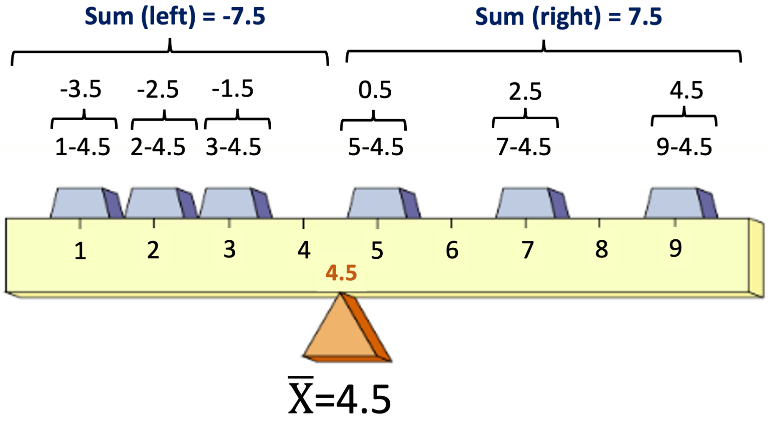
$\overline{X}$=4.5

Sum (left) + Sum (right) = -7.5 + 7.5 = 0

Recall from lecture 5: the sum of deviations from the mean is always zero, making the mean the 'center of gravity' of a distribution, i.e., the values on either side of the mean balance each other.



Observations (Y_i)	Deviations $(Y_i - \overline{Y})$
0.9	−0.475
1.2	−0.175
1.2	−0.175
1.3	−0.075
1.4	0.025
1.4	0.025
1.6	0.225
2.0	0.625
Sum	0.000



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

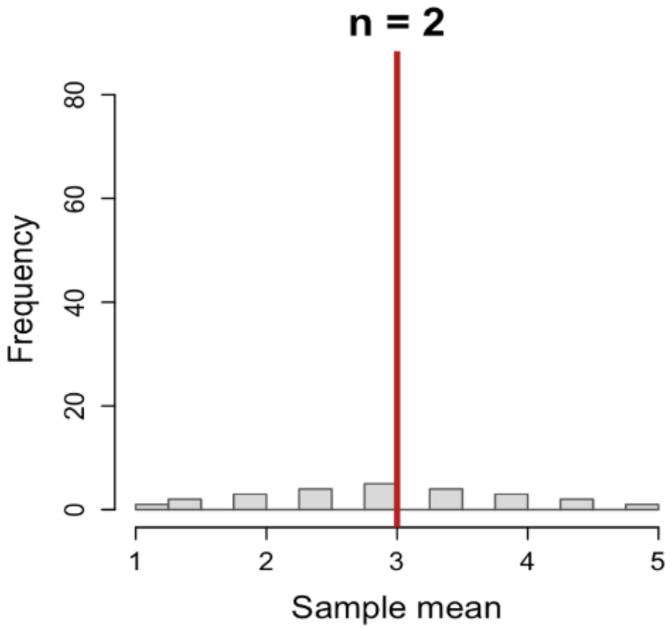
Sampling distribution: the **sample means** on either side of the **true population mean μ** balance each other, i.e., the sum = 0.

sampling error

Obs 1	Obs 2	Sample means	Sample means - μ
1	1	1.0	-2
1	2	1.5	-1.5
1	3	2.0	-1
1	4	2.5	-0.5
1	5	3.0	0
2	1	1.5	-1.5
2	2	2.0	-1
2	3	2.5	-0.5
2	4	3.0	0
2	5	3.5	0.5
3	1	2.0	-1
3	2	2.5	-0.5
3	3	3.0	0
3	4	3.5	0.5
3	5	4.0	1
4	1	2.5	-0.5
4	2	3.0	0
4	3	3.5	0.5
4	4	4.0	1
4	5	4.5	1.5
5	1	3.0	0
5	2	3.5	0.5
5	3	4.0	1
5	4	4.5	1.5
5	5	5.0	2
	MEAN	**3.0**	**0**

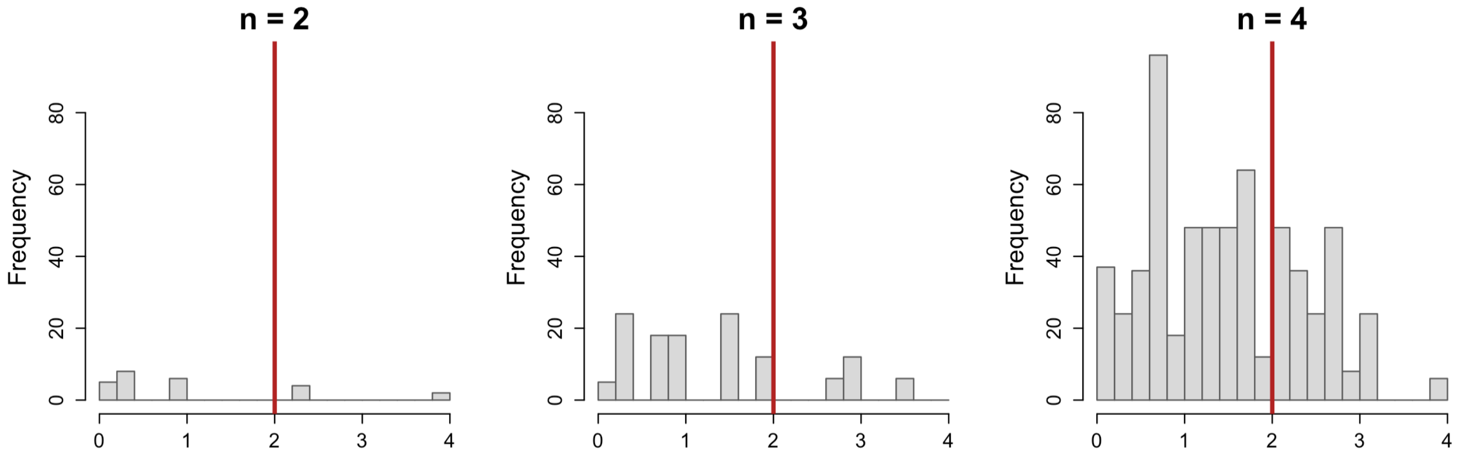
The mean of all possible sample means is exactly the population mean.

The sampling process of the mean is **unbiased**: repeated sampling does not systematically overestimate or underestimate the true population value, i.e., they balance each other out.



Sampling distribution of the **variance** obtained from the population (1, 2, 3, 4, 5), considering all possible samples drawn with replacement for sample sizes n = 2, 3, 4.

— True population variance (2.0) = mean of sample variances (2.0)



sample variances (denominator *n*-1)

sample variances (denominator *n*-1)

sample variances (denominator *n*-1)

One critical observation is that **the mean of all possible sample variance is exactly the population variance**. This is important because it tells us that, on average, the sampling process is **unbiased**: repeated sampling does not systematically overestimate or underestimate the true population value.

Even though individual samples can yield variances that are far from the population variance, especially when sample size is small, these deviations balance out across all possible samples.

[REVISED]

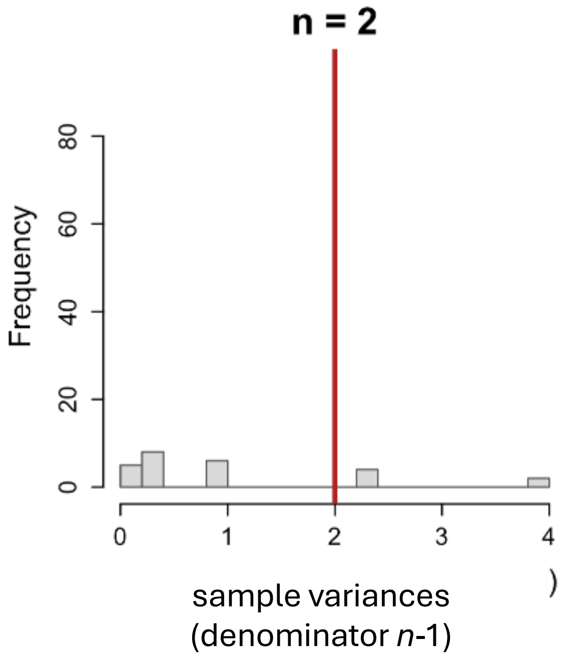
Sampling distribution: the **sample variances** on either side of the **true population variance σ^2** balance each other, i.e., the sum = 0.

sampling error

Obs 1	Obs 2	Sample variances	Sample variances - σ^2
1	1	0.000	-2
1	2	0.500	-1.5
1	3	2.000	0
1	4	4.500	2.5
1	5	8.000	6
2	1	0.500	-1.5
2	2	0.000	-2
2	3	0.500	-1.5
2	4	2.000	0
2	5	4.500	2.5
3	1	2.000	0
3	2	0.500	-1.5
3	3	0.000	-2
3	4	0.500	-1.5
3	5	2.000	0
4	1	4.500	2.5
4	2	2.000	0
4	3	0.500	-1.5
4	4	0.000	-2
4	5	0.500	-1.5
5	1	8.000	6
5	2	4.500	2.5
5	3	2.000	0
5	4	0.500	-1.5
5	5	0.000	-2
	MEAN	**2.0**	**0**

The mean of all possible sample variances is exactly the population variance.

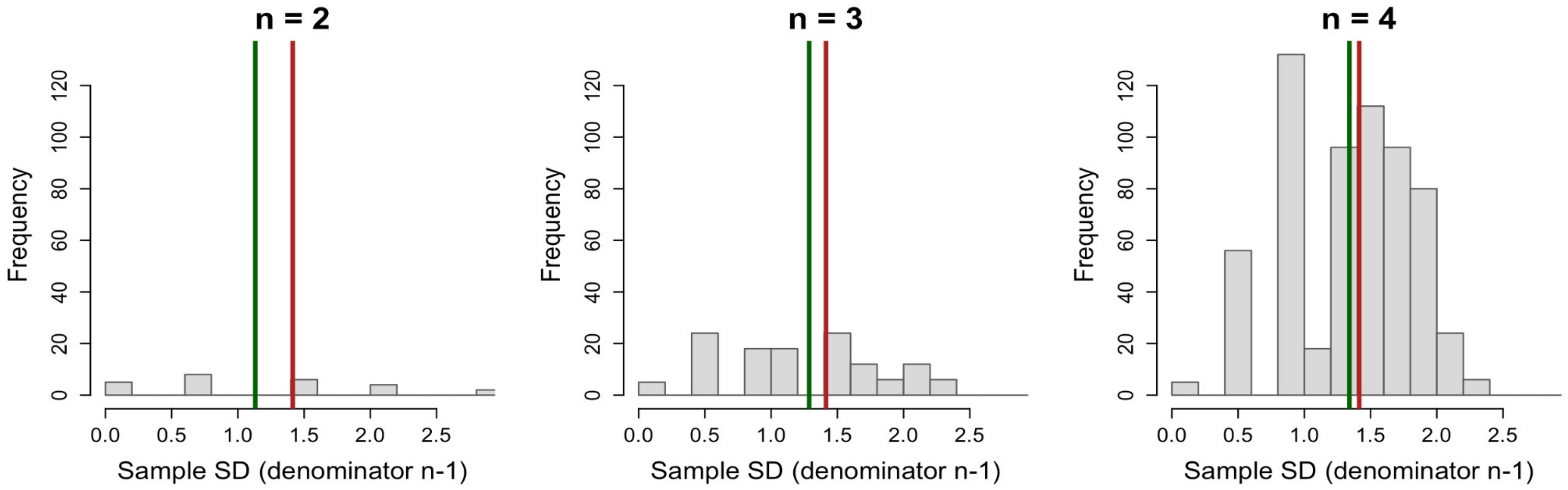
The sampling process of the variance is **unbiased:** repeated sampling does not systematically overestimate or underestimate the true population value, i.e., they balance each other out.



Sampling distribution of the **standard deviation** obtained from the population (1, 2, 3, 4, 5), considering all possible samples drawn with replacement for sample sizes n = 2, 3, 4.

🟥 True population standard deviation (1.414214)

🟩 Mean of sample standard deviations (n=2: 1.131371; n=3: 1.287381; n=4: 1.340293)



One critical observation is that **the mean of all possible sample standard deviation IS NOT exactly the population standard deviation**. This is important because it tells us that, on average, the sampling process is **BIASED**: repeated sampling ***systematically underestimate*** the true population value.

[REVISED]

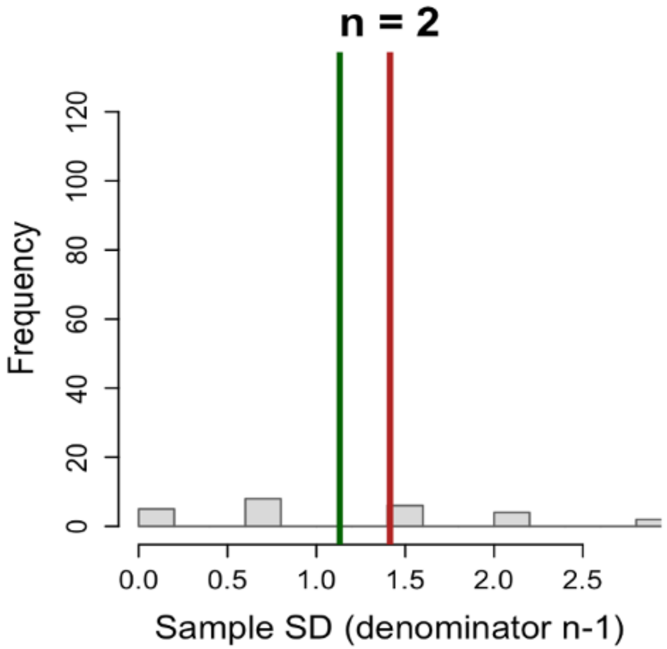
Sampling distribution: the **sample standard deviations** on either side of the **true population standard deviation σ** do not balance each other, i.e., the sum ≠ 0.

sampling error

Obs 1	Obs 2	Sample standard deviations (SD)	Sample SDs - σ
1	1	0.000	-2
1	2	0.707	-1.5
1	3	1.414	0
1	4	2.121	2.5
1	5	2.828	6
2	1	0.707	-1.5
2	2	0.000	-2
2	3	0.707	-1.5
2	4	1.414	0
2	5	2.121	2.5
3	1	1.414	0
3	2	0.707	-1.5
3	3	0.000	-2
3	4	0.707	-1.5
3	5	1.414	0
4	1	2.121	2.5
4	2	1.414	0
4	3	0.707	-1.5
4	4	0.000	-2
4	5	0.707	-1.5
5	1	2.828	6
5	2	2.121	2.5
5	3	1.414	0
5	4	0.707	-1.5
5	5	0.000	-2
	MEAN	**1.1312**	**-0.283014**

The mean of all possible sample standard deviations is not the population standard deviation.

The sampling process of the standard deviation is **biased**: repeated sampling systematically **underestimate** (sum is negative) the true population value.



sampling error
of the variance

Sample variances - σ^2

-2	-2	
-1.5	-1.5	
0		0
2.5		2.5
6		6
-1.5	-1.5	
-2	-2	
-1.5	-1.5	
0		0
2.5		2.5
0		0
-1.5	-1.5	
-2	-2	
-1.5	-1.5	
0		0
2.5		2.5
0		0
-1.5	-1.5	
-2	-2	
-1.5	-1.5	
6		6
2.5		2.5
0		0
-1.5	-1.5	
-2	-2	
SUM	**-22**	**22**

sampling error
of the standard deviation

Sample standard deviations - σ

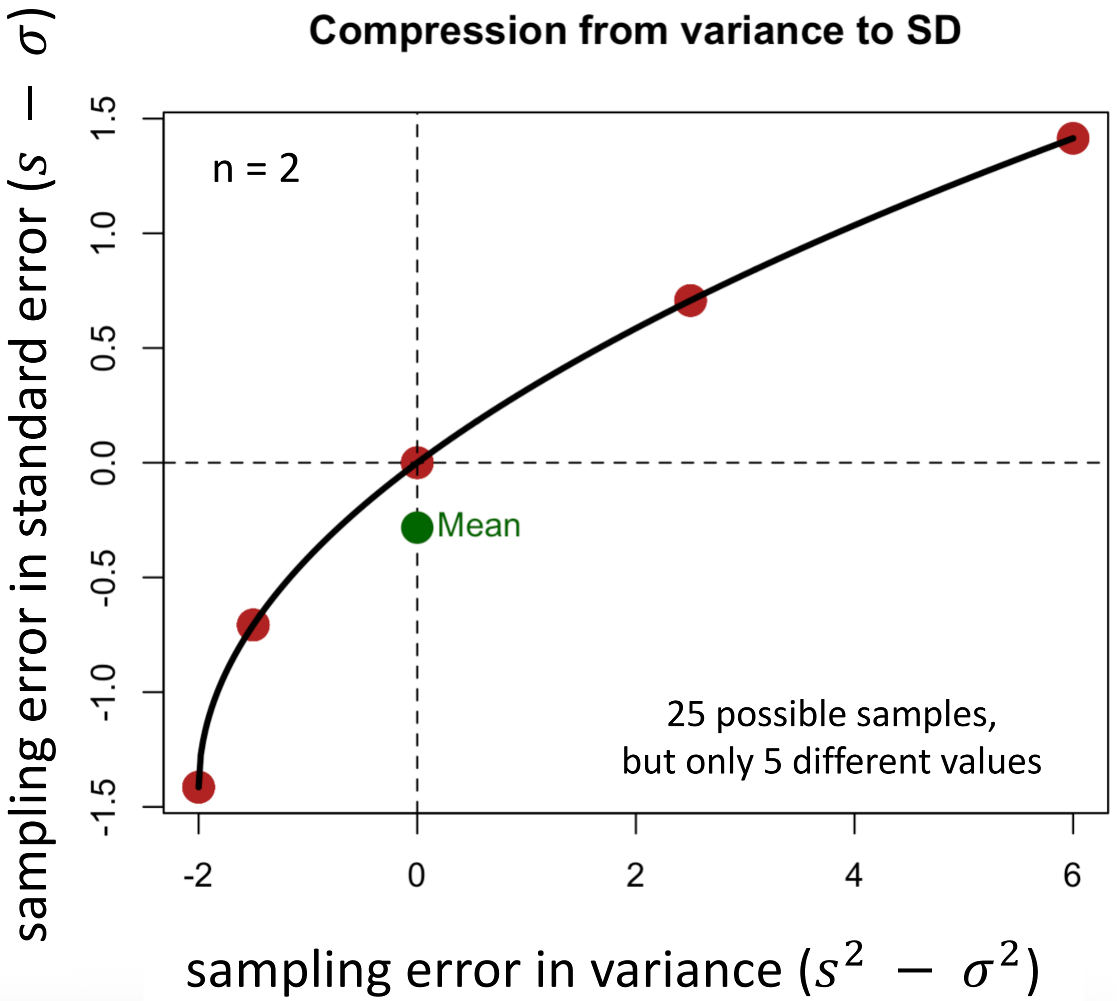
-1.41	-1.41	
-0.71	-0.71	
0.00	0.00	
0.71		0.71
1.41		1.41
-0.71	-0.71	
-1.41	-1.41	
-0.71	-0.71	
0.00	0.00	
0.71		0.71
0.00	0.00	
-0.71	-0.71	
-1.41	-1.41	
-0.71	-0.71	
0.00	0.00	
0.71		0.71
0.00	0.00	
-0.71	-0.71	
-1.41	-1.41	
-0.71	-0.71	
1.41	1.41	
0.71	0.71	
0.00	0.00	
-0.71	-0.71	
-1.41	-1.41	
SUM	**-10.61**	**3.53**

the **sample standard deviations** on either side of the **true population standard deviation ρ** do not balance each other, i.e., the sum $\neq$ 0.

[REVISED]

The square-root transformation compresses variation. Larger positive deviations are reduced more than smaller negative ones, breaking the symmetry of sampling errors (i.e., the difference between the sample value and the true population value).

This is explained by Jensen’s inequality.



sampling error	
$(s^2 - \sigma^2)$	$(s - \sigma)$
-2	-1.414214
-1.5	-0.707214
0	-0.000214
2.5	0.706786
6	1.413786
-1.5	-0.707214
-2	-1.414214
-1.5	-0.707214
0	-0.000214
2.5	0.706786
0	-0.000214
-1.5	-0.707214
-2	-1.414214
-1.5	-0.707214
0	-0.000214
2.5	0.706786
0	-0.000214
-1.5	-0.707214
-2	-1.414214
-1.5	-0.707214
6	1.413786
2.5	0.706786
0	-0.000214
-1.5	-0.707214
-2	-1.414214

The expected value ($\mathbb{E}$) is the average of the sampling distribution; that is, the mean value we would obtain if we could repeat the sampling process indefinitely.

$$\mathbb{E}(\bar{X}) = \mu$$

The expected value of the sample mean equals the population mean (unbiased).

$$\mathbb{E}(s^2) = \sigma^2$$

The expected value of the sample variance equals the population variance (biased).

$$\mathbb{E}(s) \neq \sigma$$

The expected value of the sample standard deviation IS NOT equal the population standard deviation (biased).

↓

$$\mathbb{E}\left(\frac{s}{\sqrt{n}}\right) \neq \frac{\sigma}{\sqrt{n}}$$

The expected value of the sample standard error IS NOT equal the population standard error (biased).

$\mathbb{E}$ is called the expectation operator. It represents the theoretical long-run (infinite) average of a random variable.

Although we have not yet shown how to estimate a confidence interval, we already know that it depends on the standard error of the mean, a measure of how much sample means vary due to sampling.

Given that we almost never know the population standard deviation, we estimate it with the sample value basec on the sample standard error:

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

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

$\sigma_{\bar{X}}$ = the standard deviation of the sampling distribution of means (standard error) ; σ = the standard deviation of the population.

?

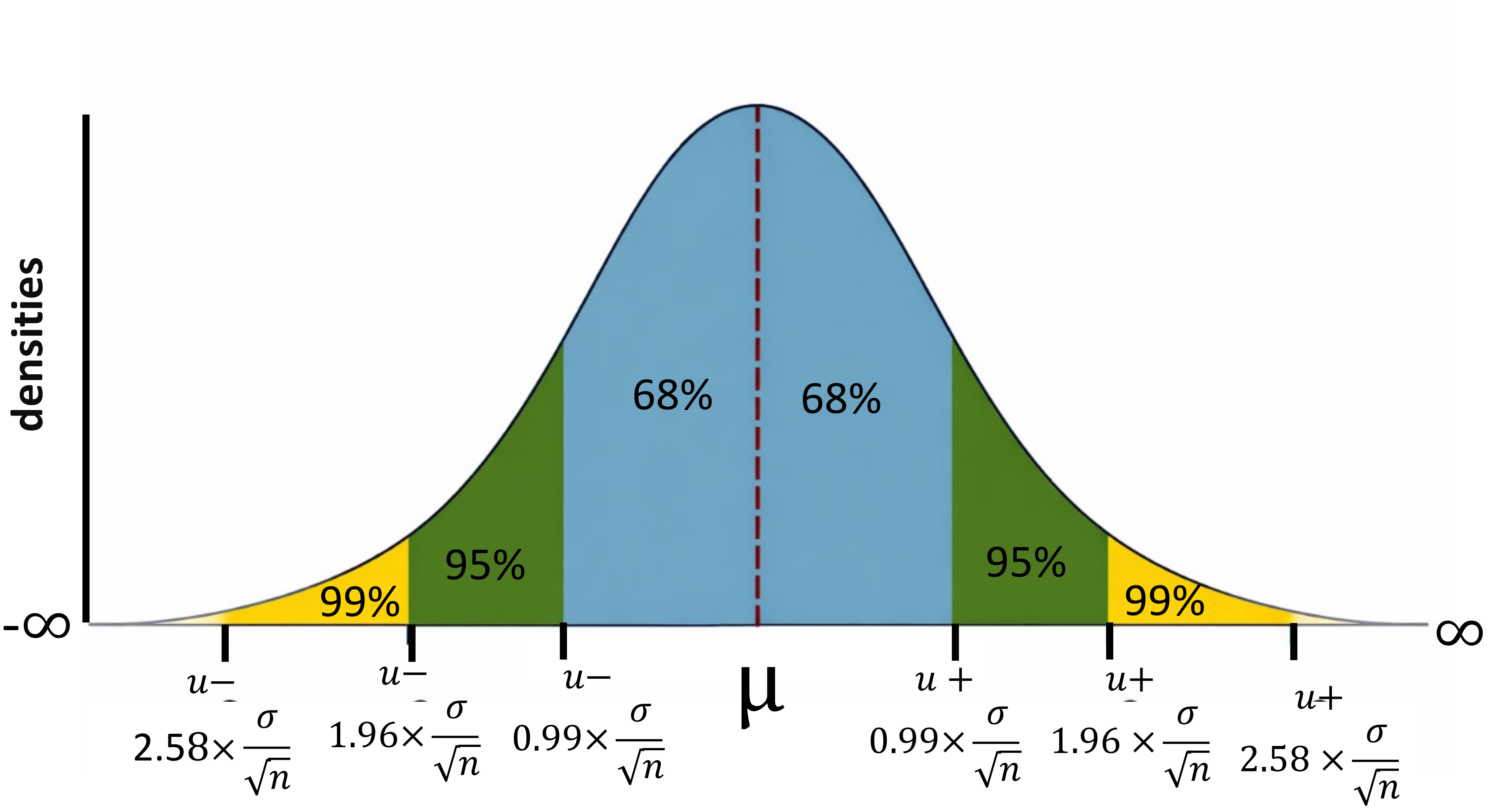
$$Confidence\ interval = \bar{X} \pm quantity * s_{\bar{X}}$$

[REVISED]

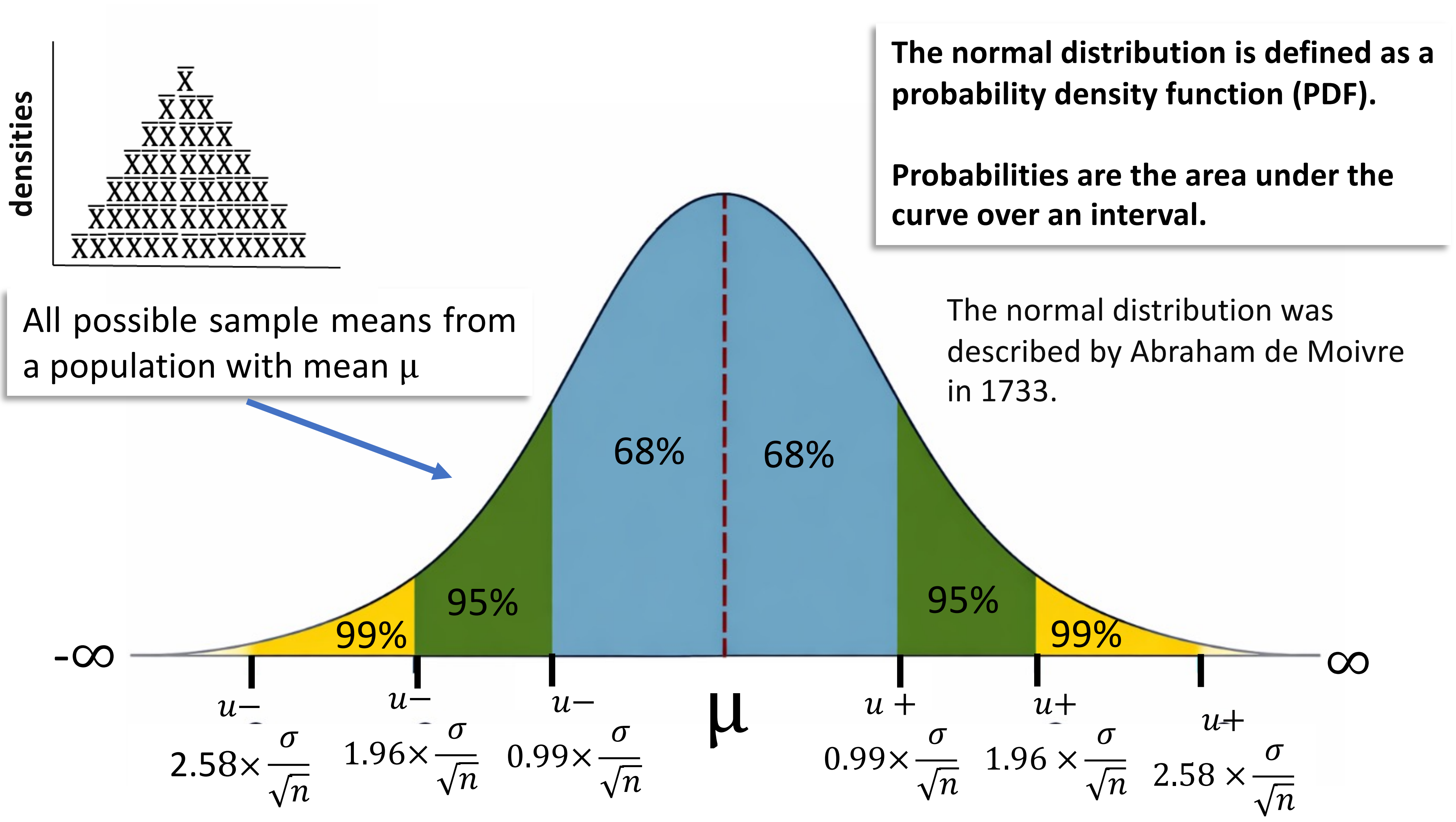
The logic behind confidence intervals before we see how the bias in the sample standard deviation is corrected and why that correction matters.

A normal distribution is a continuous probability distribution in which probabilities are represented by areas under a symmetric bell-shaped curve, and fixed percentages of the total area lie within specific distances (measured in standard deviations) from the mean.

When the population is normal (or when the sample size is large enough), the sampling distribution of the mean is normally distributed, centered at μ, with spread equal to the standard error σ/√n.


densities
-∞
∞
99%
95%
68%
68%
95%
99%
$u-2.58\times\frac{\sigma}{\sqrt{n}}$
$u-1.96\times\frac{\sigma}{\sqrt{n}}$
$u-0.99\times\frac{\sigma}{\sqrt{n}}$
μ
$u+0.99\times\frac{\sigma}{\sqrt{n}}$
$u+1.96\times\frac{\sigma}{\sqrt{n}}$
$u+2.58\times\frac{\sigma}{\sqrt{n}}$

The sampling distribution of the sample mean is ***normally distributed*** **when the population standard deviation is known and either the population itself is normally distributed or the sample size is sufficiently large**. Knowledge of the population mean is not required to determine the shape of the sampling distribution.



All possible sample means from a population with mean μ

The normal distribution is defined as a probability density function (PDF).

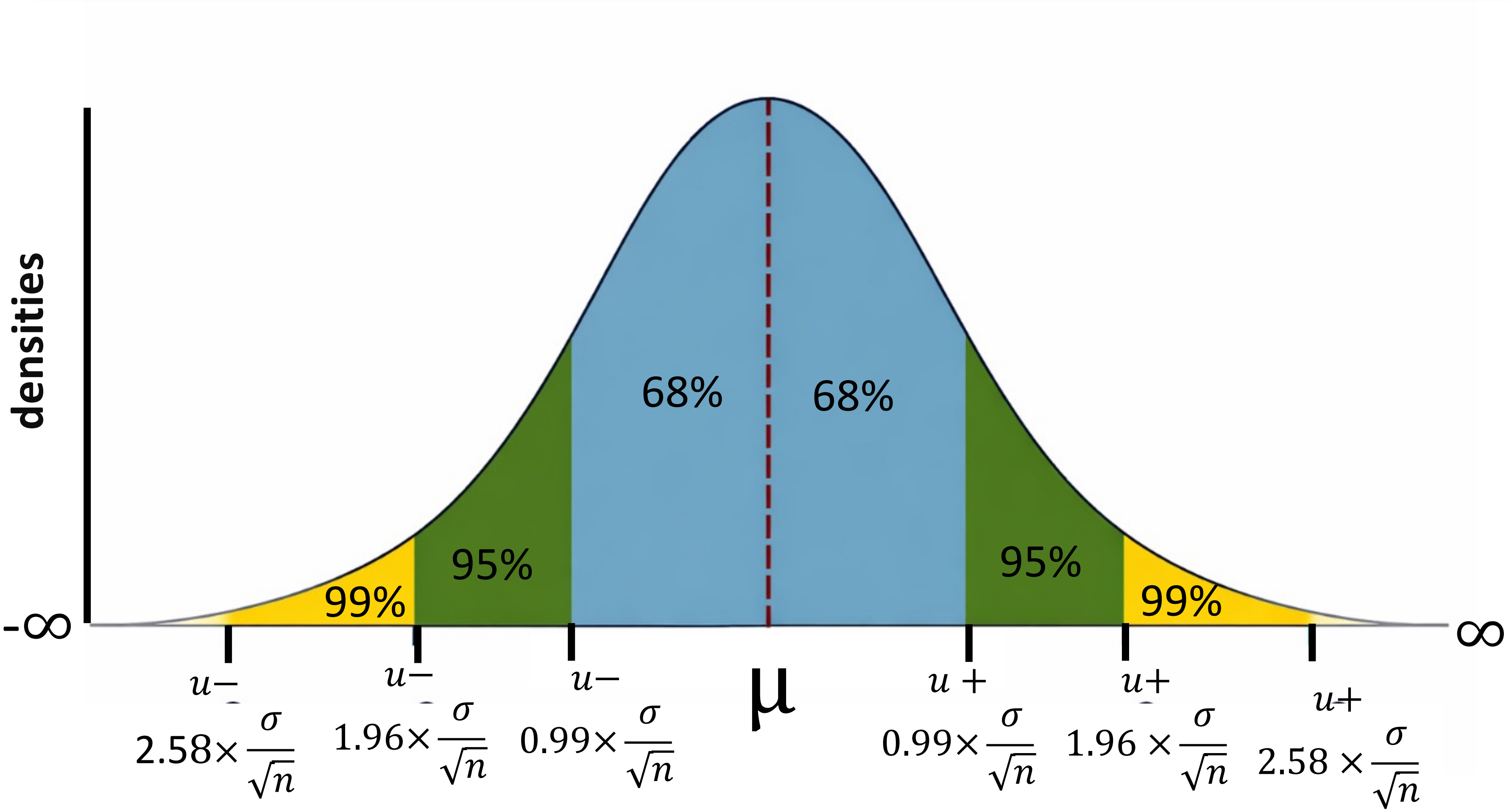
Probabilities are the area under the curve over an interval.

The normal distribution was described by Abraham de Moivre in 1733.

When the sampling distribution of the mean is normally distributed, 68% of all possible sample means lie within ±1.0 standard error (σ/√n) of the true population mean μ.

When the sampling distribution of the mean is normally distributed, 95% of all possible sample means lie within ±1.96 standard errors (σ/√n) of the true population mean μ.

When the sampling distribution of the mean is normally distributed, 99% of all possible sample means lie within ±2.576 standard errors (σ/√n) of the true population mean μ.


densities
-∞
∞
68%
68%
95%
95%
99%
99%
μ
$u- 2.58\times\frac{\sigma}{\sqrt{n}}$
$u- 1.96\times\frac{\sigma}{\sqrt{n}}$
$u- 0.99\times\frac{\sigma}{\sqrt{n}}$
$u + 0.99\times\frac{\sigma}{\sqrt{n}}$
$u+ 1.96\times\frac{\sigma}{\sqrt{n}}$
$u+ 2.58\times\frac{\sigma}{\sqrt{n}}$

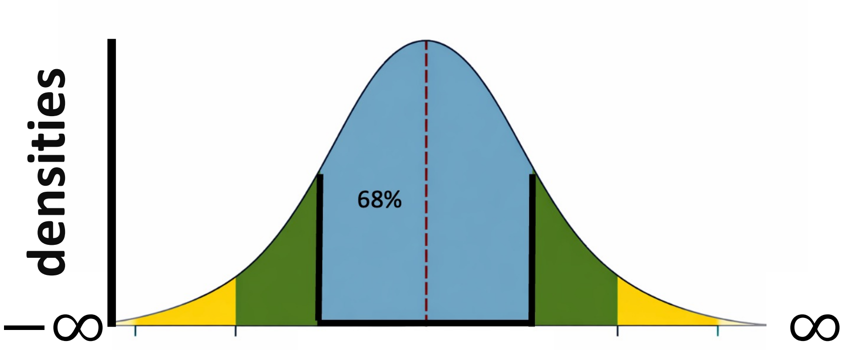
The sampling distribution of the means goes from $-\infty$ to ∞

How many possible samples of 100 trees out of 100000 trees?

1e+15 (zeros)

How many possible samples of 100 trees out of 1000000?

10768272362e+432 (zeros)

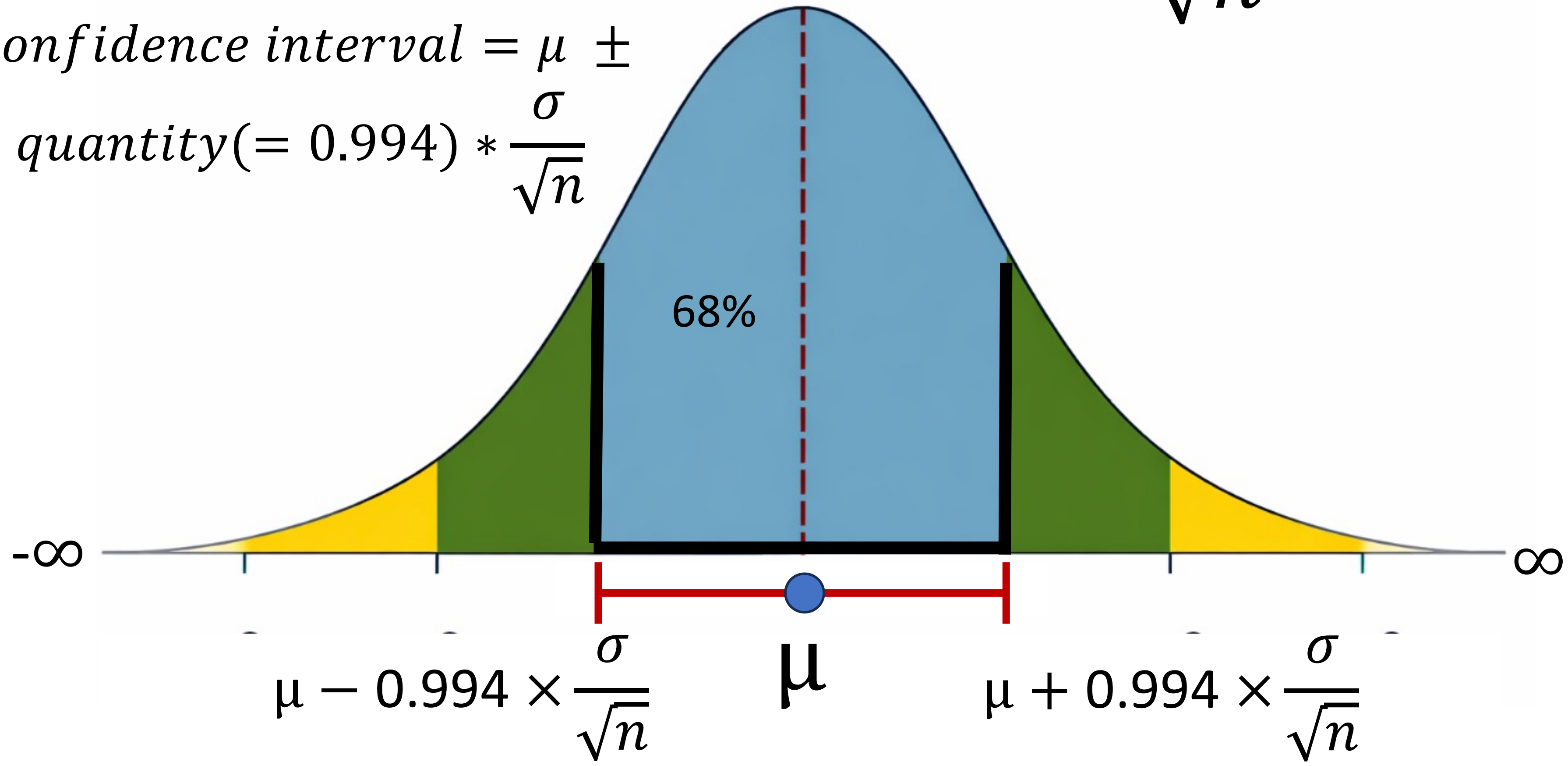


The **human body** consists of some 37.2 trillion **cells** (3.72e+13 zeros)

68% of all possible sample means (i.e., 68% of the area under the sampling distribution of the mean) lies within $\mu \pm 0.994 \times \sigma$, i.e., there is a 68% probability that a randomly drawn sampling from a population with mean μ and standard deviation σ will be within $0.994 \times \frac{\sigma}{\sqrt{n}}$ of the true mean, i.e., within the interval.

$$68\%\ \text{CI:}\ \mu \pm 0.994 \times \frac{\sigma}{\sqrt{n}}$$

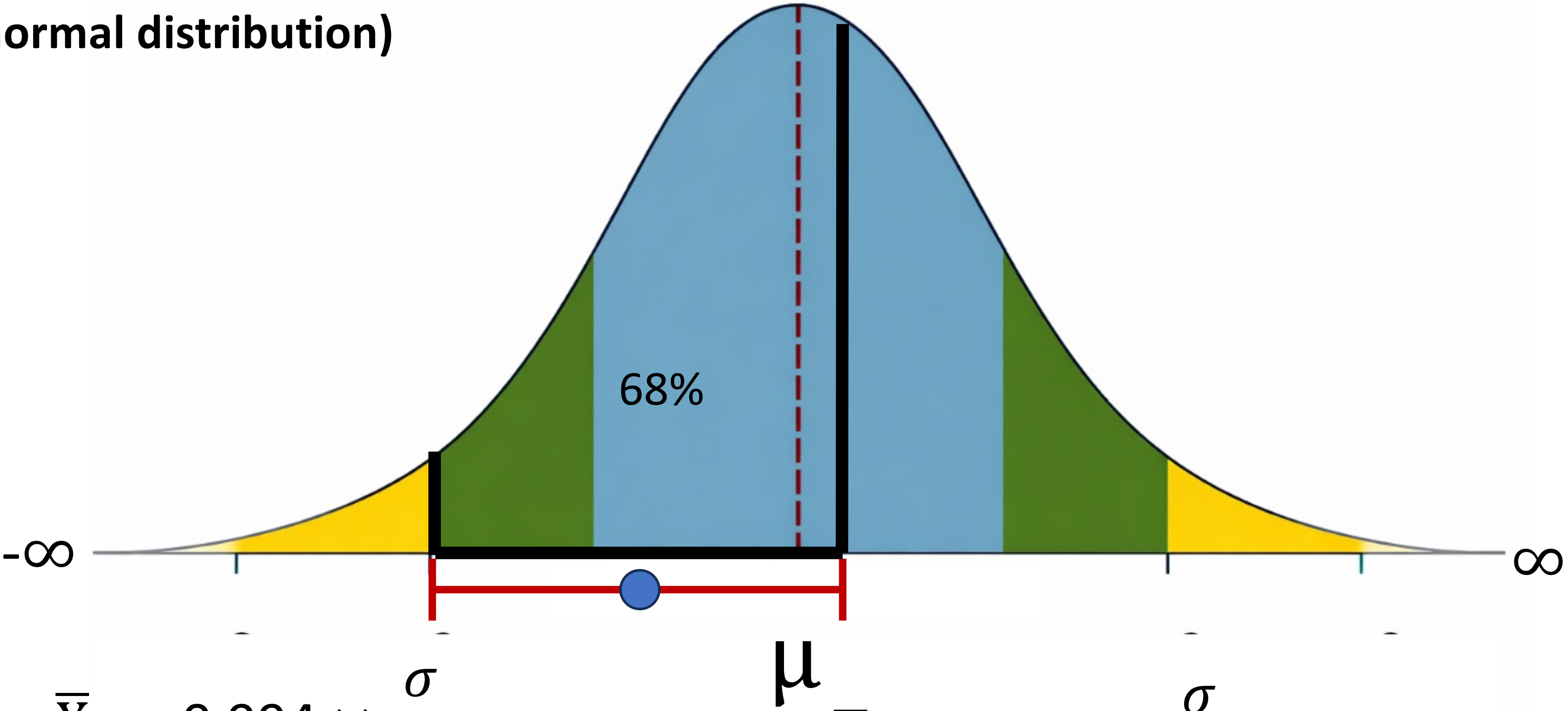
$$Confidence\ interval = \mu \pm quantity(= 0.994) * \frac{\sigma}{\sqrt{n}}$$


68%
-∞
∞
$\mu - 0.994 \times \frac{\sigma}{\sqrt{n}}$
μ
$\mu + 0.994 \times \frac{\sigma}{\sqrt{n}}$

68% of all possible sample means (i.e., 68% of the area under the sampling distribution of the mean) lies within $\overline{\mathrm{X}}_i \pm 0.994 \times \sigma, \mathrm{i.e.,}$ there is a 68% probability that a randomly drawn sampling from a population with mean μ and standard deviation σ will be within $0.994 \times \frac{\sigma}{\sqrt{n}}$ of the true mean, i.e., within the interval.

Invariance of a pivotal quantity under algebraic rearrangement (or simply symmetry of the normal distribution)

$$68\%\ \mathrm{CI:}\ \bar{X}_i \pm 0.994 \times \frac{\sigma}{\sqrt{n}}$$



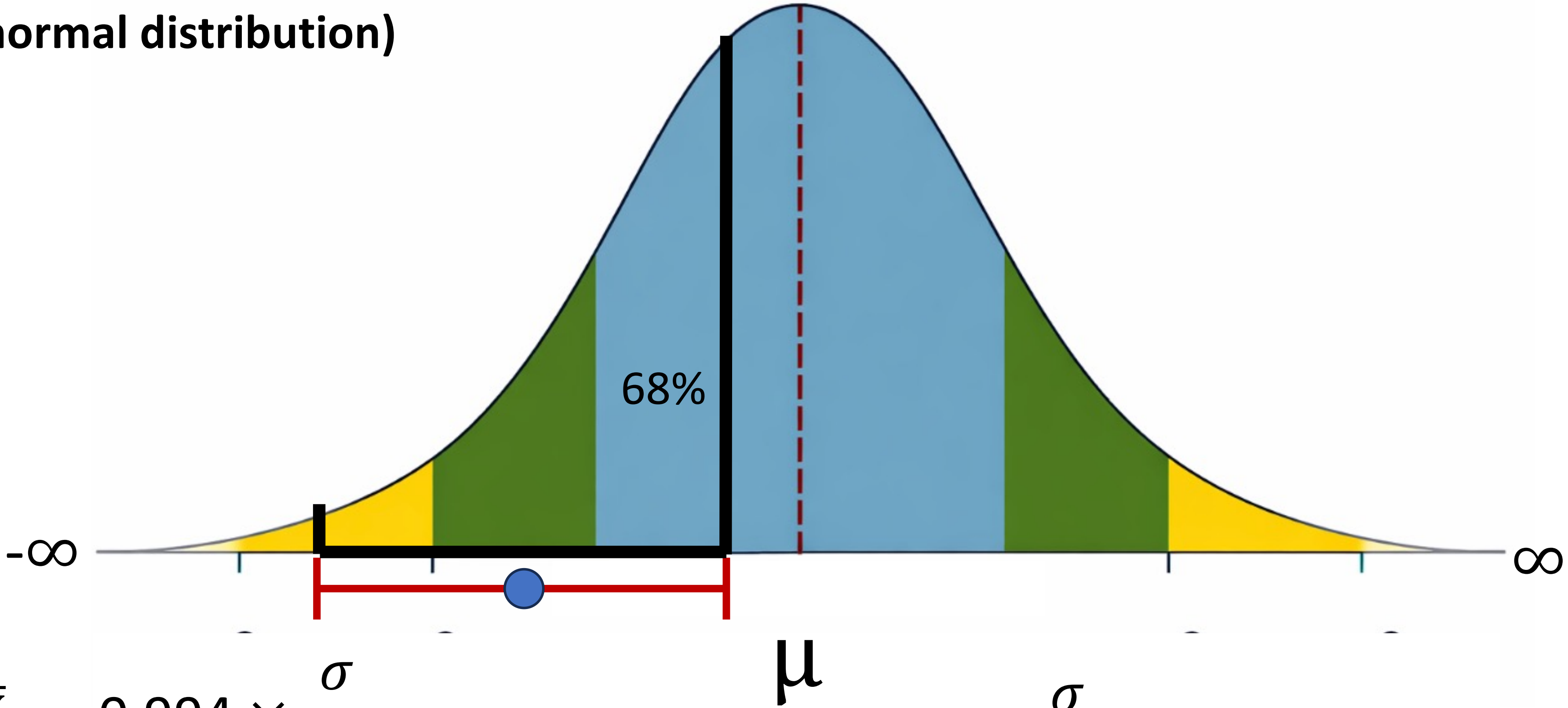
$$\overline{\mathrm{X}}_i - 0.994 \times \frac{\sigma}{\sqrt{n}}$$

$$\overline{\mathrm{X}}_i + 0.994 \times \frac{\sigma}{\sqrt{n}}$$

68% of all possible sample means (i.e., 68% of the area under the sampling distribution of the mean) lies within $\bar{\mathrm{X}}_i \pm 0.994 \times \sigma, \mathrm{i.e.},$ there is a 68% probability that a randomly drawn sampling from a population with mean μ and standard deviation σ will be within $0.994 \times \frac{\sigma}{\sqrt{n}}$ of the true mean, i.e., within the interval.

Invariance of a pivotal quantity under algebraic rearrangement (or simply symmetry of the normal distribution)

$$68\%\ \mathrm{CI:}\ \bar{X}_i \pm 0.994 \times \frac{\sigma}{\sqrt{n}}$$



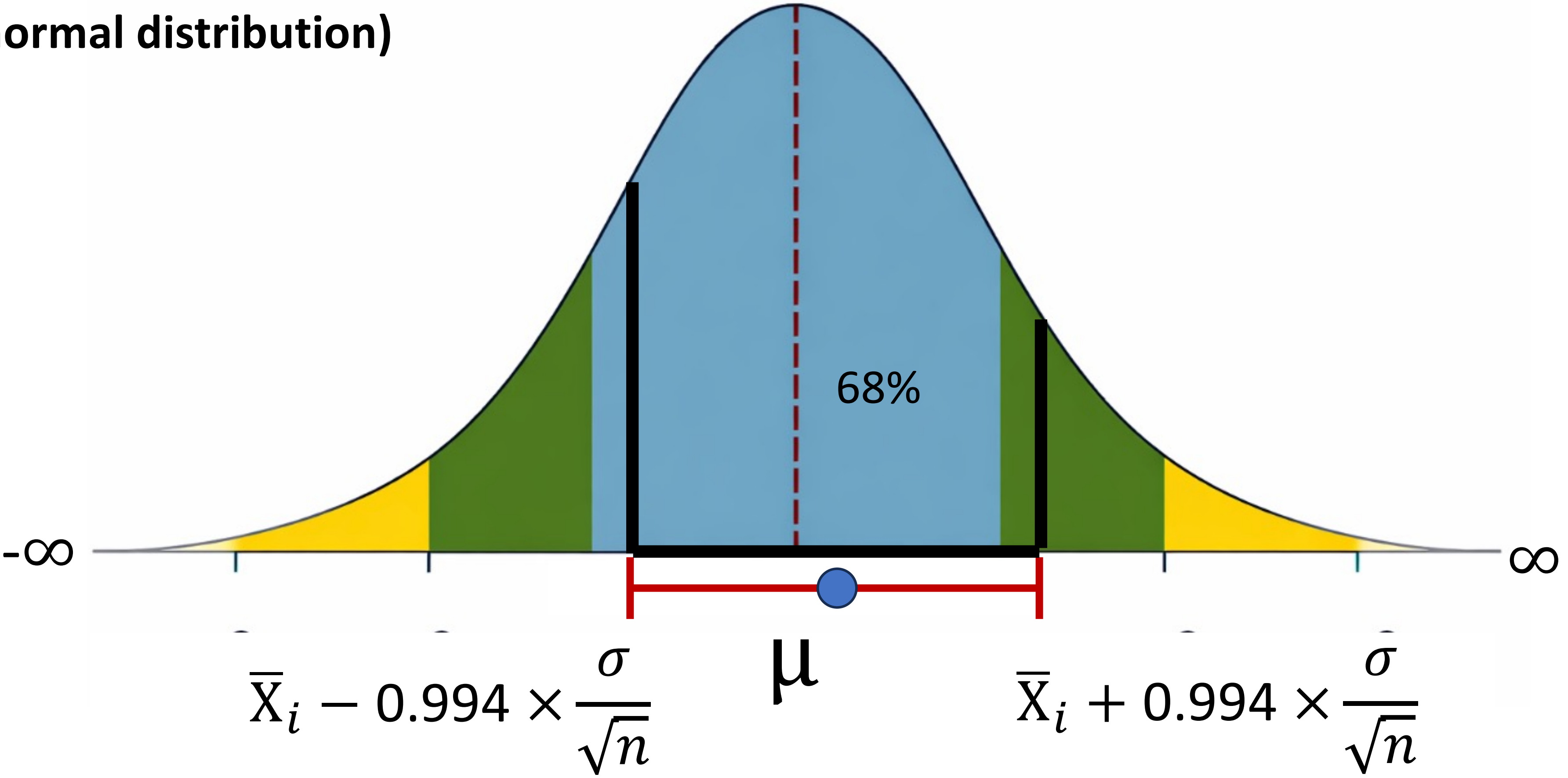
$$\bar{\mathrm{X}}_i - 0.994 \times \frac{\sigma}{\sqrt{n}}$$

$$\bar{\mathrm{X}}_i + 0.994 \times \frac{\sigma}{\sqrt{n}}$$

68% of all possible sample means (i.e., 68% of the area under the sampling distribution of the mean) lies within $\bar{\mathrm{X}}_i \pm 0.994 \times \sigma, \mathrm{i.e.,}$ there is a 68% probability that a randomly drawn sampling from a population with mean μ and standard deviation σ will be within $0.994 \times \frac{\sigma}{\sqrt{n}}$ of the true mean, i.e., within the interval.

Invariance of a pivotal quantity under algebraic rearrangement (or simply symmetry of the normal distribution)

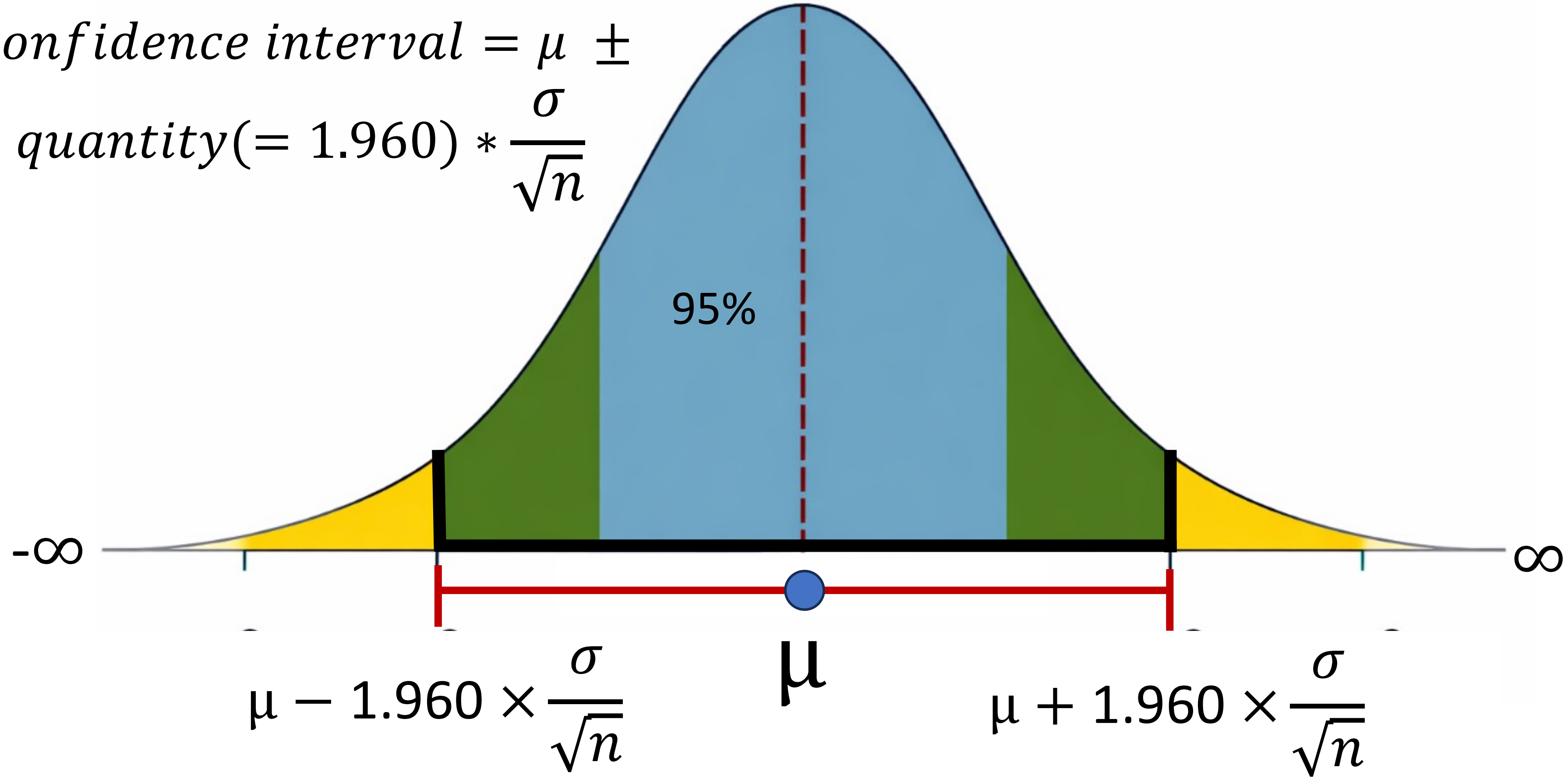
$$68\%\ \mathrm{CI:}\ \bar{X}_i \pm 0.994 \times \frac{\sigma}{\sqrt{n}}$$


68%
-∞
∞
$\bar{\mathrm{X}}_i - 0.994 \times \frac{\sigma}{\sqrt{n}}$
μ
$\bar{\mathrm{X}}_i + 0.994 \times \frac{\sigma}{\sqrt{n}}$

95% of all possible sample means (area under the curve) is within μ ± 1.960× σ

$$95\%\ \mathrm{CI}:\ \mu \pm 1.960 \times \frac{\sigma}{\sqrt{n}}$$

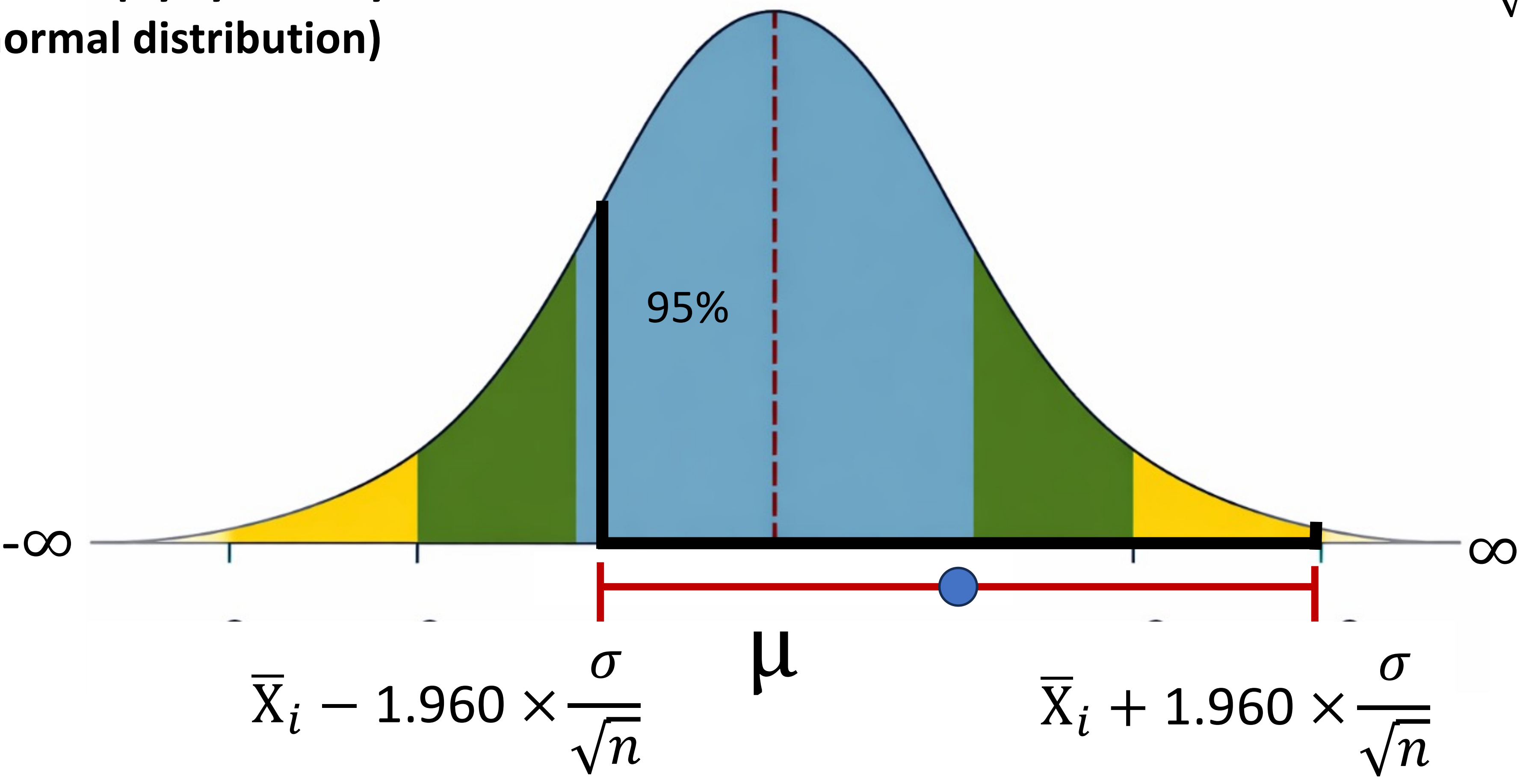
$$Confidence\ interval = \mu \pm quantity(= 1.960) * \frac{\sigma}{\sqrt{n}}$$


-∞
95%
∞
$\mu - 1.960 \times \frac{\sigma}{\sqrt{n}}$
μ
$\mu + 1.960 \times \frac{\sigma}{\sqrt{n}}$

95% of all possible sample means (i.e., 95% of the area under the sampling distribution of the mean) lies within $\overline{\mathrm{X}}_i \pm 1.960 \times \sigma$, i.e., there is a 95% probability that a randomly drawn sampling from a population with mean μ and standard deviation σ will be within $1.960\times \frac{\sigma}{\sqrt{n}}$ of the true mean, i.e., within the interval.

Invariance of a pivotal quantity under algebraic rearrangement (or simply symmetry of the normal distribution)

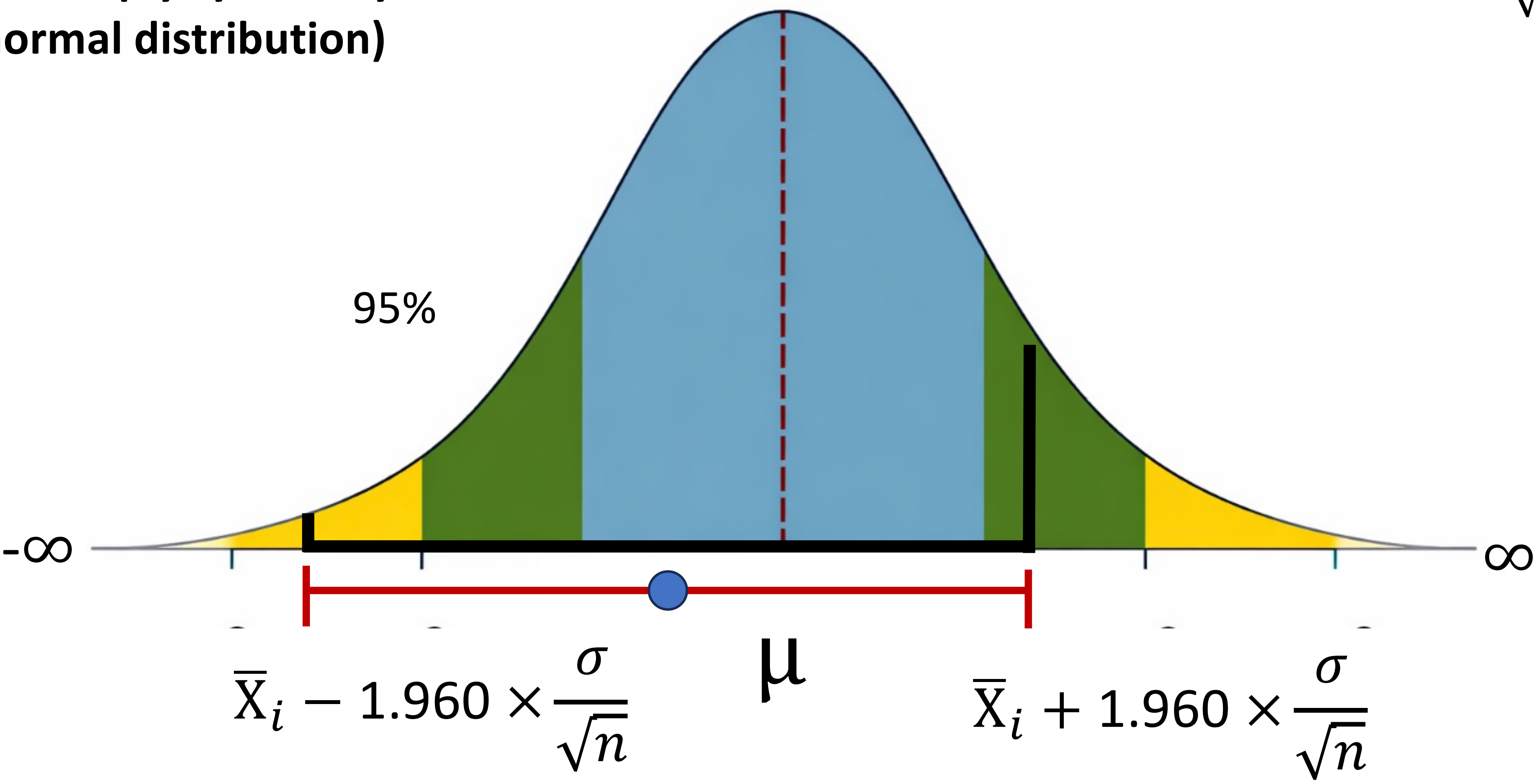
95% CI: $\overline{\mathrm{X}}_i \pm 1.960 \times \frac{\sigma}{\sqrt{n}}$


95%
-∞
∞
μ
$\overline{\mathrm{X}}_i - 1.960 \times \frac{\sigma}{\sqrt{n}}$
$\overline{\mathrm{X}}_i + 1.960 \times \frac{\sigma}{\sqrt{n}}$

95% of all possible sample means (i.e., 95% of the area under the sampling distribution of the mean) lies within $\overline{\mathrm{X}}_i \pm 1.960 \times \sigma$, i. e., there is a 95% probability that a randomly drawn sampling from a population with mean μ and standard deviation σ will be within $1.960\times \frac{\sigma}{\sqrt{n}}$ of the true mean, i.e., within the interval.

Invariance of a pivotal quantity under algebraic rearrangement (or simply symmetry of the normal distribution)

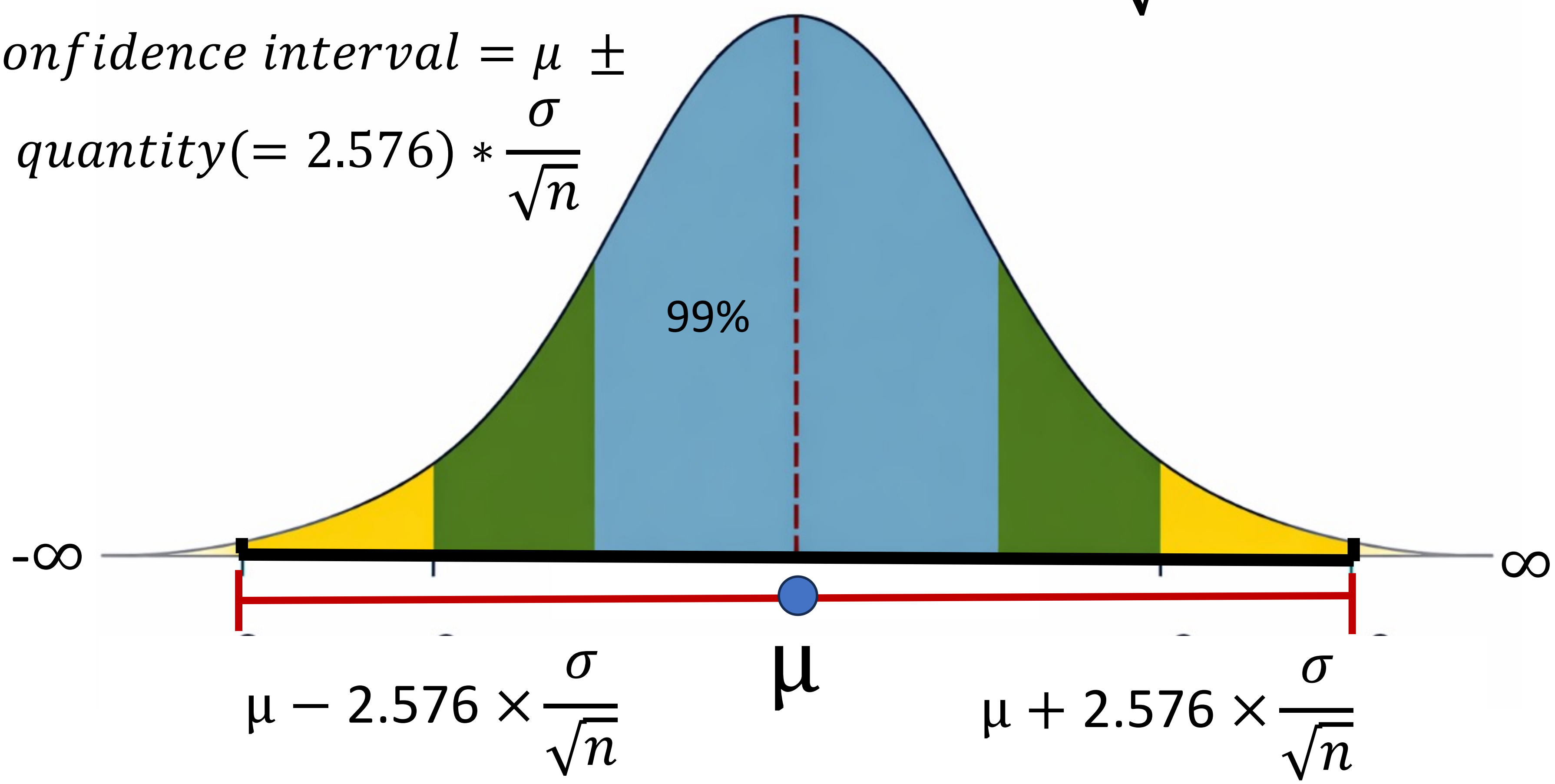
95% CI: $\overline{\mathrm{X}}_i \pm 1.960 \times \frac{\sigma}{\sqrt{n}}$


95%
-∞
∞
$\overline{\mathrm{X}}_i - 1.960 \times \frac{\sigma}{\sqrt{n}}$
μ
$\overline{\mathrm{X}}_i + 1.960 \times \frac{\sigma}{\sqrt{n}}$

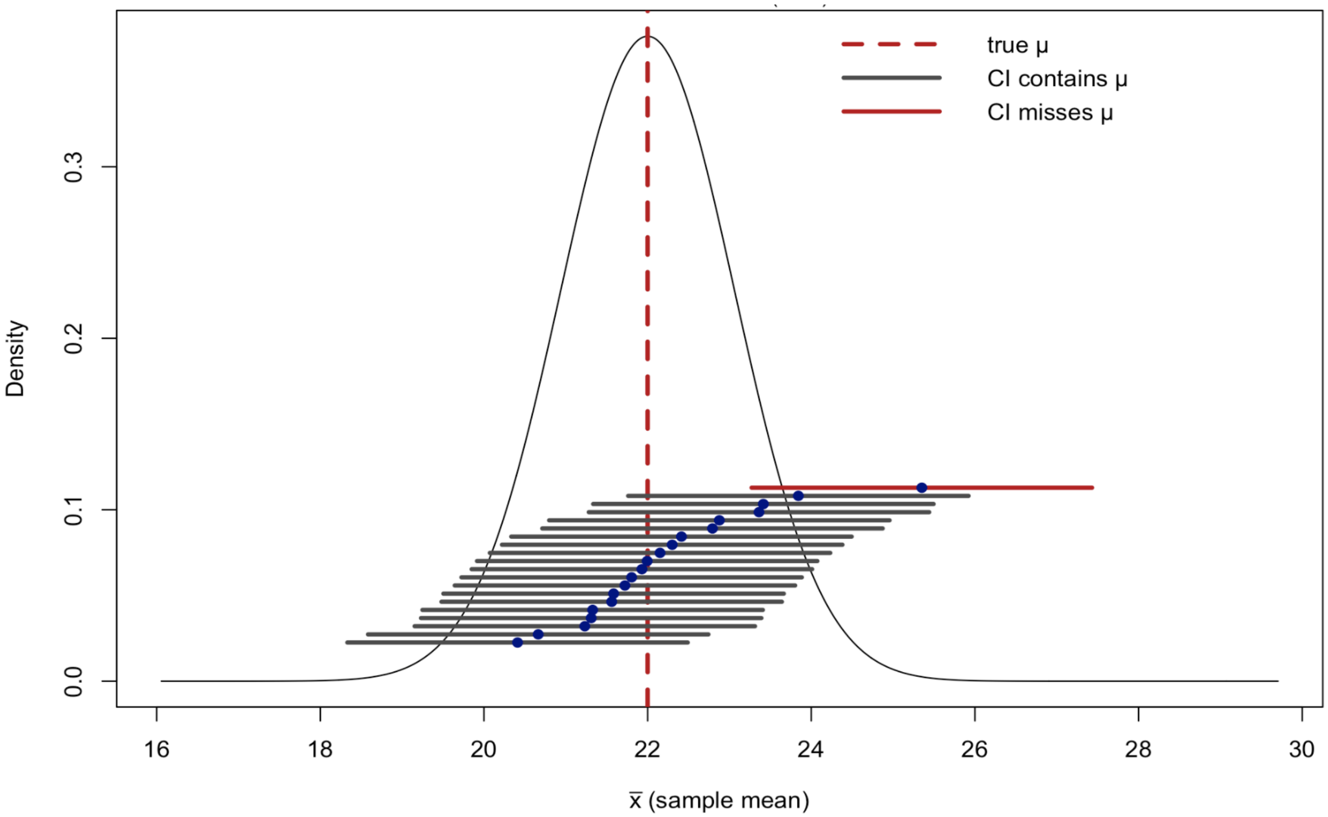
99% of all possible sample means (area under the curve) is within μ $\pm 2.576 \times \sigma$

$$99\%\ \text{CI: } \mu \pm 2.576 \times \frac{\sigma}{\sqrt{n}}$$

$$Confidence\ interval = \mu \pm quantity(= 2.576) * \frac{\sigma}{\sqrt{n}}$$


99%
-∞
∞
$\mu - 2.576 \times \frac{\sigma}{\sqrt{n}}$
μ
$\mu + 2.576 \times \frac{\sigma}{\sqrt{n}}$

95% of all possible sample means (i.e., 95% of the area under the sampling distribution of the mean) lies within $\overline{\mathrm{X}}_i \pm 1.960 \times \sigma, \mathrm{i.e.,}$ there is a 95% probability that a randomly drawn sampling from a population with mean μ and standard deviation σ will be within $1.960\times \frac{\sigma}{\sqrt{n}}$ of the true mean, i.e., within the interval.



Note that because $1.960\times \frac{\sigma}{\sqrt{n}}$ doesn't change, the interval is contant (same size); position changes as a function of the sample mean only.

Note that because of the pivot symmetry, these intervals were calculated for sample means, i.e., we don't need to know the population mean μ but we need to know the population standard deviation σ.

[REVISED]

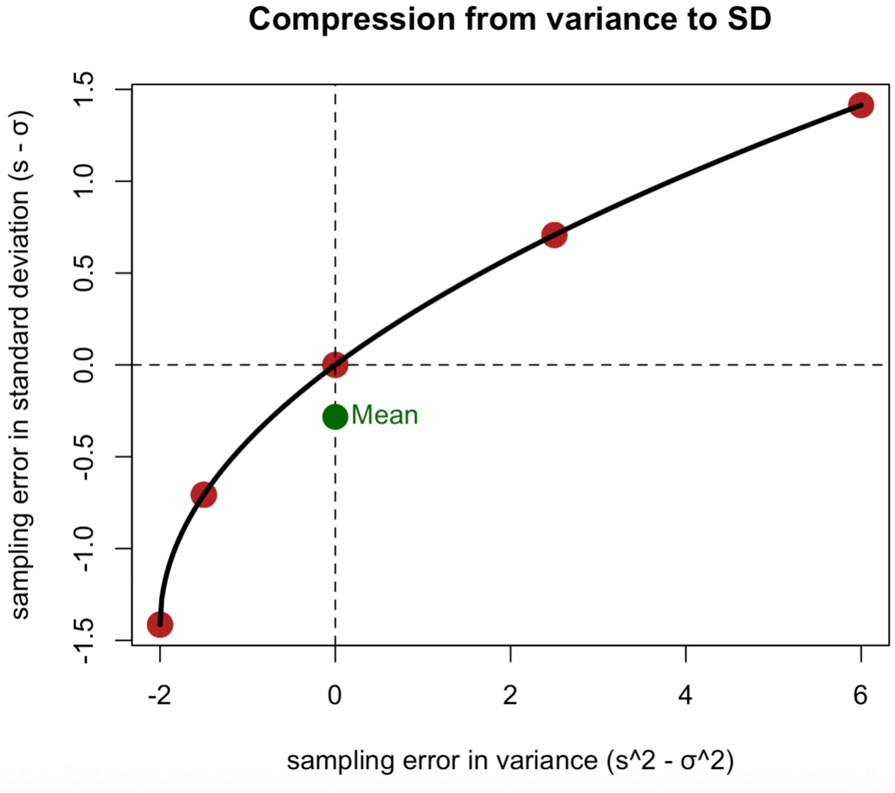
CONCLUSION: Because of the pivot symmetry, we don’t need to know the population mean μ but we need to know the population standard deviation σ; BUT the standard deviation of the sample is biased due to the Jensen’s inequality.

The interval below would work, but we don’t know the true population standard deviation.

$$\mathrm{CI} = \mu \pm 1.960 * \frac{\sigma}{\sqrt{n}}$$

The interval below won’t work because the sample-based standard deviation is biased.

$$\mathrm{CI} = \mu \pm 1.960 * \frac{s_i}{\sqrt{n}}$$

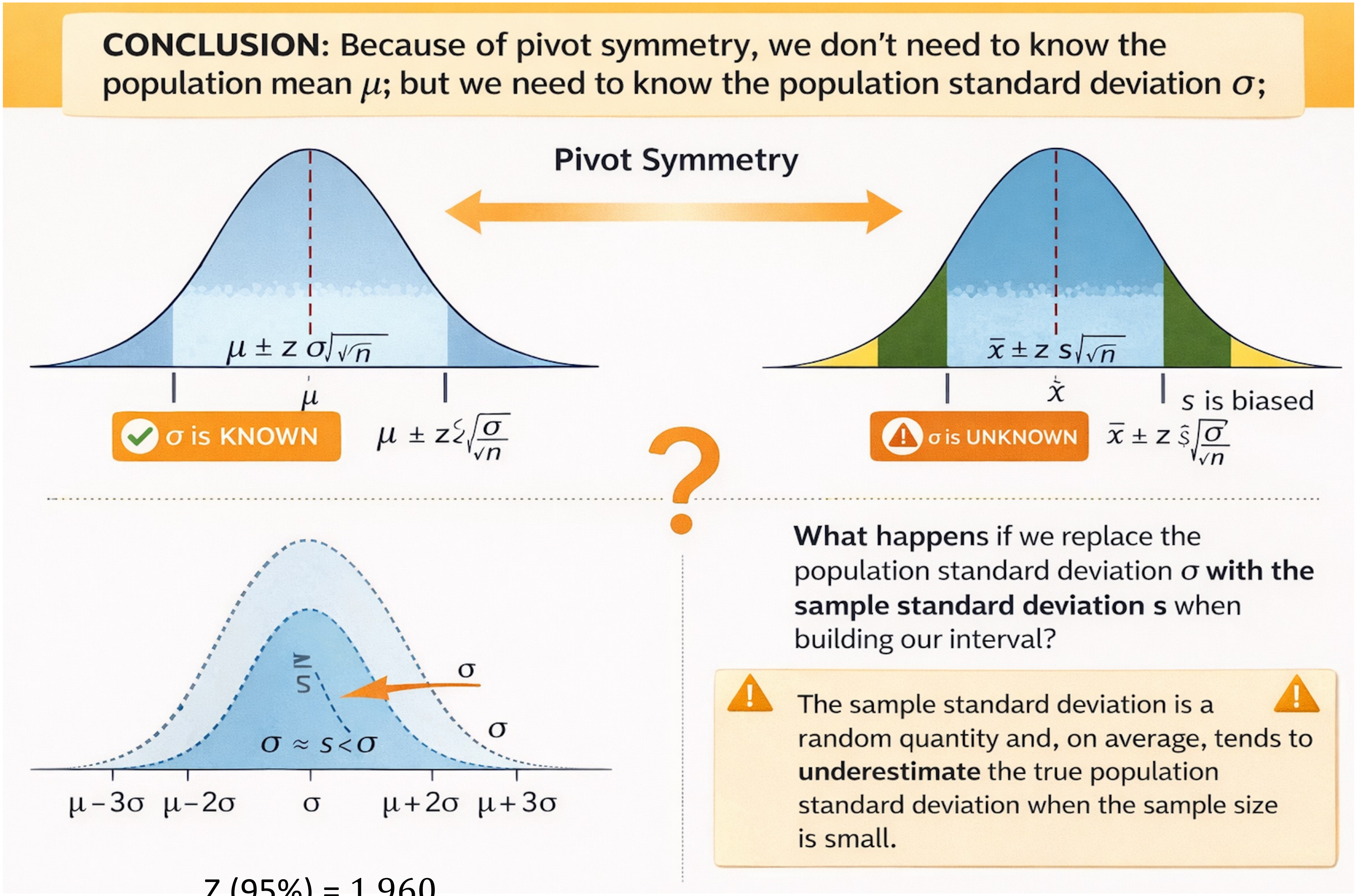

Compression from variance to SD

Remember that the sample standard deviation is a random quantity and, on average, tends to underestimate the true population standard deviation when the sample size is small.

So, what happens if we replace the population standard deviation with the sample standard deviation when building our interval?????

[REVISED]

So, what happens if we replace the population standard deviation with the sample standard deviation when building our interval?????



CONCLUSION: Because of pivot symmetry, we don't need to know the population mean μ; but we need to know the population standard deviation σ;

What happens if we replace the population standard deviation σ **with the sample standard deviation s** when building our interval?

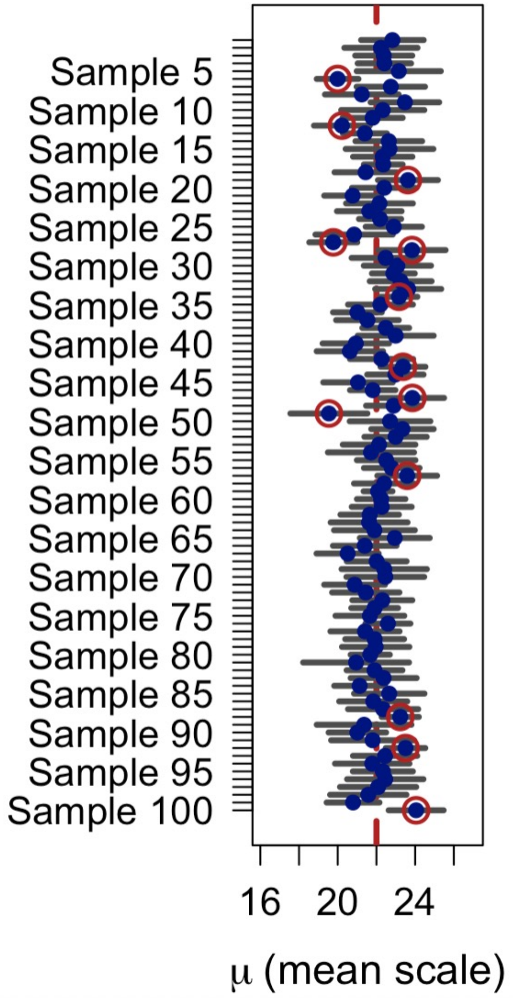
The sample standard deviation is a random quantity and, on average, tends to **underestimate** the true population standard deviation when the sample size is small.

Pivot Symmetry
$\mu \pm z\,\sigma\sqrt{\sqrt{n}}$
μ
σ is KNOWN
$\mu \pm z\xi\sqrt{\frac{\sigma}{\sqrt{n}}}$
$\bar{x} \pm z\,s\sqrt{\sqrt{n}}$
$\bar{x}$
s is biased
σ is UNKNOWN
$\bar{x} \pm z\,\hat{s}\sqrt{\frac{\sigma}{\sqrt{n}}}$
?
$s \leq$
σ
σ
$\sigma \approx s < \sigma$
μ−3σ
μ−2σ
σ
μ+2σ
μ+3σ

Z (95%) = 1.960

Remember that the sample standard deviation that, on average, tends to underestimate the true population standard deviation. It underestimates even more when sample sizes are smaller (Jenzen's inequality).

So, what happens if we replace the population standard deviation with the sample standard deviation when building our interval still assuming a normal distribution for the sampling distribution of the mean?



👎 $$\text{CI} = \mu \pm 1.960 * \frac{s_i}{\sqrt{n}}$$

This interval won't work because the sample-based standard deviation is biased.

For a 95% confidence interval, we expect that only 5% of intervals will fail to include the true population mean.

However, if the sample standard deviation underestimates the true population standard deviation, especially when the sample size is small (Jenzen's inequality), the resulting intervals will, on average, be too narrow.

As a consequence, more than 5% of such intervals will fail to contain the true population mean.

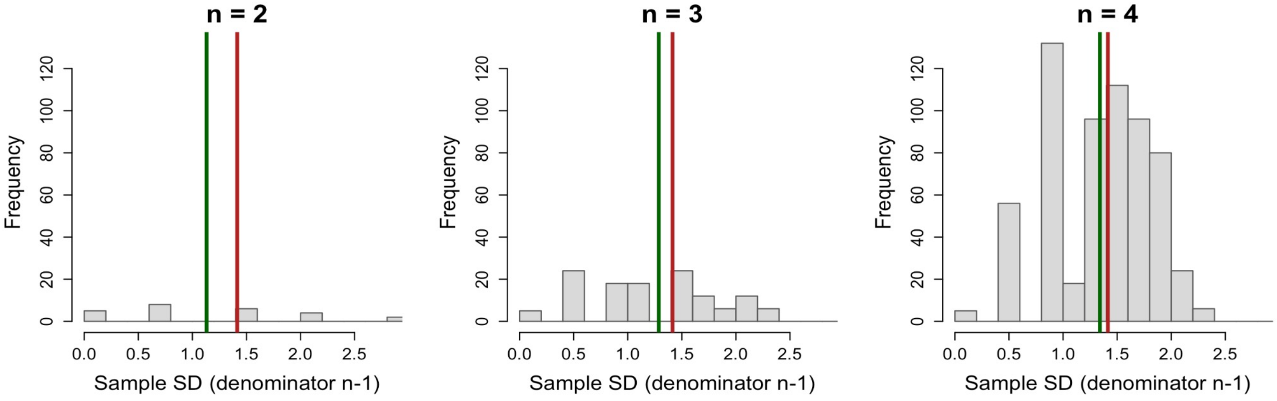
[REVISED]

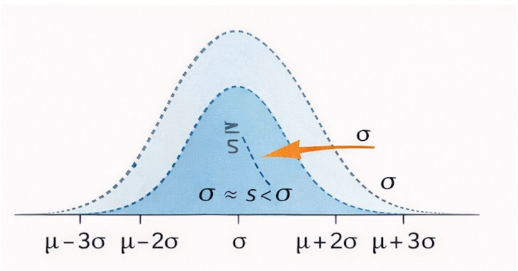
However, if the sample standard deviation underestimates the true population standard deviation, especially when the sample size is small (Jenzen's inequality), the resulting intervals will, on average, be too narrow.

As a consequence, more than 5% of such intervals will fail to contain the true population mean.

🟥 True population standard deviation (1.414214)

🟩 Mean of sample standard deviations (n=2: 1.131371; n=3: 1.287381; n=4: 1.340293)





$$\mathbb{E}\left[\mathbf{width}_{\mathbf{sample.based}}\right] < \mathbb{E}\left[\mathbf{width}_{\mathbf{population.based}}\right]$$

Sample-based (unknown σ)

Population-based (known σ)

$$\mathrm{CI} = \bar{X}_i \pm 1.960 * \frac{s_i}{\sqrt{n}} \xrightarrow{?????} \mathrm{CI} = \bar{X}_i \pm 1.960 * \frac{\sigma}{\sqrt{n}}$$

[REVISED]

How to adjust confidence intervals for the biases in sample-based standard deviations?

Now imagine the following computational approach (which is approximated with calculus in practice):

1) We repeatedly (10,000,000) take samples of size *n* from a normally distributed population

2) For each sample, we calculate the sample mean $\bar{X}_i$ the sample standard error of the mean $s_{\bar{X}_i} = \frac{s_i}{\sqrt{n}}$.

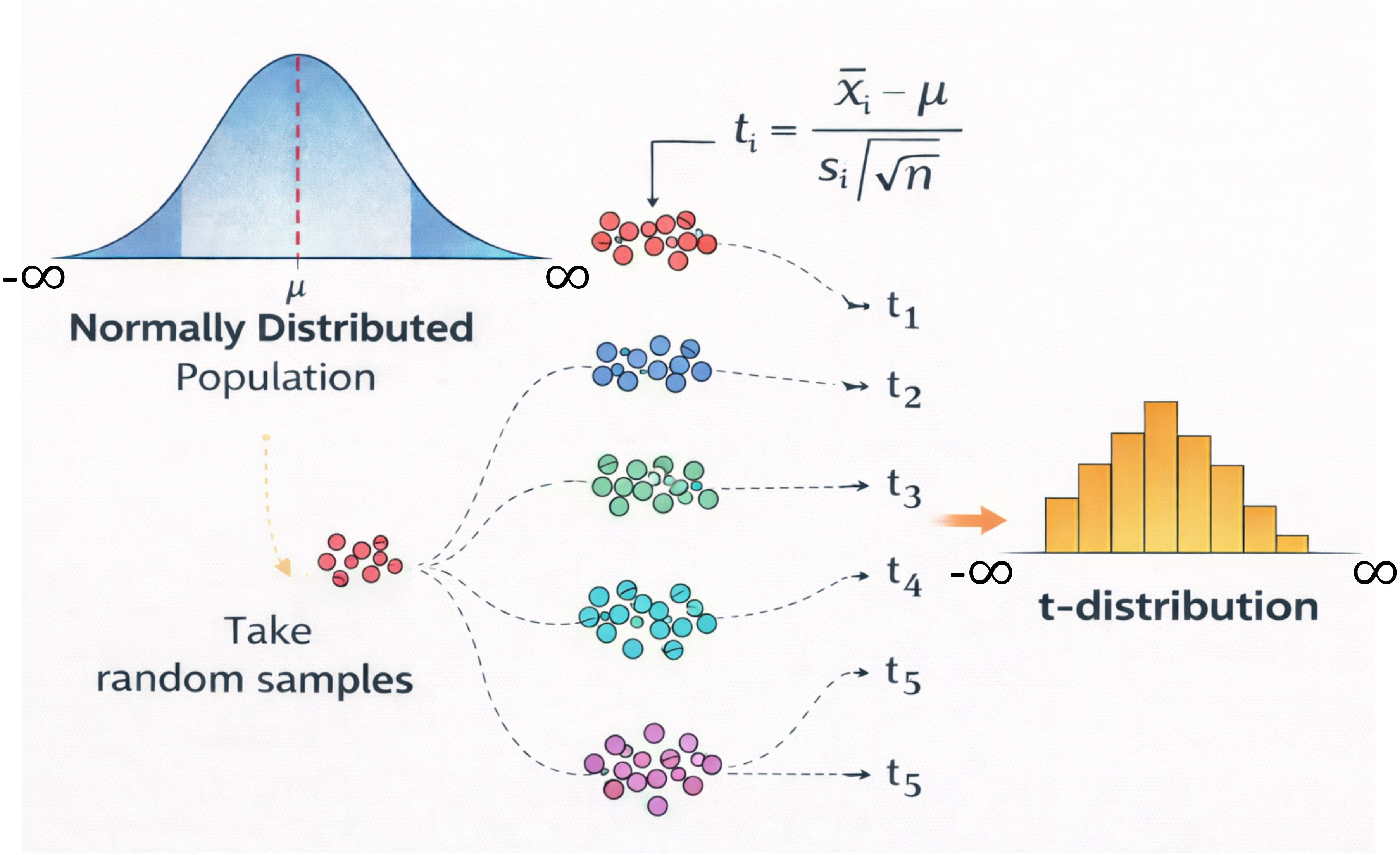
3) We then subtract the sample $\bar{X}_i$ by the population mean μ and divide the result by the sample standard error of the mean.

4) Let's call this quantity the t value for the i^{th} sample: $t_i = \frac{\bar{X}_i - \mu}{\frac{s_i}{\sqrt{n}}}$

5) Combine (pool) all 10,000,000 computed t-values, one from each generated sample, and plot their distribution as a probability density function (e.g., a histogram scaled to density or a smooth density curve). The resulting density is the t-distribution (with degrees of freedom (later in the course) = n-1, because s_i is estimated as:

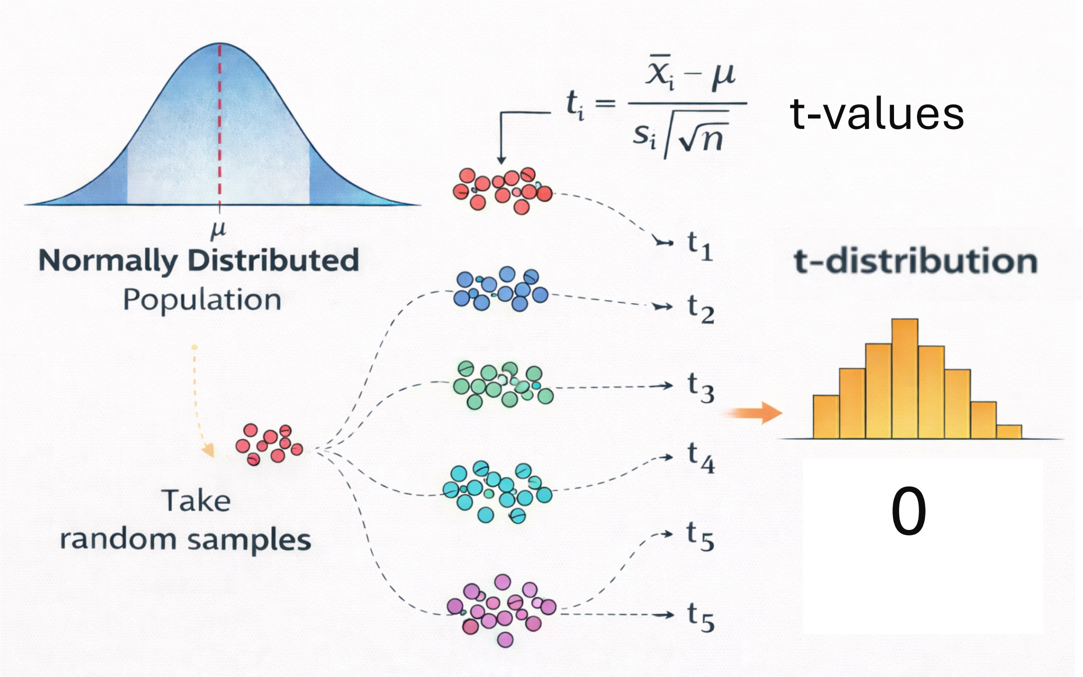
$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

The t-distribution - developed analytically by *student* (William Sealy Gosset) in 1908.



Some t-values are negative (when sample mean is smaller than the population mean μ) and others are positive (when sample mean is greater than than μ).

What is the mean of the t-distribution?


$t_i = \frac{\bar{x}_i - \mu}{s_i / \sqrt{n}}$
t-values
μ
Normally Distributed Population
Take random samples
t_1
t_2
t_3
t_4
t_5
t_5
t-distribution
0

Obs 1	Obs 2	Sample means	Sample means - μ
1	1	1.0	-2
1	2	1.5	-1.5
1	3	2.0	-1
1	4	2.5	-0.5
1	5	3.0	0
2	1	1.5	-1.5
2	2	2.0	-1
2	3	2.5	-0.5
2	4	3.0	0
2	5	3.5	0.5
3	1	2.0	-1
3	2	2.5	-0.5
3	3	3.0	0
3	4	3.5	0.5
3	5	4.0	1
4	1	2.5	-0.5
4	2	3.0	0
4	3	3.5	0.5
4	4	4.0	1
4	5	4.5	1.5
5	1	3.0	0
5	2	3.5	0.5
5	3	4.0	1
5	4	4.5	1.5
5	5	5.0	2
	MEAN	**3.0**	**0**

For any given sample *i*, its confidence interval is:

$$\mathrm{CI}_i = \bar{X}_i \pm t * \frac{s_i}{\sqrt{n}}$$

[REVISED]

Because we subtract the true population mean, the t-statistic measures the deviation of the sample mean from μ, regardless of the numerical value of μ.

→ **This removes location: the distribution is centered at zero.**

Because we divide by the sample standard error $\frac{s_i}{\sqrt{n}}$, the deviation is expressed in units of estimated variability.

→ **This removes scale: the statistic becomes unit-free and comparable across populations.**

What remains is a standardized quantity whose distribution depends only on sample size (or more precisely, degrees of freedom), not on μ or σ.

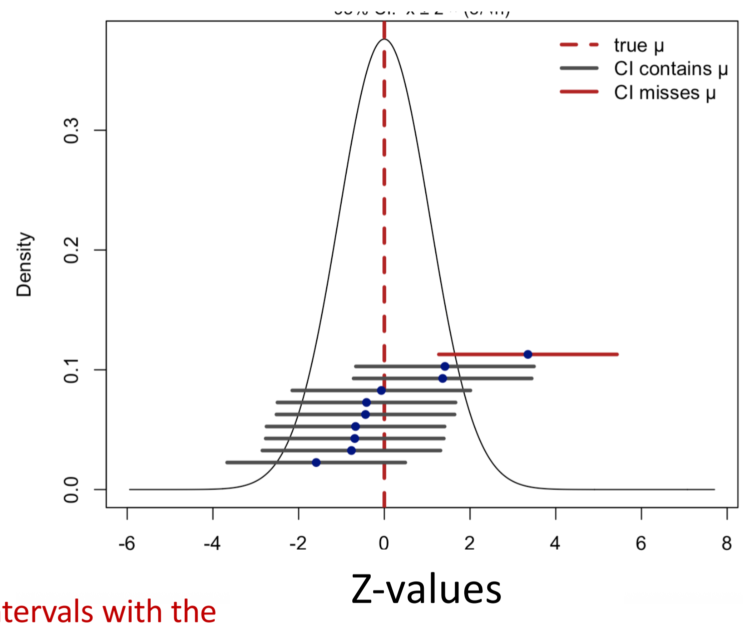
The same t value (e.g., ±2.365 for df=7) can be used regardless of whether μ = 10, 100, or 10,000 or what the standard deviation is.

→ **The t-distribution is UNIVERSAL and only varies as a function of sample size!**

The sampling distribution of means that varies as a function of the sample size (here df = degrees of freedom; df = n - 1) is t-distributed.

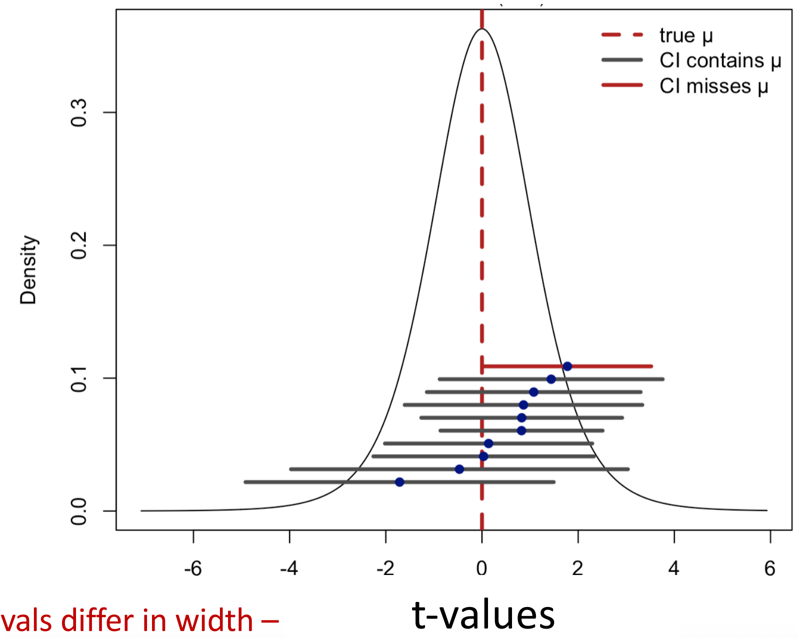
The exact multiplier applied to the sample standard error to construct a 95% confidence interval based on the t-distribution changes with sample size (df).

Sampling distribution of the mean
(normally distributed – infinite sample sizes)



All intervals with the same width – single σ

Sampling distribution of the mean
(t- distributed – non-infinite sample sizes)



Intervals differ in width – sample-based

$$\text{XX\% CI: } \overline{\mathrm{X}}_i \pm \mathrm{Z} \times \frac{\sigma}{\sqrt{n}}$$

$$\text{95\% CI (Z = 1.960): } \overline{\mathrm{X}}_i \pm \mathbf{1.960} \times \frac{\sigma}{\sqrt{n}}$$

$$\text{XX\% CI: } \overline{\mathrm{X}}_i \pm t.df \times \frac{s_i}{\sqrt{n}}$$

$$\text{95\% CI: } \overline{\mathrm{X}}_i \pm \boldsymbol{t.df} \times \frac{s_i}{\sqrt{n}}$$

[REVISED]

By now, we have seen in this lecture three types of confidence intervals:

👍 $$\mathrm{CI}(95\%) = \bar{X}_i \pm 1.960 * \frac{\sigma}{\sqrt{n}}$$

95% Confidence interval (CI) for the population mean based on any sample mean *i* $\bar{X}_i$ with known population standard deviation σ. This is correct but we never really know σ in applied situations.

👎 $$\mathrm{CI}(95\%) = \bar{X}_i \pm 1.960 * \frac{s_i}{\sqrt{n}}$$

95% Confidence interval (CI) for the population mean based on any sample mean *i* $\bar{X}_i$ with unknown population standard deviation σ and, therefore, based on its (sample(standard deviation s_i. This is incorrect leading to small coverage than anticipated (i.e., less than 95% of the CIs will cover the true population mean. This was shown for pedagogical purposes only, i.e., to show that we need another solution because this won't work. v

$$\mathbb{E}\left[\mathbf{width}_{\mathbf{sample.based}}\right] < \mathbb{E}\left[\mathbf{width}_{\mathbf{population.based}}\right]$$

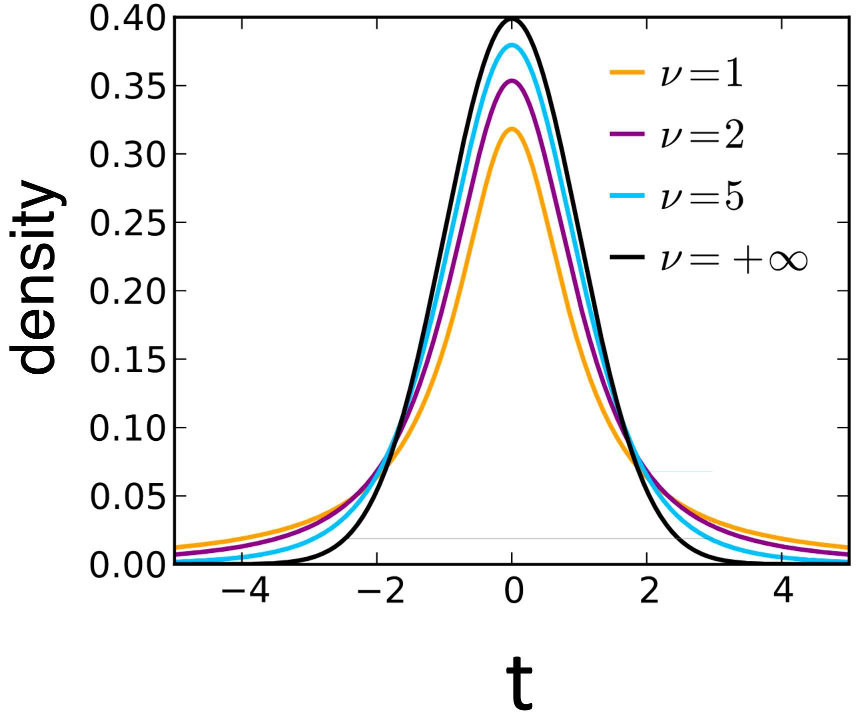
👍 $$\mathrm{CI}(95\%) = \bar{\mathrm{X}}_i \pm \boldsymbol{t.df} \times \frac{s_i}{\sqrt{n}}$$

95% Confidence interval (CI) for the population mean based on any sample mean *i* $\bar{X}_i$ with unknown population standard deviation σ and, therefore, based on its (sample(standard deviation s_i. This is correct because uses the t-distribution and, as such, won't be biased when using the sample standard deviation s_i.

[NEW]

The exact multiplier applied to the sample standard error to construct a 95% confidence interval depends on the sample size.

The sampling distribution of means that varies as a function of the sample size (here v = degrees of freedom; v = n - 1) is t-distributed.

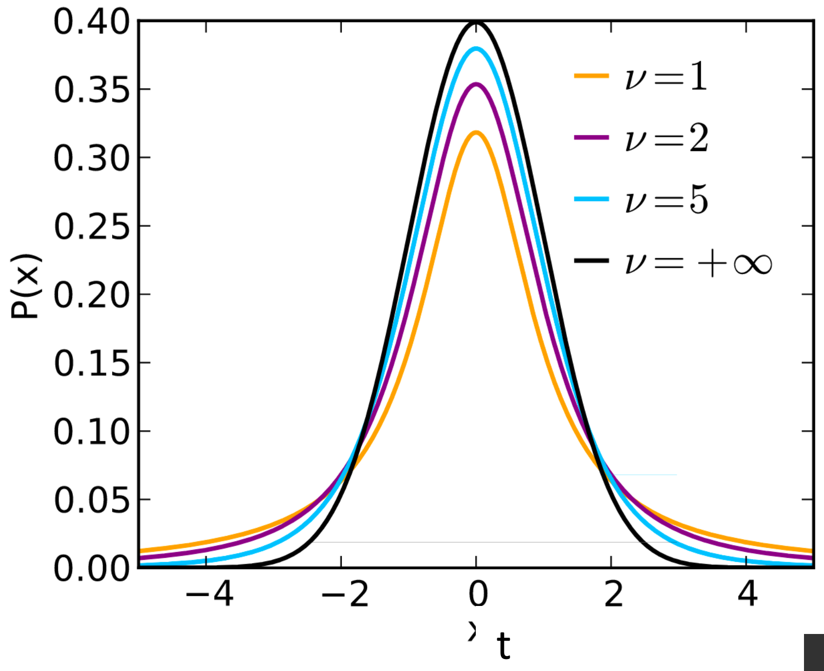


v = ∞ → t = Z → the t distribution becomes normally distributed when sample size is infinite.

$$95\%\ \text{CI}: \overline{\text{X}}_i \pm \boldsymbol{t}_{df} \times \frac{\sigma}{\sqrt{n}}$$

Extreme t-values occur more often than they would under the normal model. Graphically, the curve looks slightly flatter in the center and thicker in the tails. That is what “heavier tails” means: a higher probability of observing values far from zero compared to the normal distribution.

[NEW]



v = ∞ → t = Z → the t distribution becomes normally distributed when sample size is infinite.

$$95\%\ \text{CI}: \overline{\text{X}}_i \pm \boldsymbol{t_{df}} \times \frac{\sigma}{\sqrt{n}}$$

$$95\%\ \text{CI}\ (\text{Z} = 1.960): \overline{\text{X}}_i \pm \mathbf{1.960} \times \frac{\sigma}{\sqrt{n}}$$

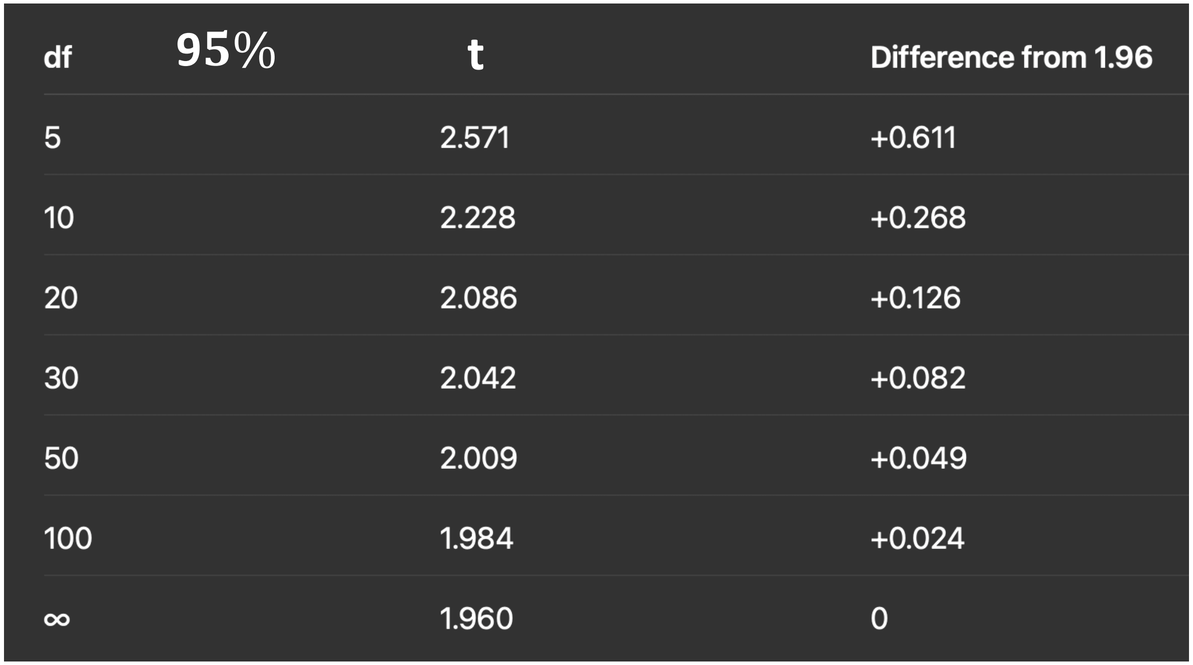
Interval width

$$95\%\ \text{CI}: \overline{\text{X}}_i \pm \mathbf{2.571} \times \frac{s}{\sqrt{n}}$$

$$95\%\ \text{CI}: \overline{\text{X}}_i \pm \mathbf{2.086} \times \frac{s}{\sqrt{n}}$$

$$95\%\ \text{CI}: \overline{\text{X}}_i \pm \mathbf{2.009} \times \frac{s}{\sqrt{n}}$$

$$95\%\ \text{CI}: \overline{\text{X}}_i \pm \mathbf{1.960} \times \frac{s}{\sqrt{n}}$$



df	95%	t	Difference from 1.96
5		2.571	+0.611
10		2.228	+0.268
20		2.086	+0.126
30		2.042	+0.082
50		2.009	+0.049
100		1.984	+0.024
∞		1.960	0

As the sample size becomes very large, the t-distribution converges to the normal distribution.

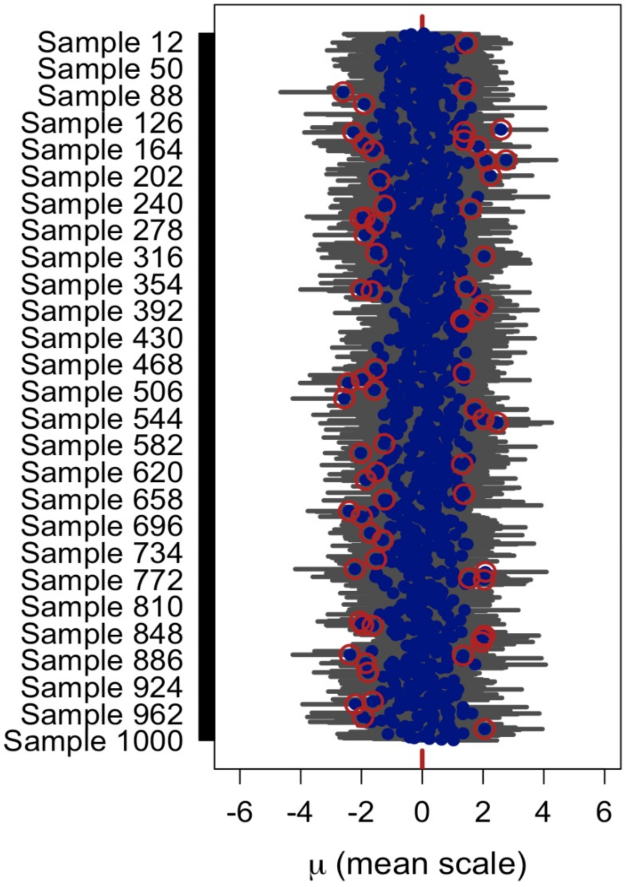
[REVISED]

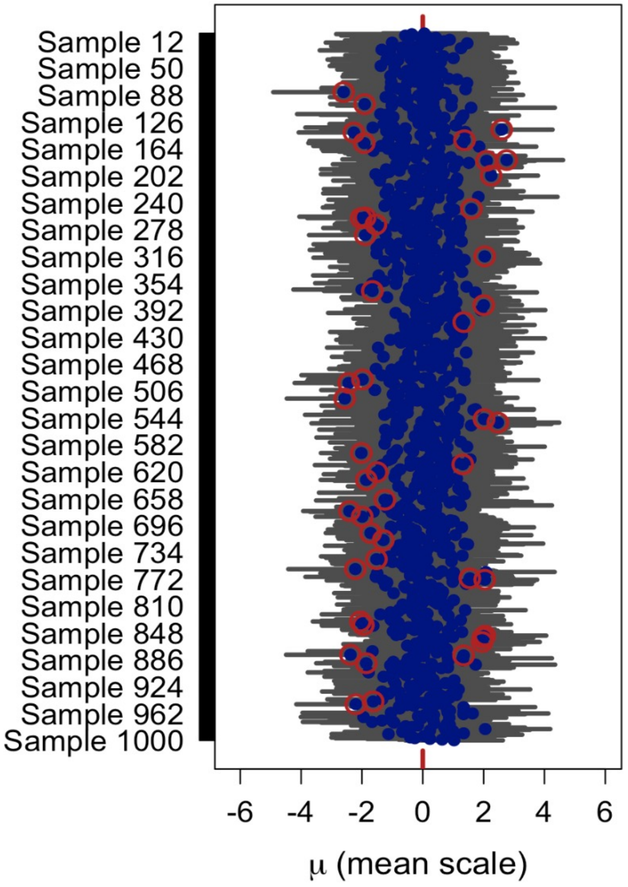
Using the wrong distribution affects coverage: If we use the normal distribution instead of t (when σ is unknown), intervals are too narrow and more than 5% of 95% CIs will miss the true mean (undercoverage). With the correct approach: t-based CIs account for extra uncertainty and 5% will miss the true mean in the long run, as intended.

$$\mathbb{E}[\mathbf{width}_{\mathbf{Normal.based}}] < \mathbb{E}[\mathbf{width}_{\boldsymbol{t.based}}]$$

$$\text{Normal.based:vCI} = \mu \ \pm 1.960 * \frac{s_i}{\sqrt{n}}$$

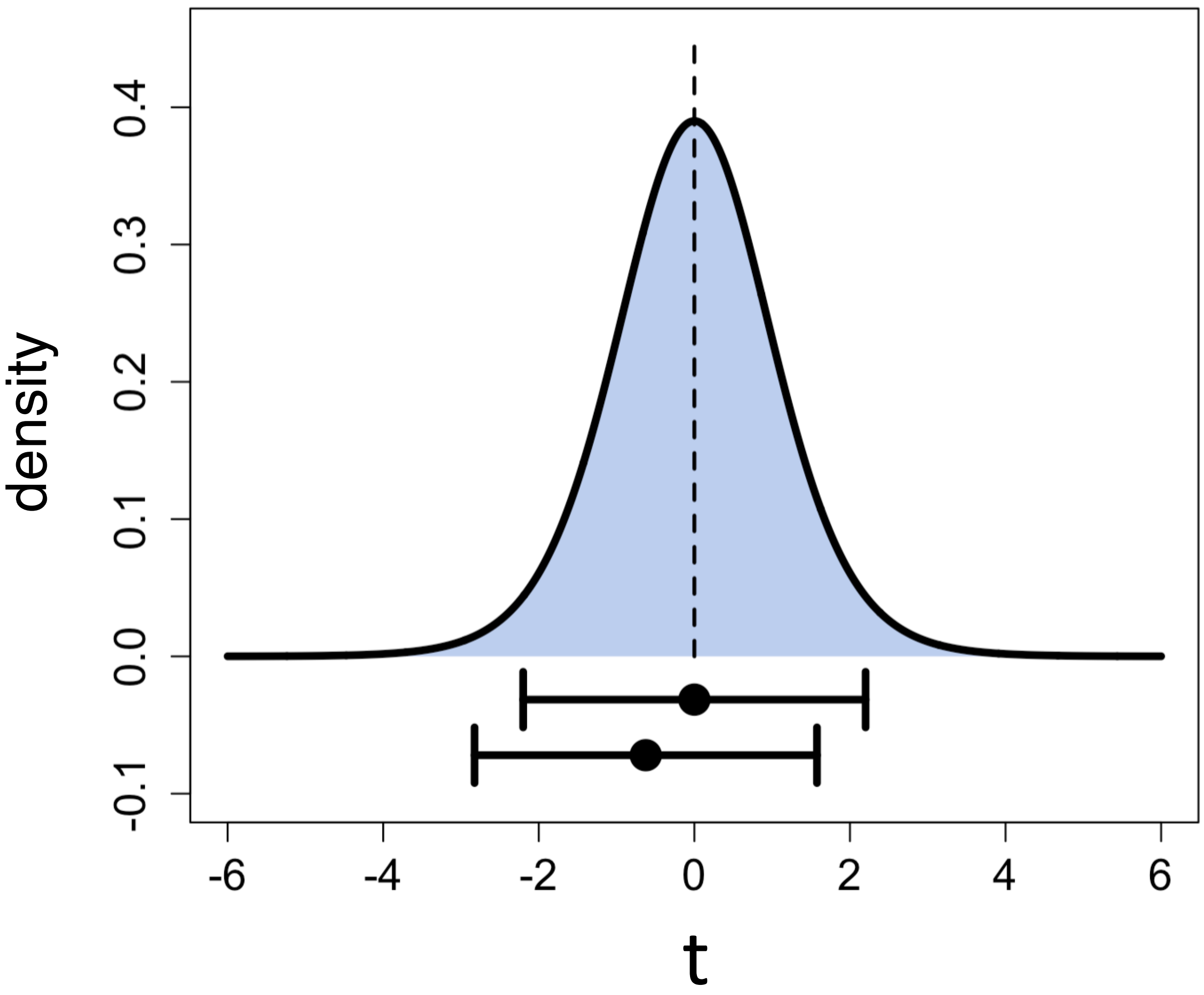
$$\text{t. based: CI} = \mu \ \pm t * \frac{s_i}{\sqrt{n}}$$





[REVISED]

And we can still plot the confidence interval around the sample and the central region interval (not confidence in the population case) around the population value (0) for the t-based distribution. The true population value in the t-distribution is always zero (due to centering - see previous slides).



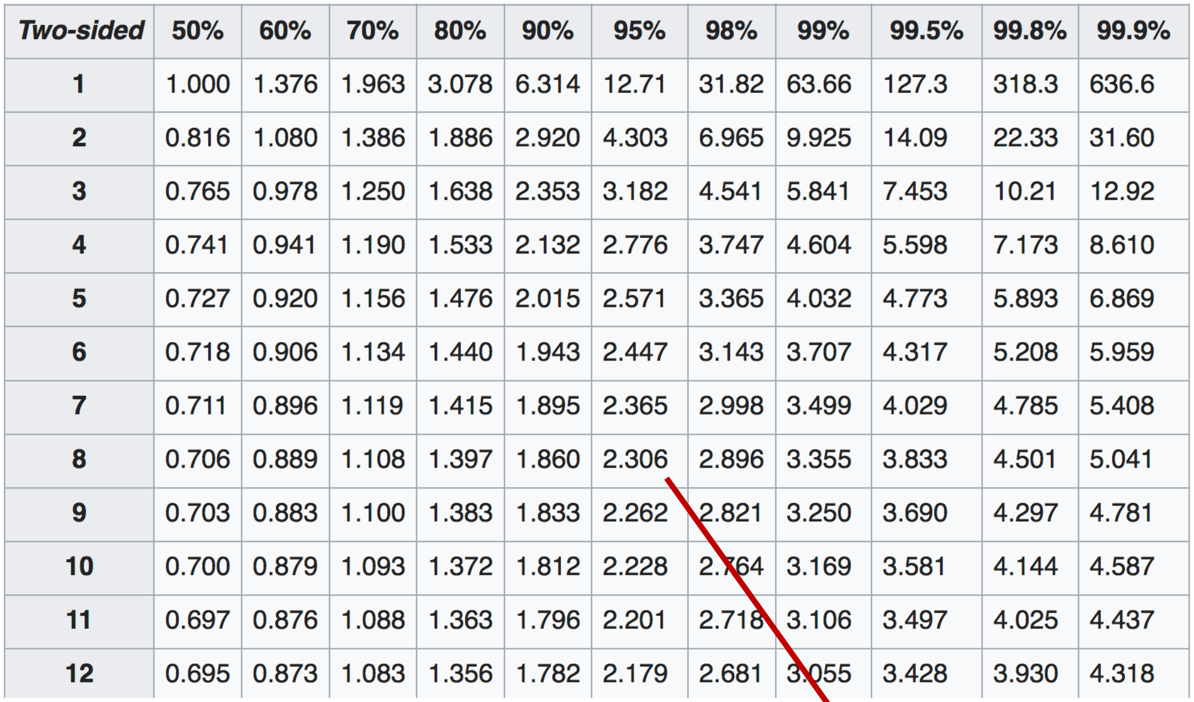
[REVISED]

Let’s take a power break – 1 minute

How to find the appropriate values of t?

the old days of tables allow to understand the principle – in practice (today) we use software (e.g., R).

Degrees of freedom (n-1)



Two-sided	50%	60%	70%	80%	90%	95%	98%	99%	99.5%	99.8%	99.9%
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.080	1.386	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318

Assume a sample size of n = 9, then the degrees of freedom would be 8 for the t value to calculate the confidence interval for the sample mean.

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

$$\bar{X}_i \pm t_{df}\frac{s_i}{\sqrt{n}} \therefore X_i \pm 2.306\frac{s_i}{\sqrt{9}}$$

Degrees of freedom are commonly denoted as df (or sometimes v, the Greek letter nu).

[REVISED]

Let’s consider a biological example: The stalk-eyed fly – the span in millimeters of nine male individuals are as follows:

8.69 8.15 9.25 9.45 8.96 8.65 8.43 8.79 8.63   n=9 flies

Let’s estimate the 95% confidence interval for the population mean of eyes spans:

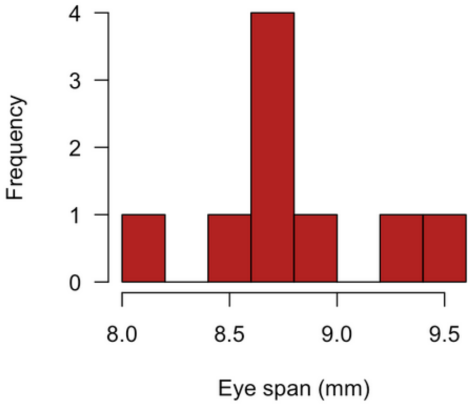
$$\bar{X} = 8.778 \; s = 0.398$$

$$\mathrm{SE}_{\bar{X}} \frac{0.398}{\sqrt{9}} = 0.133$$

$$t_{0.05(2),8} = 2.306$$

$$\bar{\mathrm{X}} \; - 2.306 \times 0.133 < \; \mu < \bar{\mathrm{X}} \; + 2.306 \times 0.133$$

$$8.47 \text{ mm} < \; \mu < 9.08 \text{ mm}$$



In practice (modern days), we use software

Let's estimate the 95% confidence interval for the population mean of eyes spans:

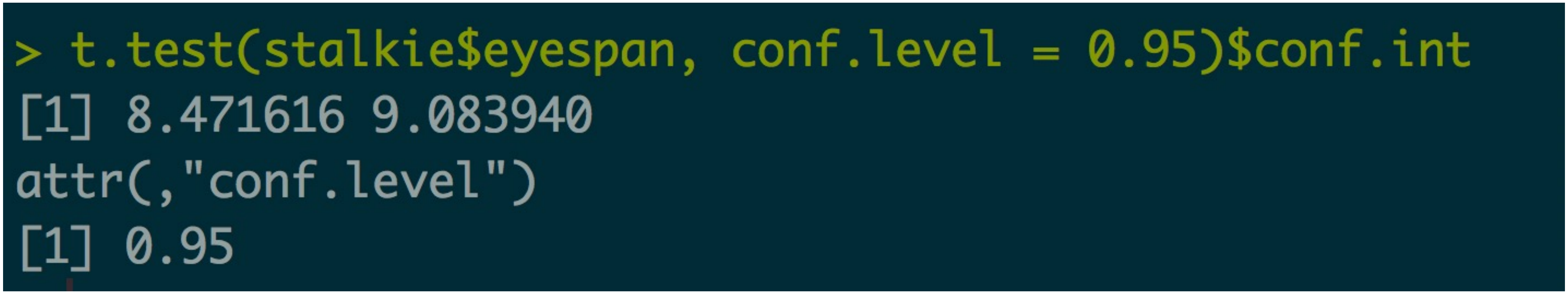
$$\bar{X} = 8.778 \; s = 0.398$$

$$\mathrm{SE}_{\bar{X}} \frac{0.398}{\sqrt{9}} = 0.133$$

$$t_{0.05(2),8} = 2.306$$

$$\bar{X} - 2.306 \times 0.133 < \mu < \bar{X} + 2.306 \times 0.133$$

$$8.47 \text{ mm} < \mu < 9.08 \text{ mm}$$



```
> t.test(stalkie$eyespan, conf.level = 0.95)$conf.int
[1] 8.471616 9.083940
attr(,"conf.level")
[1] 0.95
```

In practice (modern days), we use software

Let’s estimate the 99% confidence interval for the population mean of eyes spans:

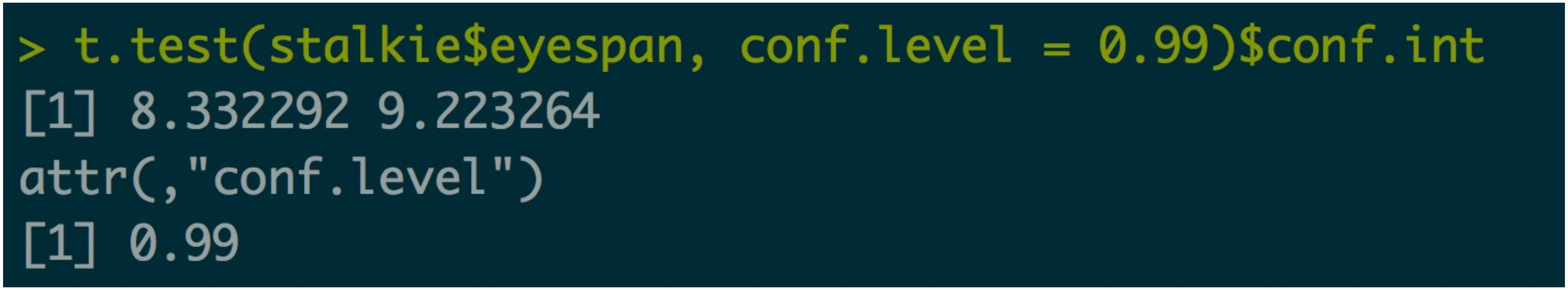
$$\bar{Y} = 8.778 \; s = 0.398$$

$$\mathrm{SE}_{\bar{Y}} \frac{0.398}{\sqrt{9}} = 0.133$$

$$t_{0.05(2),8} = 3.355$$

$$\bar{\mathrm{Y}} \; - 3.355 \times 0.133 < \; \mu < \bar{\mathrm{Y}} \; + 3.355 \times 0.133$$

$$8.33 \text{ mm} < \; \mu < 9.22 \text{ mm}$$

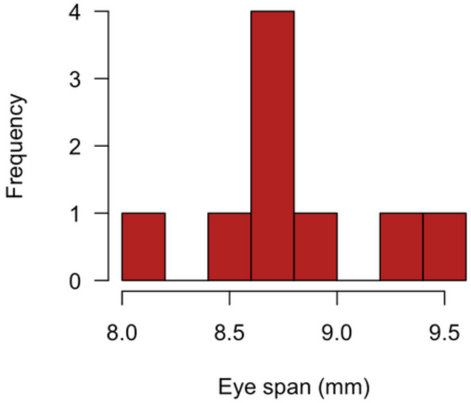


```
> t.test(stalkie$eyespan, conf.level = 0.99)$conf.int
[1] 8.332292 9.223264
attr(,"conf.level")
[1] 0.99
```

Comparing the two intervals, the implications for precision depend on the biological context and the level of certainty required. In practice, we most often report the 95% confidence interval, which provides a balance between coverage and precision—offering less coverage than a 99% interval but greater precision (narrower width).

95% confidence interval: $8.47 \text{ mm} < \mu < 9.08 \text{ mm}$

99% confidence interval: $8.33 \text{ mm} < \mu < 9.22 \text{ mm}$



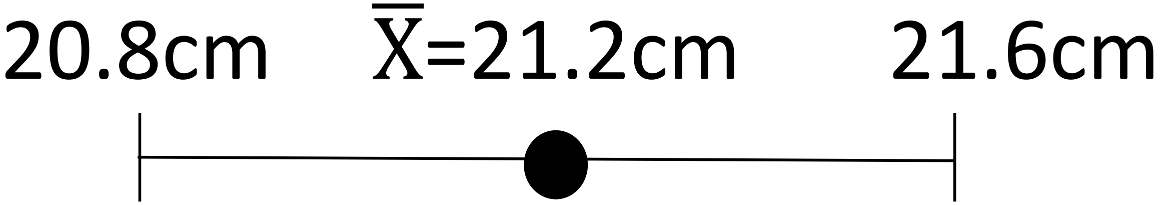
Confidence intervals can be used in many situations; e.g., decision making

The optimal average body length for a healthy brook trout population is 24 cm or greater. According to management policy, if the true population mean length is below 24 cm, the lake must be closed to fishing.

At the beginning of the fishing season, a government biologist surveys a heavily exploited lake, where strict regulation is particularly important. The biologist captures, anesthetizes, and measures 200 brook trout. The sample mean length is 21.3 cm, with a sample standard deviation of 3.2 cm.

Should the lake be opened for fishing?

$$\bar{\mathrm{Y}} \pm \; t \times \mathrm{SE}_{\bar{Y}_s} \therefore 21.2 \pm 1.971957 \times \frac{3.2}{\sqrt{200}} = 21.2 \; \pm 0.4462$$


20.8cm $\bar{\mathrm{X}}$=21.2cm 21.6cm