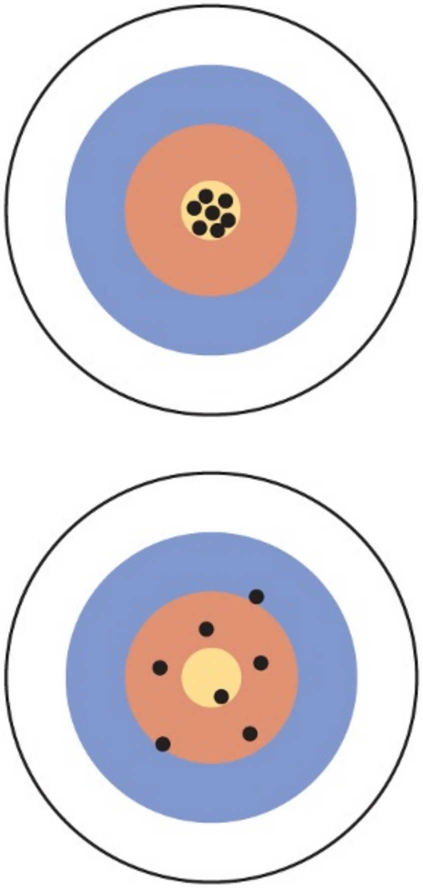
Lecture 11 – Understanding biases in estimators

Estimating with Confidence: Understanding and Quantifying Uncertainty

For statistical inference to be reliable, we must be able to trust our sample estimators. At a minimum, this means they should be unbiased; that is, they should not systematically overestimate or underestimate the true population parameter across repeated samples.



The bullseye is the population mean μ and each dot is a sample mean $\bar{X}$.

Under random sampling, the sample mean $\bar{X}$ is an unbiased estimator of the population mean μ.

This means that the expected value of the sampling distribution of $\bar{X}$ equals μ:

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

In practical terms, if we repeatedly draw samples of the same size from the population and compute their means, those sample means will fluctuate due to sampling variation, but they will be centered on the true population mean.

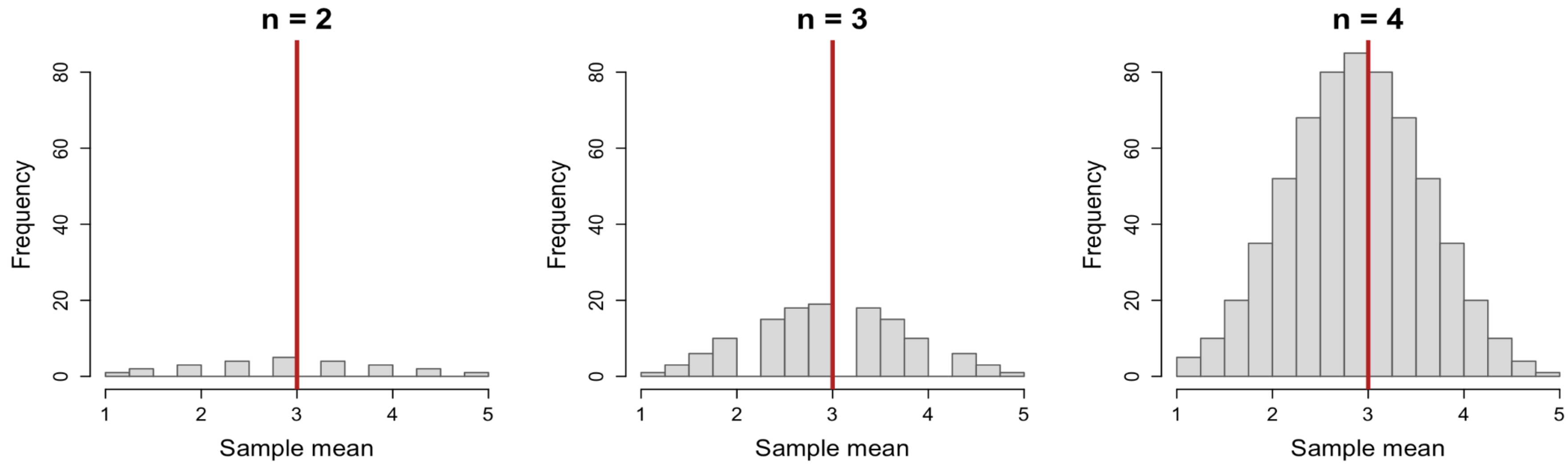
Unbiasedness therefore does not mean that any single sample mean equals μ. It means that, across many repeated samples, the estimator does not systematically overestimate or underestimate the true value.

The shape of a population’s frequency distribution does not necessarily resemble the sampling distribution of a statistic (such as the distribution of sample means).

Regardless of the population’s shape, even if it is uniform or skewed, the sample mean remains an unbiased estimator of the population mean, provided sampling is random.

{1,2,3,4,5} (uniform)

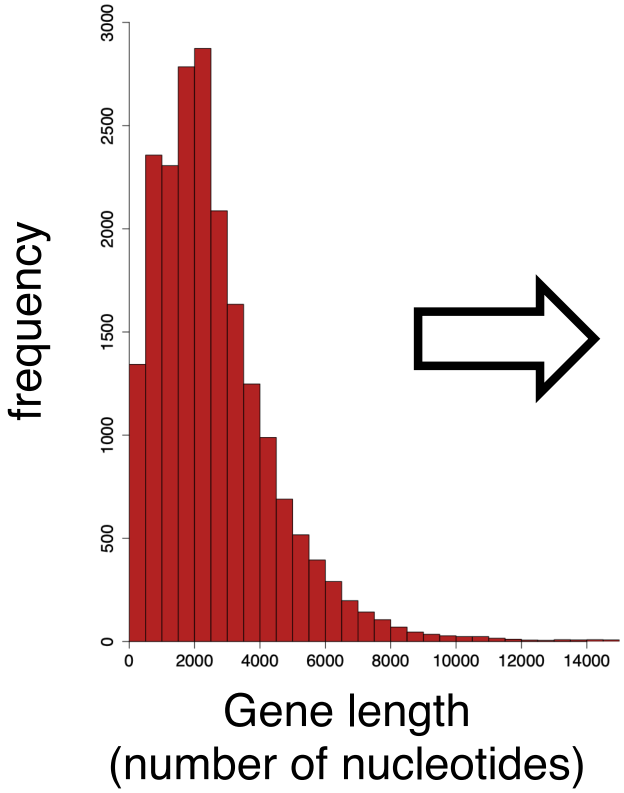
▬ True population mean = mean of sample means



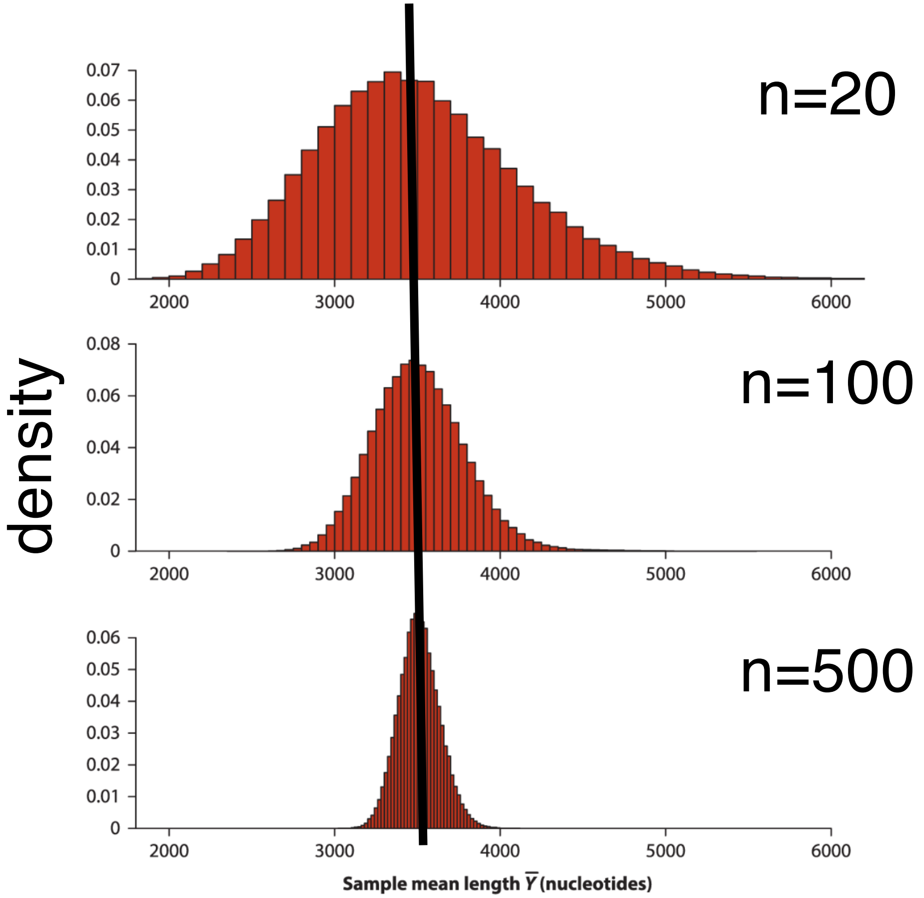
The shape of a population’s frequency distribution does not necessarily resemble the sampling distribution of a statistic (such as the distribution of sample means).

Regardless of the population’s shape, even if it is uniform or skewed, the sample mean remains an unbiased estimator of the population mean, provided sampling is random.

Frequency distribution of the gene Population



Sampling distributions for the sample means of the gene population!

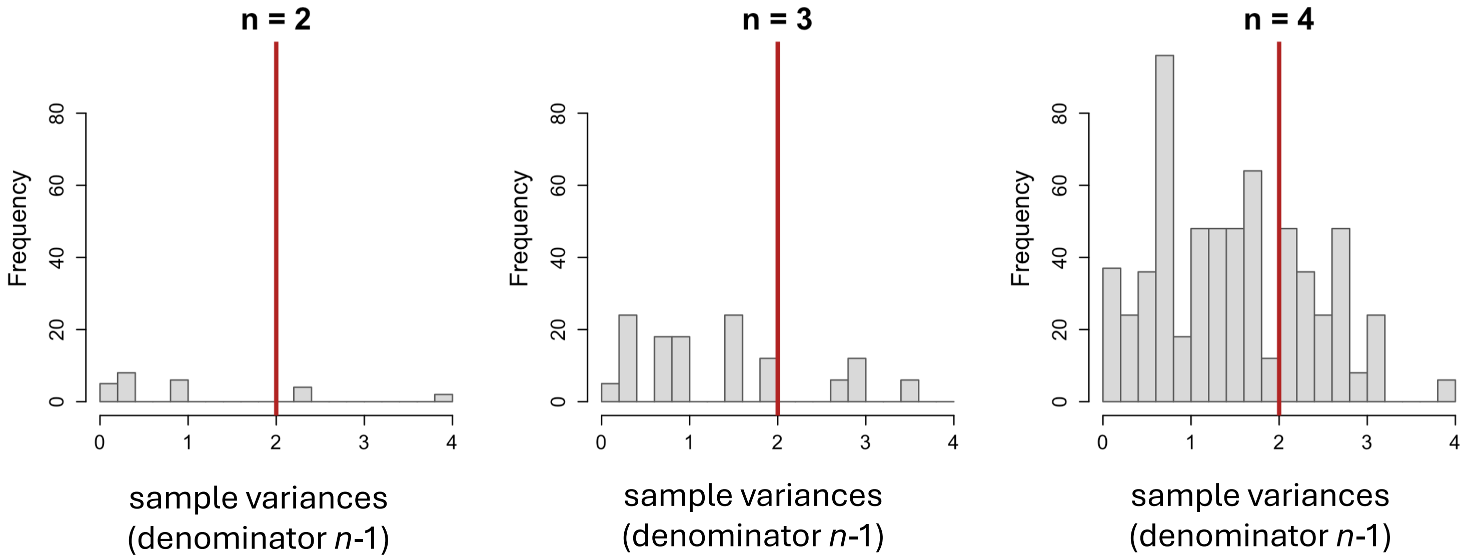


Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

Sample mean length $\bar{X}$ (nucleotides)

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.

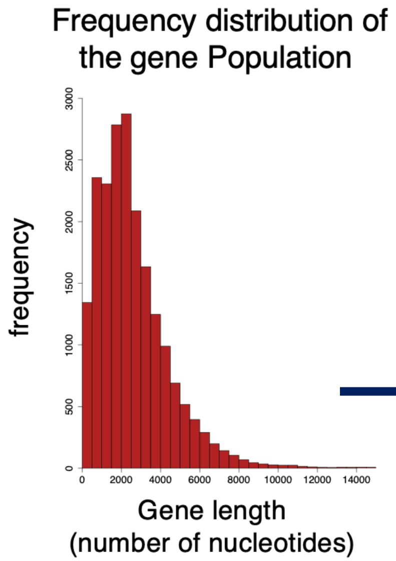
▬ True population variance (2.0) = mean of sample variances (2.0)



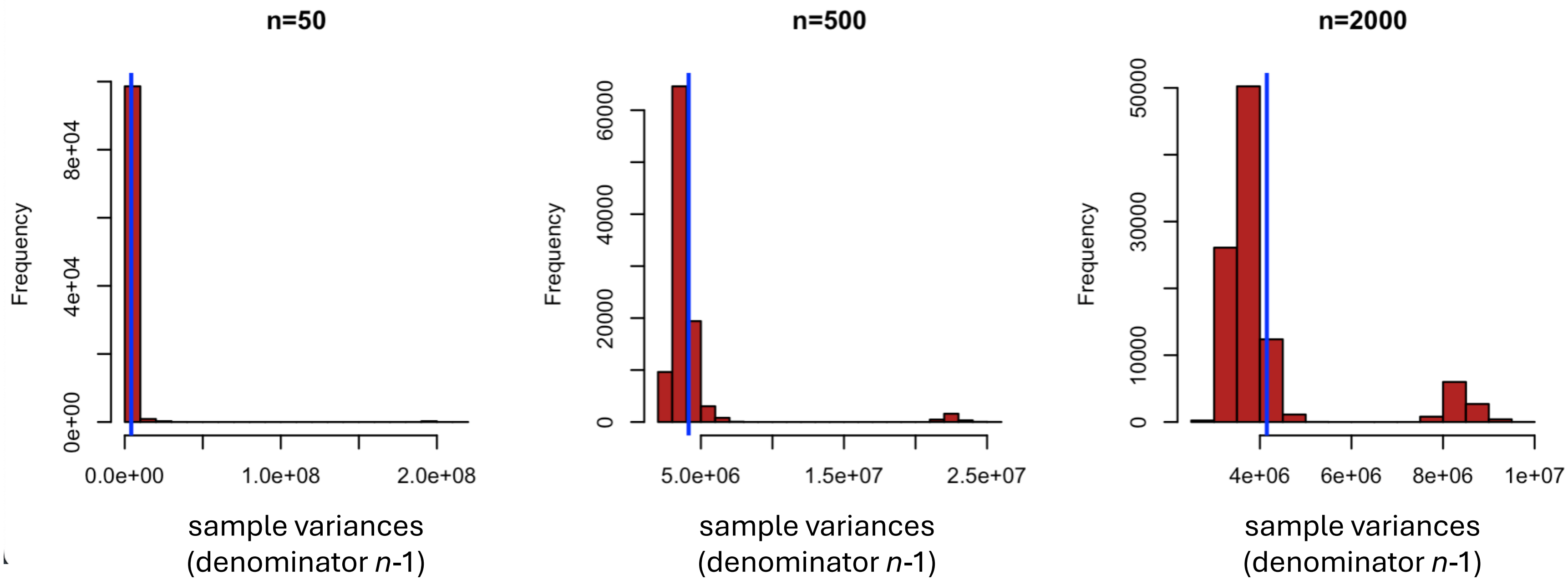
One critical observation is that **the mean of all possible sample variance is exactly the population variance**. This is important because it tells us that, on average, the sampling process is **unbiased**: repeated sampling does not systematically overestimate or underestimate the true population value.

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.

Even for a highly asymmetric population, the variance is an unbiased estimator



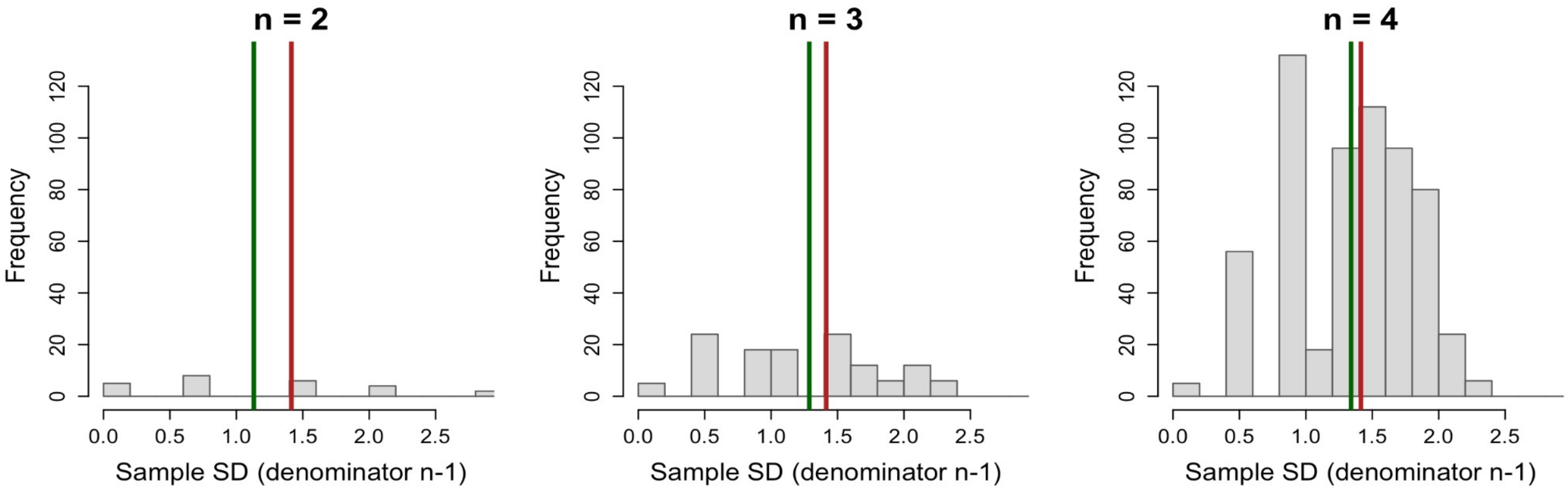
4,149,235 number of nucleotides^2



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.

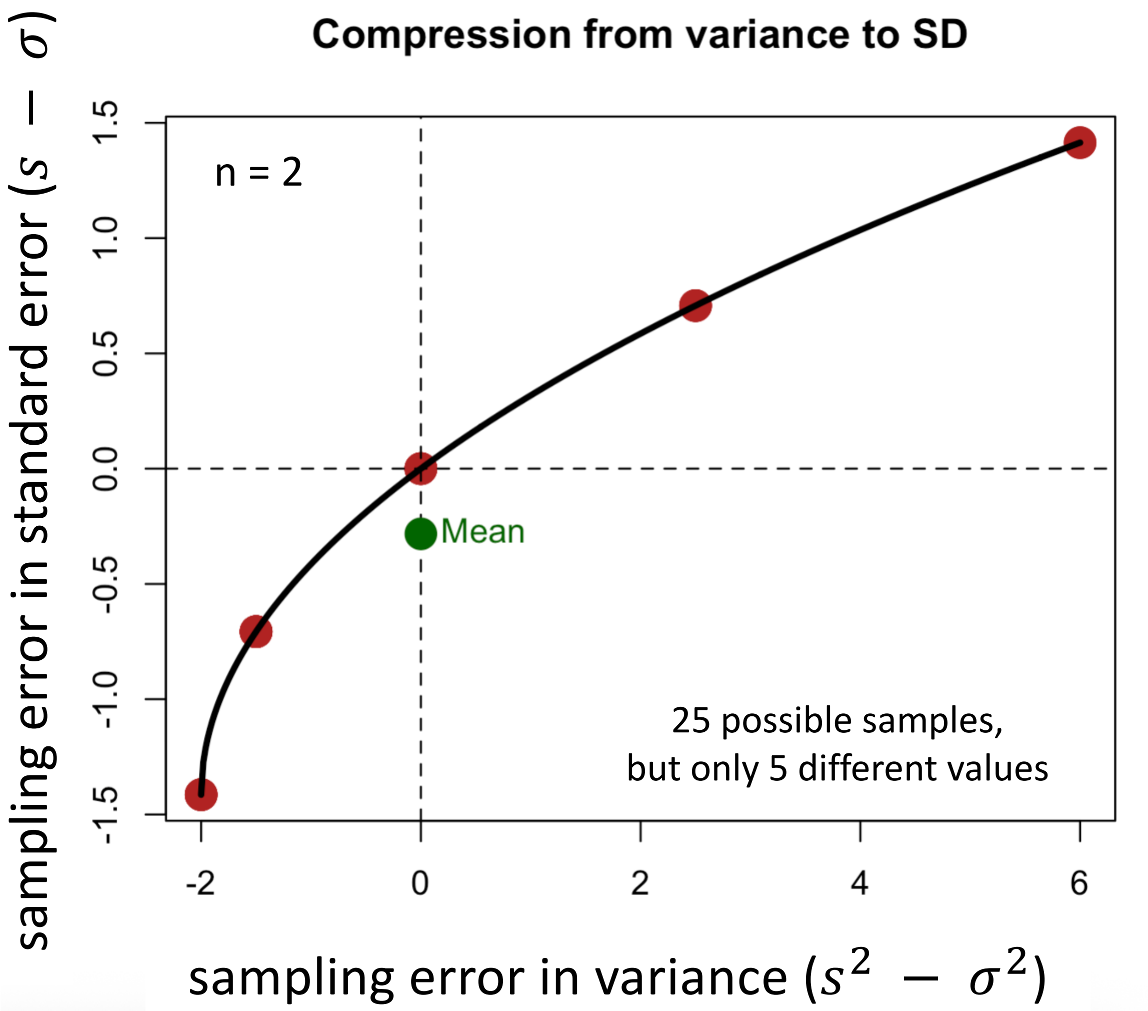
🟥 True population standard deviation (1.414214)

🟩 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.

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.



25 possible samples,
but only 5 different values

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

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.

?

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

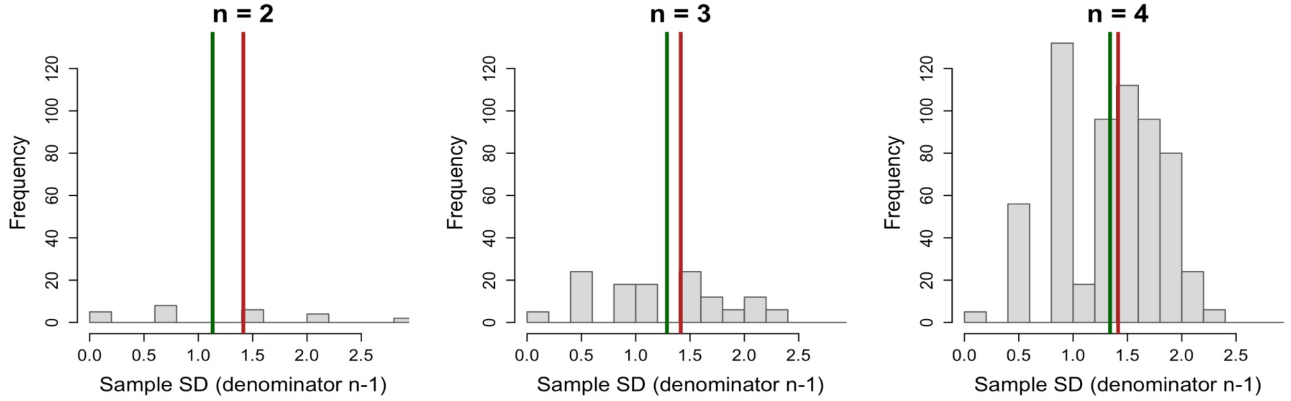
[REVISED]

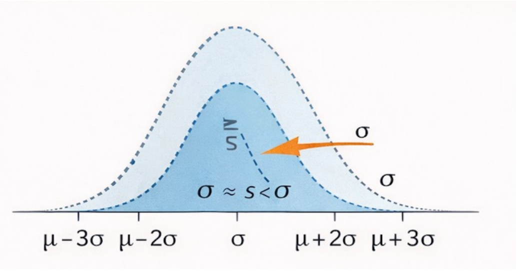
However, if the sample standard deviation underestimates the true population standard deviation, especially when the sample size is small (Jenzen's inequality), the resulting intervals will, on average, be too narrow.

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)





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

Sample-based
(unknown σ)

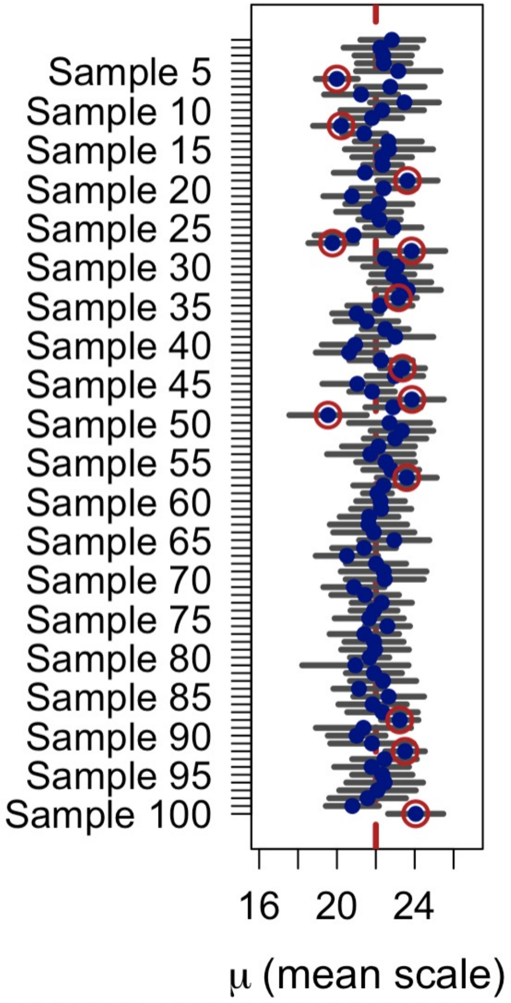
Population-based
(known σ)

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

[REVISED]

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?



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

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

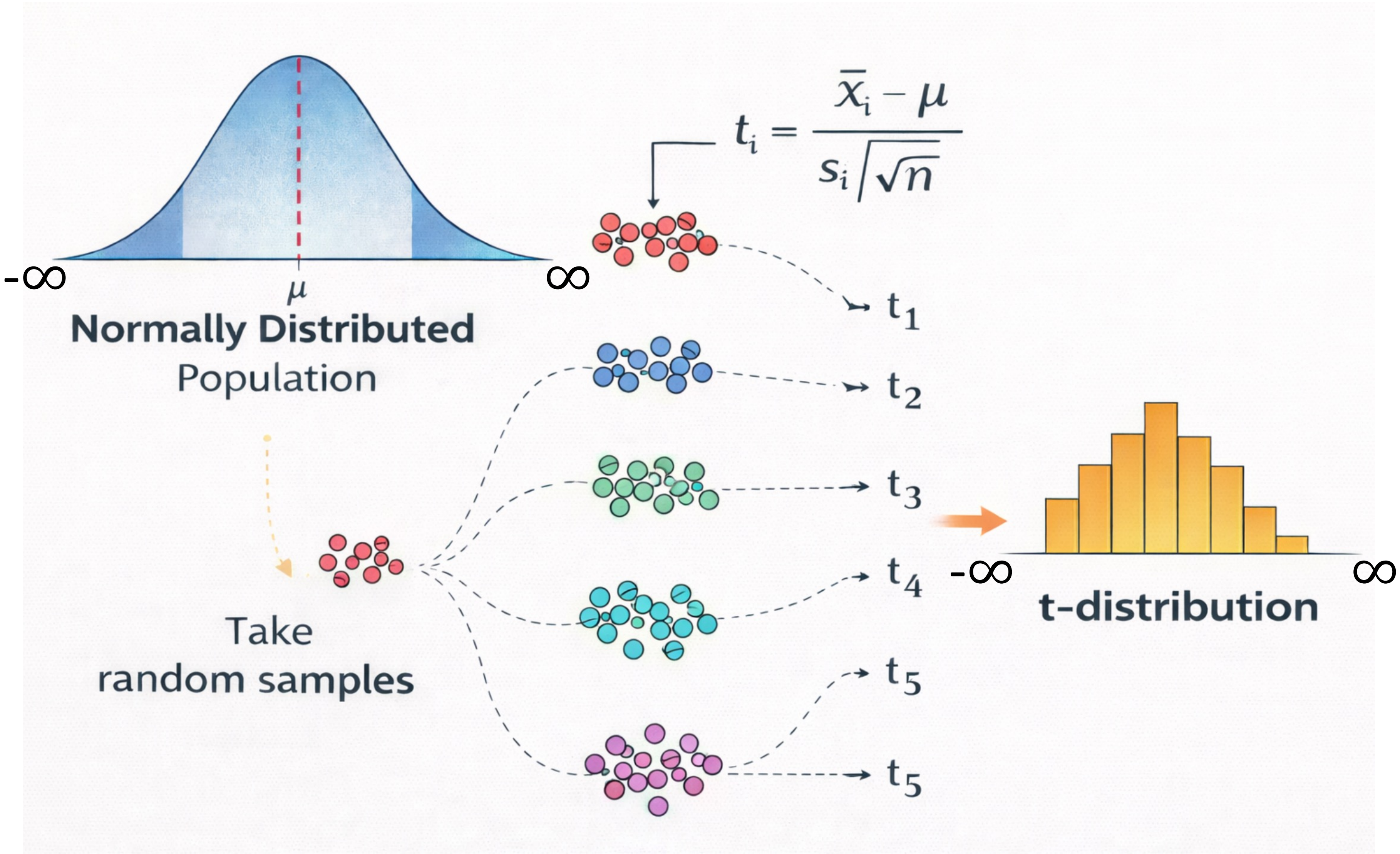
For a 95% confidence interval, we expect that only 5% of intervals will fail to include the true population mean.

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]

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



By now, we have seen 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

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

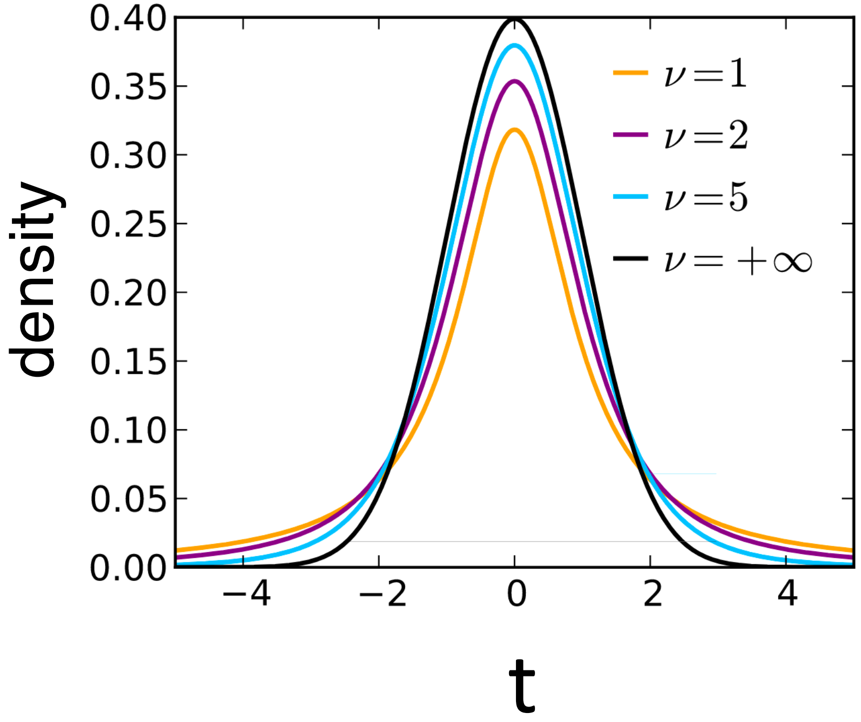
👍 $$\mathrm{CI(95\%)} = \bar{\mathrm{X}}_i \pm \boldsymbol{t.df} \times \frac{s_i}{\sqrt{n}}$$

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.

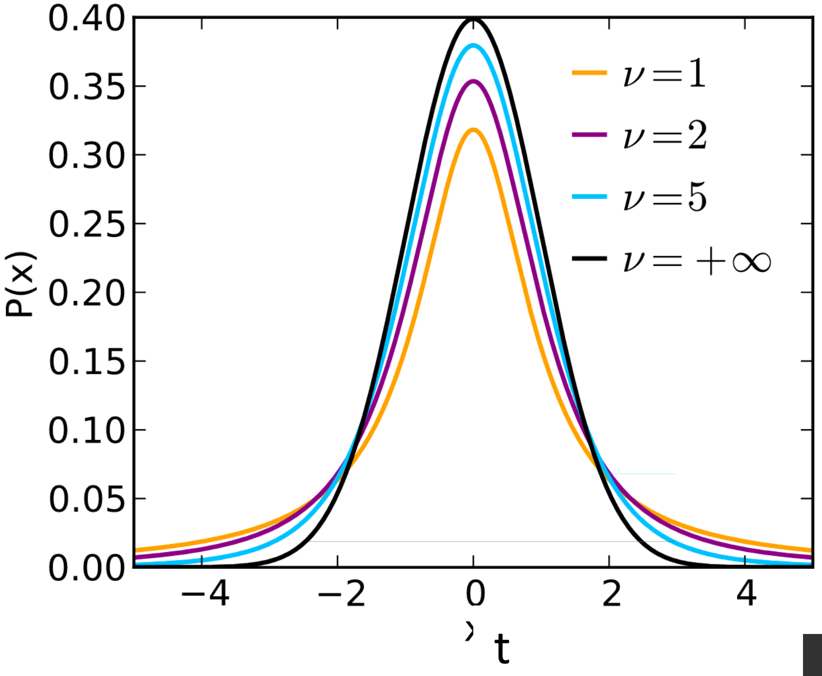


v = ∞ → t = Z → the t distribution becomes normally distributed when sample size is infinite.

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

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.

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

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

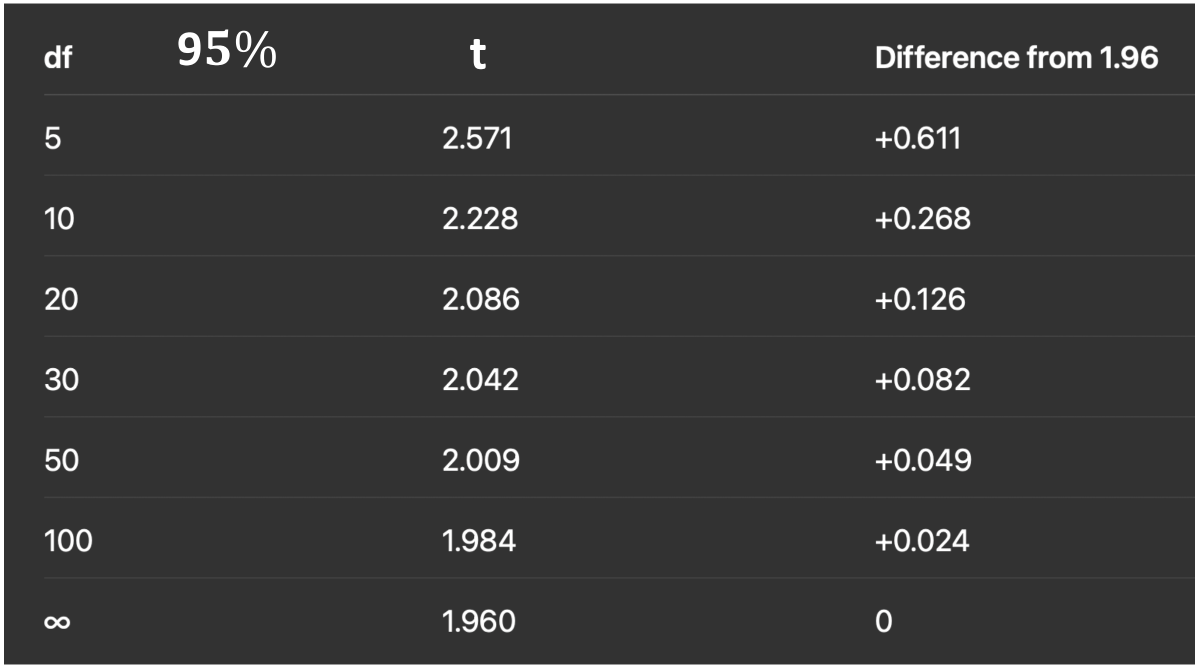
Interval width

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

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

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

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



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]

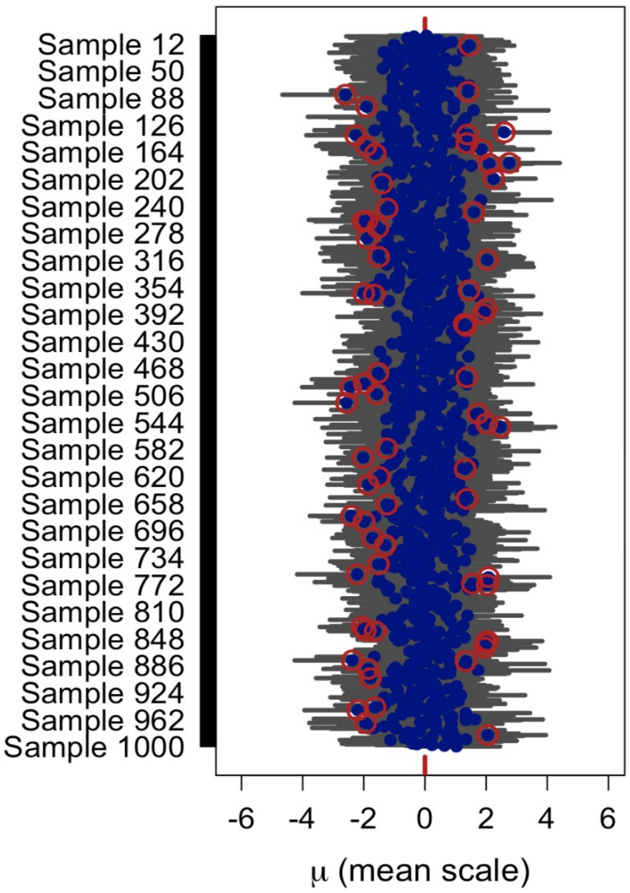
Using the wrong distribution affects coverage: If we use the normal distribution instead of t (when σ is unknown), intervals are too narrow and more than 5% of 95% CIs will miss the true mean (undercoverage). With the correct approach: t-based CIs account for extra uncertainty and 5% will miss the true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

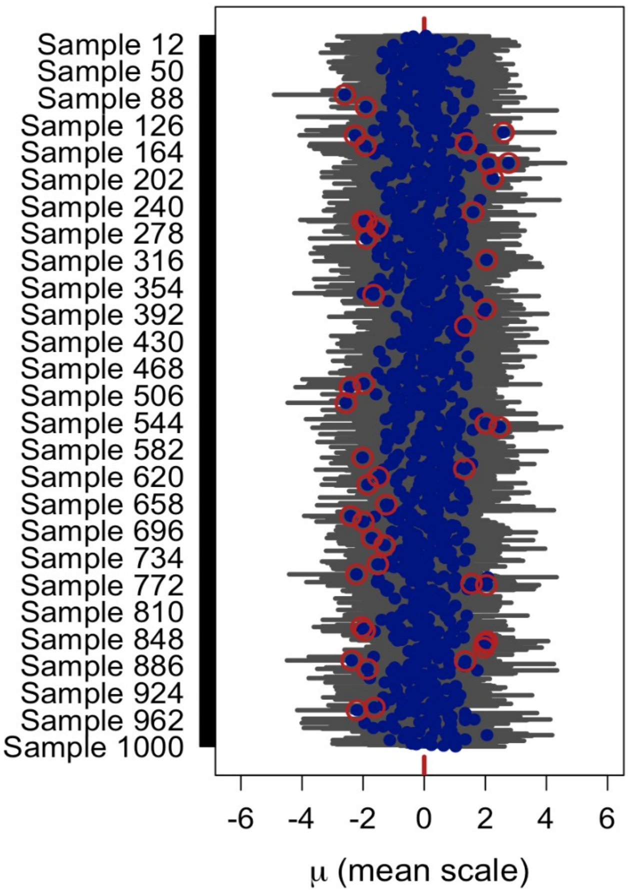
$$\mathbb{E}[\mathbf{width}_{\mathbf{Normal.based}}] < \mathbb{E}[\mathbf{width}_{\boldsymbol{t.based}}]$$

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

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

👎





👍

[REVISED]

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.

$$\sigma^2 = \mathbb{E}\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\boldsymbol{n-1}}\right) = 2$$

Suppose, hypothetically, that we knew the true population mean \mu (which in practice we do not). In that case, the natural sample-based estimator of the population variance from a single sample would be:

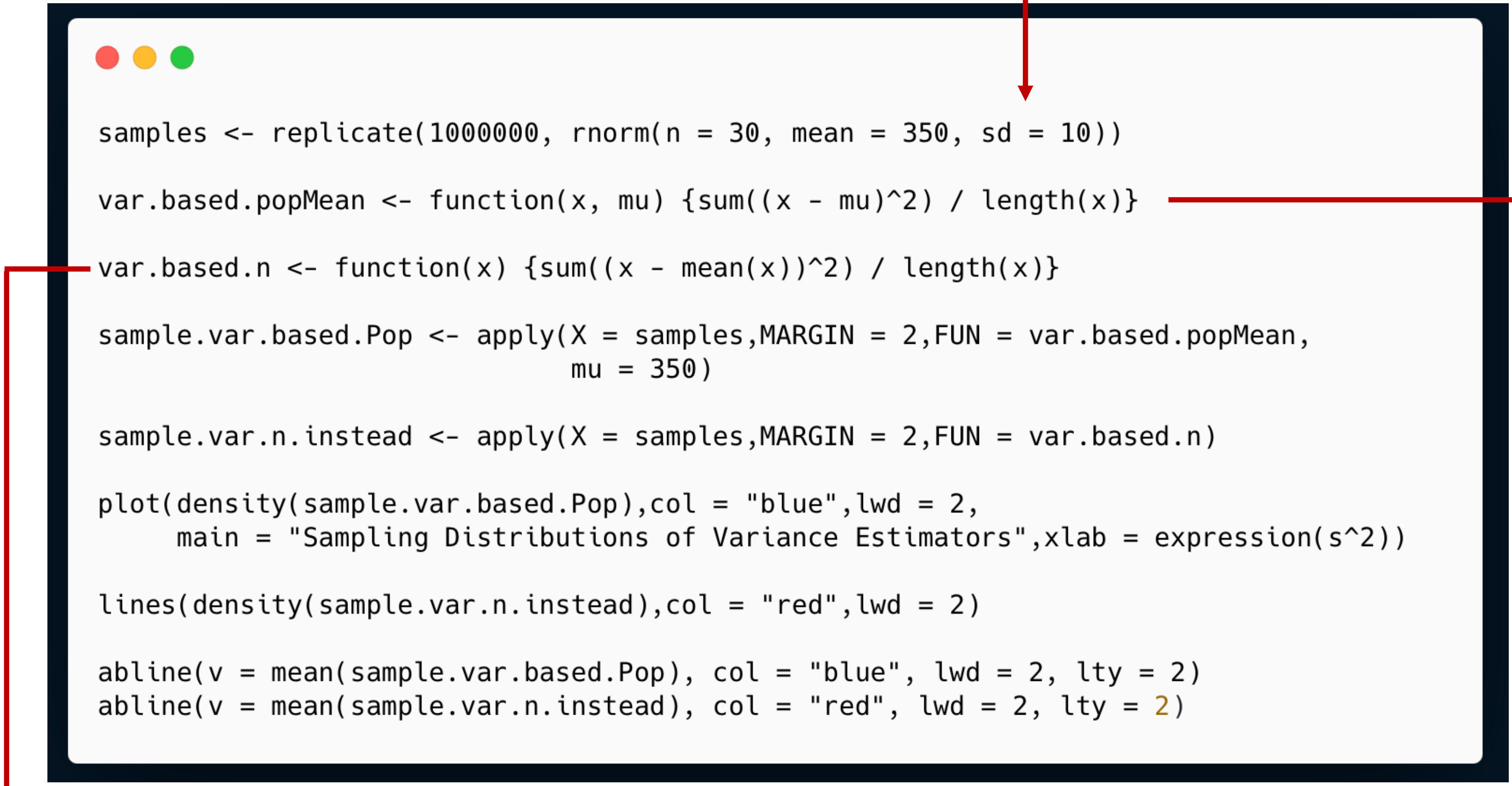
$$s_{\mu}^{2} = \frac{\sum_{i=1}^{n}(X_i - \boldsymbol{\mu})^2}{n} \therefore \sigma^2 = \mathbb{E}(s_{\mu}^{2})$$

Since we almost never know the true population mean μ, let us examine what happens when we replace μ with the sample mean $\bar{X}$ as its estimate:

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

Let's use a computational approach to evaluate the accuracy of these two sample-based estimators.:

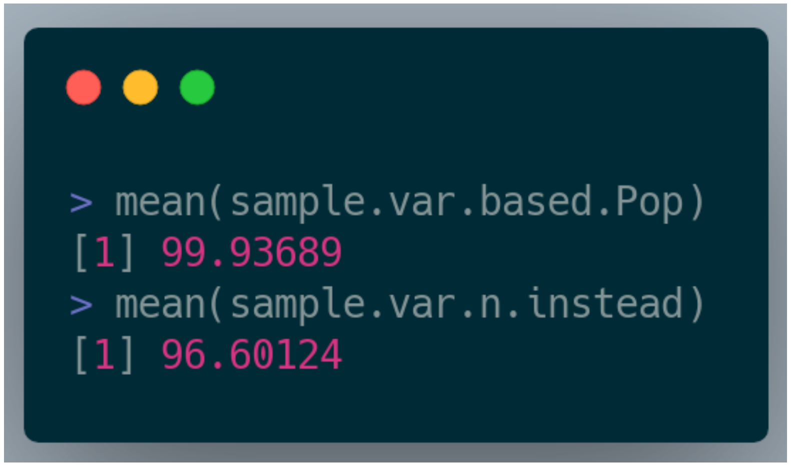
$$n = 30; \sigma^2 = 100; \sigma = 10$$



```
samples <- replicate(1000000, rnorm(n = 30, mean = 350, sd = 10))

var.based.popMean <- function(x, mu) {sum((x - mu)^2) / length(x)}

var.based.n <- function(x) {sum((x - mean(x))^2) / length(x)}

sample.var.based.Pop <- apply(X = samples,MARGIN = 2,FUN = var.based.popMean,
                              mu = 350)

sample.var.n.instead <- apply(X = samples,MARGIN = 2,FUN = var.based.n)

plot(density(sample.var.based.Pop),col = "blue",lwd = 2,
     main = "Sampling Distributions of Variance Estimators",xlab = expression(s^2))

lines(density(sample.var.n.instead),col = "red",lwd = 2)

abline(v = mean(sample.var.based.Pop), col = "blue", lwd = 2, lty = 2)
abline(v = mean(sample.var.n.instead), col = "red", lwd = 2, lty = 2)
```

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

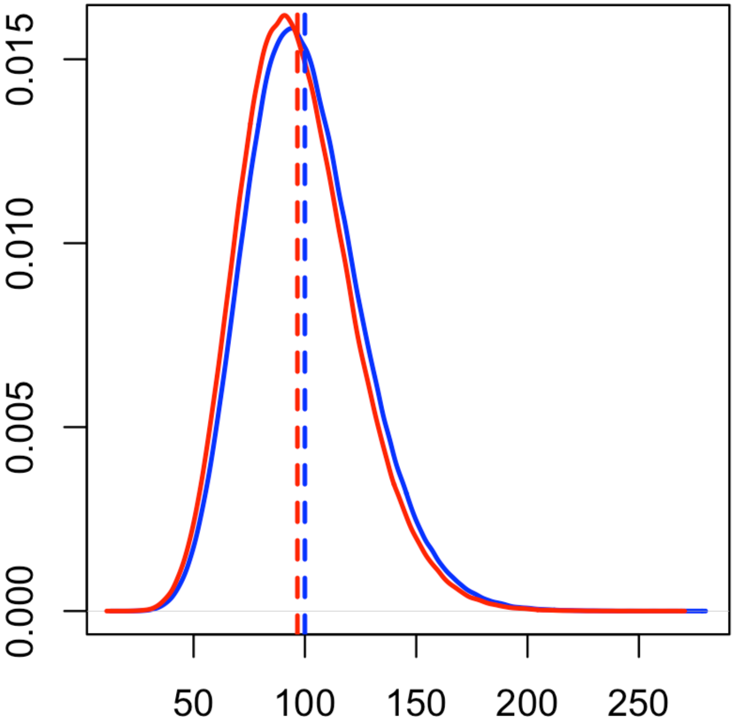
$$s^2_{\mu} = \frac{\sum_{i=1}^{n}(X_i - \boldsymbol{\mu})^2}{n}$$

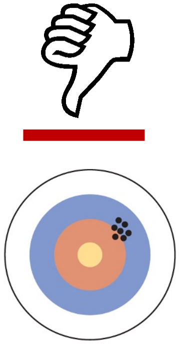


```
> mean(sample.var.based.Pop)
[1] 99.93689
> mean(sample.var.n.instead)
[1] 96.60124
```

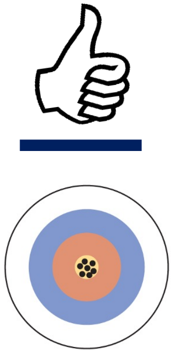
The mean of s^2 for the estimator based on the population mean μ divided by *n* is unbiased (i.e., matched the population σ^2; it would have exactly equalled σ^2 =100 based on infinite sampling). However, the estimator based on the sample mean $\bar{X}$ divided by *n* is biased.

```
Bias <- sample.var.based.Pop - sample.var.n.instead
```



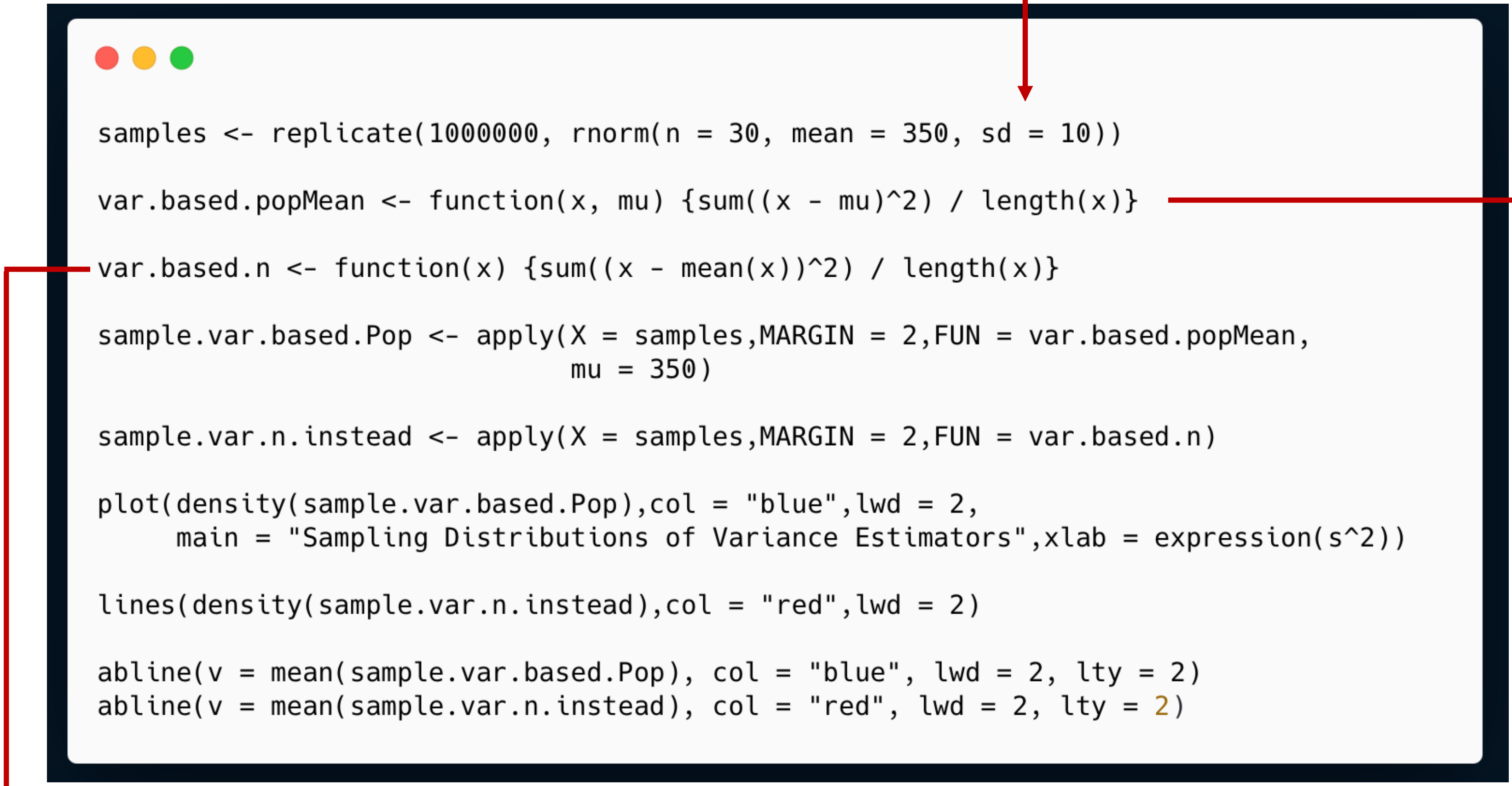


$$\tilde{s}^2_{\bar{X}} = \frac{\sum_{i=1}^{n}(X_i \ - \bar{\mathbf{X}})^2}{n}$$



$$s^2_{\mu} = \frac{\sum_{i=1}^{n}(X_i \ - \boldsymbol{\mu})^2}{n}$$

$$n = 30; \sigma^2 = 100; \sigma = 10$$



```
samples <- replicate(1000000, rnorm(n = 30, mean = 350, sd = 10))

var.based.popMean <- function(x, mu) {sum((x - mu)^2) / length(x)}

var.based.n <- function(x) {sum((x - mean(x))^2) / length(x)}

sample.var.based.Pop <- apply(X = samples,MARGIN = 2,FUN = var.based.popMean,
                              mu = 350)

sample.var.n.instead <- apply(X = samples,MARGIN = 2,FUN = var.based.n)

plot(density(sample.var.based.Pop),col = "blue",lwd = 2,
     main = "Sampling Distributions of Variance Estimators",xlab = expression(s^2))

lines(density(sample.var.n.instead),col = "red",lwd = 2)

abline(v = mean(sample.var.based.Pop), col = "blue", lwd = 2, lty = 2)
abline(v = mean(sample.var.n.instead), col = "red", lwd = 2, lty = 2)
```

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

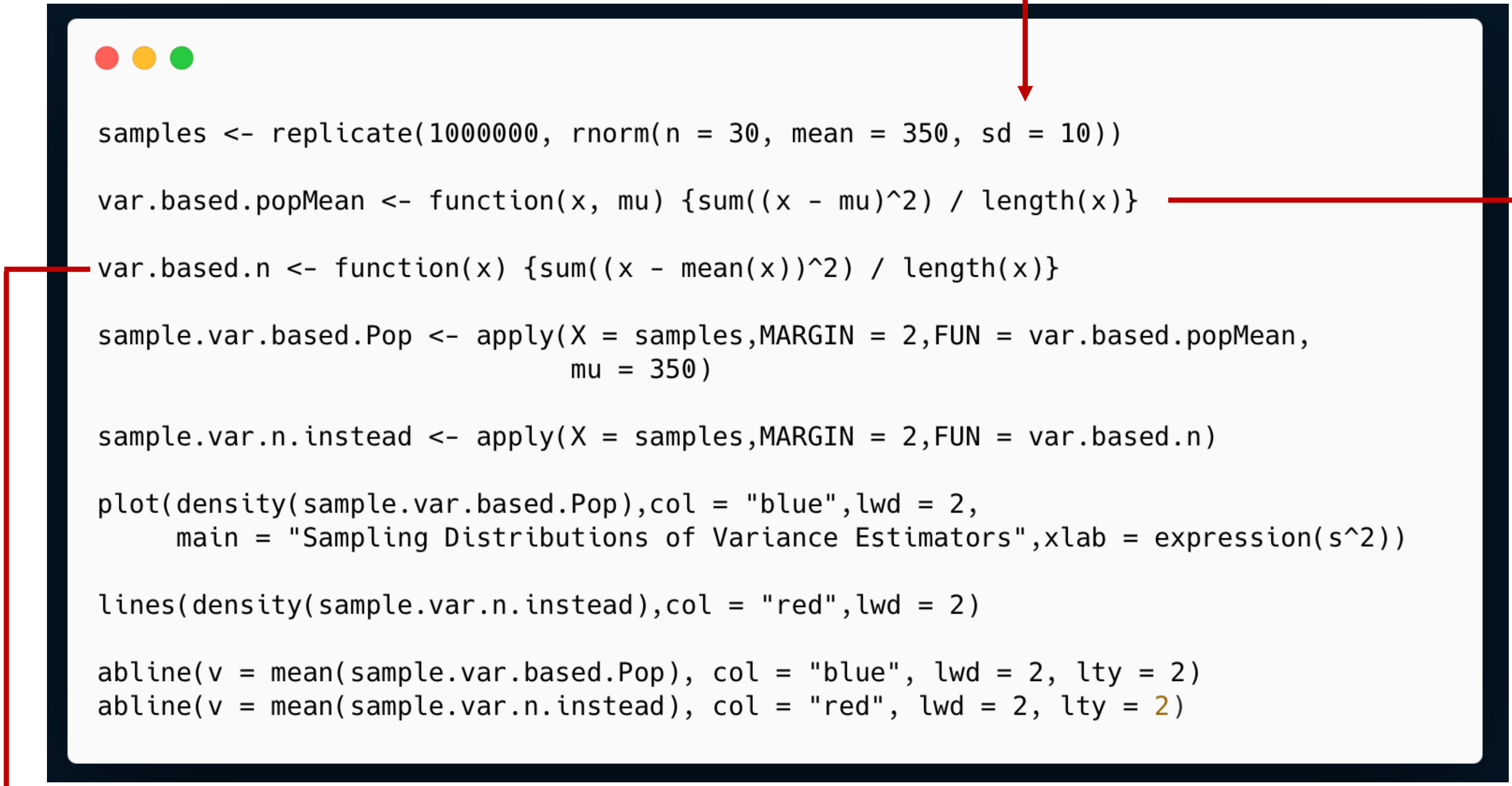
$$s^2_{\mu} = \frac{\sum_{i=1}^{n}(X_i - \boldsymbol{\mu})^2}{n}$$

Biases <- sample.var.based.Pop - sample.var.n.instead

Average(Biases) <- 3.333506

true.var / n <- 100/n <- 3.333333 -> Bias = s^2_{μ} $(true\ variance)$ - sample.var.n.instead

$$n = 30; \mu = 350; \sigma^2 = 100; \sigma = 10$$



```
samples <- replicate(1000000, rnorm(n = 30, mean = 350, sd = 10))

var.based.popMean <- function(x, mu) {sum((x - mu)^2) / length(x)}

var.based.n <- function(x) {sum((x - mean(x))^2) / length(x)}

sample.var.based.Pop <- apply(X = samples,MARGIN = 2,FUN = var.based.popMean,
                              mu = 350)

sample.var.n.instead <- apply(X = samples,MARGIN = 2,FUN = var.based.n)

plot(density(sample.var.based.Pop),col = "blue",lwd = 2,
     main = "Sampling Distributions of Variance Estimators",xlab = expression(s^2))

lines(density(sample.var.n.instead),col = "red",lwd = 2)

abline(v = mean(sample.var.based.Pop), col = "blue", lwd = 2, lty = 2)
abline(v = mean(sample.var.n.instead), col = "red", lwd = 2, lty = 2)
```

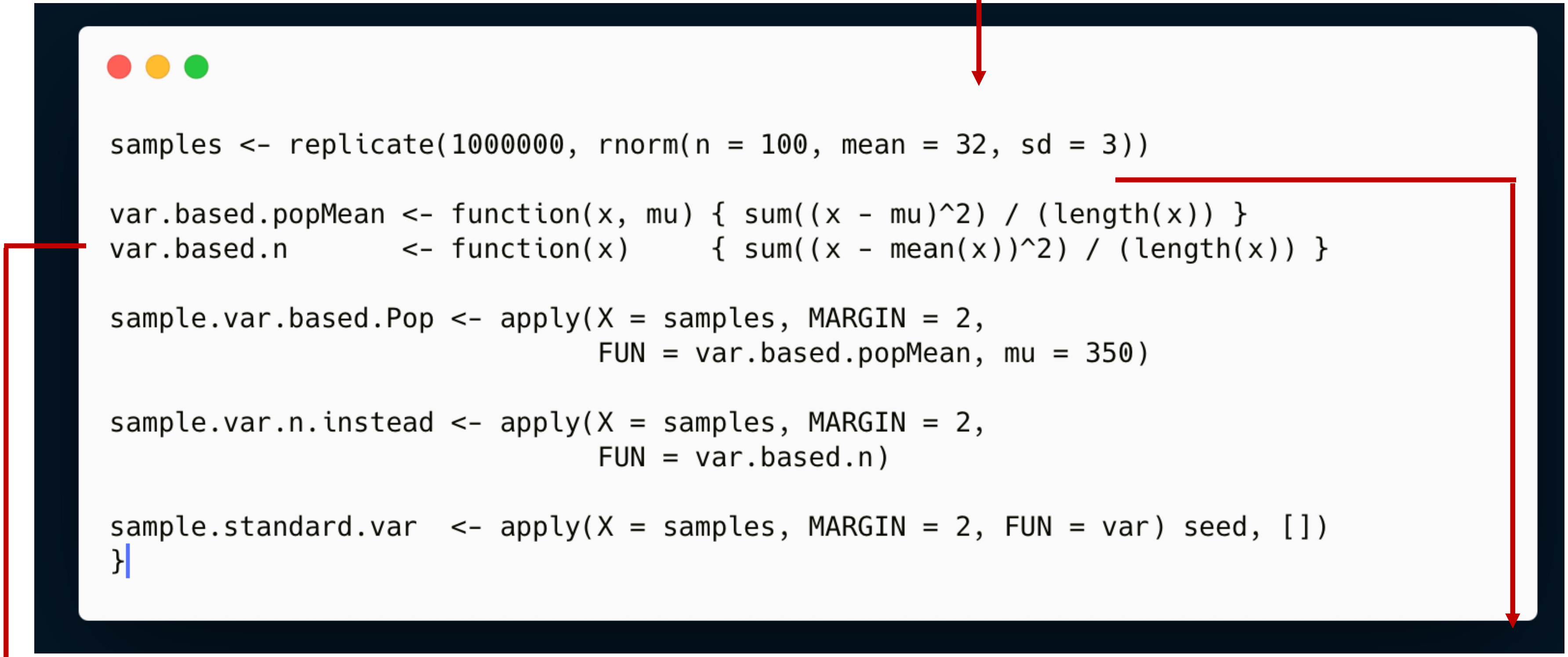
$$\tilde{s}^2_{\bar{X}} = \frac{\sum_{i=1}^{n}(X_i - \bar{\mathbf{X}})^2}{n}$$

$$s^2_{\mu} = \frac{\sum_{i=1}^{n}(X_i - \boldsymbol{\mu})^2}{n}$$

Bias <- sample.var.based.Pop - sample.var.n.instead (1,000,000)
Average(Bias) <- 3.333506
true.var / n <- 100/30 <- 3.333333 -> Average(Bias) = true.var / n

BEFORE: $n = 30; \mu = 350; \sigma^2 = 100; \sigma = 10$

$n = 100; \mu = 32; \sigma^2 = 9; \sigma = 3$



```
samples <- replicate(1000000, rnorm(n = 100, mean = 32, sd = 3))

var.based.popMean <- function(x, mu) { sum((x - mu)^2) / (length(x)) }
var.based.n       <- function(x)     { sum((x - mean(x))^2) / (length(x)) }

sample.var.based.Pop <- apply(X = samples, MARGIN = 2,
                              FUN = var.based.popMean, mu = 350)

sample.var.n.instead <- apply(X = samples, MARGIN = 2,
                              FUN = var.based.n)

sample.standard.var  <- apply(X = samples, MARGIN = 2, FUN = var) seed, [])
}
```

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

$$s^2_{\mu} = \frac{\sum_{i=1}^{n}(X_i - \boldsymbol{\mu})^2}{n}$$

Bias <- sample.var.based.Pop - sample.var.n.instead (1,000,000)
Average(Bias) <- 0.89999
true.var / n <- 9/100 <- 0.09 -> Average(Bias) = true.var / n

Bias in estimating sample variance based on n (and not n-1)

$n = 30; \mu = 350; \sigma^2$=*100; σ=10*

Bias <- sample.var.based.Pop - sample.var.n.instead (1,000,000)
Average(Bias) <- 3.333506
true.var / n <- 100/30 <- 3.333333 -> **Average(Bias) = true.var / n**

$n = 100; \mu = 32;\ \sigma^2$=9; σ=3

Biases <- sample.var.based.Pop - sample.var.n.instead
Average(Biases) <- 0.89999
true.var / n <- 9/100 <- 0.09 -> **Average(Bias) = true.var / n**

the bias equals σ²/n in the long run 1,000,0000 (infinite samples).

$$\left.\begin{array}{l} \tilde{s}^2_{\bar{X}} = \dfrac{\sum_{i=1}^{n}(X_i - \bar{\mathbf{X}})^2}{n} \\ s^2_{\mu} = \dfrac{\sum_{i=1}^{n}(X_i - \boldsymbol{\mu})^2}{n} \end{array}\right\}$$

sample variance (n)

population variance

$$\text{average}(\tilde{s}^2_{\bar{X}}) = \text{average}\left(s^2_{\mu}\right) - \sigma^2/n$$

$$\downarrow$$

$$\mathbb{E}(\tilde{\boldsymbol{s}}^2_{\bar{X}}) = \mathbb{E}\left(\boldsymbol{s}^2_{\mu}\right) - \boldsymbol{\sigma}^2/\boldsymbol{n}$$

The variance estimator that divides by *n* ($\tilde{s}^2_{\bar{X}}$) is smaller than the true variance by exactly σ^2/n on average.

Correction bias using n-1

$$\mathbb{E}(\tilde{s}^2_{\bar{X}}) = \mathbb{E}\left(s^2_{\mu}\right) - \sigma^2/n$$

Since $\mathbb{E}\left(s^2_{\mu}\right) = \sigma^2$, this becomes:

$$\mathbb{E}(\tilde{s}^2_{\bar{X}}) = \sigma^2 - \sigma^2/n$$

Step 1 − Factor σ^2:

$$\mathbb{E}(\tilde{s}^2_{\bar{X}}) = \sigma^2 - (1 - 1/n)$$

Step 2 − Simplify the parenthesis:

$\left(1 - \frac{1}{n}\right) = \left(\frac{n-1}{n}\right)$, **So:** $\mathbb{E}(\tilde{s}^2_{\bar{X}}) = \frac{n-1}{n}\sigma^2$

That′s the biased variance formula

Correction bias using n-1

$$\mathbb{E}(\tilde{s}^2_{\bar{X}}) = \mathbb{E}\left(s^2_{\mu}\right) - \sigma^2/n$$

Since $\mathbb{E}\left(s^2_{\mu}\right) = \sigma^2$, this becomes:

$$\mathbb{E}(\tilde{s}^2_{\bar{X}}) = \sigma^2 - \sigma^2/n$$

Step 1 – Factor σ^2:

$$\mathbb{E}(\tilde{s}^2_{\bar{X}}) = \sigma^2 - (1 - 1/n)$$

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

Step 2 – Simplify the parenthesis:

$$\left(1 - \frac{1}{n}\right) = \left(\frac{n-1}{n}\right), \text{So: } \mathbb{E}(\tilde{s}^2_{\bar{X}}) = \frac{n-1}{n}\sigma^2$$

That′s the same biased variance formula but ready for correction:

Step 3 – Multiply both sides by $\frac{n}{n-1}$:

$$\frac{n}{n-1}\,\mathbb{E}(\tilde{s}^2_{\bar{X}}) = \frac{n-1}{n}\sigma^2\,\frac{n}{n-1} \longrightarrow \frac{n}{n-1}\,\mathbb{E}(\tilde{s}^2_{\bar{X}}) = \sigma^2 \longrightarrow \mathbb{E}\left(\frac{n}{n-1}\,\tilde{s}^2_{\bar{X}}\right) = \sigma^2$$

$$\sigma^2 = \mathbb{E}\left(\frac{n}{n-1}\,\tilde{s}^2_{\bar{X}}\right)$$

Correction bias using n-1

$$\boldsymbol{\sigma}^2 = \mathbb{E}\left(\frac{\boldsymbol{n}}{\boldsymbol{n-1}}\ \tilde{\boldsymbol{s}}^{\mathbf{2}}_{\bar{\boldsymbol{X}}}\right)$$

Recall the definition of the biased estimator:

$$\tilde{s}^2_{\bar{X}} = \frac{\sum_{i=1}^{n}(X_i - \bar{\mathrm{X}})^2}{n} = \frac{\mathrm{SS}_{\bar{\mathrm{X}}}}{n}$$

$\mathrm{SS}_{\bar{\mathrm{X}}}$-> sum of squares of the mean

Substitute into the expression:

$$\boldsymbol{\sigma}^2 = \mathbb{E}\left(\frac{\boldsymbol{n}}{\boldsymbol{n-1}} \cdot \frac{\mathrm{SS}_{\bar{\mathrm{X}}}}{n}\right)$$

Simplify inside the expectation and *n* cancels:

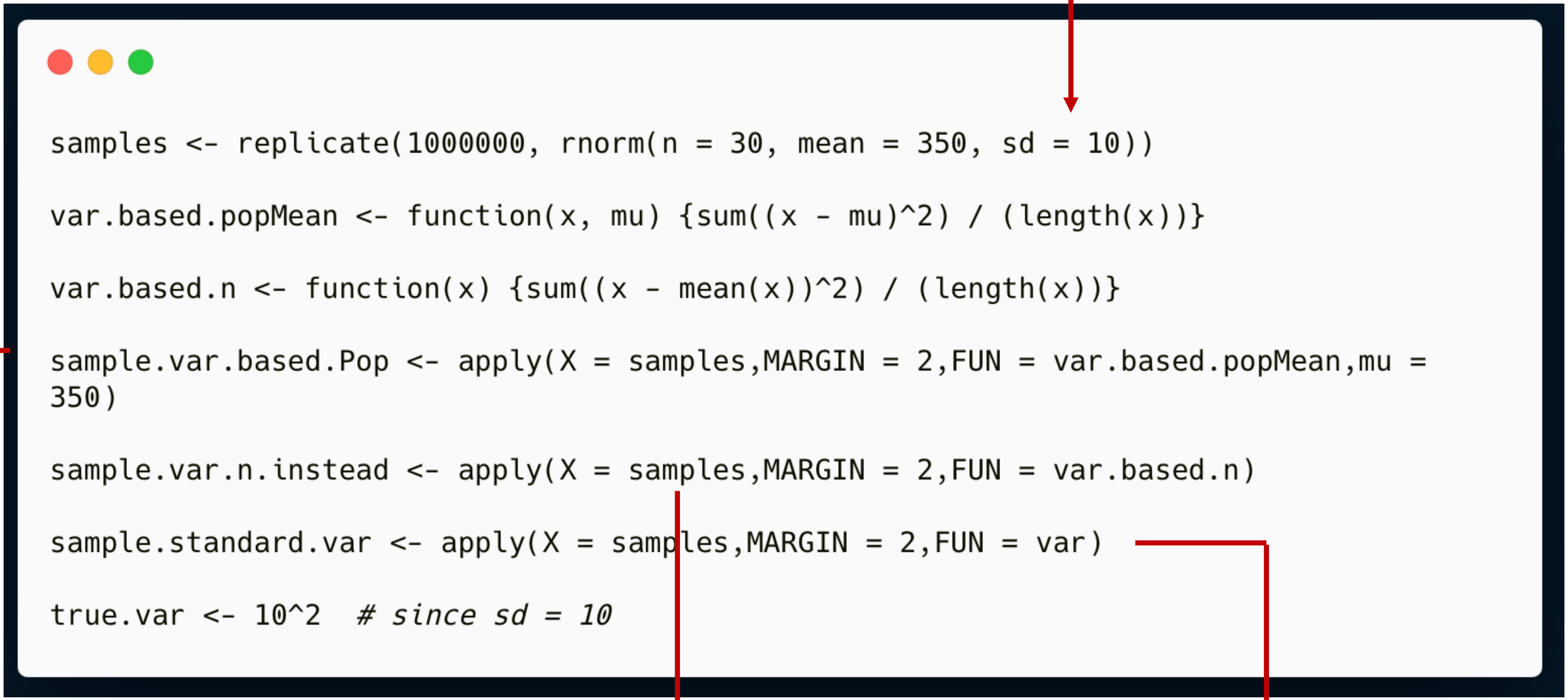
$$\frac{\boldsymbol{n}}{\boldsymbol{n-1}} \cdot \frac{\mathrm{SS}_{\bar{\mathrm{X}}}}{n} = \frac{\mathrm{SS}_{\bar{\mathrm{X}}}}{n-1}$$

So now we have:

$$\boldsymbol{\sigma}^2 = \mathbb{E}\left(\frac{\mathrm{SS}_{\bar{\mathrm{X}}}}{\boldsymbol{n-1}}\right) \longrightarrow s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\boldsymbol{n-1}}$$

Let's use a computational approach to evaluate the accuracy of these two sample-based estimators.:

σ=10 ∴ σ^2=100

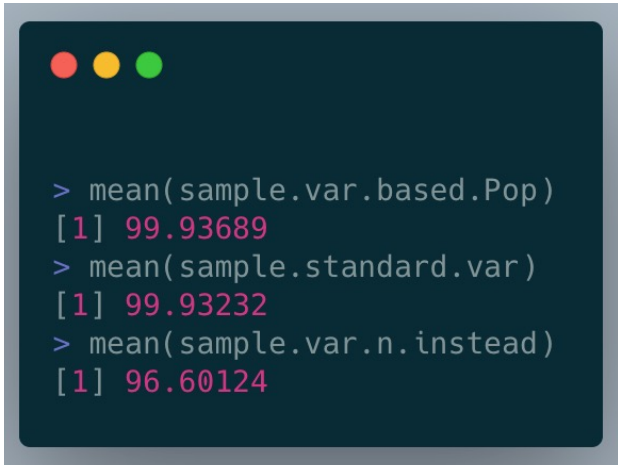


```
samples <- replicate(1000000, rnorm(n = 30, mean = 350, sd = 10))

var.based.popMean <- function(x, mu) {sum((x - mu)^2) / (length(x))}

var.based.n <- function(x) {sum((x - mean(x))^2) / (length(x))}

sample.var.based.Pop <- apply(X = samples,MARGIN = 2,FUN = var.based.popMean,mu =
350)

sample.var.n.instead <- apply(X = samples,MARGIN = 2,FUN = var.based.n)

sample.standard.var <- apply(X = samples,MARGIN = 2,FUN = var)

true.var <- 10^2  # since sd = 10
```

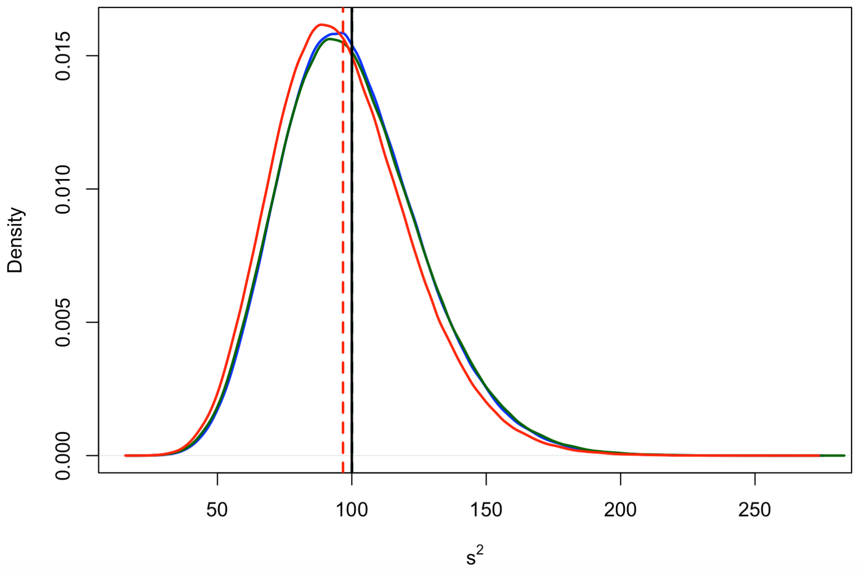
$$s_{\mu}^2 = \frac{\sum_{i=1}^{n}(X_i - \boldsymbol{\mu})^2}{n}$$

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

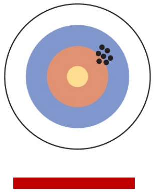
$$s_{\bar{X}}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\boldsymbol{n-1}}$$



```
> mean(sample.var.based.Pop)
[1] 99.93689
> mean(sample.standard.var)
[1] 99.93232
> mean(sample.var.n.instead)
[1] 96.60124
```

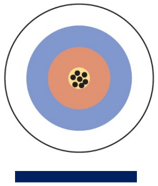


The mean of s^2 for the estimator based on the population mean μ divided by *n* and the one based on the sample mean $\bar{X}$ were unbiased (i.e., matched the population σ^2; it would have exactly equalled σ^2 =100 based on infinite sampling). However, the estimator based on the sample mean $\bar{X}$ divided by *n* is biased.



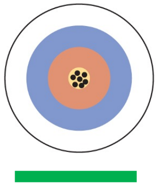
$$\tilde{s}^2_{\bar{X}} = \frac{\sum_{i=1}^{n}(X_i \; - \bar{\mathbf{X}})^2}{n}$$

👎



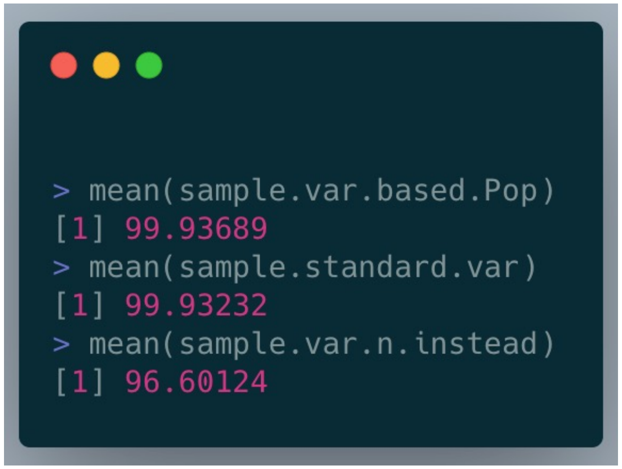
$$s^2_{\mu} = \frac{\sum_{i=1}^{n}(X_i \; - \boldsymbol{\mu})^2}{n}$$

👍

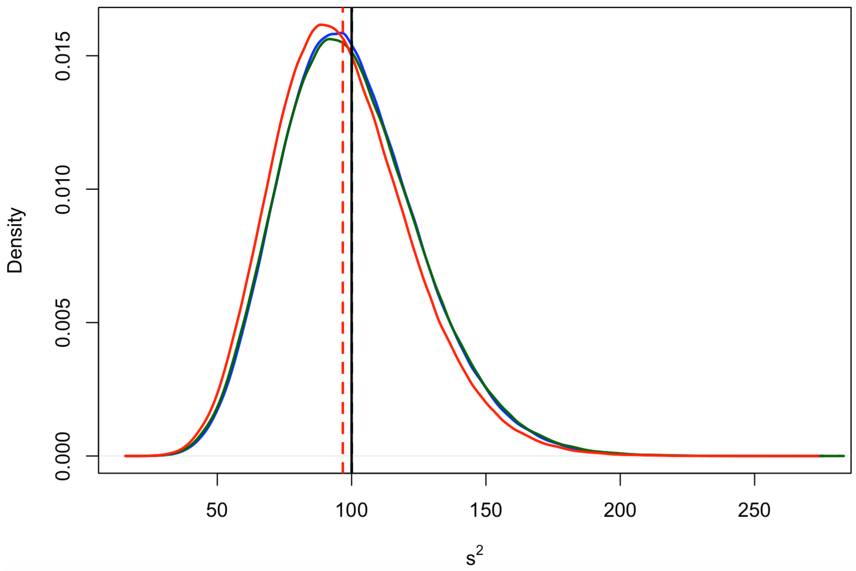


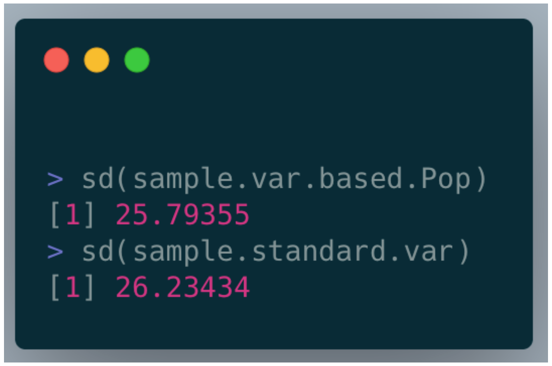
$$s^2_{\bar{X}} = \frac{\sum_{i=1}^{n}(X_i \; - \bar{X})^2}{\boldsymbol{n-1}}$$

👍



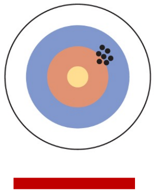
```
> mean(sample.var.based.Pop)
[1] 99.93689
> mean(sample.standard.var)
[1] 99.93232
> mean(sample.var.n.instead)
[1] 96.60124
```



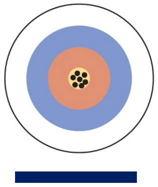


```
> sd(sample.var.based.Pop)
[1] 25.79355
> sd(sample.standard.var)
[1] 26.23434
```

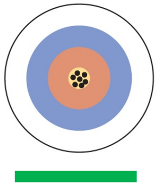
The mean of s^2 for the estimator based on the population mean μ divided by *n* and the one based on the sample mean $\bar{X}$ were unbiased (i.e., matched the population σ^2; it would have exactly equalled σ^2 =100 based on infinite sampling). However, the estimator based on the sample mean $\bar{X}$ divided by *n* is biased.



$$\tilde{s}_{\bar{X}}^2 = \frac{\sum_{i=1}^{n}(X_i \ - \bar{\mathbf{X}})^2}{n}$$ 👎



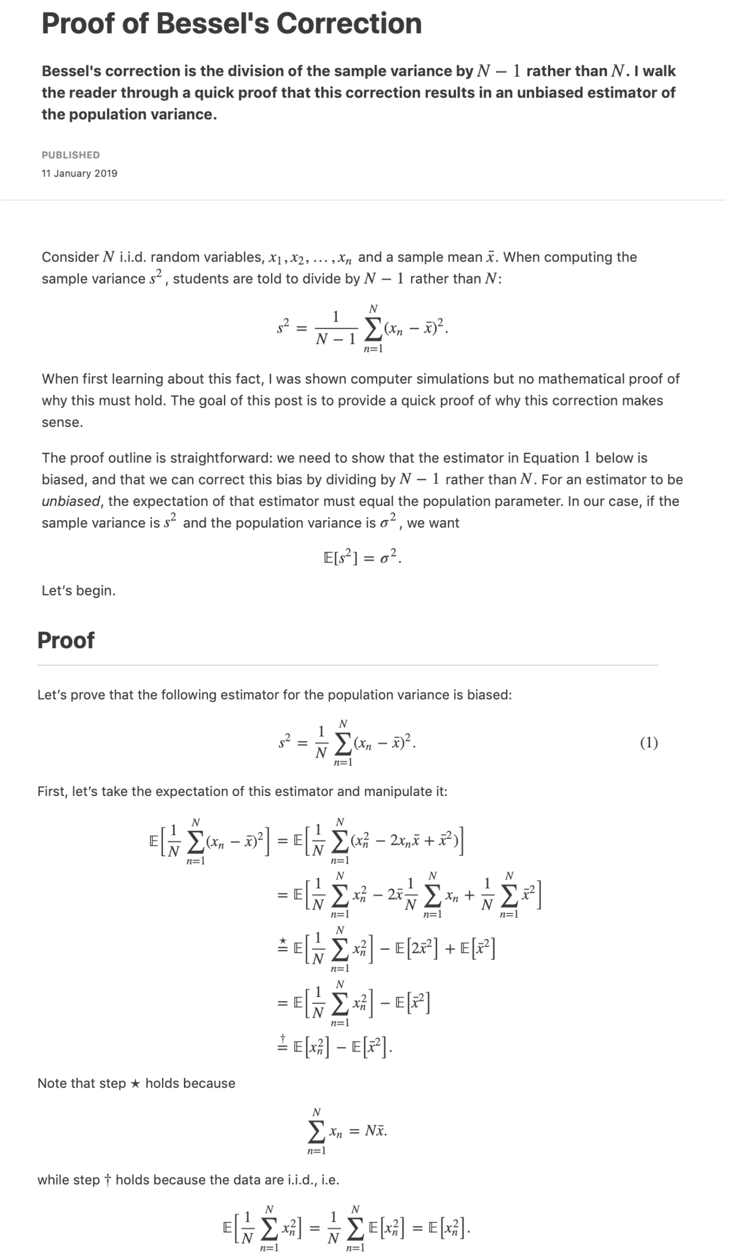
$$s_{\mu}^2 = \frac{\sum_{i=1}^{n}(X_i \ - \boldsymbol{\mu})^2}{n}$$ 👍



$$s_{\bar{X}}^2 = \frac{\sum_{i=1}^{n}(X_i \ - \bar{X})^2}{\boldsymbol{n-1}}$$ 👍

The precision of $s_{\mu}^2 > s_{\bar{X}}^2$

The math way: called the Bessel's correction – no need to know the math



Proof of Bessel's Correction

Bessel's correction is the division of the sample variance by $N-1$ rather than N. I walk the reader through a quick proof that this correction results in an unbiased estimator of the population variance.

PUBLISHED
11 January 2019

Consider N i.i.d. random variables, $x_1, x_2, \dots, x_n$ and a sample mean $\bar{x}$. When computing the sample variance s^2, students are told to divide by $N-1$ rather than N:

$$s^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})^2.$$

When first learning about this fact, I was shown computer simulations but no mathematical proof of why this must hold. The goal of this post is to provide a quick proof of why this correction makes sense.

The proof outline is straightforward: we need to show that the estimator in Equation 1 below is biased, and that we can correct this bias by dividing by $N-1$ rather than N. For an estimator to be *unbiased*, the expectation of that estimator must equal the population parameter. In our case, if the sample variance is s^2 and the population variance is σ^2, we want

$$\mathbb{E}[s^2] = \sigma^2.$$

Let's begin.

Proof

Let's prove that the following estimator for the population variance is biased:

$$s^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2. \tag{1}$$

First, let's take the expectation of this estimator and manipulate it:

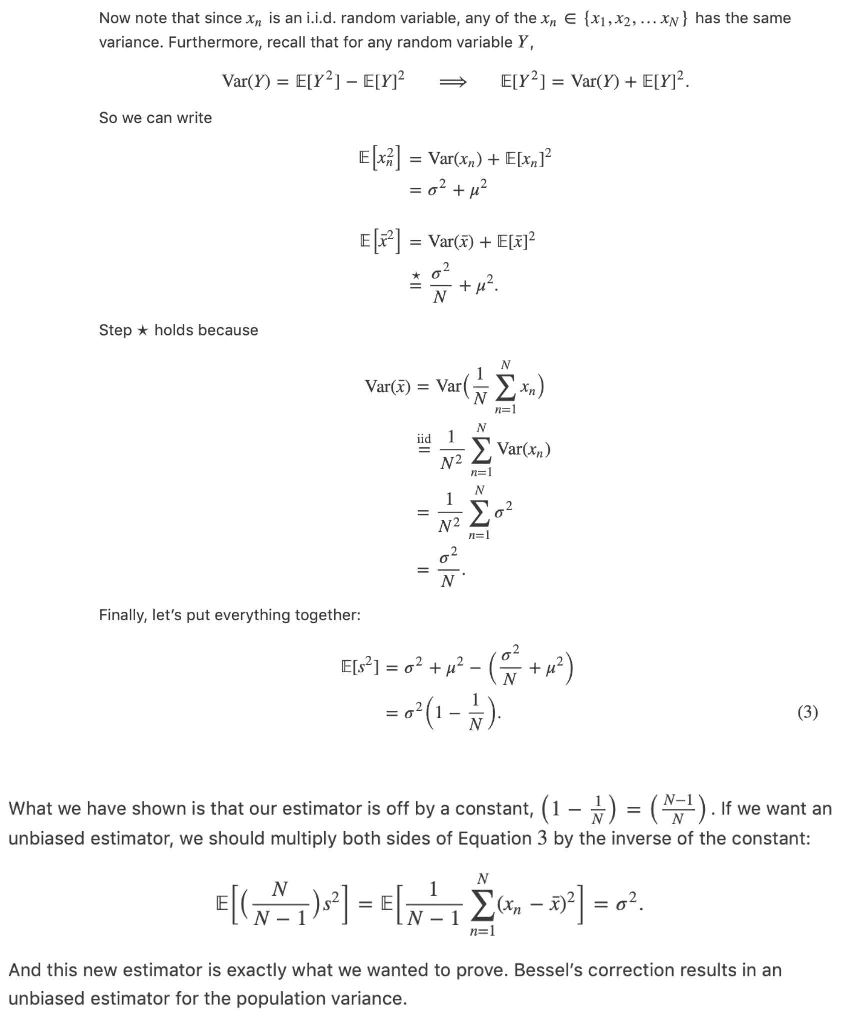
$$\begin{aligned}
\mathbb{E}\Big[\frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2\Big] &= \mathbb{E}\Big[\frac{1}{N}\sum_{n=1}^{N}(x_n^2 - 2x_n\bar{x} + \bar{x}^2)\Big] \\
&= \mathbb{E}\Big[\frac{1}{N}\sum_{n=1}^{N}x_n^2 - 2\bar{x}\frac{1}{N}\sum_{n=1}^{N}x_n + \frac{1}{N}\sum_{n=1}^{N}\bar{x}^2\Big] \\
&\overset{\star}{=} \mathbb{E}\Big[\frac{1}{N}\sum_{n=1}^{N}x_n^2\Big] - \mathbb{E}\big[2\bar{x}^2\big] + \mathbb{E}\big[\bar{x}^2\big] \\
&= \mathbb{E}\Big[\frac{1}{N}\sum_{n=1}^{N}x_n^2\Big] - \mathbb{E}\big[\bar{x}^2\big] \\
&\overset{\dagger}{=} \mathbb{E}\big[x_n^2\big] - \mathbb{E}\big[\bar{x}^2\big].
\end{aligned}$$

Note that step ★ holds because

$$\sum_{n=1}^{N}x_n = N\bar{x}.$$

while step † holds because the data are i.i.d., i.e.

$$\mathbb{E}\Big[\frac{1}{N}\sum_{n=1}^{N}x_n^2\Big] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\big[x_n^2\big] = \mathbb{E}\big[x_n^2\big].$$



Now note that since x_n is an i.i.d. random variable, any of the $x_n \in \{x_1, x_2, \dots x_N\}$ has the same variance. Furthermore, recall that for any random variable Y,

$$\text{Var}(Y) = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 \quad\implies\quad \mathbb{E}[Y^2] = \text{Var}(Y) + \mathbb{E}[Y]^2.$$

So we can write

$$\begin{aligned}
\mathbb{E}\big[x_n^2\big] &= \text{Var}(x_n) + \mathbb{E}[x_n]^2 \\
&= \sigma^2 + \mu^2
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\big[\bar{x}^2\big] &= \text{Var}(\bar{x}) + \mathbb{E}[\bar{x}]^2 \\
&\overset{\star}{=} \frac{\sigma^2}{N} + \mu^2.
\end{aligned}$$

Step ★ holds because

$$\begin{aligned}
\text{Var}(\bar{x}) &= \text{Var}\Big(\frac{1}{N}\sum_{n=1}^{N}x_n\Big) \\
&\overset{\text{iid}}{=} \frac{1}{N^2}\sum_{n=1}^{N}\text{Var}(x_n) \\
&= \frac{1}{N^2}\sum_{n=1}^{N}\sigma^2 \\
&= \frac{\sigma^2}{N}.
\end{aligned}$$

Finally, let's put everything together:

$$\begin{aligned}
\mathbb{E}[s^2] &= \sigma^2 + \mu^2 - \Big(\frac{\sigma^2}{N} + \mu^2\Big) \\
&= \sigma^2\Big(1 - \frac{1}{N}\Big).
\end{aligned} \tag{3}$$

What we have shown is that our estimator is off by a constant, $\left(1 - \frac{1}{N}\right) = \left(\frac{N-1}{N}\right)$. If we want an unbiased estimator, we should multiply both sides of Equation 3 by the inverse of the constant:

$$\mathbb{E}\Big[\Big(\frac{N}{N-1}\Big)s^2\Big] = \mathbb{E}\Big[\frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})^2\Big] = \sigma^2.$$

And this new estimator is exactly what we wanted to prove. Bessel's correction results in an unbiased estimator for the population variance.

Source: http://gregorygundersen.com/blog/2019/01/11/bessel/

Let’s take a small break – 1 minute

But why is the variance (or standard deviation) biased when divided by n instead of n-1?

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

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

👍 👎

No Math then! Let’s try a more accessible way to understand the need for a correction [“a gentle introduction to degrees of freedom”]

To understand why we use n - 1 rather than n, we need to recognize a key constraint: observations in a sample can vary freely around the true population mean $\boldsymbol{\mu}$, but once we calculate the sample mean $\bar{X}$, the deviations from $\bar{X}$ are no longer independent, i.e., they must sum to zero.

$$s_{\bar{X}}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\boldsymbol{n-1}} \qquad \tilde{s}_{\bar{X}}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{\mathbf{X}})^2}{n}$$

Free to vary **Not free to vary**

To understand why we use n - 1 rather than n, we need to recognize a key constraint: observations in a sample can vary freely around the true population mean $\boldsymbol{\mu}$, but once we calculate the sample mean $\bar{X}$, the deviations from $\bar{X}$ are no longer independent, i.e., they must sum to zero.

Let's say we have a set of 6 numbers, but one number is hidden. If we know the sample mean $\bar{X}$, we can use it to find the missing number: $1, 5, 7, ???, 9, 12 \;\; \bar{Y} = 7$

$$\frac{1 + 5 + 7 + ??? + 9 + 12}{6} = 7 \;\; \therefore 34 + ??? = 6 \times 7$$

6×7

??? = 42 - 34 = 8

So, there is always one number that is not free to vary around the sample mean $\bar{Y}$

To understand why we use n - 1 rather than n, we need to recognize a key constraint: observations in a sample can vary freely around the true population mean $\boldsymbol{\mu}$, but once we calculate the sample mean $\bar{X}$, the deviations from $\bar{X}$ are no longer independent, i.e., they must sum to zero.

Value	Deviation from the mean
1	1-7=-6
5	5-7=-2
7	7-7=0
8	8-7=1
9	9-7=2
12	12-7=5
SUM	0

-2 + 0 + 1 + 2 + 5= 6 (missing value is -6)

-6 -2 + 1 + 2 + 5= 0 (missing value is 0)

-6 -2 + 0 + 1 + 2 = -5 (missing value is 5)

So, only n - 1 deviations from the sample mean are free to vary independently.

Let’s assume we know the population mean $\boldsymbol{\mu} = 6$ (though, in reality, this is usually unknown - this is to illustrate the point).

Based on the sample mean $\bar{\mathbf{X}}$:

$$s_{\bar{\mathbf{X}}}^2 = \frac{(1-7)^2+(5-7)^2+(7-7)^2+(8-7)^2+(9-7)^2+(12-7)^2}{n}$$

$$= \frac{70}{6} = 11.7$$

Based on the population mean $\boldsymbol{\mu}$

$$s_{\mu}^2 = \frac{(1-6)^2+(5-6)^2+(7-6)^2+(8-6)^2+(9-6)^2+(12-6)^2}{n}$$

$$= \frac{76}{6} = 12.7$$

Note that the sample-based values were smaller than the population-based values. *This occurs because the sample mean, in average, underestimate variability compared to the true population mean.*

Now we can trust our estimates, let’s calculate confidence intervals in practice

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

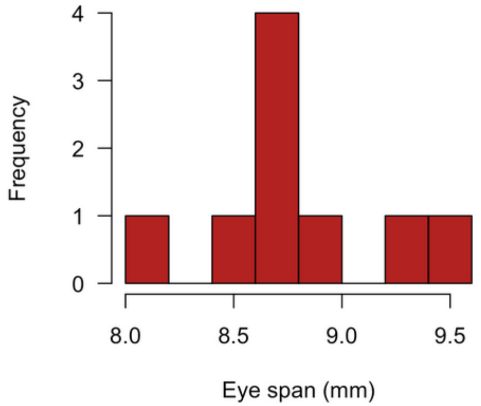
Let’s estimate the 95% confidence interval for the population mean

$$\bar{Y} = 8.778 \text{ mm } s = 0.398 \text{ mm}$$

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

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

“symmetric” (we can “trust” estimates)



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

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

Now we can trust our estimates, let's calculate confidence intervals in practice

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

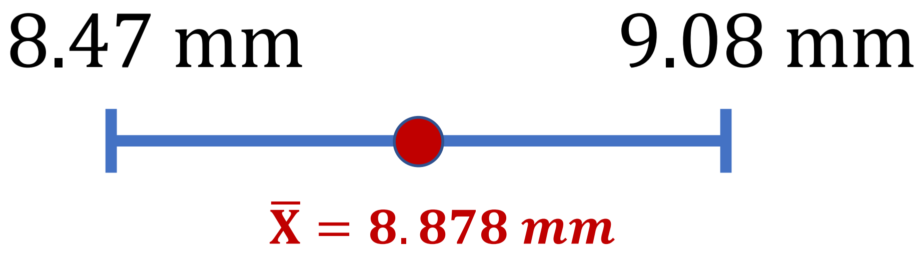
$$\mathrm{SE}_{\bar{Y}} \frac{0.398}{\sqrt{9}} = 0.133$$

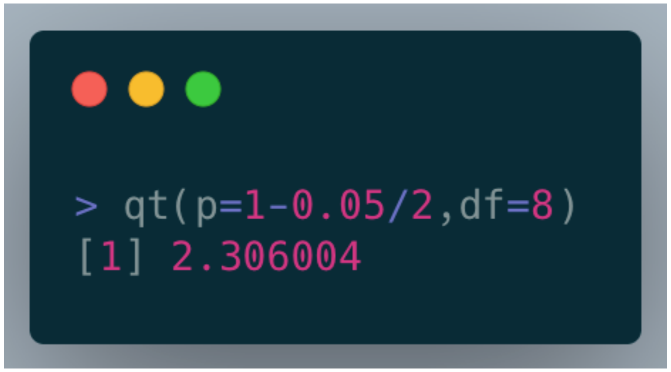
$$t_{0.05(2),8} = 2.306$$

Degrees of freedom
(v, df)

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

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





```
> qt(p=1-0.05/2,df=8)
[1] 2.306004
```

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

In practice (today) we use software (e.g., R).

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

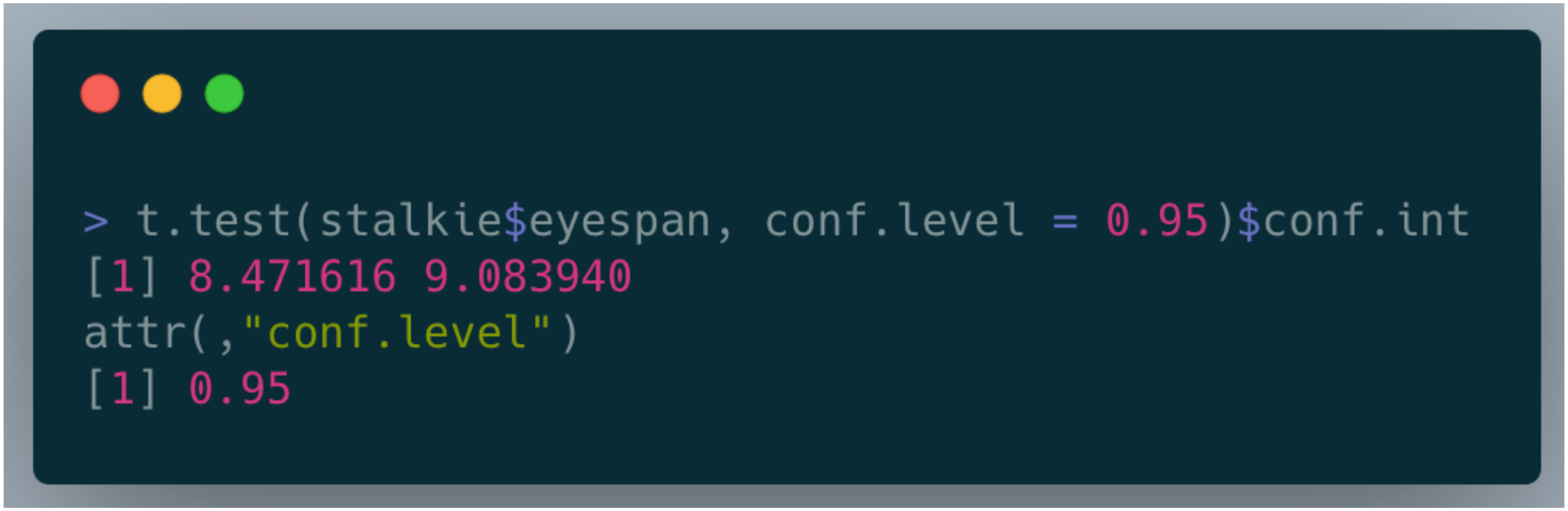
$$\mathrm{SE}_{\bar{Y}} \frac{0.398}{\sqrt{9}} = 0.133$$

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

$$\bar{Y} - 2.306 \times 0.133 < \mu < \bar{Y} + 2.31 \times 0.133$$

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

© melvyn yeo



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

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

Let’s estimate the 99% confidence interval for the population mean

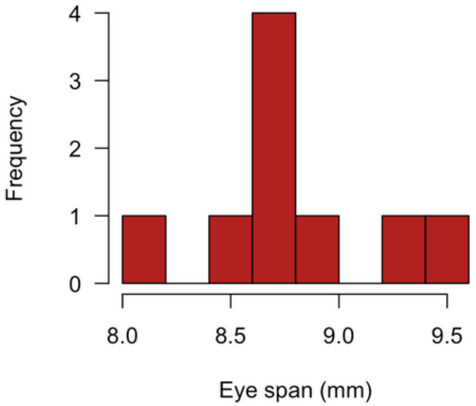
$$\bar{Y} = 8.778\ s = 0.398$$

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

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

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

$$8.33\ \mathrm{mm} < \mu < 9.22\ \mathrm{mm}$$



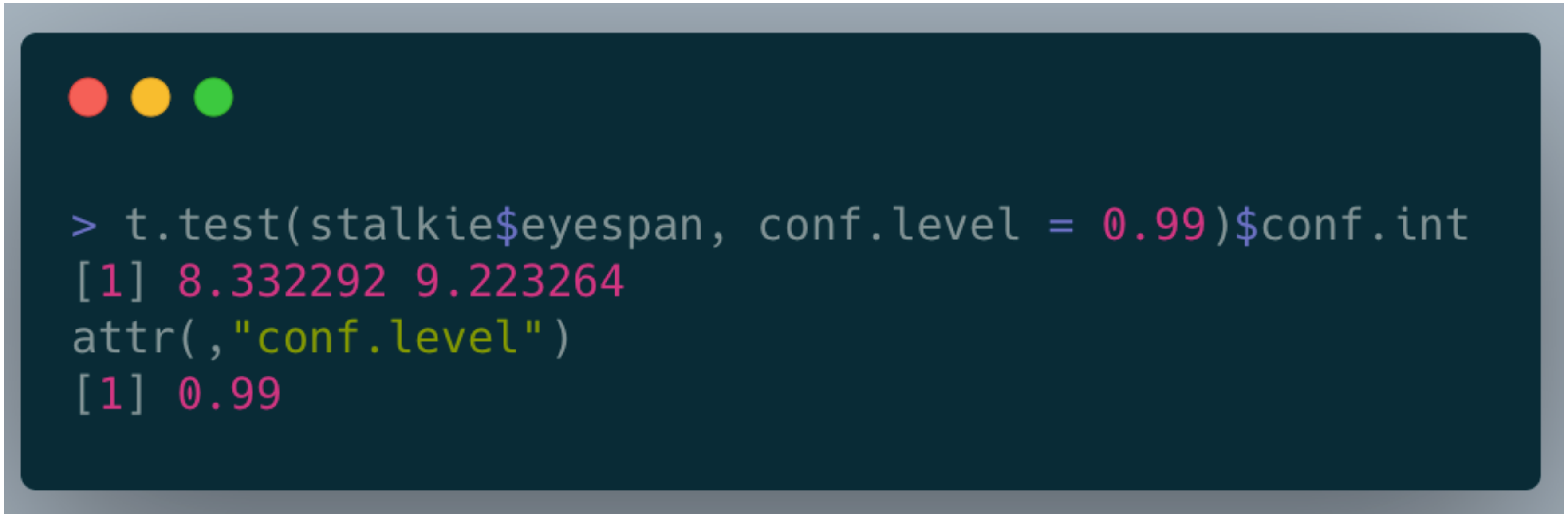
$$\bar{Y} = 8.778 \; s = 0.398$$

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

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

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

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



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

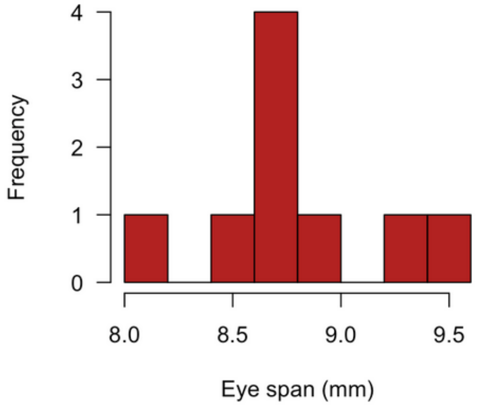
In most cases, however, we report the 95% confidence interval.

95% confidence interval:

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

99% confidence interval:

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



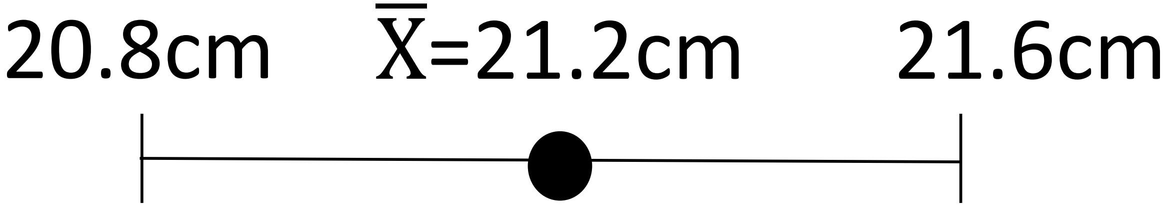
Confidence intervals can be used in many situations; e.g., decision making

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