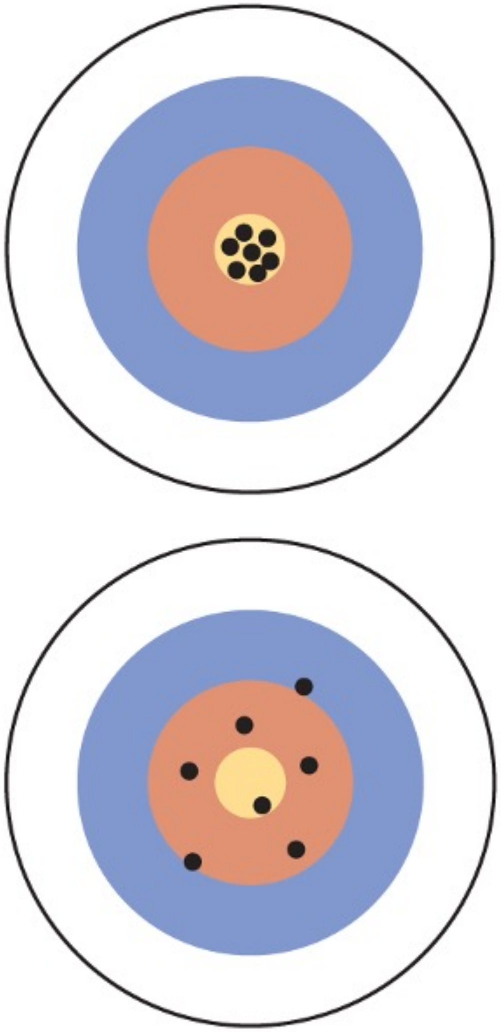
Lecture 11 - estimation is not as easy as we think: challenges and solutions from 100 years ago that are mainstream statistics!

Building long-term statistical intuition & knowledge

The statistical road: estimate with uncertainty but with confidence

WE NEED TO TRUST OUR SAMPLE ESTIMATORS FOR STATISTICS TO WORK, I.E., SAMPLES ESTIMATORS NEED TO BE UNBIASED

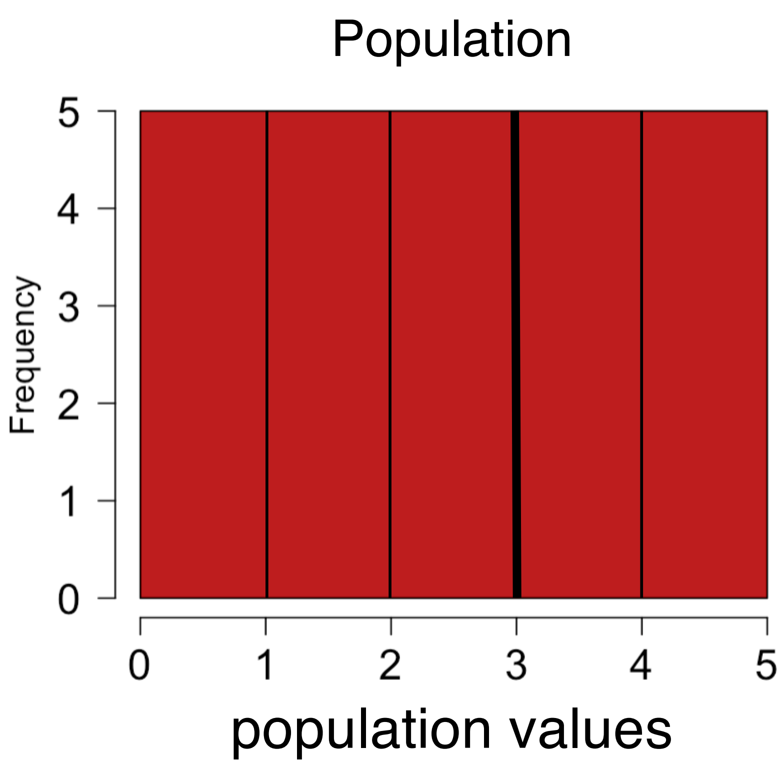


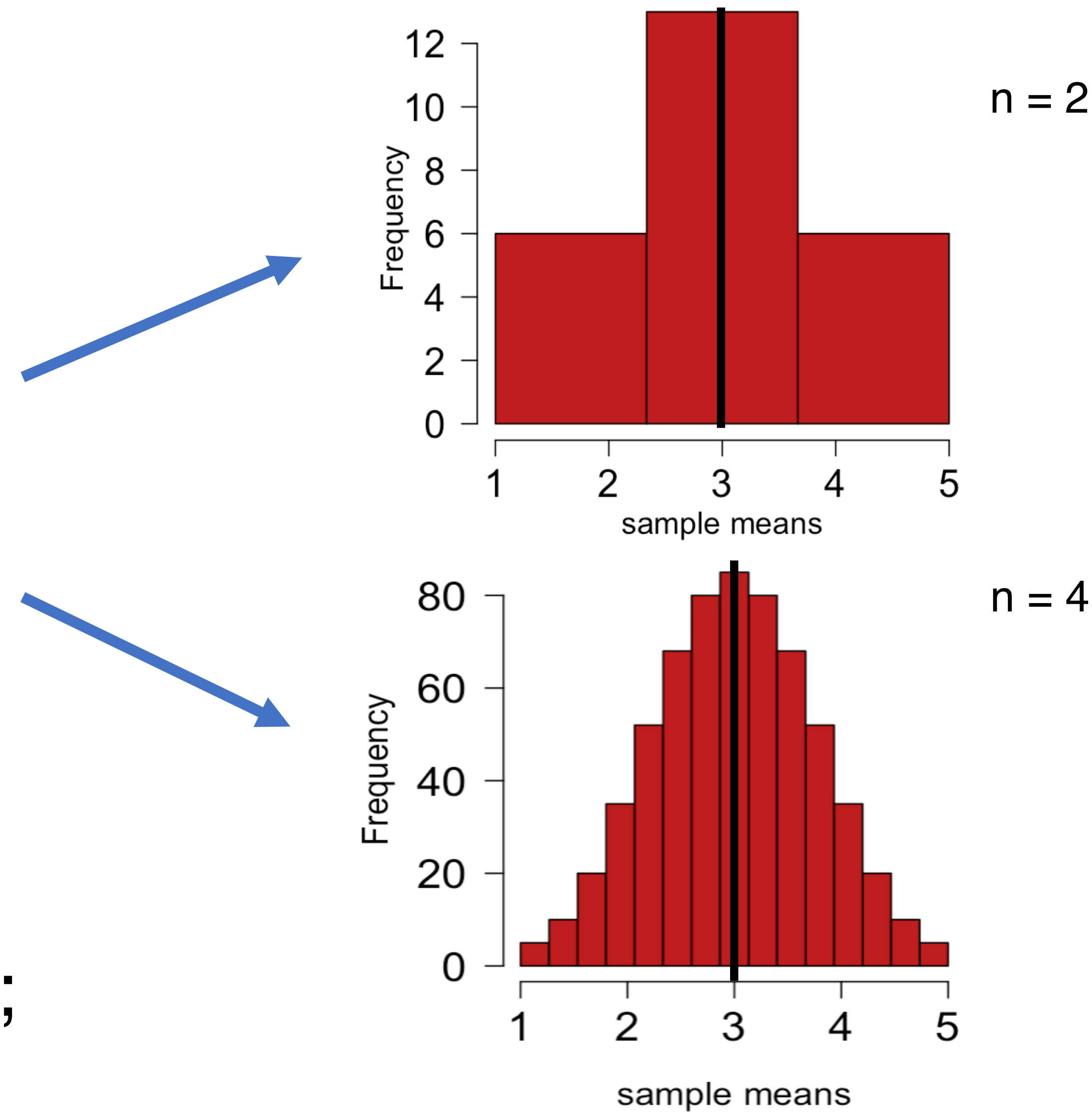
The bullseye is the population mean μ and
each dot is a sample mean $\bar{X}$.

We know that under random sampling, the sample mean is an unbiased estimator of the population mean μ.

This is because the mean of all sample means equal the population mean.

The shape of the frequency distribution of the population is not necessarily similar to the frequency distribution of the sample estimates (e.g., here the distribution of sample mean values) from the population. And regardless of the what the shape of the population distribution is (here uniform), the sample mean is an unbiased estimator of the population mean.



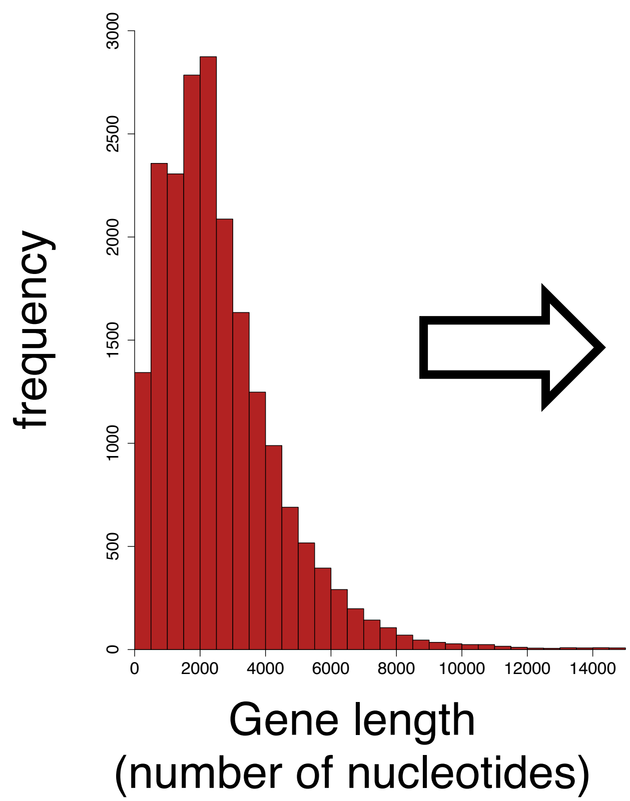


population: 1,2,3,4,5;

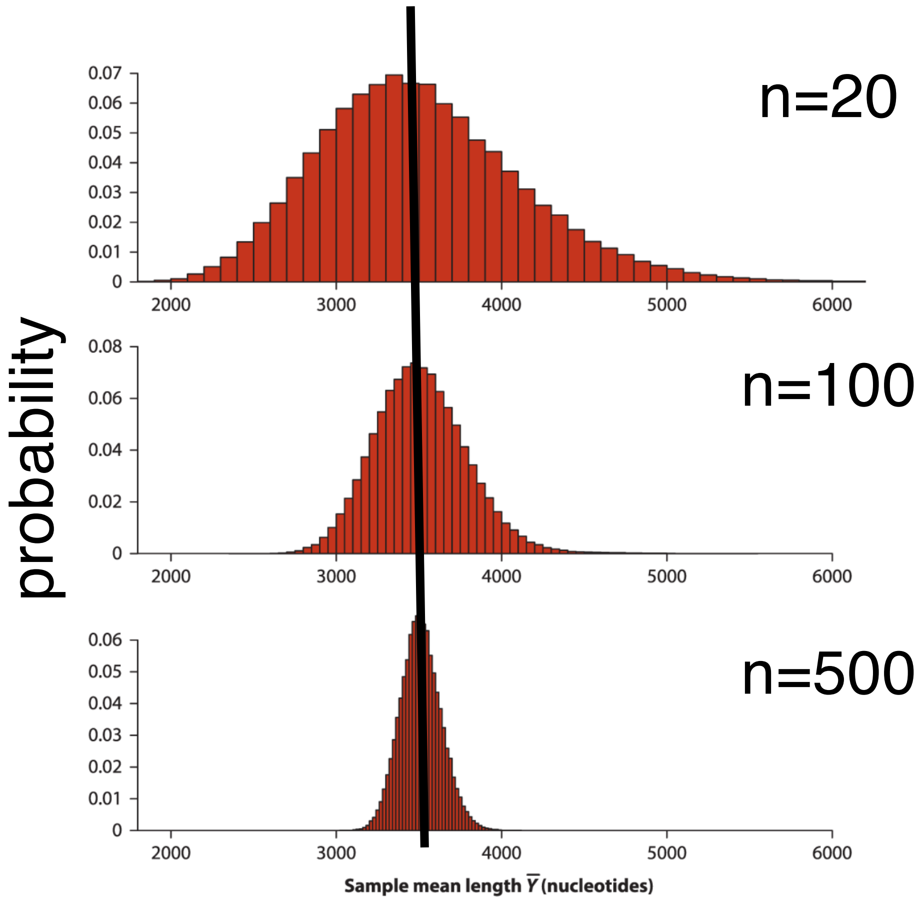
μ = **3.0**

The shape of the frequency distribution of the population is not necessarily similar to the frequency distribution of the sample estimates (e.g., here the distribution of sample mean values) from the population. And regardless of the what the shape of the population distribution is (here uniform), the sample mean is an unbiased estimator of the population mean.

Frequency distribution of the gene Population



Sampling distributions for the sample means of the gene population!

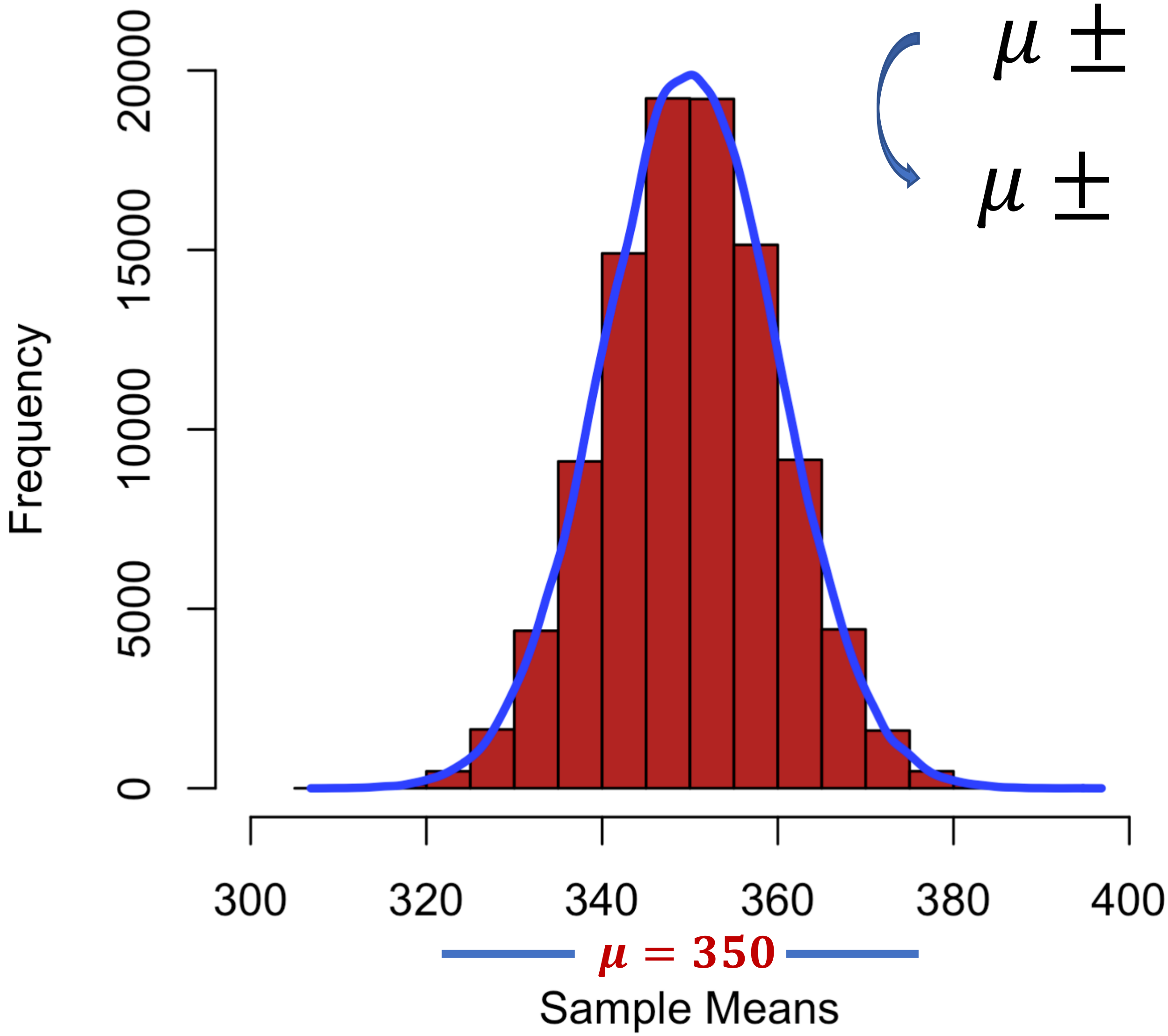


Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

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

Sampling distribution of the means for a normally distributed population follows a t-distribution (we say "is t-distributed").

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



$\mu = 350$

$$\mu \pm\ 2\times\sigma_{\bar{Y}}$$

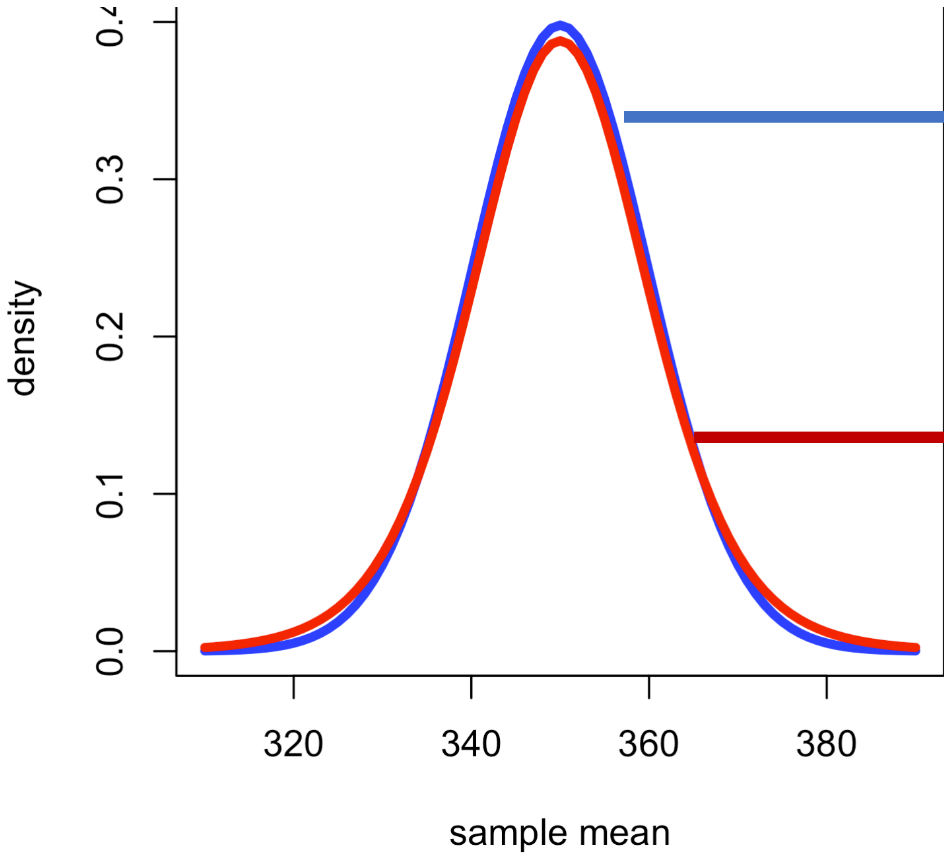
$$\mu \pm\ 1.96\times\sigma_{\bar{Y}}$$

Student's 1908 Article "The Probable Error of a Mean"

William Sealy Gosset (pseudonym: Student)

Sampling distribution of the means for a normally distributed population follows a t-distribution (we say "is t-distributed").

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



$$n = 100$$

$$\mu \pm\ 1.96 \times \sigma_{\bar{Y}}$$

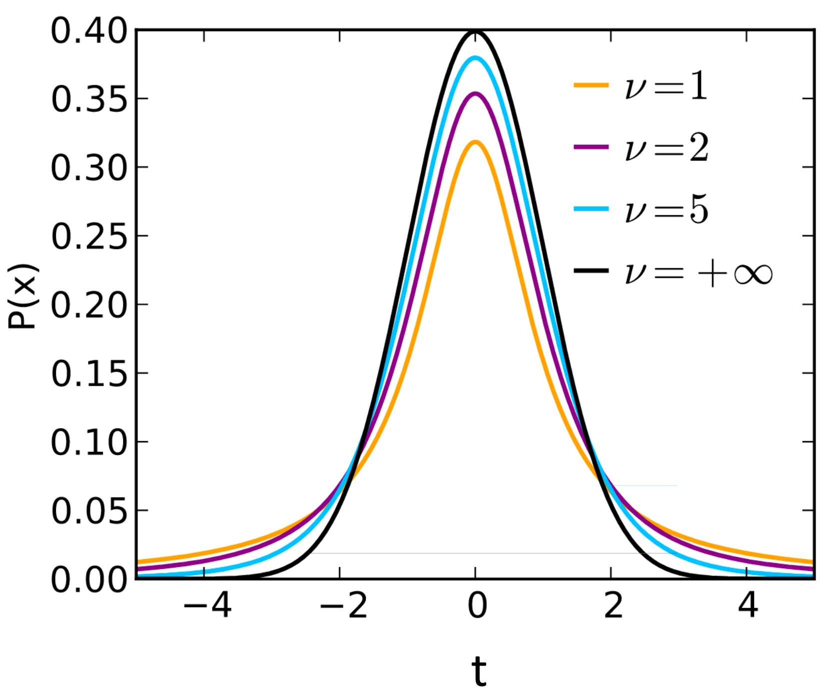
$$n = 10$$

$$\mu \pm\ 2.26 \times \sigma_{\bar{Y}}$$

$$t = \frac{\overline{X}_i - \mu}{SE_{\overline{X}_i}} \longrightarrow \overline{X}_i \pm\ t \times SE_{\overline{X}_i}$$

By now, you should suspect that one of the "inconveniences" is that the exact value needed to be multiplied by SE to create 95% confidence intervals changes as a function of sample size.

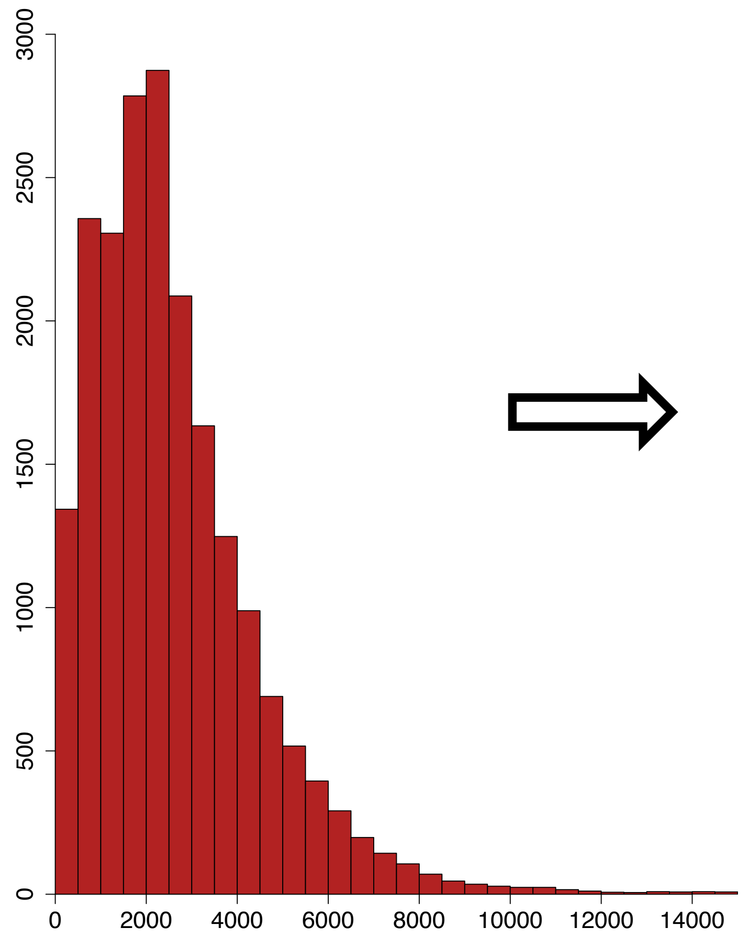
The sampling distribution of means that varies as a function of the sample size (here v = degrees of freedom; v = n - 1).

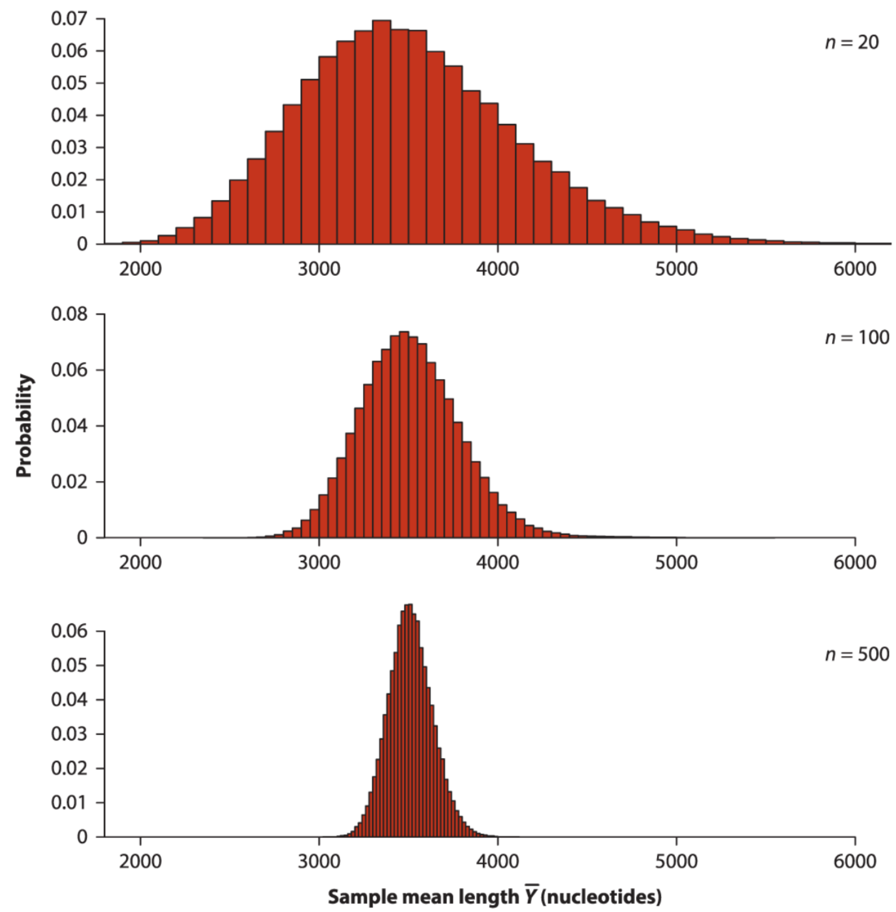


This *t* distribution (standardized) is a sampling distribution of the the number of sample standard errors away from the mean (now always 0 after the standardization) necessary to produce a confidence interval of the desired coverage (e.g., 95%).

$$t = \frac{\bar{X}_i - \mu}{SE_{\bar{X}_i}} \longrightarrow \bar{X}_i \pm t \times SE_{\bar{X}_i}$$

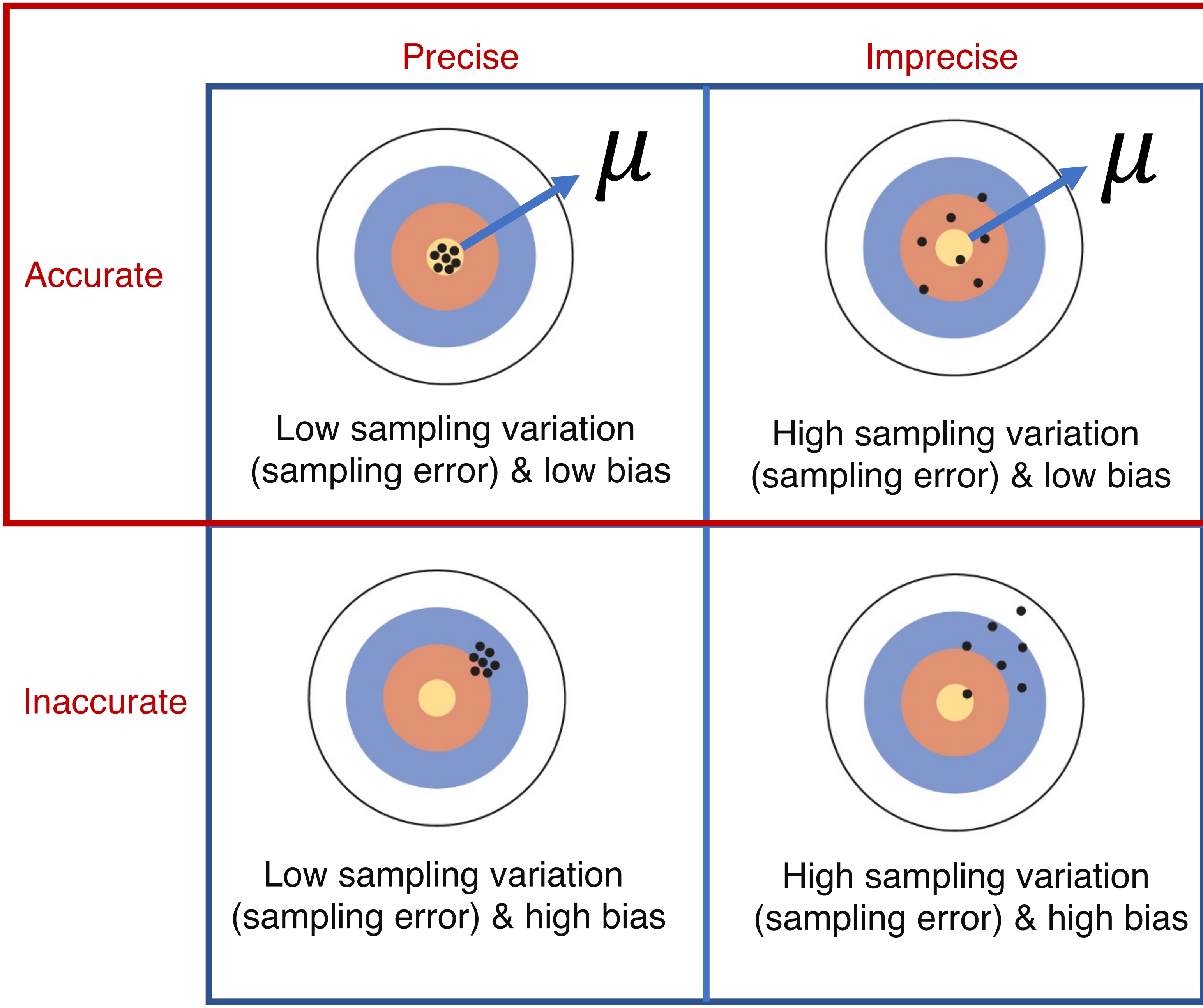
Even though the distribution of the population is asymmetric, the sampling distribution of means tend to me symmetric. This is an important property because it allows us to generalize sampling distributions based on standard distributions such as the t-distribution. I





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

WE NEED TO TRUST OUR SAMPLE ESTIMATORS FOR STATISTICS TO WORK, I.E., SAMPLES ESTIMATORS NEED TO BE UNBIASED

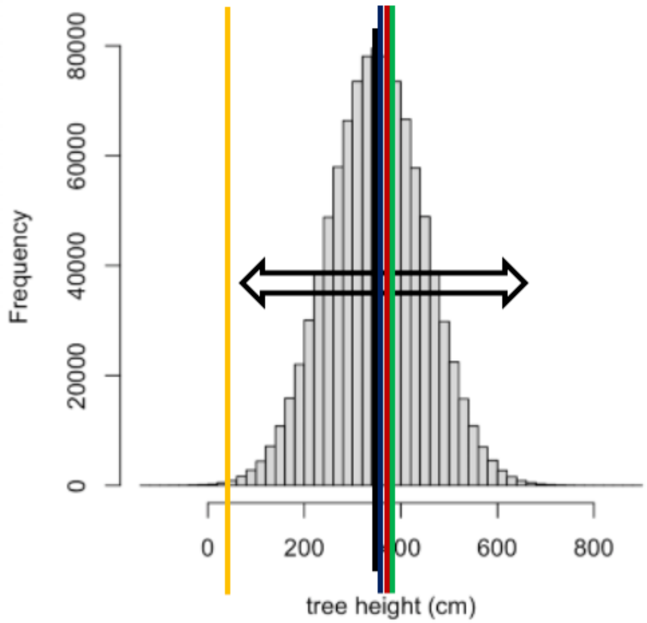


The mean is an unbiased estimator under random sampling because the mean of all sample means equal the population mean.

The variation among observations within samples (standard deviation) can inform us about how far sample means in general might be from the true population mean (estimate how wrong one could be).

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

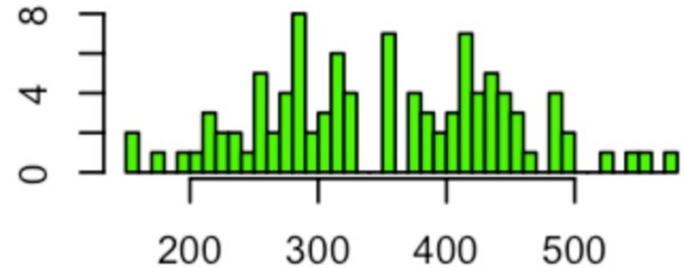
$\bar{X} = 352.3\ cm; s = 94.0\ cm$



$\mu = 350\ cm$

Variation among samples

sampling



Variation within samples

Variation within samples (among observations) can estimate certainty about uncertainty (variation among sample means)

The ability to estimate variation within a sample to estimate variation among samples (standard error) is key to statistics and not only to confidence interval.

The ability to estimate variation within a sample to estimate variation among samples (standard error) is key to statistics and not only to confidence interval.

Sampling error - the difference between sample means and the population mean. The estimate of this error is the standard deviation of the sampling distribution, i.e., the average difference between all sample means and the true mean:

$$\sigma_{\bar{Y}} = \sqrt{\sum_{i=i}^{\infty} \frac{(\bar{Y}_i - \mu)^2}{\infty}}$$

The number of samples is so large that can be considered infinite (∞)

The standard deviation of the sampling distribution of the mean $\sigma_{\bar{Y}}$ is called standard error and is exactly the standard deviation of the population σ divided by $\sqrt{n}$:

$$= \quad \sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$$

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

$$\mathrm{SE}_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

But can we trust the sample standard deviation s? Is it an unbiased estimator of σ ?

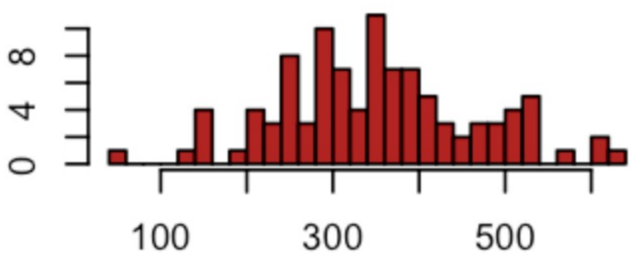
As we are going to see, the mean and standard deviation are key sample statistics used in pretty much all standard statistical analysis (i.e., not only confidence intervals)

$$\bar{Y} \pm \ t_n \times SE_{\bar{Y}} \ \therefore \ SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

Margin of error

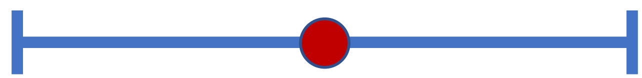
$\bar{X} = 351.5\ cm; s = 114.2\ cm$

$$351.5c \pm t_n \times \frac{114.2}{\sqrt{100}}$$



328.66 cm

374.34 cm



$\bar{X} = 351.5\ cm$

But can we trust the sample standard deviation s?
Is it an unbiased estimator of σ ?

Today we will study the case of the sample standard deviation as an estimator of the true population standard deviation.

This has three goals:

- Develop stronger knowledge and intuition about statistics.

- How statisticians work to develop statistics that we can trust.

- Acquire greater knowledge about how the other statistical frameworks we will learn in BIOL322 were developed. We won't revisit all sample estimators, but the type of work that was done for the standard deviation can be "generalized" to most sample statistics.

But can we trust the sample standard deviation s?
Is it an unbiased estimator of σ ?

1) The importance of corrections for creating unbiased sample estimators for any statistic of interest [the case of degrees of freedom].

2) The importance of the distribution of the population for creating unbiased sample estimators for any statistic of interest [the case of assumptions].

3) The importance of [data transformation] for making biased sample estimators unbiased.

1) The importance of corrections for creating unbiased sample estimators for any statistic of interest [the case of degrees of freedom].

Why is the sample standard deviation calculated by dividing the sum of the squared deviations from the mean divided by *n* – 1 and not *n*?

$$s = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\boldsymbol{n-1}}} \qquad s = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\boldsymbol{n}}}$$

But why?

Let's switch to variance s^2 (hang in there with me); and after all $s = \sqrt{s^2}$. If it were possible to know the true mean of the population μ, than the best estimator (i.e., sampled based) for the variance of the entire population based on a single sample would be:

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

But we almost never know the population mean μ, so we could try using the sample mean value $\bar{Y}$:

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{\mathbf{Y}})^2}{n}$$

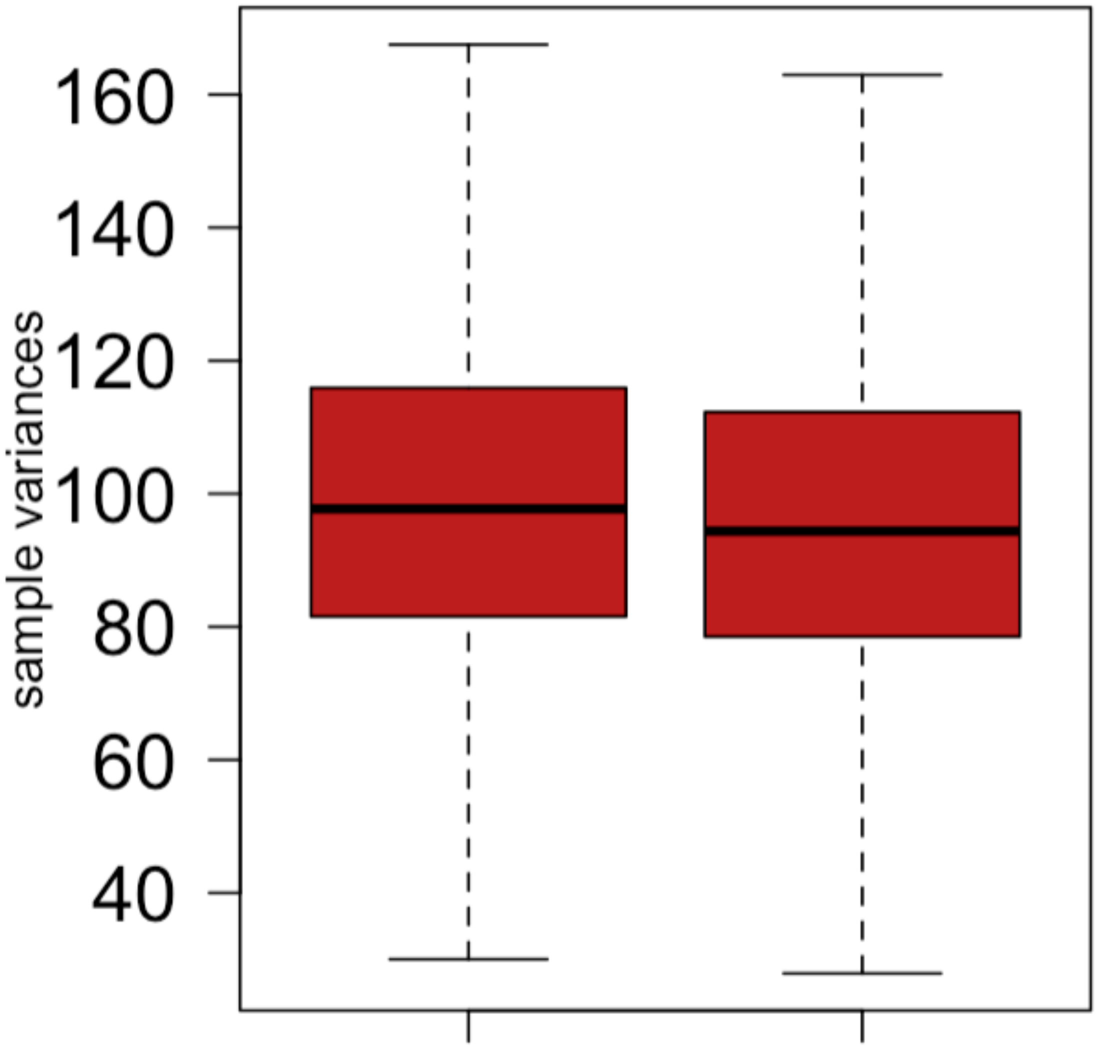
Let's use a computational approach to verify the quality of these two estimators (i.e., sample based):

σ^2=100; σ=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)

boxplot(sample.var.based.Pop,sample.var.n.instead,
        outline=FALSE,col="firebrick",
        cex.axis=1.5,las=1,ylab="sample variances")
```

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

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n}$$




```
> 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* was unbiased (i.e., pretty much the population σ^2; would had been exactly σ^2 =100 with infinite sampling); whereas the estimator based on the sample $\bar{Y}$ divided by *n* is biased.

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

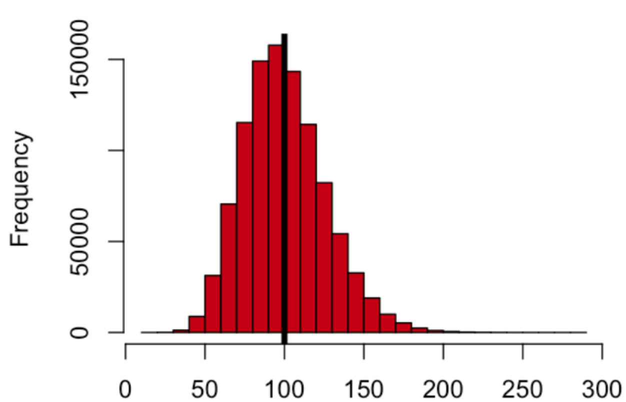
$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \boldsymbol{\bar{Y}})^2}{n}$$

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

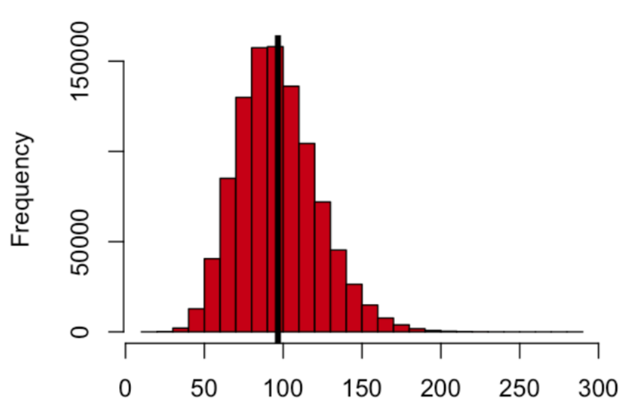
```
> mean(sample.var.based.Pop)
[1] 99.93689
> mean(sample.var.n.instead)
[1] 96.60124
```

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n}$$

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



sample variances

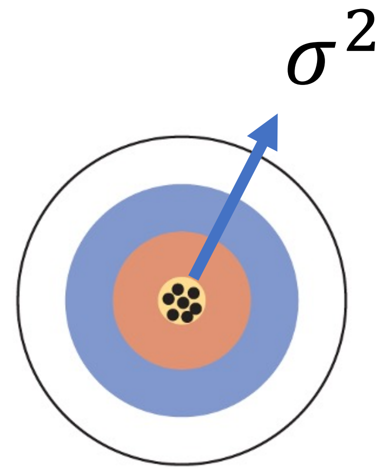


sample variances

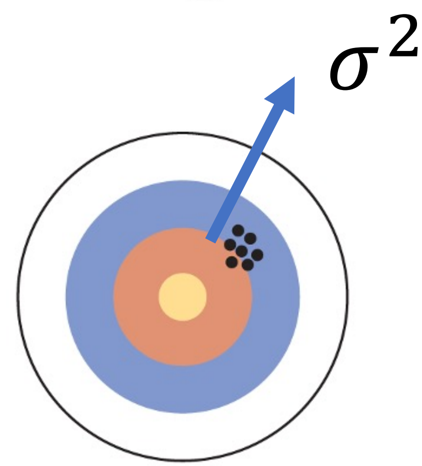
Note the asymmetry of the sampling distribution of variances; hence the median is not exactly equal to the mean. But the variance is unbiased when based on $\boldsymbol{\mu}$ but biased when based on $\bar{Y}$. Remember: unbiased expectations are based on means and not medians.

But in most (if not all) cases, one doesn't know the parameter value μ (true population mean).

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



$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{\mathbf{Y}})^2}{n}$$



There is a correction factor for the sample bias in s^2 called Bessel's correction (but seems that Gauss 1823 came up with it first)

$$s^2 = \frac{\sum_{i=1}^{n}(\mathrm{Y_i} - \boldsymbol{\mu})^2}{n} \cong \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\boldsymbol{n-1}}$$

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\boldsymbol{n}}$$

https://mathworld.wolfram.com/BesselsCorrection.html

Let's use a computational approach to verify the quality of the three estimators (i.e., sample based):

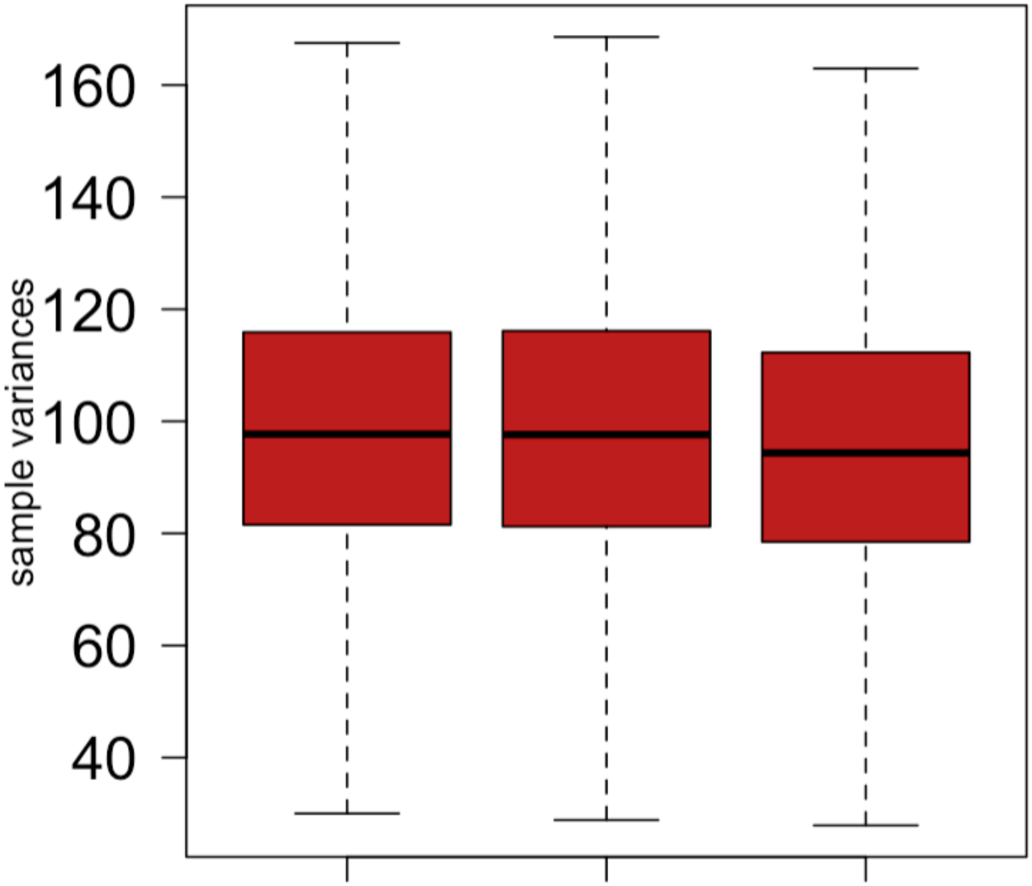
$\sigma=10 \therefore \sigma^2=100$

```
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)

boxplot(sample.var.based.Pop,sample.standard.var,sample.var.n.instead,
        outline=FALSE,col="firebrick",cex.axis=1.5,
        las=1,ylab="sample variances")
```

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

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{\mathbf{Y}})^2}{\boldsymbol{n-1}}$$

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{\mathbf{Y}})^2}{\boldsymbol{n}}$$



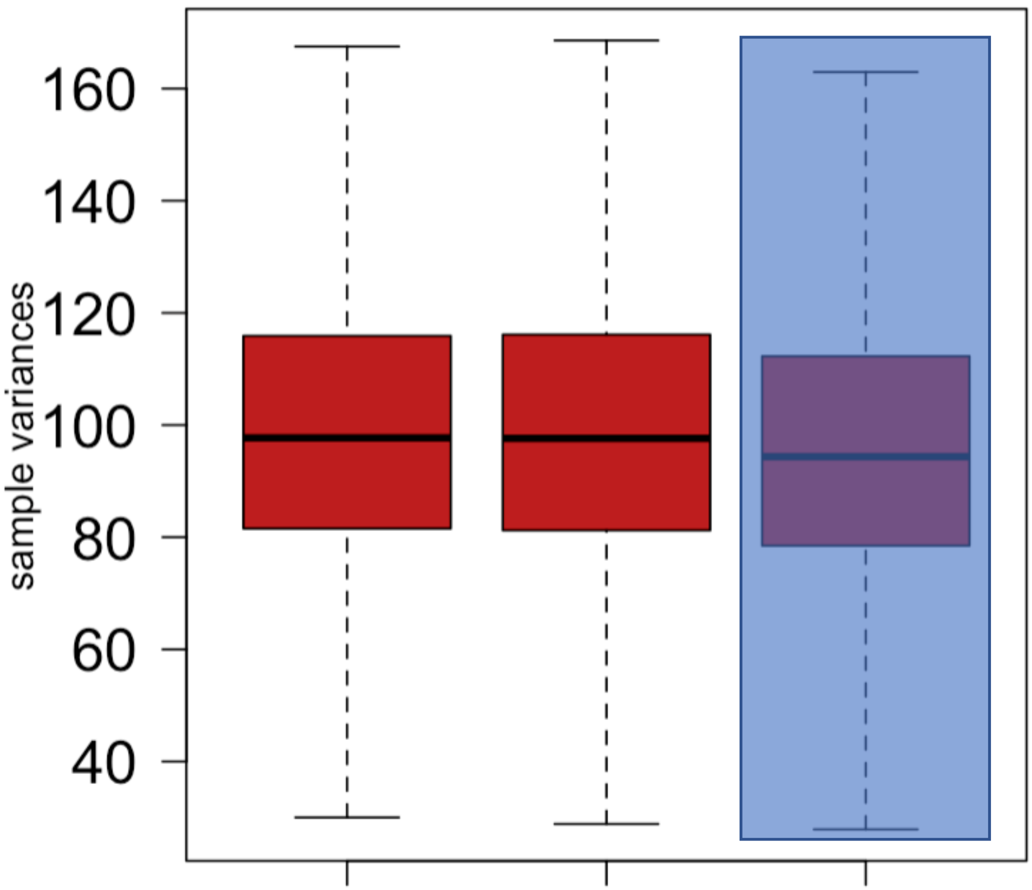

```
> mean(sample.var.based.Pop)
[1] 99.93689
> mean(sample.standard.var)
[1] 99.93232
> mean(sample.var.n.instead)
[1] 96.60124
```

The sample based on the sample mean divided by n-1 is unbiased!

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

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{\mathbf{Y}})^2}{\boldsymbol{n-1}}$$

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{\boldsymbol{Y}})^2}{\boldsymbol{n}}$$




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

Note though that:

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

is slightly more precise then:

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{\mathbf{Y}})^2}{\boldsymbol{n-1}}$$

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

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{\mathbf{Y}})^2}{\boldsymbol{n-1}}$$

Let’s take a small break – 2 minutes

BUT WHY does this bias happen???

But why is variance (or standard deviation) biased when based on n instead of n-1?

$$s = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\boldsymbol{n-1}}}$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\boldsymbol{n}}}$$

But why?

Obviously, you don't need to know the "math" but good to know that someone did it for us!

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:

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

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,

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

So we can write

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

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

Step ★ holds because

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

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/

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 instead of *n*, we need first to understand that values in a sample **are free** to vary around the population mean $\boldsymbol{\mu}$ but values in a sample **are not free** to vary around the sample mean $\bar{Y}$.

$$s = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\boldsymbol{n-1}}} \qquad s = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\boldsymbol{n}}}$$

Free to vary **Not free to vary**

To understand why we use *n*-1 instead of *n*, we need first to understand that values in a sample **are free** to vary around the population mean μ but values in a sample **are not free** to vary around the sample mean $\bar{Y}$.

Let's say we have a series of 6 numbers and we hide one number but we know the sample mean $\bar{Y}$ and want to know the missing number: 1, 5, 7, ???, 9, 12 $\bar{Y} = 7$

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

6×7

??? = 42 - 34 = 8

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

Suppose we know the population mean $\boldsymbol{\mu} = 6$ (but this is just to make the point here as usually you don't know it).

$$\bar{\mathbf{Y}} = \frac{1 + 5 + 7 + 8 + 9 + 12}{6} = 7$$

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

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

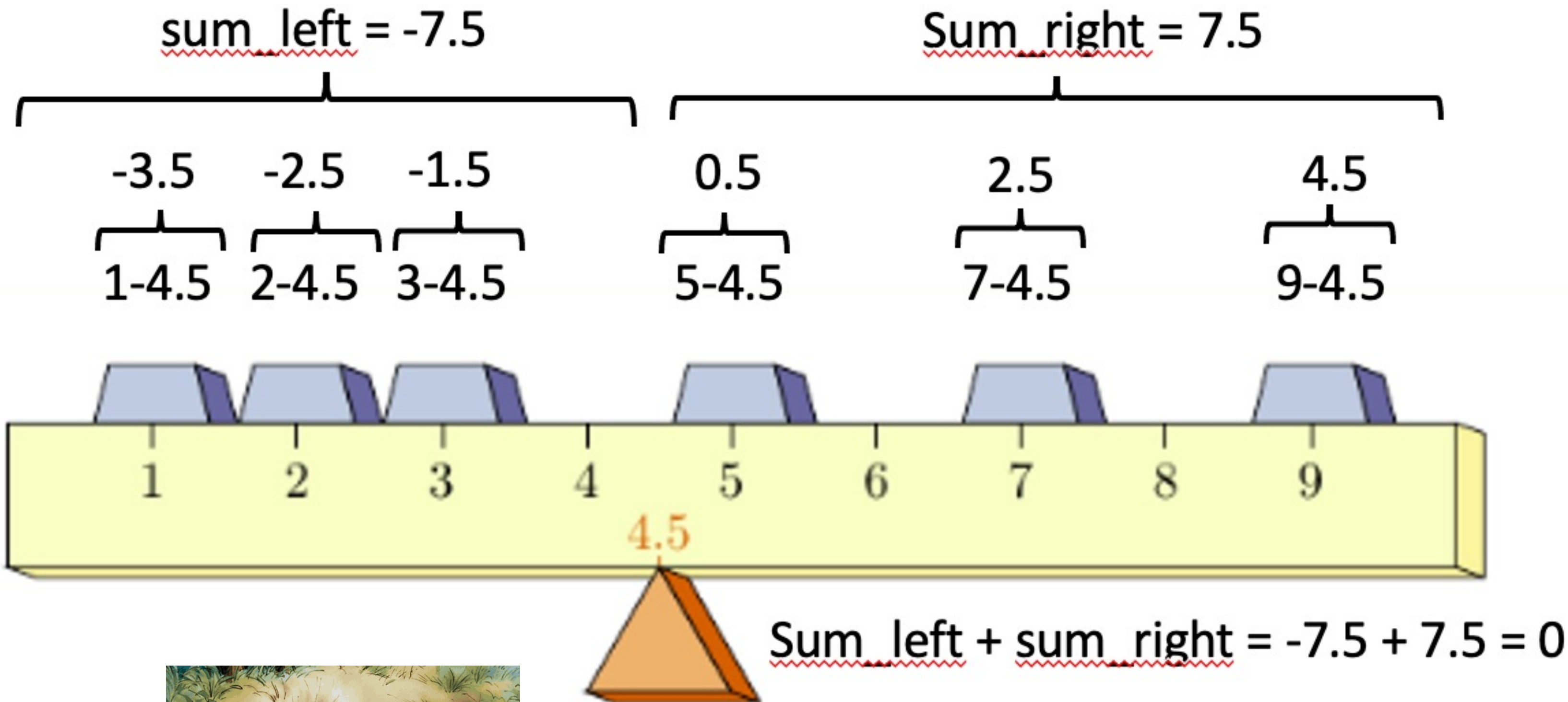
$$= \frac{70}{6} = 11.7$$

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

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

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

Note that the sample-based values was smaller than population-based value



$\bar{X}$=4.5

μ

Remember that the sample values will be always centered around the sample mean; but that's not the case for the population mean, which is free to vary within the sample values.

The sample sum-of-squares is expected to be smaller (i.e., in average) than the population sum-of-squares because the sample mean $\bar{Y}$ is within the range of values of the sample but this is not necessarily the case of μ which can be anywhere within or outside the range of sample values.

$$\underbrace{1, 5, 7, 8, 9, 12}_{} \qquad \bar{Y} = 7$$

The sample mean (7 here) is always within the range of values of the sample, but the population value is free to vary and can be within the sample, smaller or greater than any of these values (i.e., outside of the range of the sample values).

If we use the population mean μ rather than the sample mean $\bar{Y}$ to calculate the sum-of-squares, we will always get a larger sum-of-squares than if we had used the sample mean $\bar{Y}$. So, the sample mean-based sum-of-squares is always smaller than the population mean based sum-of-squares (unless $\bar{Y} = \boldsymbol{\mu}$; which is not very probable)

$$\underbrace{\sum_{i=1}^{n}(Y_i - 7)^2 = 70}_{\text{Based on the original sample mean}} < \underbrace{\sum_{i=1}^{n}(Y_i - 6)^2 = 76}_{\text{Based on the population mean}}$$

From our lecture on variance and standard deviation

Observations (Y_i)	Deviations $(Y_i - \overline{Y})$	Squared deviations $(Y_i - \overline{Y})^2$
0.9	−0.475	0.225625
1.2	−0.175	0.030625
1.2	−0.175	0.030625
1.3	−0.075	0.005625
1.4	0.025	0.000625
1.4	0.025	0.000625
1.6	0.225	0.050625
2.0	0.625	0.390625
Sum	0.000	0.735

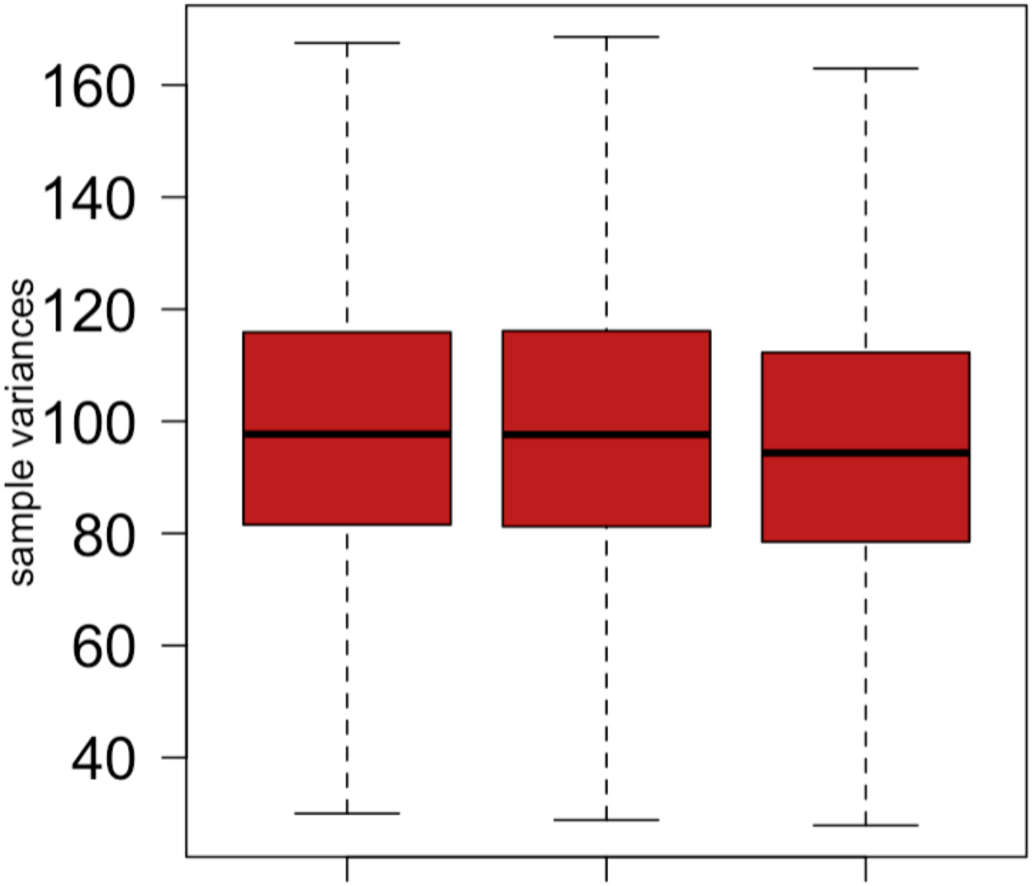
$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} = \frac{0.735}{8-1} = 0.11\ \mathrm{Hz}^2$$

From our lecture on variance and standard deviation

Observations (Y_i)	Deviations $(Y_i - \overline{Y})$		Squared deviations $(Y_i - \overline{Y})^2$
0.9	−0.475		0.225625
1.2	−0.175		0.030625
1.2	−0.175		0.030625
1.3	−0.075		0.005625
1.4	0.025	Because sum of	0.000625
1.4	0.025	deviations is zero,	0.000625
1.6	0.225	this impacts the	0.050625
2.0	0.625	sum of square	0.390625
Sum	0.000	$\longrightarrow$	0.735

$\sum_{i=1}^{n}(Y_i - \bar{Y}) = 0$ (this sum is always zero

But greater or smaller than zero when the population mean is used instead; as such the squared deviations for the sample will be always smaller than the population value)



$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 <= \sum_{i=1}^{n}(Y_i - \mu)^2$$

Bessel's demonstrated that using n-1 in the denominator, the sample standard deviation based on n would be corrected. We say that the sample standard deviation looses 1 degree of freedom.

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

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{\mathbf{Y}})^2}{\boldsymbol{n-1}}$$

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{\boldsymbol{Y}})^2}{\boldsymbol{n}}$$

The average of all infinite *s* based on *n-1* provides an unbiased estimator because the mean of all sample *s* values equals the population value σ.

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$$

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n}$$

Why is the sample standard deviation calculated by dividing the sum of the squared deviations from the mean divided by *n* – 1 and not *n*? **NOW YOU KNOW!**

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$$

How did Bessel find that n – 1 would be the value that would work and not n – 2 or n – 3, for example? That needs some math and it's often the task of statisticians to find if estimates of statistics are biased and how to make them unbiased!

The Statistical Road!!

Sample variance is not biased.
How about the sample standard deviation?

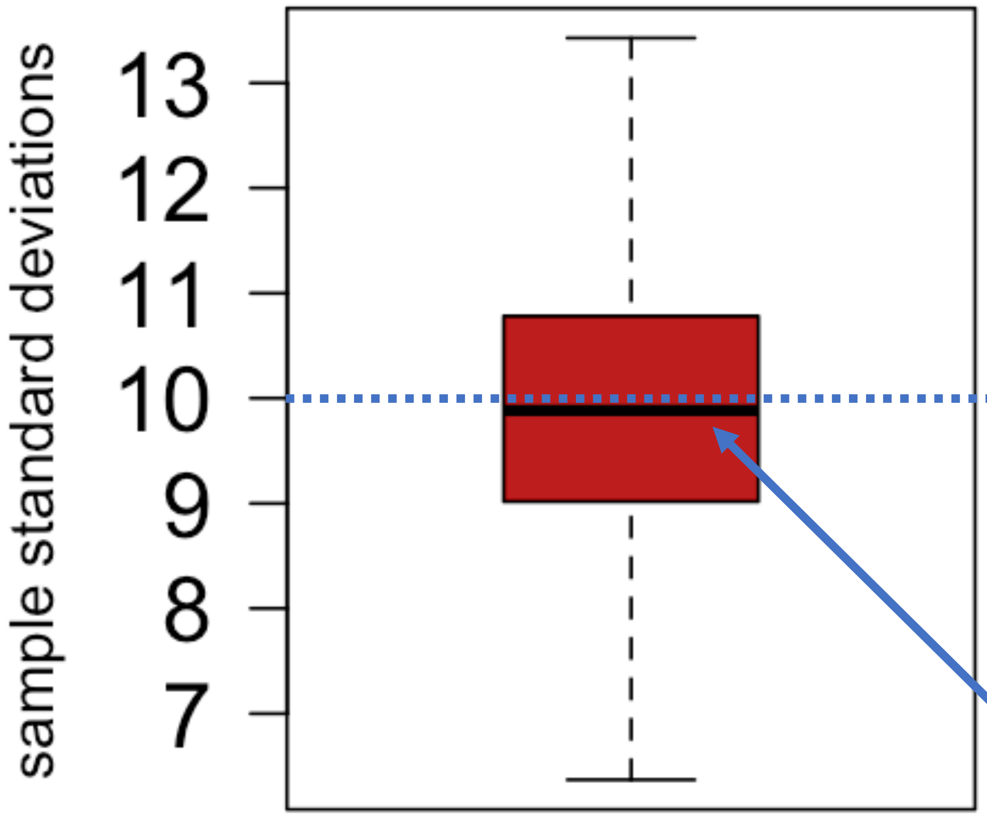
$$s = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}$$

Population standard deviation

$$\sigma = 10$$

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

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

boxplot(sample.sd, outline=FALSE,col="firebrick",
        cex.axis=1.5, las=1,
        ylab="sample standard deviations")
```

Sample variance is not biased.
How about the sample standard deviation? **IT IS A BIT BIASED!**



Population
standard deviation

$\sigma = 10$

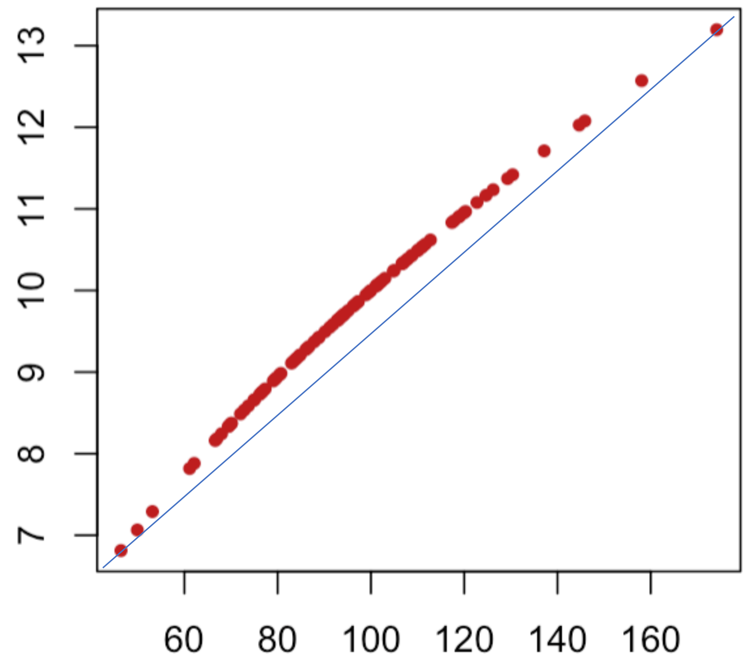
```
mean(sample.sd)
[1] 9.914211
```

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

The sample standard deviation **IS A BIT BIASED!**

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

plot(sample.var[1:100],sample.sd[1:100],pch=16,col="firebrick",
     ylab="",xlab="",cex=0.8)
```

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



$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$$

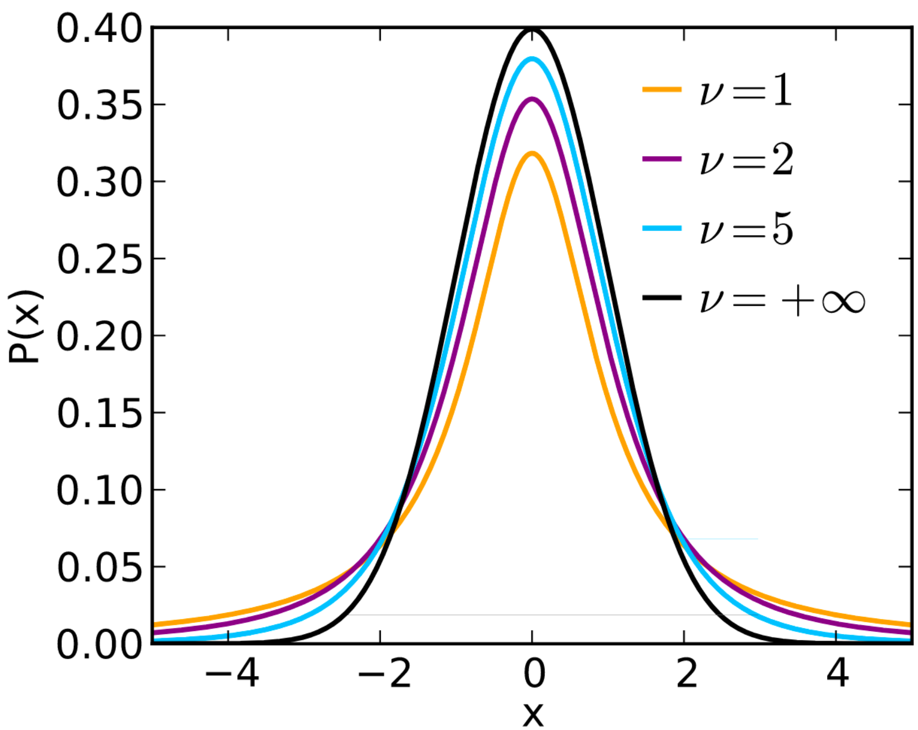
This bias relates to the square root transformation of the variance.

Difficult to establish a general unbiased procedure for the standard deviation as it will change with sample sizes; but there are correction.

Only 100 samples are plotted; 1000000 would have been too many!

The sample standard deviation **IS A BIT BIASED!**

Although there are correction for this bias for normally distributed population, the bias "has little relevance to applications of statistics since its need is avoided by standard procedures”. One example is the t-distribution used to calculate confidence intervals and so many other important statistical analysis (starting next lecture).



$$t = \frac{\bar{X} - \mu}{SE_{\bar{X}}}$$

$$t = \frac{\bar{X} - \mu}{\frac{\sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}}{\sqrt{n}}}$$

Because the t-distribution is based on the sample standard deviation, it incorporates this bias directly in its distribution, so that won’t cause issue with statistical analysis based on the sample standard deviation.

But can we trust the sample standard deviation s?
Is it an unbiased estimator of σ ?

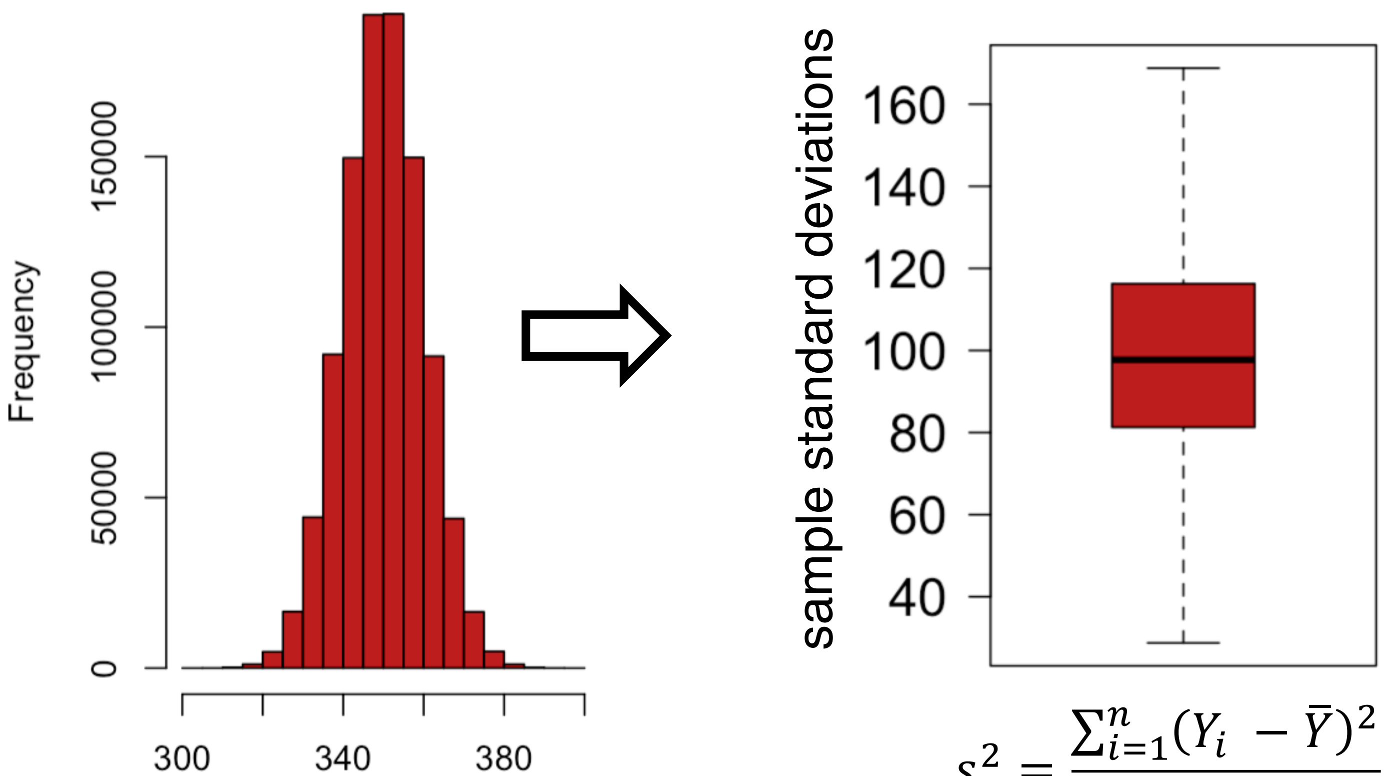
1) The importance of corrections for creating unbiased sample estimators for any statistic of interest [the case of degrees of freedom].

2) **The importance of the distribution of the population for creating unbiased sample estimators for any statistic of interest [the case of assumptions].**

3) The importance of [data transformation] for making biased sample estimators unbiased.

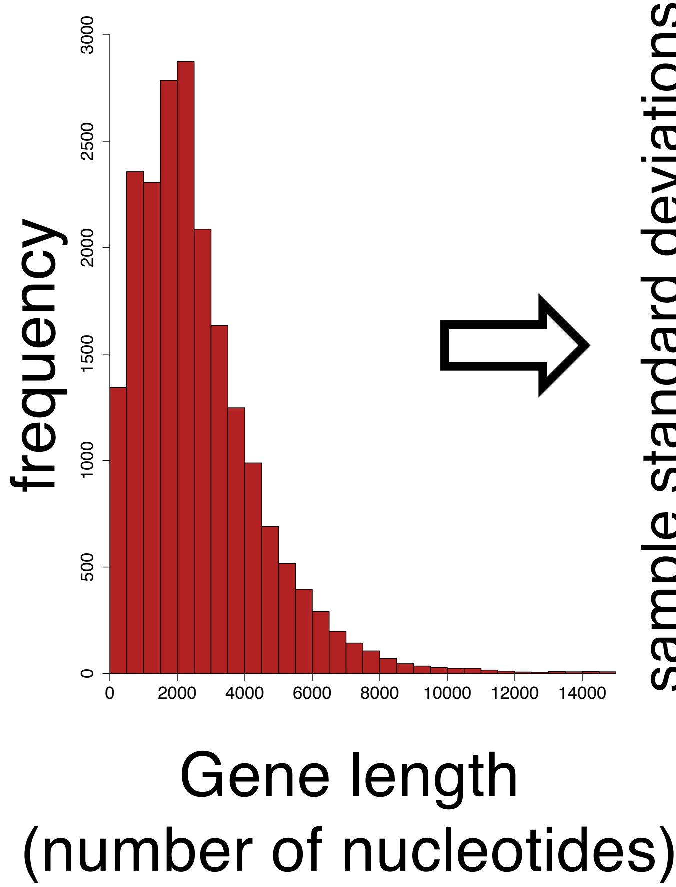
Can we trust the sample estimator for variance when the population is non-normal? So far we assumed normality!

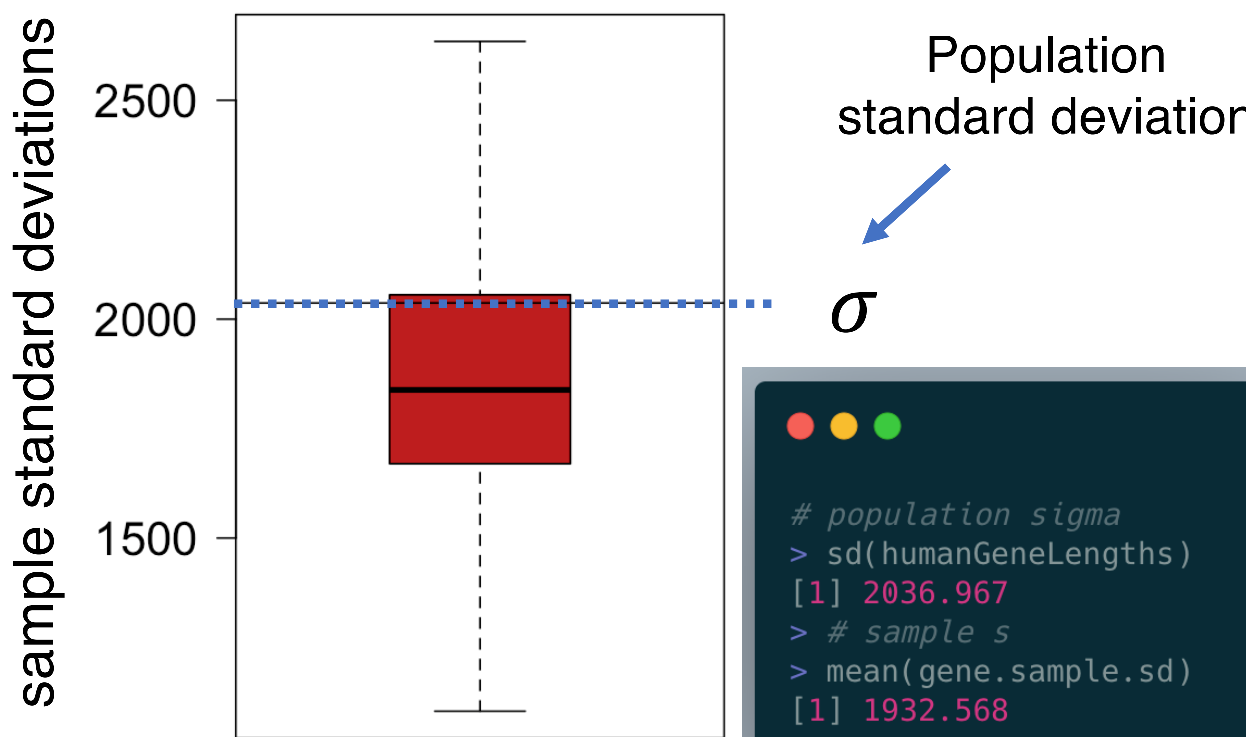
```r
samples <- replicate(1000000, rnorm(n=30,mean=350,sd=10))
```



$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$$

```r
humanGeneLengths <- as.matrix(read.csv("chap04e1HumanGeneLengths.csv"))
geneSample100 <- replicate(100000, sample(humanGeneLengths, size = 100))
gene.sample.var <- apply(geneSample100,MARGIN=2,FUN=var)
boxplot(gene.sample.var, outline=FALSE,col="firebrick",cex.axis=1.5,
        las=1, ylab="sample standard deviations")
```



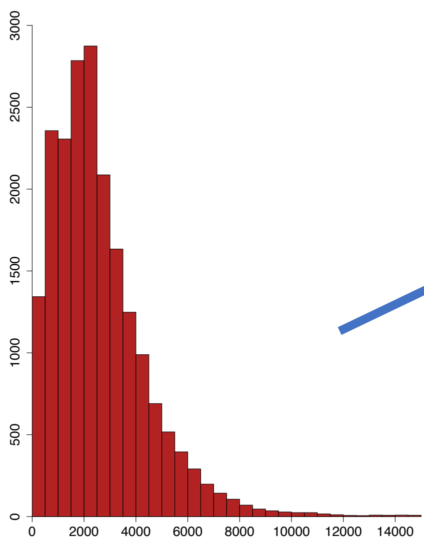


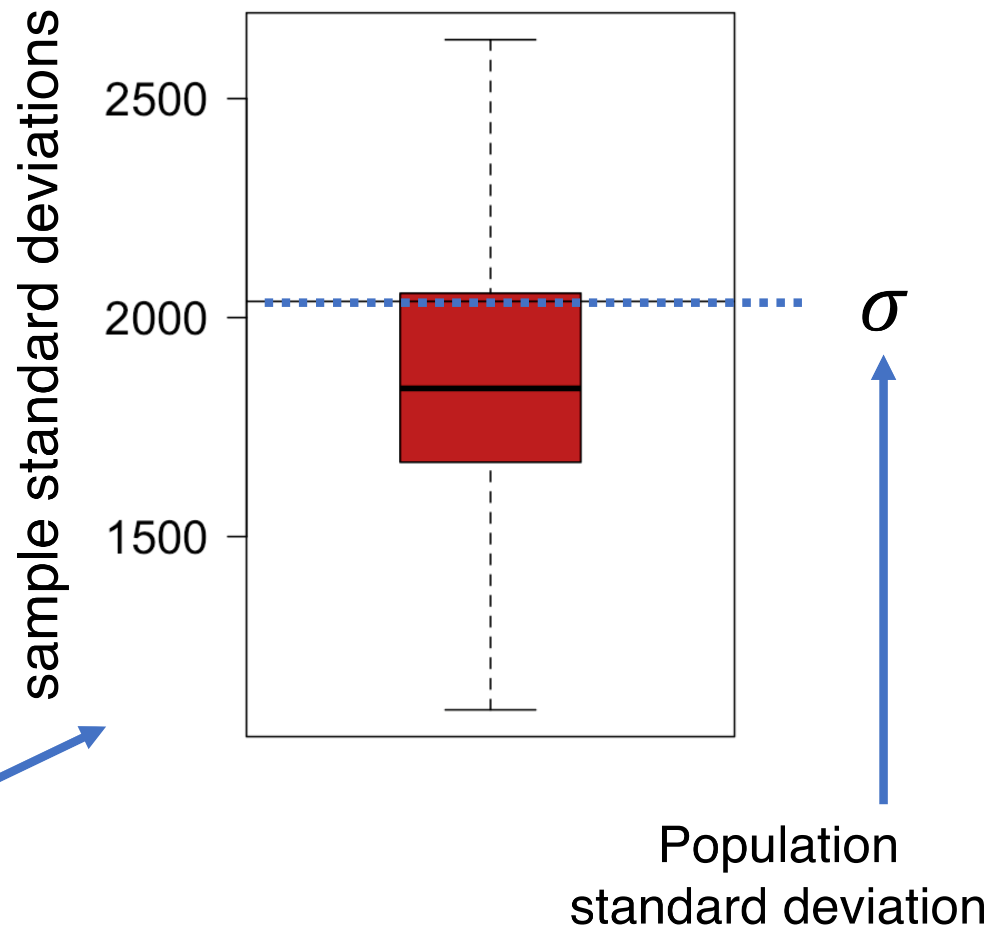

```r
# population sigma
> sd(humanGeneLengths)
[1] 2036.967
> # sample s
> mean(gene.sample.sd)
[1] 1932.568
```

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

Can we trust the sample estimator for variance when the population is non-normal? IN MANY CASES WE CAN'T!

```
# population sigma
> sd(humanGeneLengths)
[1] 2036.967
> # sample s
> mean(gene.sample.sd)
[1] 1932.568
```





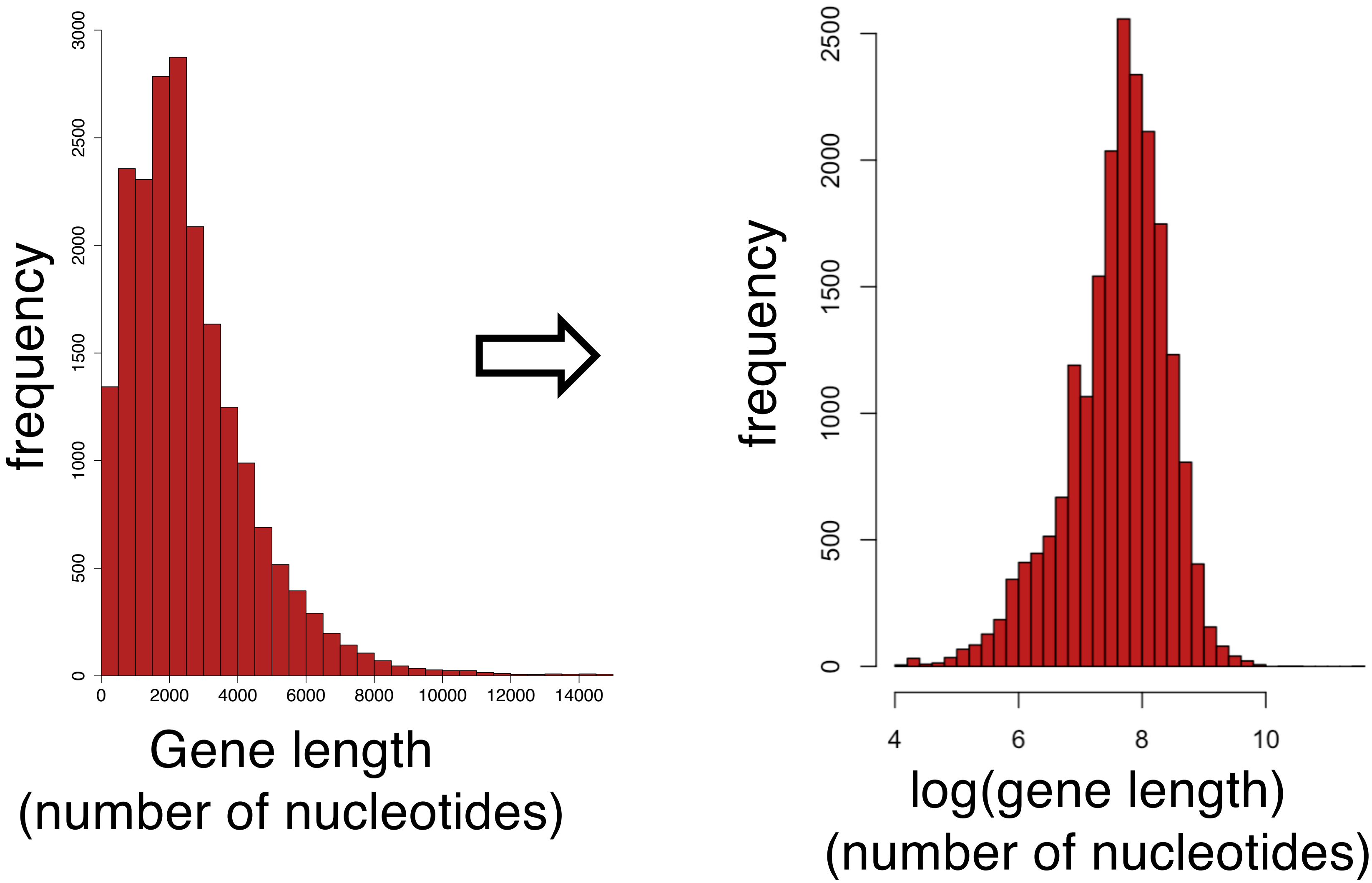
But can we trust the sample standard deviation s?
Is it an unbiased estimator of σ ?

1) The importance of corrections for creating unbiased sample estimators for any statistic of interest [the case of degrees of freedom].

2) The importance of the distribution of the population for creating unbiased sample estimators for any statistic of interest [the case of assumptions].

3) **The importance of [data transformation] for making biased sample estimators unbiased.**

The log transformation make asymmetric distributions more symmetric.



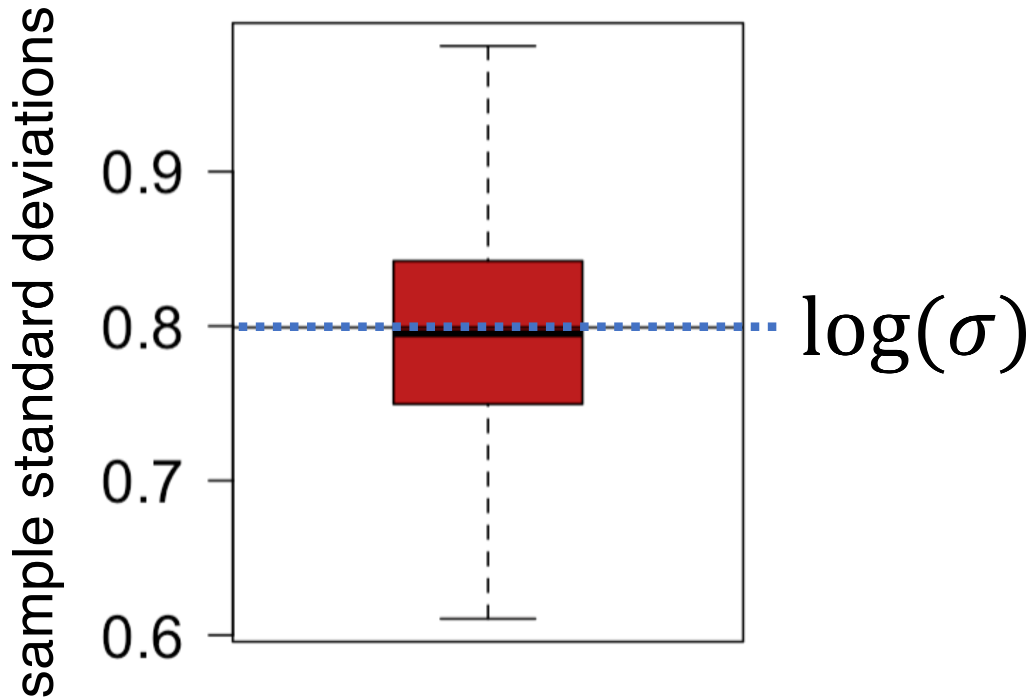
The log transformation make asymmetric distributions more symmetric.

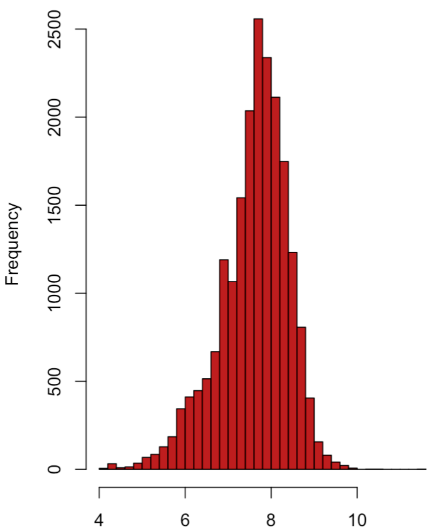
```
humanGeneLengths <- as.matrix(read.csv("chap04e1HumanGeneLengths.csv"))
geneSample100 <- replicate(100000, sample(humanGeneLengths, size = 100))
gene.sample.sd <- apply(log(geneSample100),MARGIN=2,FUN=sd)
boxplot(gene.sample.sd, outline=FALSE,col="firebrick",cex.axis=1.5,
        las=1, ylab="")
```

Samples were log-transformed here

Can we trust the sample estimator for variance when the population is non-normal? IN MANY CASES WE TRUST THEM WHEN SAMPLE DATA ARE TRANSFORMED!

```
> sd(log(humanGeneLengths))
[1] 0.7991254
> # sample s
> mean(gene.sample.sd)
[1] 0.79644
>
```





Develop stronger knowledge and intuition about statistics

1) The importance of corrections for creating unbiased sample estimators for any statistic of interest [the case of degrees of freedom].

2) The importance of the distribution of the population for creating unbiased sample estimators for any statistic of interest [the case of assumptions]. We often assume normality because we know whether estimators are biased or not (i.e., and how to remove their biases using corrections, often called degrees of freedom).

3) The importance of [data transformation] for making biased sample estimators unbiased.

Key goals today

- Develop stronger knowledge and intuition about statistics.

- Understand via the standard deviation case the work that statisticians do so that you can trust the “standard statistics” (i.e., most used) you will use and apply in most of your future professional careers.

- Acquire greater (sophisticated) knowledge about how the other statistical frameworks we will learn in BIOL322 were developed. We won’t revisit all sample estimators, but the type of work that was done for the standard deviation can be “generalized” to most sample statistics.

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

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

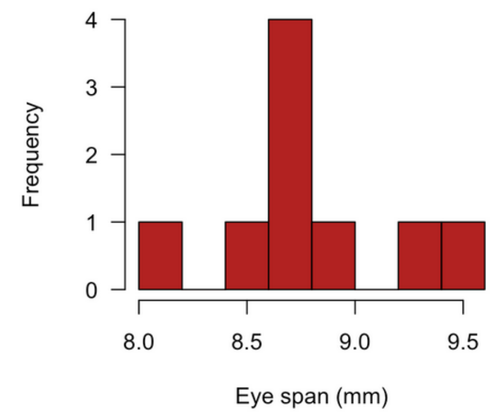
$SE_{\bar{Y}} \frac{0.398}{\sqrt{9}} = 0.133$ mm

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

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

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

"symmetric" (we can "trust" estimates)



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

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

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

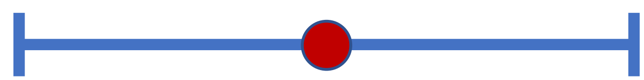
$t_{0.05(2),8} = 2.306$

Degrees of freedom
(v, df)

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

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

8.47 mm 9.08 mm



$\bar{X} = 8.878 \; mm$

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

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

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

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

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

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

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

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


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

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

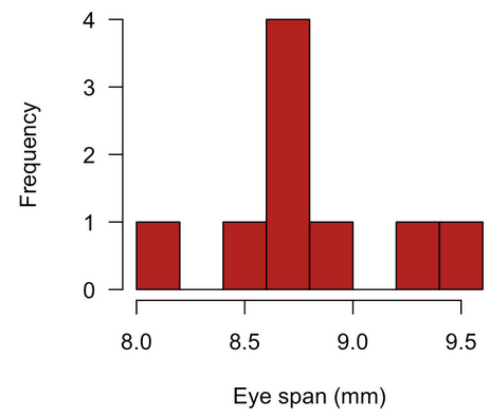
$$\bar{Y} = 8.778 \; s = 0.398$$

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

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

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

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



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

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

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

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

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


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

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:

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