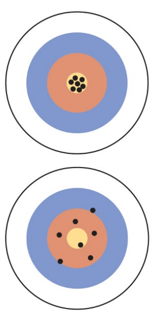
Lecture 11 - Estimation is more complex than it seems: the challenges and solutions from 100 years ago have become integral to mainstream statistics!

Building long-term statistical intuition & knowledge

The statistical road: embrace uncertainty while estimating with confidence

1

For statistics to be reliable, we must trust our sample estimators, meaning they need to be unbiased.



The bullseye is the population mean μ and each dot is a sample mean X.

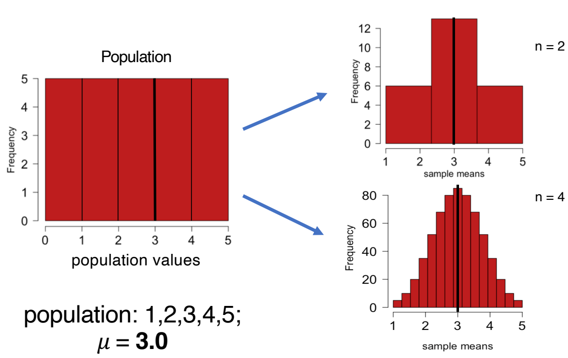
We know that under random sampling, the sample mean is an unbiased estimator of the population mean μ.

This is because the mean of all possible sample means equals the population mean.

In other words, across many repeated samples, the sample mean accurately reflects the true population mean on average, without systematic error.

2

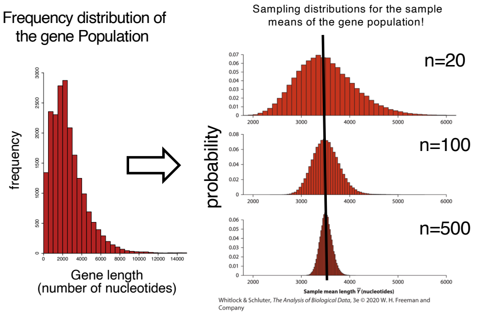
The shape of the population's frequency distribution does not necessarily resemble the frequency distribution of sample estimates (such as the distribution of sample means).
Regardless of the population's distribution shape (e.g., even if it's uniform), the sample mean remains an unbiased estimator of the population mean when sampling is random.



population: 1,2,3,4,5;
μ = **3.0**

3

The shape of the population's frequency distribution does not necessarily resemble the frequency distribution of sample estimates (such as the distribution of sample means).
Regardless of the population's distribution shape (e.g., even if it's uniform), the sample mean remains an unbiased estimator of the population mean when sampling is random.

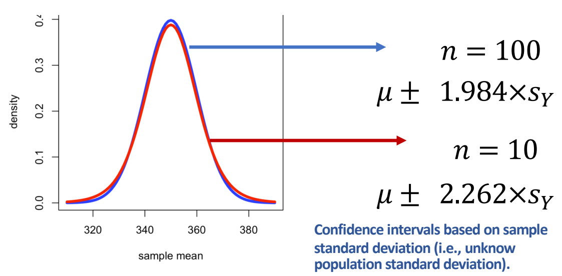


Sample mean length Y (nucleotides)

4

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

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



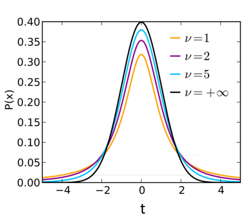
Confidence intervals based on sample standard deviation (i.e., unknow population standard deviation).

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

5

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) is called t when based on the sample standard error (i.e., estimate of the true standard error of the sampling distribution).

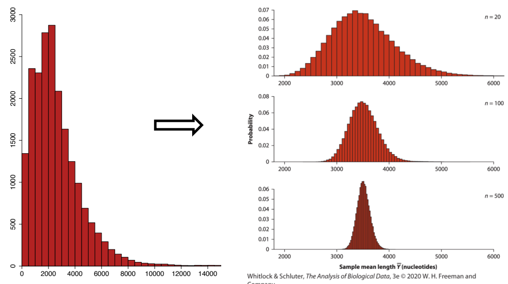


This *t* distribution (standardized) is a sampling distribution of the the number of sample standard errors away from the mean (now always 0 after the standardization) necessary to produce a confidence interval of the desired coverage (e.g., 95%).

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

6

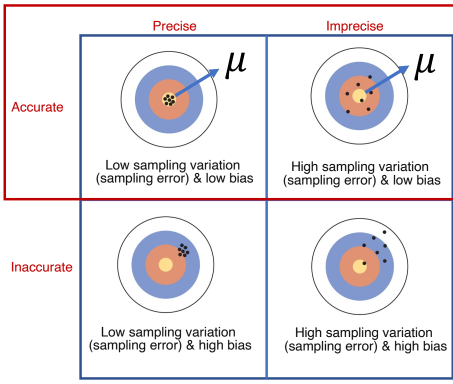
Even though the distribution of the population is asymmetric, the sampling distribution of means tend to me symmetric. This is an important property because it allows us to generalize sampling distributions based on standard distributions such as the t-distribution (not always but often).



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

7

We must rely on our sample estimators for statistical methods to be valid, meaning they need to be unbiased.



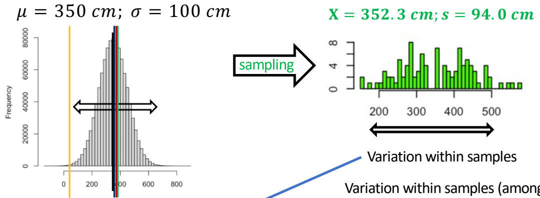
The sample mean is an unbiased estimator under random sampling because the average of all sample means equals the population mean.

8

The variation within a sample (standard deviation) can be used to estimate how far the sample means might be from the true population mean, giving us an idea of the potential error in our estimate.

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

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



$\mu = 350\ cm$

Variation among samples

Variation within samples (among observations) can be used to estimate the uncertainty in the sample means.

Estimating variation within a sample to assess variation among samples (standard error, i.e., uncertainty around sample means) is fundamental to statistics, not just to constructing confidence intervals.

9

The ability to estimate variation within a sample to assess variation among samples (standard error) is crucial to statistics, not just for confidence intervals

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

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

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

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

Since we almost never know the population standard deviation, we estimate it using the sample standard deviation:

$$\mathrm{SE}_Y = \frac{s}{\sqrt{n}}$$

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

10

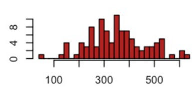
As we will see, the mean and standard deviation are fundamental sample statistics used in nearly all standard statistical analyses, not just for confidence intervals.

$$\mathrm{Y} \pm \mathrm{t}_n \times \mathrm{SE}_Y \therefore \mathrm{SE}_Y = \frac{s}{\sqrt{n}}$$

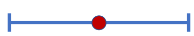
Margin of error

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

$$351.5c \pm \mathrm{t}_n \times \frac{114.2}{\sqrt{100}}$$



328.66 cm   374.34 cm



$\mathbf{X = 351.5\ cm}$

11

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

Today, we will explore the sample standard deviation as an estimator of the true population standard deviation.

Our goals are threefold:

Build a deeper understanding and intuition about statistical concepts.

Learn how statisticians develop reliable statistical measures.

Gain insight into how the other statistical methods we will learn in BIOL322 were created.

Note: While we won't revisit every sample estimator, the process used for standard deviation can be generalized to most sample statistics.

12

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

1) The significance of applying corrections to create unbiased sample estimators for any statistic of interest [the case of degrees of freedom].

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

3) The importance of [data transformation] in converting biased sample estimators into unbiased ones.

13

1) The significance of applying corrections to create 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 squared deviations from the mean b $n - 1$ 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?

14

Let's switch to variance s^2 (hang in there with me); after all $s = \sqrt{s^2}$.
If we knew (but we don't really) the true population mean μ, the best sample-based estimator for the population variance using a single sample would be:

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

Since we almost never know the population mean μ, let's see what happens when we use the sample mean value Yas an estimate of μ:

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

15

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

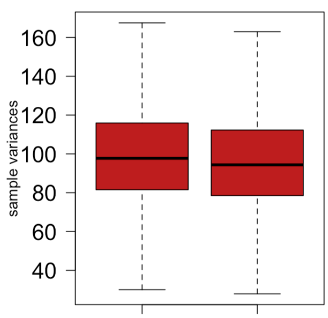
σ^2=100; σ=10

```
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")
```

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

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

16




```
> 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., it closely 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 Y divided by n is biased.

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

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

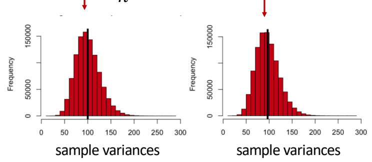
17

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

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

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

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



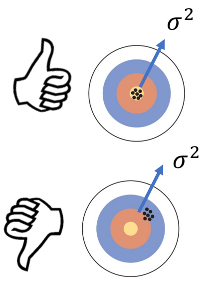
Note the asymmetry in the sampling distribution of variances, i.e., the median doesn't equal the mean. The variance is unbiased when based on $\boldsymbol{\mu}$ but biased when based on $\boldsymbol{Y}$. Remember: unbiased expectations are based on means and not medians.

18

In most cases, the parameter value μ (the true population mean) is unknown.

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

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



19

There is a correction factor for the sample bias in s^2 called Bessel's correction (although it appears that Gauss first introduced it in 1823).

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

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

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

20

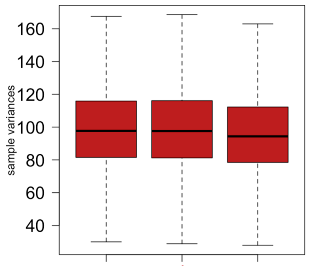
Let's use a computational approach to evaluate the accuracy of these two sample-based estimators.:

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)

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")
```

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

21




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

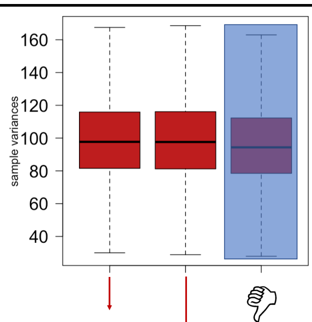
The sample based on the sample mean divided by n-1 is unbiased!

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

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

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

22




```
> 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:

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

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

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

23

Let's take a small break – 1 minute

24

BUT WHY does this bias occur???

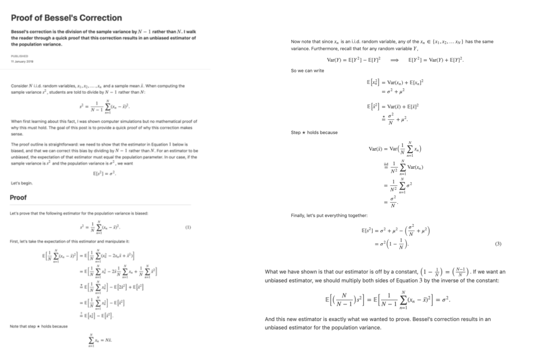
But why is the variance (or standard deviation) biased when divided by n instead of n-1?

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

But why?

25

Obviously, you don't need to know the math, but it's reassuring to know that someone worked it out 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:

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

Note that step $\star$ holds because

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

while step $\dagger$ holds because the data are i.i.d., i.e.

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

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,

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

So we can write

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

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

Step $\star$ holds because

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

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/

26

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 μ but values in a sample **are not free** to vary around the sample mean 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**

27

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

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

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

6×7

??? = 42 - 34 = 8

So, there is always one number that is not free to vary around the sample mean Y

28

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 **Y**:

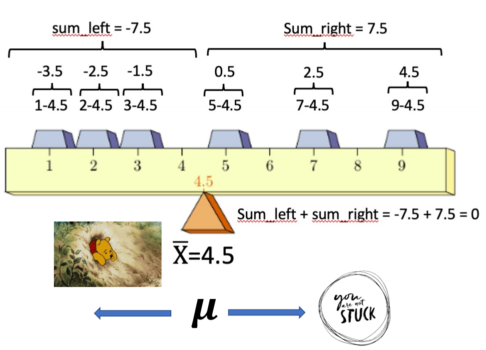
$$s^2 = \frac{(1-7)^2+(5-7)^2+(7-7)^2+(8-7)^2+(9-7)^2+(12-7)^2}{n} = \frac{70}{6} = 11.7$$

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 were smaller than the population-based values. This occurs because the sample mean tends to underestimate variability compared to the true population mean. This is why corrections, like dividing by n-1, are necessary to provide an unbiased estimate of the population parameters.

29



Remember that the sample values will always be centered around the sample mean, but this is not true for the population mean, which can vary freely within the range of the sample values.

30

The sample sum of squares is typically smaller, on average, than the population sum of squares because the sample mean (Y) lies within the range of the sample values, whereas the population mean (μ) can be located anywhere, either within or outside the sample range.

1, 5, 7, 8, 9, 12 $Y = 7$

The sample mean (7 in this case) always falls within the range of the sample values, but the population mean is free to vary—it can lie within the sample values or be smaller or larger than any of them (i.e., outside the range of the sample values).

If we use the population mean (μ) instead of the sample mean (Y) to calculate the sum of squares, the result will almost always be larger than if we had used the sample mean. This is because the sample mean minimizes the sum of squared deviations within the sample. Therefore, the sum of squares based on the sample mean will always be smaller than that based on the population mean, unless the two means happen to be equal (which is unlikely).

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

31

From our lecture on variance and standard deviation

Observations (Y_i)	Deviations ($Y_i - \bar{Y}$)	Squared deviations $(Y_i - \bar{Y})^2$
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

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

32

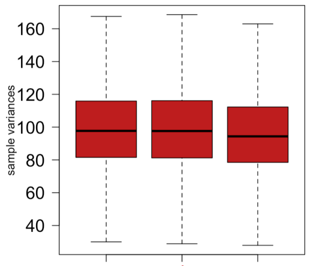
From our lecture on variance and standard deviation

Observations (Y_i)	Deviations ($Y_i - \bar{Y}$)	Squared deviations $(Y_i - \bar{Y})^2$
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

Because sum of deviations is zero, this impacts the sum of square

$\sum_{i=1}^{n}(Y_i - Y) = 0$ (this sum is always zero when using the sample mean. However, when the population mean is used instead, the sum can be either greater or smaller than zero. Consequently, the squared deviations from the sample will be always smaller than those from the population mean).

33



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

Bessel demonstrated that by using n-1 in the denominator, the sample standard deviation based on n observations is corrected. This adjustment accounts for the fact that the sample loses 1 degree of freedom when estimating the population standard deviation.

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

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

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

34

The average of all possible sample standard deviations calculated with n-1 in the denominator provides an unbiased estimator, as the mean of all sample standard deviation values equals the population standard deviation (σ).

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

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

35

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? This requires some mathematical work, and it's often the role of statisticians to determine whether estimates of statistics are biased and how to adjust them to make them unbiased.

36

The Statistical Road!!

37

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

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

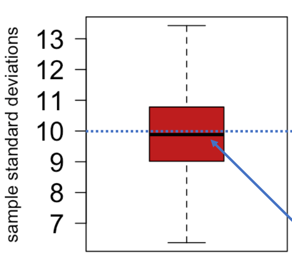
Population standard deviation

$\sigma = 10$

```
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")
```

38

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
```

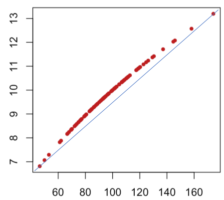
$$s = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}$$

39

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

```
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)
```

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



This bias arises from the square root transformation of the variance.

It's challenging to establish a general unbiased procedure for the standard deviation, as it varies with sample size, but there are correction methods available.

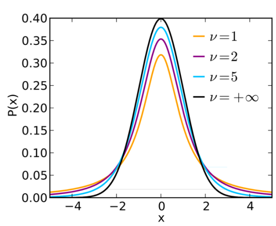
Only 100 samples are plotted; 1000000 would have been too many!

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

40

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

Although corrections for this bias exist for normally distributed populations, the bias itself 'has little relevance to applications of statistics,' as it is generally avoided through standard procedures. For instance, the t-distribution, which is used to calculate confidence intervals and perform many other important statistical analyses (to be covered in the next lecture), effectively addresses this issue.



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

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

Since the t-distribution is based on the sample standard deviation, it inherently accounts for this bias in its distribution, ensuring that it does not pose any issues for statistical analyses that rely on the sample standard deviation.

41

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

1) The significance of applying corrections to create unbiased sample estimators for any statistic of interest [the case of degrees of freedom].

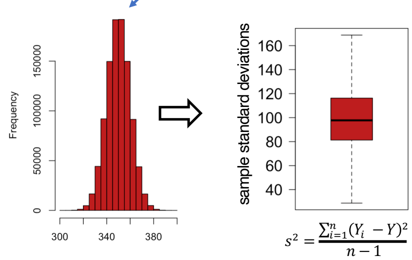
2) The role of population distribution in creating unbiased sample estimators for any statistic of interest [the case of assumptions].

3) The importance of [data transformation] in converting biased sample estimators into unbiased ones.

42

Can we rely on the sample estimator for variance when the population is non-normal? Up until now, we've been assuming normality!

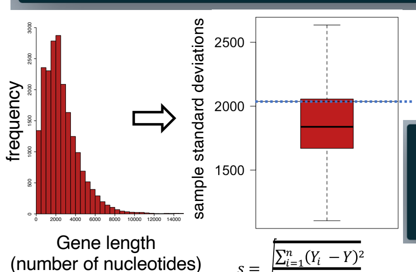
```
samples <- replicate(1000000, rnorm(n=30,mean=350,sd=10))
```



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

43

```
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")
```



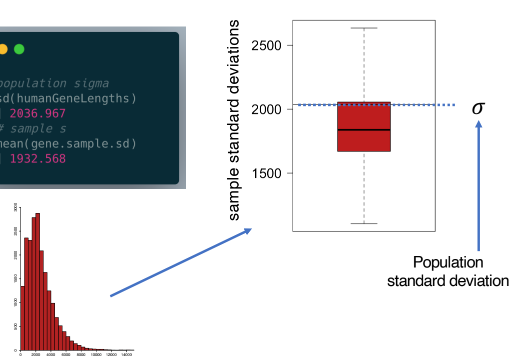

```
# population sigma
> sd(humanGeneLengths)
[1] 2036.967
> # sample s
> mean(gene.sample.sd)
[1] 1932.568
```

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

44

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
```



45

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

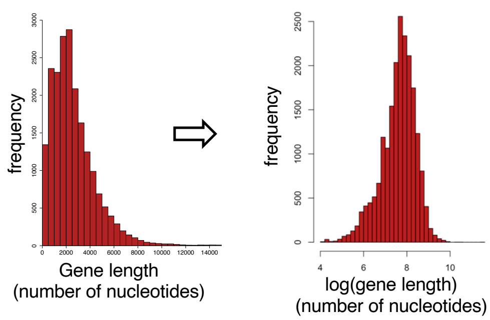
1) The significance of applying corrections to create unbiased sample estimators for any statistic of interest [the case of degrees of freedom].

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

3) The importance of [data transformation] in converting biased sample estimators into unbiased ones.

46

Log transformation helps to reduce skewness, making asymmetric distributions more symmetric.



47

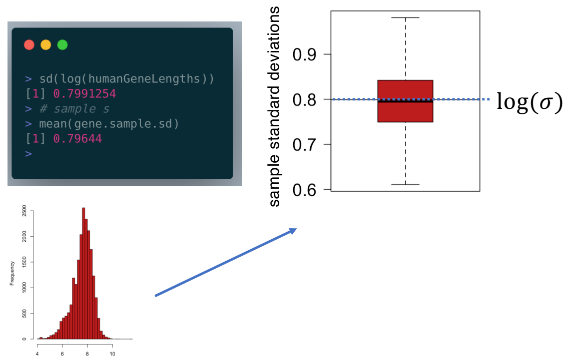
Log transformation helps to reduce skewness, making asymmetric distributions more symmetric.

```
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

48

Can we trust the sample estimator for variance when the population is non-normal? In many cases, we can trust them when the sample data have been transformed!



49

Develop stronger knowledge and intuition about statistics

1) The significance of applying corrections to create unbiased sample estimators for any statistic of interest [the case of degrees of freedom].

2) The role of population distribution in 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] in converting biased sample estimators into unbiased ones.

50

Key goals today

- Develop a stronger understanding and intuition about statistics.
- By exploring the case of the standard deviation, gain insight into the work statisticians do, allowing you to trust the 'standard statistics' (i.e., the most commonly used methods) that you will apply in your future professional careers.
- Acquire deeper knowledge about how the other statistical frameworks we will cover in BIOL322 were developed. While we won't revisit every sample estimator, the principles applied to the standard deviation can be generalized to most sample statistics.

51

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:

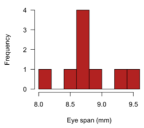
8.69 8.15 9.25 9.45 8.96 8.65 8.43 8.79 8.63

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

$Y = 8.778$ mm $s = 0.398$ mm

$SE_Y \frac{0.398}{\sqrt{9}} = 0.133$ mm

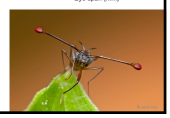
"symmetric" (we can "trust" estimates)



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

$Y - 2.306 \times 0.133 < \mu < Y + 2.306 \times 0.133$

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



52

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

$Y = 8.778$ $s = 0.398$

$SE_Y \frac{0.398}{\sqrt{9}} = 0.133$

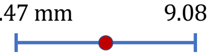
$t_{0.05(2),8} = 2.306$

Degrees of freedom (v, df)

$Y - 2.306 \times 0.133 < \mu < Y + 2.306 \times 0.133$

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

8.47 mm 9.08 mm



$X = 8.878\ mm$

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
```

53

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

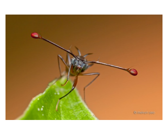
$Y = 8.778$ $s = 0.398$

$SE_Y \frac{0.398}{\sqrt{9}} = 0.133$

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

$Y - 2.306 \times 0.133 < \mu < Y + 2.31 \times 0.133$

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



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

54

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

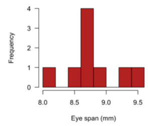
$Y = 8.778 \; s = 0.398$

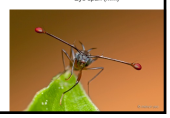
$\mathrm{SE}_Y \frac{0.398}{\sqrt{9}} = 0.133$

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

$Y - 3.355 \times 0.133 < \mu < Y + 3.355 \times 0.133$

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





55

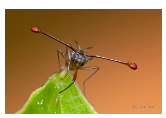
$Y = 8.778 \; s = 0.398$

$\mathrm{SE}_Y \frac{0.398}{\sqrt{9}} = 0.133$

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

$Y - 3.355 \times 0.133 < \mu < Y + 3.355 \times 0.133$

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



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

56

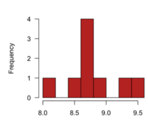
In most cases, however, we report the 95% confidence interval.

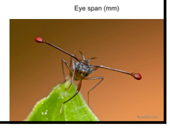
95% confidence interval:

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

99% confidence interval:

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





57