Mean and variance are fundamental: they describe central tendency and variability and form the basis of most inferential tools in Biology (e.g., t-tests, ANOVA, regression, confidence intervals).

mean

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n-1}$$

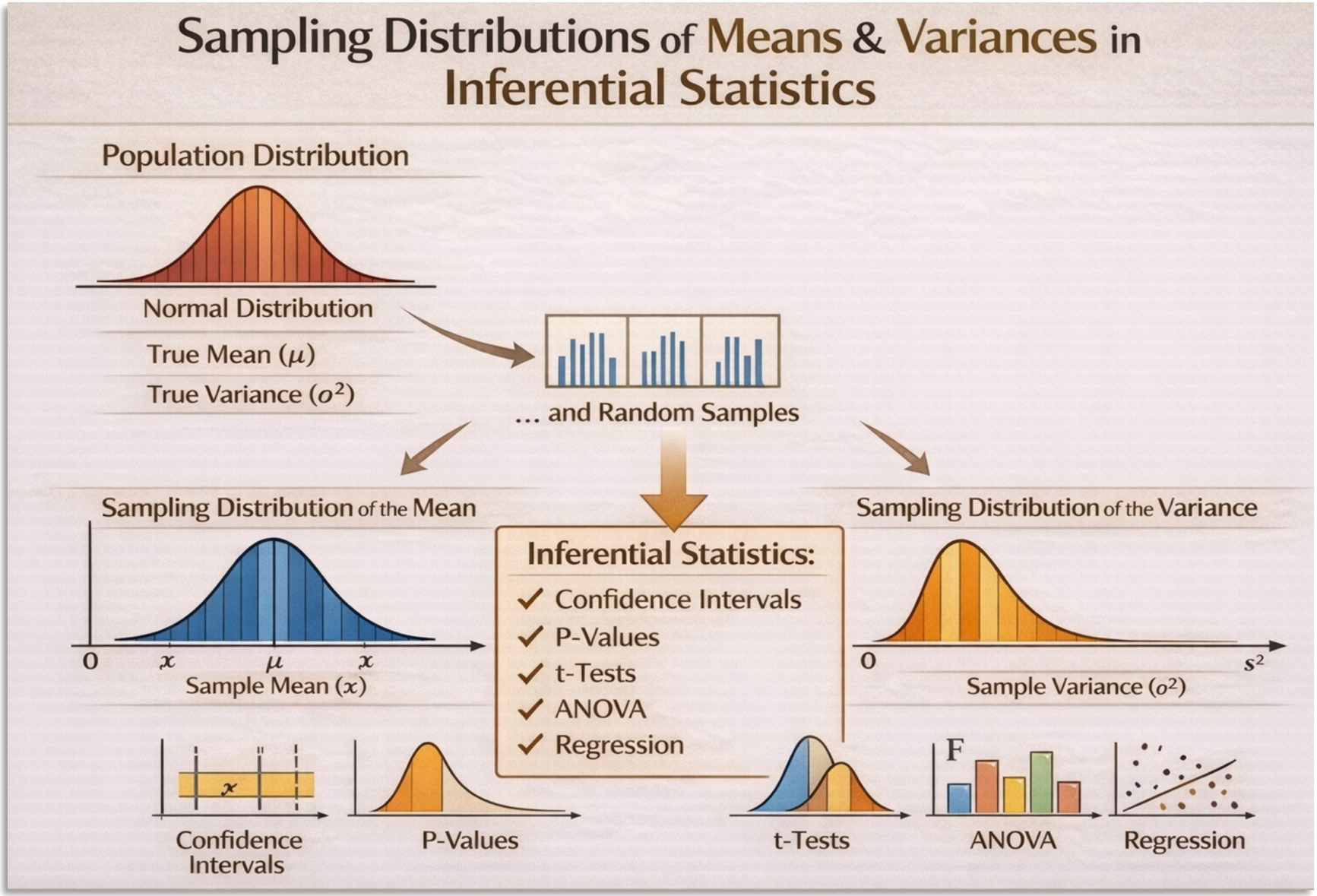
variance

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

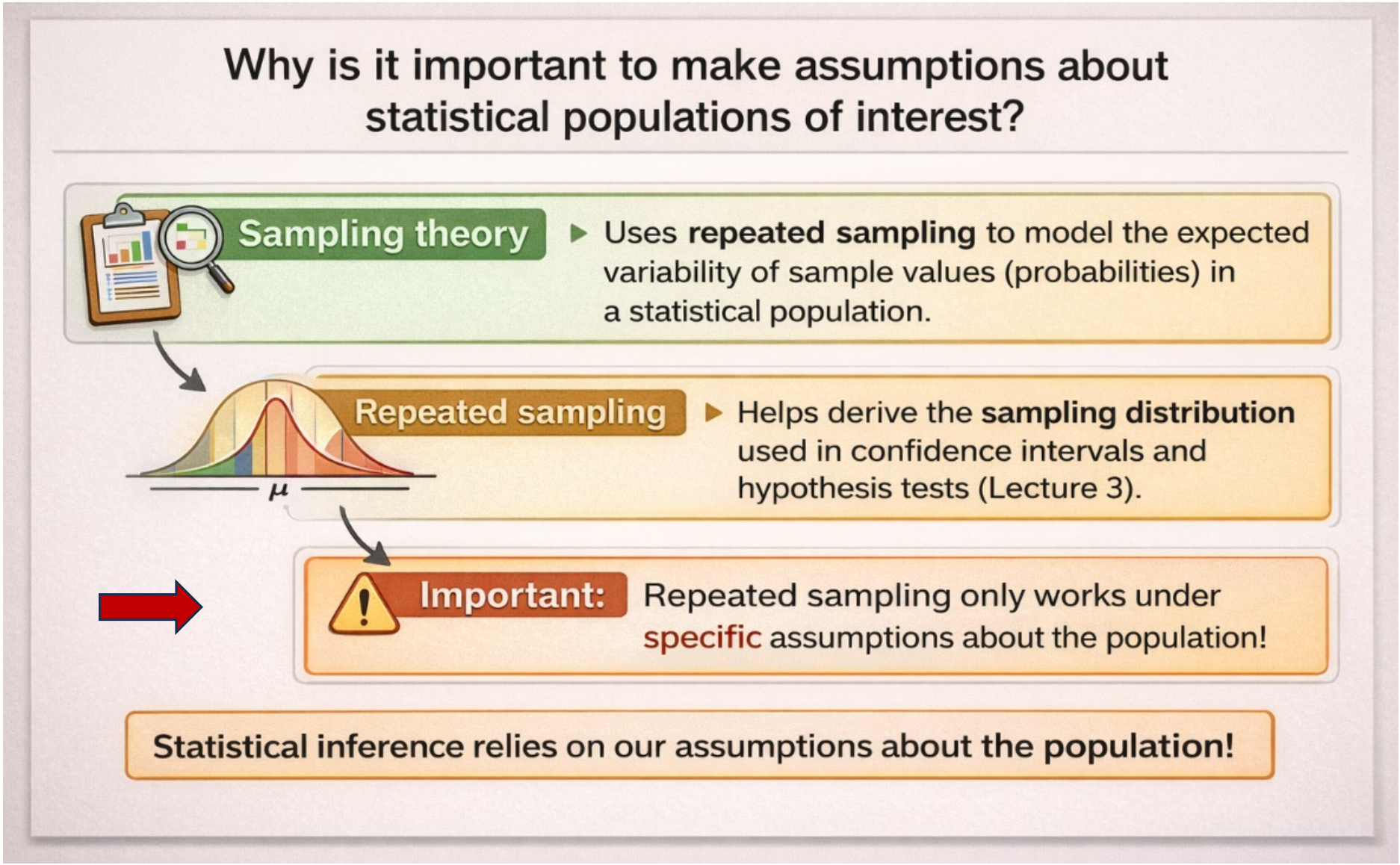
Standard deviation

$$s = \sqrt{s^2}$$

To make inferences about a population while accounting for sampling uncertainty, we use the sampling distribution rather; not the original population.



The role of normality in biology: We often work with continuous variables that are assumed to be “normally” distributed (or other types of distribution) to derive the sampling distribution of the statistic of interest (mean, variance, etc).



To make inferences about a population while accounting for sampling uncertainty, we use the sampling distribution rather; not the original population.

Property of the Mean as an Estimator

The **mean of all possible sample means** (the sampling distribution) **always equals the population mean** — regardless of the **population's shape.**

Population

Values: 1, 2, 3, 4, 5

Population mean: $\mu = 3.0$

Key result

$$\frac{1.0 + 2.0 + 3.0 + 4.0 + 1.5 + 2.0 + 2.5 + 3.0 + 2.5 + 3.0 + 3.5}{15}$$

All possible samples of size $n = 2$ (with replacement)

(Order does not matter: (1, 2) = (2,1))

Sample	Mean	Sample	Mean	Sample	Mean
(1,1) 1.0	1.5	(2,3) 2.5	2.5	(2,3) 2.5	4.0
(2,2) 2.0	2.0	(1,3) 2.0	2.5	(2,4) 3.0	3.5
(3,3) 3.0	2.5	(1,4) 2.5	3.5	(3,4) 3.5	4.0
(4,4) 4.0	3.0	(1,5) 3.0		(3,5) 4.0	4.5

Total number of samples: 15

$$1.0 + 2.0 + 3.0 + 4.0 + 5.0 + 1.5 + 2.0 + 2.5 + 3.0 + 2.5 + 3.0 + 3.5$$

Why this works ➡ The overestimates and underestimates balance perfectly.

- 6 sample means < μ (red)
- 3 sample means = μ (black)
- 6 sample means > μ (green)

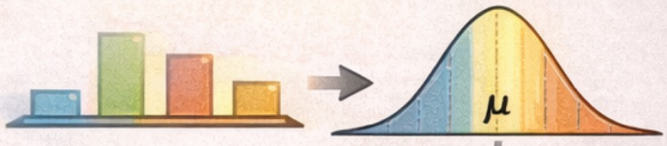
The sample mean is an unbiased estimator of the population mean.

To make inferences about a population while accounting for sampling uncertainty, we use the sampling distribution rather; not the original population.

Why sampling properties of estimators are important?

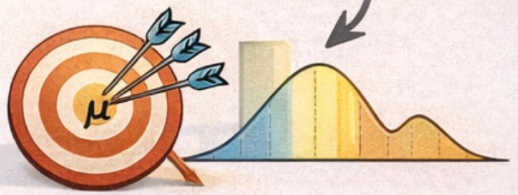
The mean of all possible sample means always equals the population mean, regardless of the population's original distribution.

(normal, uniform, asymmetric, etc.)



The sample mean is an unbiased ("honest") estimator of the true population mean.

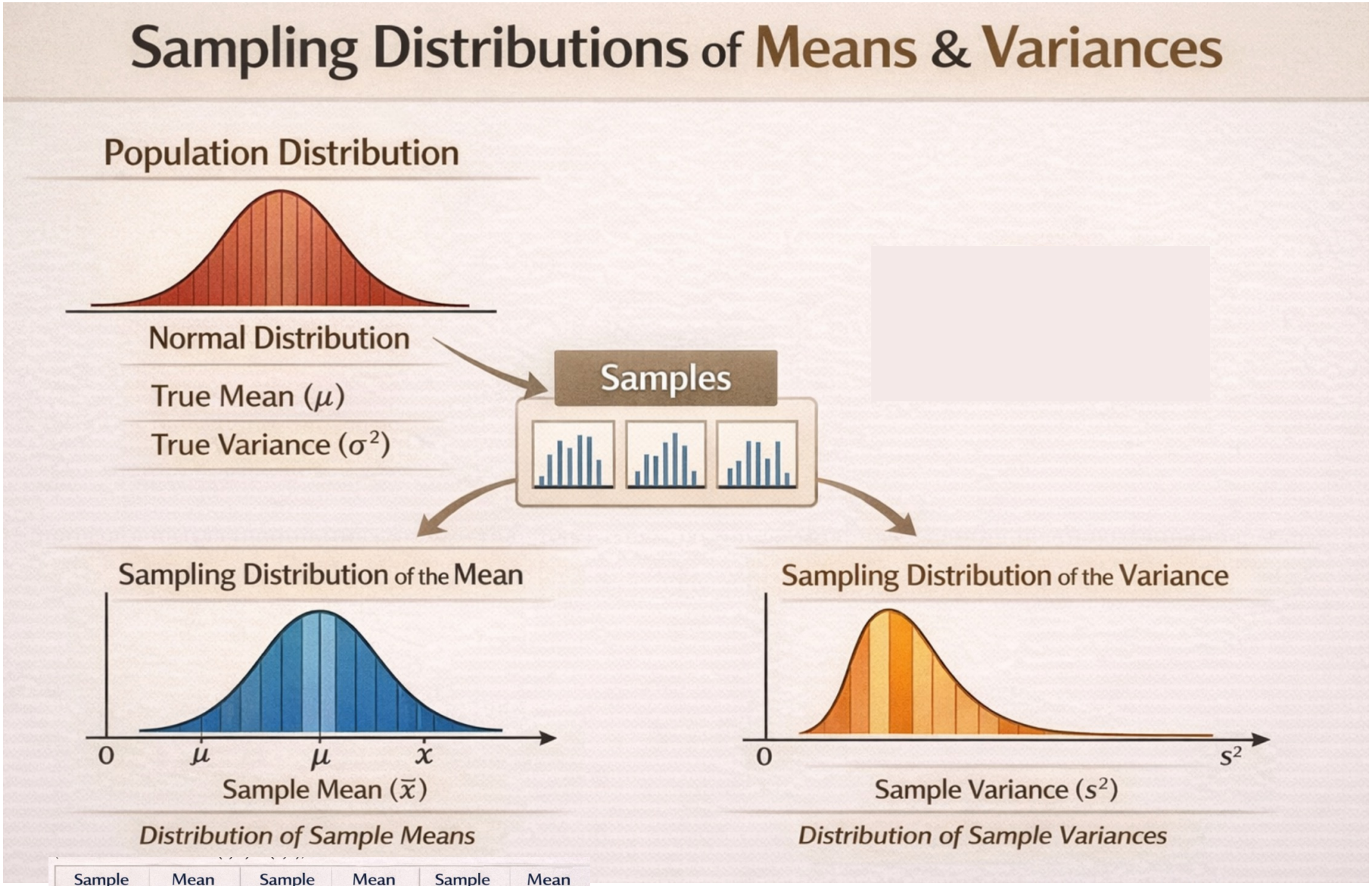
"Honest" means equals the true value on average.



Critical for inference: In the long run, conclusions based on sample means are centered on the truth, avoiding systematic bias.

An unbiased estimator provides accurate estimates on average!

The ***mean*** and the ***variance*** are fundamental statistics



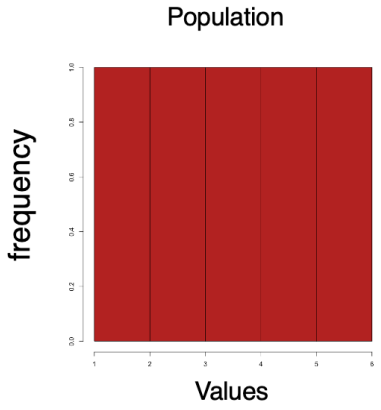
Sample	Mean	Sample	Mean	Sample	Mean
(1,1) 1.0	1.5	(2,3) 2.5	2.5	(2,3) 2.5	4.0
(2,2) 2.0	2.0	(1,3) 2.0	2.5	(2,4) 3.0	3.5
(3,3) 3.0	2.5	(1,4) 2.5	3.5	(3,4) 3.5	4.0
(4,4) 4.0	3.0	(1,5) 3.0		(3,5) 4.0	4.5

Properties of the mean: the sample mean is an unbiased estimator of the population mean, and robust to departure of population distribution

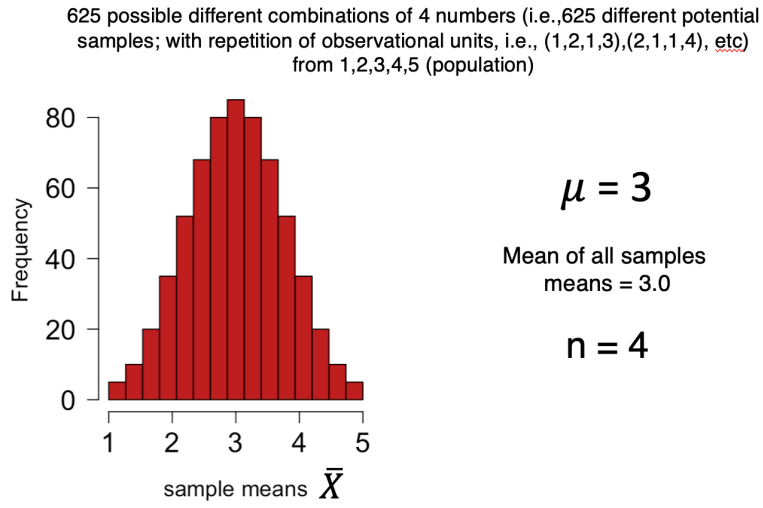
Assuming random sampling, the ***sample mean*** is an ***unbiased estimator*** of the population mean, regardless of the shape of the population distribution (normal or not).

When the population is normally distributed, the ***sampling distribution of the mean*** is normal for any sample size.

When the population is not normal, the sampling distribution of the mean becomes approximately normal as sample size increases.



Original distribution (uniform) of the population; it's called marginal distribution



Sample distribution (normal) of means

Properties of the variance: **the sample variance is an unbiased estimator of the population variance, but it is fragile to departures from the assumed population distribution.**

For normally distributed populations and assuming random sampling, the mean of all sample variances is the variance of the population (provided we use the **unbiased sample variance)**.

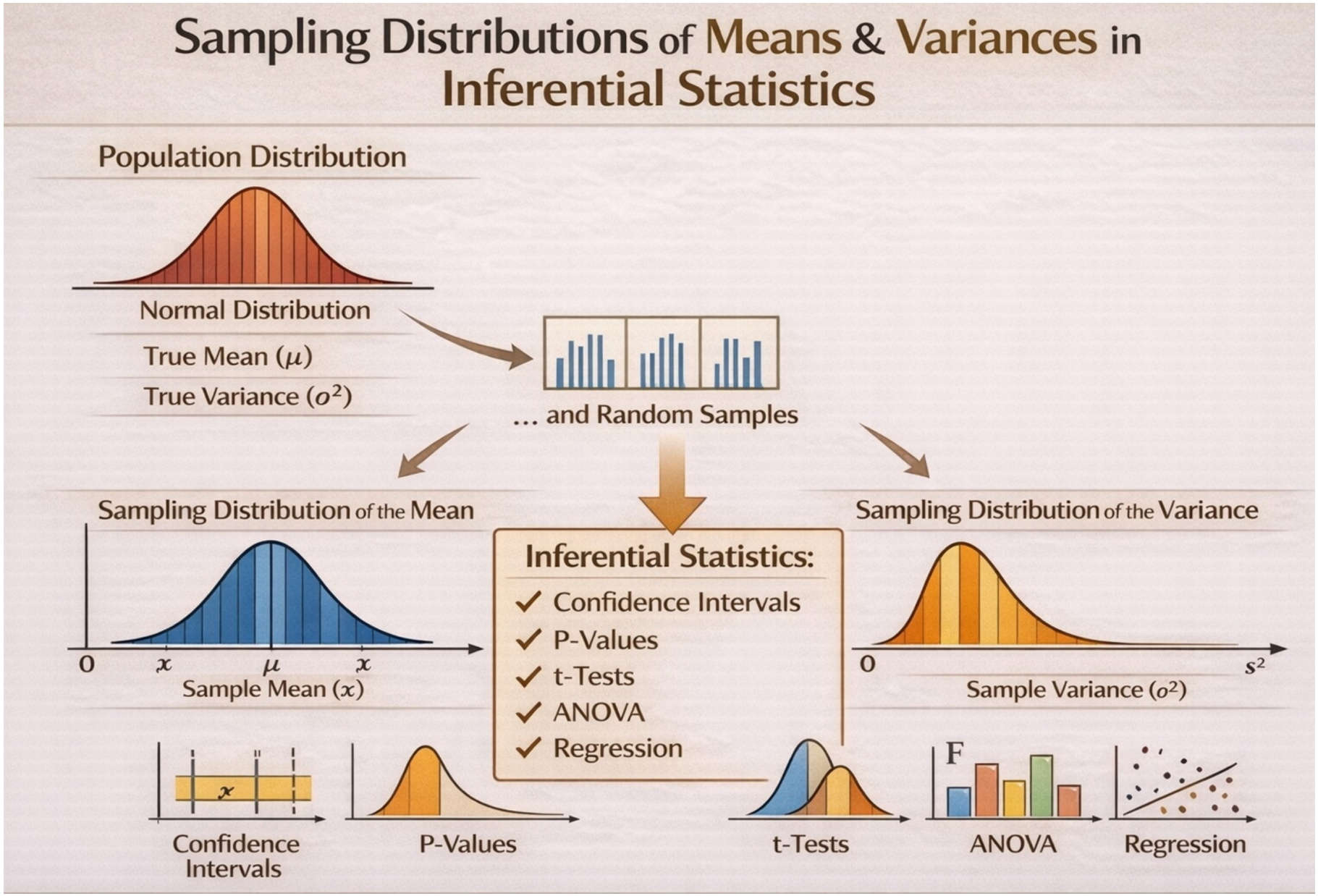
The ***sample variance*** is an ***unbiased estimator*** of the population variance, regardless of the shape of the population distribution (normal or not).

When the population is normal, a scaled version (not shown here) of the sample variance follows a chi-squared distribution.

When the population is not normal, the sampling distribution of the variance has no simple form though often remains skewed, even for large sample sizes. And this behaviour affects inference based on samples.

When populations depart strongly from normality, confidence intervals for the mean (which requires sampling distribution of variance) may not achieve their nominal coverage (e.g., 95% may become 93% or 97%), and hypothesis tests may exhibit Type I error rates that differ from the chosen significance level (e.g., 0.05 or 5% may become 0.03 or 0.07).

To make inferences based on uncertainty, we need the sampling distribution – not the original population



Why sampling properties of estimators are important?

The sample variance is unbiased even under moderate departures from normality.

In practice we rarely know when this robustness holds. For this reason, many statistical procedures that rely on variability (e.g., t-tests, ANOVA, and regression) explicitly assume normality.

Why sampling properties of estimators are important?

The sample variance is unbiased even under moderate departures from normality.

In practice we rarely know when this robustness holds. For this reason, many statistical procedures that rely on variability (e.g., t-tests, ANOVA, and regression) explicitly assume normality.

Normality ensures that test statistics (such as t and F) based on sample values can be validly compared to their theoretical sampling distributions (e.g., under H_0 or around sample values) allowing p-values to be correctly calculated.

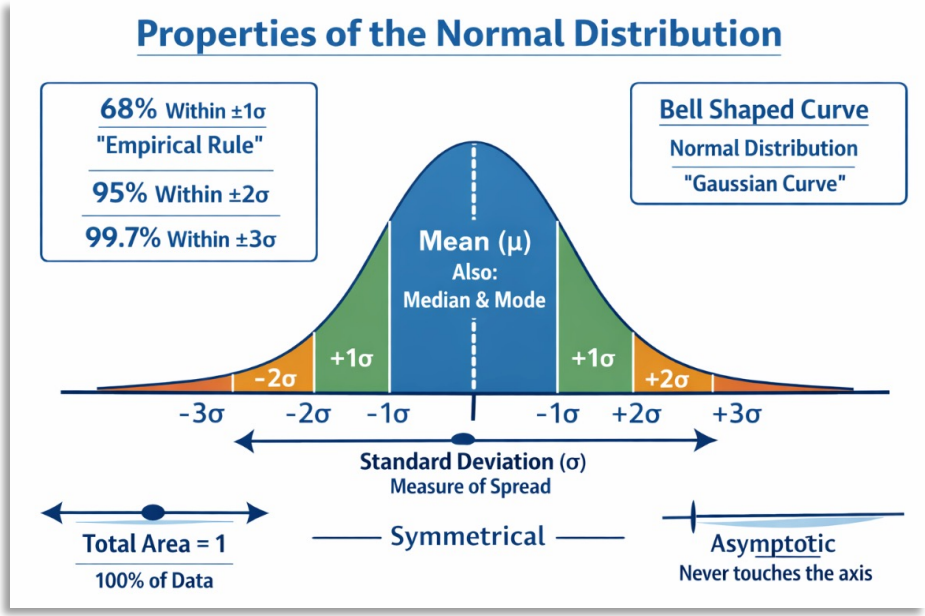
While some simple statistics (e.g., proportions) rely on a single observed value for inference (confidence intervals and p-values), others (notably means, e.g., t-tests, ANOVA, regression) depend on the proper behaviour of the sampling distributions of both the mean and the variance.

How common is the normal distribution in nature?

"Normality is a myth: there never has, and never will be, a normal distribution." Roy C. Geary (1896 - 1983).

This is because a normal distribution is a mathematical idealization:
it is continuous, perfectly symmetric, and defined over an infinite range, whereas real biological data are finite, bounded, discrete, and shaped by biological constraints.

But many biological variables are approximately normally distributed (i.e., their distribution is close enough to normal for standard statistical inference to work well).



How common is the normal distribution in nature?

The normal distribution is a statistical model used to derive sampling distributions, not a literal description of most real populations.

There is only one way to be perfectly normal, but infinitely many ways to deviate from normality. Nevertheless, the normal distribution often provides a good approximation for many biological variables.

Importantly, key estimators such as the sample mean and variance are often robust to moderate departures from normality.

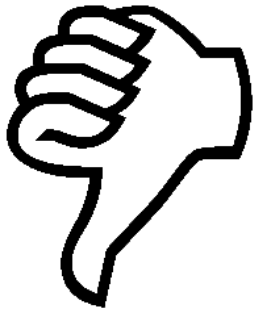
The above makes normal-based methods effective for approximately normal populations, but this robustness cannot be assumed in all cases.

Degrees of Freedom: Why They Matter for Unbiased Estimation

Why Corrections Are Needed for Unbiased Estimation: The Role of Degrees of Freedom

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

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



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*?

Let's use a computational approach to examine the performance of two estimators of the population variance

μ below is the population mean (often unknown)

σ^2=100; σ=10

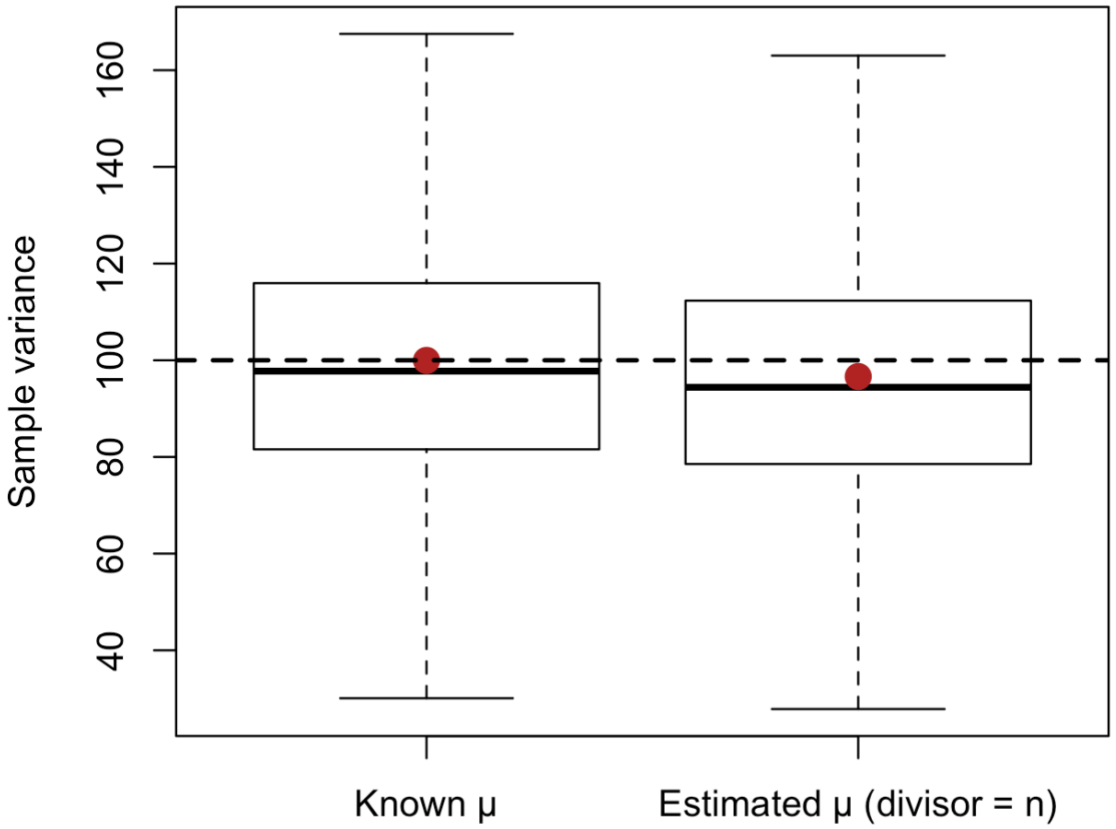
```
# Generate many samples
samples <- replicate(1e6, rnorm(n = 30, mean = 350, sd = 10))

# Two variance estimators
var_pop_mu <- function(x, mu) sum((x - mu)^2) / length(x)
var_sample_mu <- function(x) sum((x - mean(x))^2) / length(x)

# Apply estimators
v1 <- apply(samples, 2, var_pop_mu, mu = 350)
v2 <- apply(samples, 2, var_sample_mu)
```

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

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

Complete code (if you're curious)

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

# Variance using the TRUE population mean (μ known)
var.based.popMean <- function(x, mu) {sum((x - mu)^2) / length(x)}

# Variance using the sample mean, divisor = n (biased)
var.based.n <- function(x) {sum((x - mean(x))^2) / length(x)}

# Apply estimators across samples

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",
  xaxt = "n",ylab = "Sample variance",main = "Bias in variance estimation")

axis(1,at = c(1, 2),labels = c("Known μ", "Estimated μ (divisor = n)"))

# Add means
points(x = c(1, 2),y = c(mean(sample.var.based.Pop), mean(sample.var.n.instead)),
  pch = 19,cex = 1.5,col = "black"
)

# True population variance (σ² = 100)
abline(h = 100, lty = 2, lwd = 2)

legend("topright",legend = c("Mean of estimator", "True variance"),pch = c(19, NA),
  lty = c(NA, 2),lwd = c(NA, 2),bty = "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; it would have been exactly σ^2 =100 with infinite sampling); whereas the estimator based on the sample $\bar{Y}$ divided by *n* was biased.

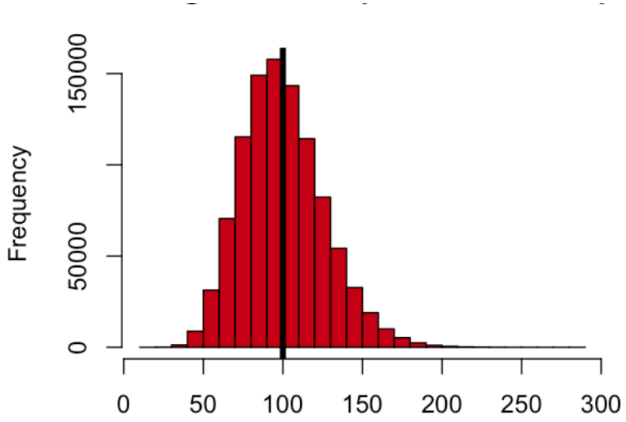
$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \boldsymbol{\mu})^2}{n} \qquad s^2 = \frac{\sum_{i=1}^{n}(Y_i - \boldsymbol{\bar{Y}})^2}{n}$$

Note the asymmetry of the sampling distribution of variances.

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.

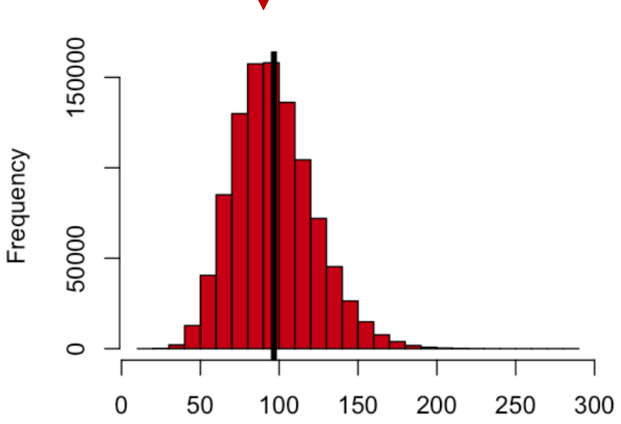
```
> mean(sample.var.based.Pop)
[1] 99.93689
> mean(sample.var.n.instead)
[1] 96.60124
```

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

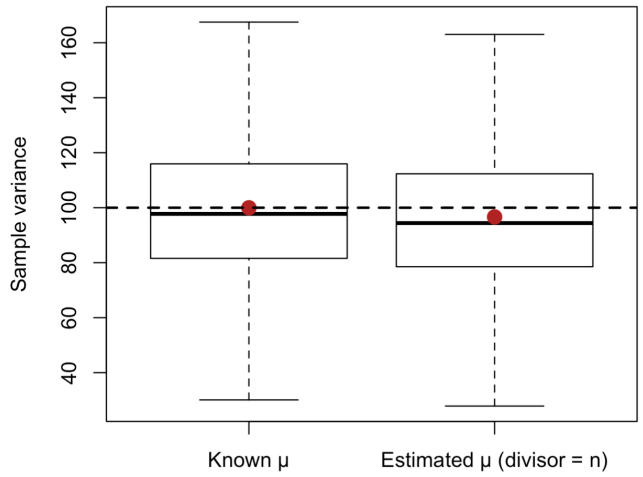


sample variances

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

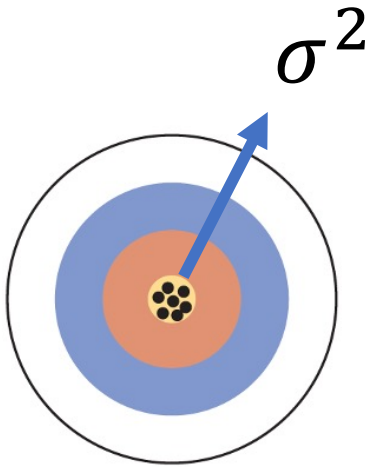


sample variances

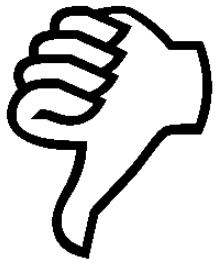


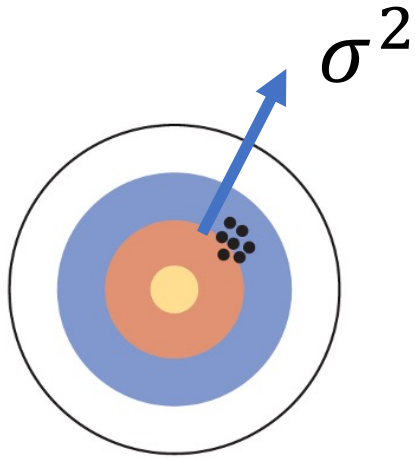
But in most (if not all) cases one doesn't know the parameter value μ (true population mean).

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



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





There is a correction factor for the sample bias in s^2 called Bessel's correction (It appears that Gauss had already introduced this idea in 1823).

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

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

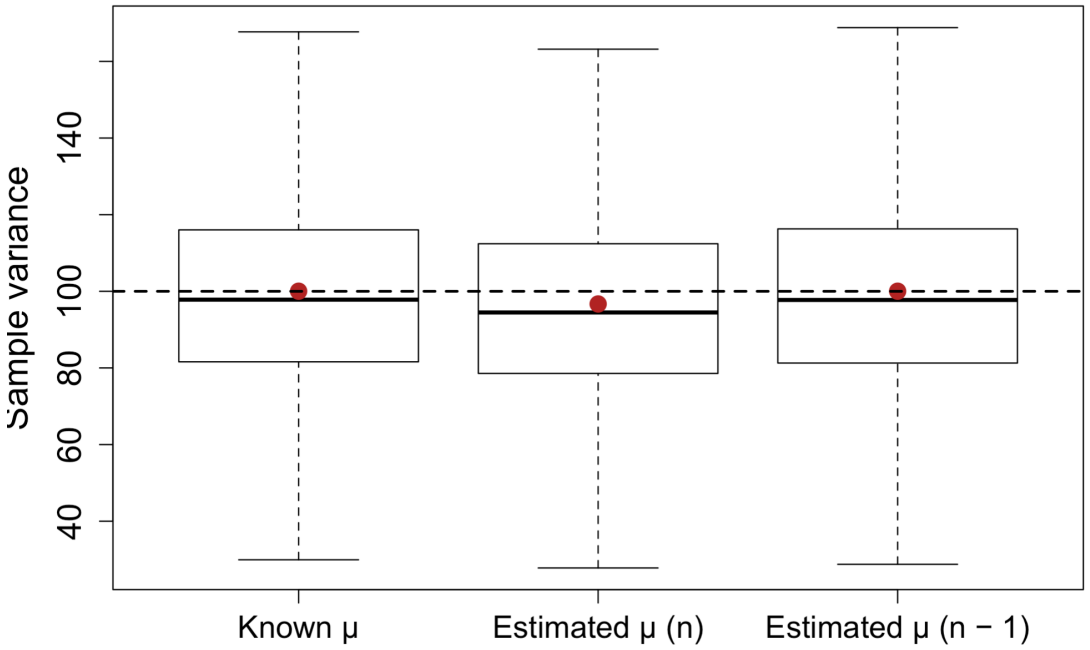
https://mathworld.wolfram.com/BesselsCorrection.html

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

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

```r
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")
```

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




```
> 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!

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

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

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

BUT WHY???

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?

You don’t need to know the mathematical details—but it’s useful to know that this theory has been worked 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:

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

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,

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

So we can write

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

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

Step ⋆ holds because

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

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, $\Big(1 - \frac{1}{N}\Big) = \Big(\frac{N-1}{N}\Big)$. If we want an unbiased estimator, we should multiply both sides of Equation 3 by the inverse of the constant:

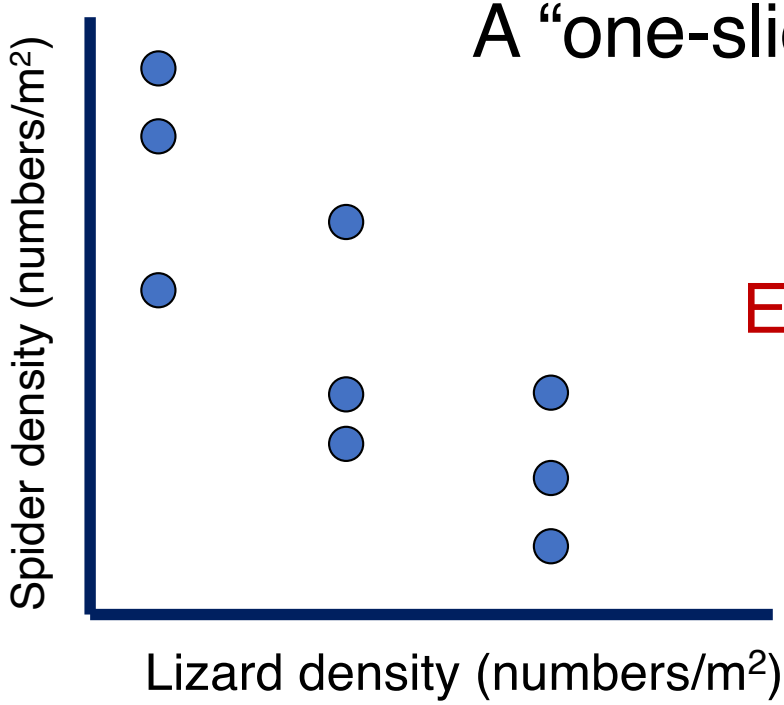
$$\mathbb{E}\Big[\Big(\frac{N}{N-1}\Big)s^2\Big] = \mathbb{E}\Big[\frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})^2\Big] = \sigma^2.$$

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

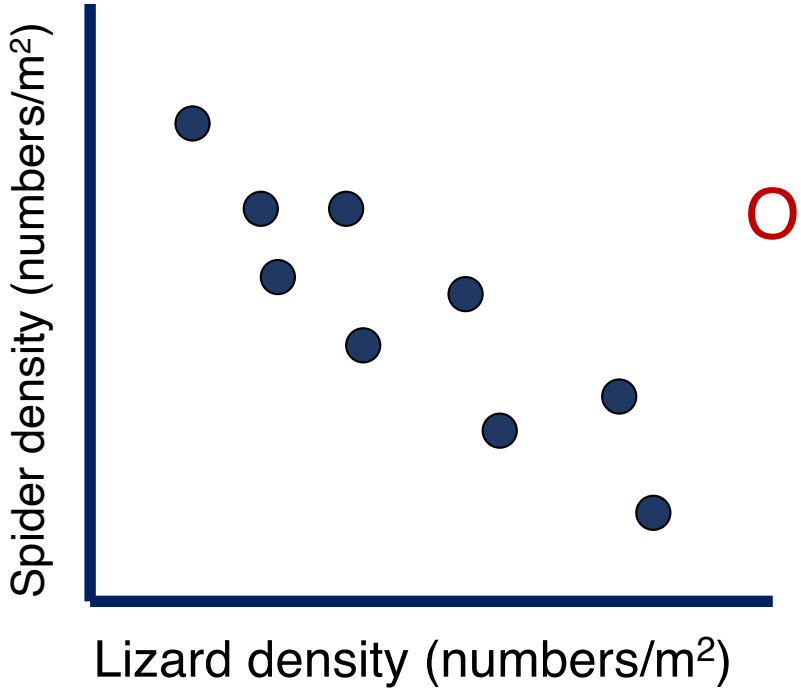
Source: http://gregorygundersen.com/blog/2019/01/11/bessel/

Let’s take a break - 1 minute

A “one-slide” discussion on experimental versus observational studies



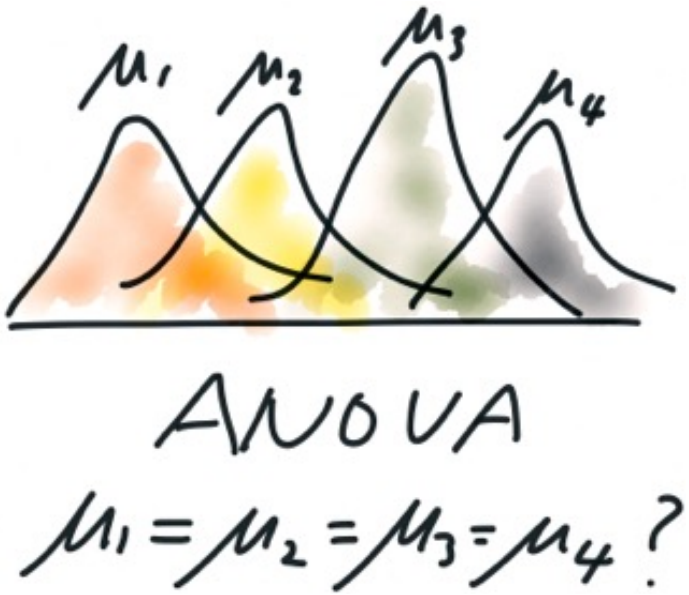
Experimental study



Observational study

COMPARING THE MEANS OF THREE OR MORE GROUPS (often called treatments or levels in experiments)

A REALLY QUICK REVIEW OF THE ANALYSIS OF VARIANCE (ANOVA)



THE ANALYSIS OF VARIANCE (ANOVA) for comparing multiple sample means (groups)

The problem about "The knees who say night"

By Whitlock and Schluter (2009)

OR

"Bright light behind the knees is just bright light behind the knees"

http://www.genomenewsnetwork.org/articles/08_02/bright_knees.shtml

Extraocular Circadian Phototransduction in Humans

Scott S. Campbell* and Patricia J. Murphy

Physiological and behavioral rhythms are governed by an endogenous circadian clock. The response of the human circadian clock to extraocular light exposure was monitored by measurement of body temperature and melatonin concentrations throughout the circadian cycle before and after light pulses presented to the popliteal region (behind the knee). A systematic relation was found between the timing of the light pulse and the magnitude and direction of phase shifts, resulting in the generation of a phase response curve. These findings challenge the belief that mammals are incapable of extraretinal circadian phototransduction and have implications for the development of more effective treatments for sleep and circadian rhythm disorders.

SCIENCE • VOL. 279 • 16 JANUARY 1998

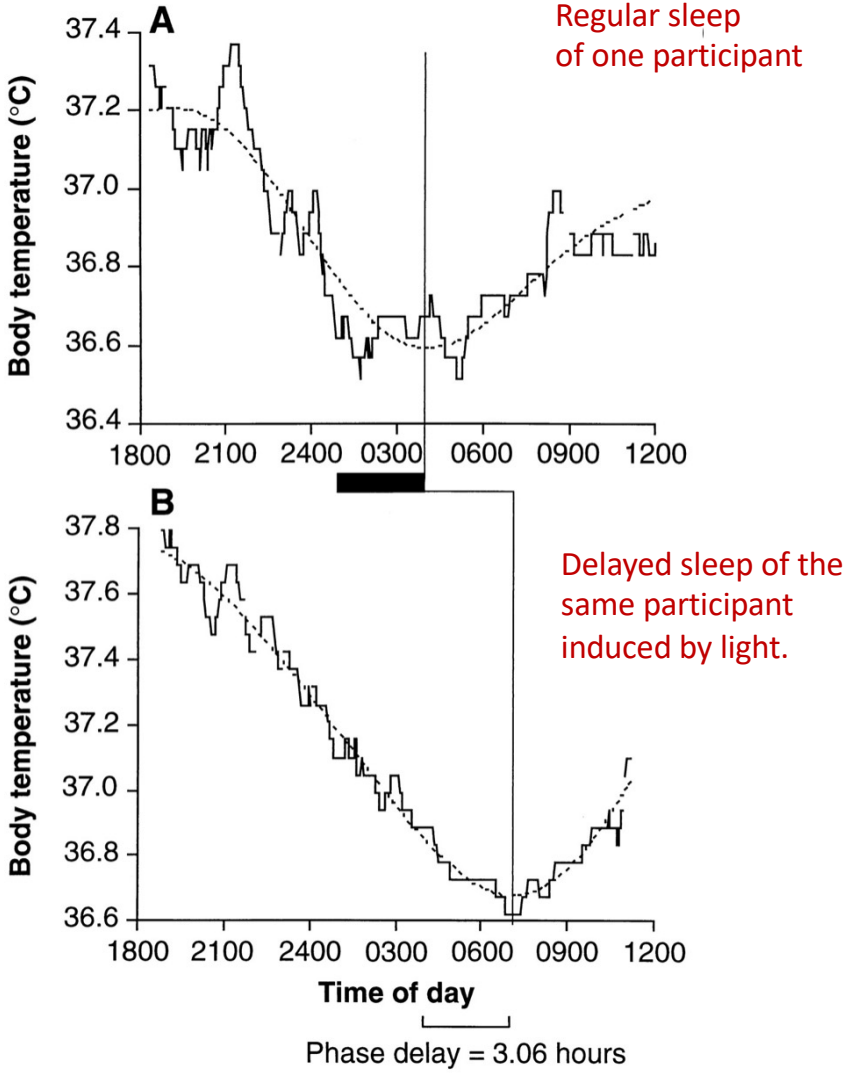
Data challenged as subjects were exposed to light while knees being illuminated

Our core body temperature is around 37°C but it fluctuates by about 1°C or so throughout the night.

The drop in temperature starts about two hours before you go to sleep, coinciding with the release of the sleep hormone melatonin.

Extraocular Circadian Phototransduction in Humans

Scott S. Campbell* and Patricia J. Murphy



Regular sleep of one participant

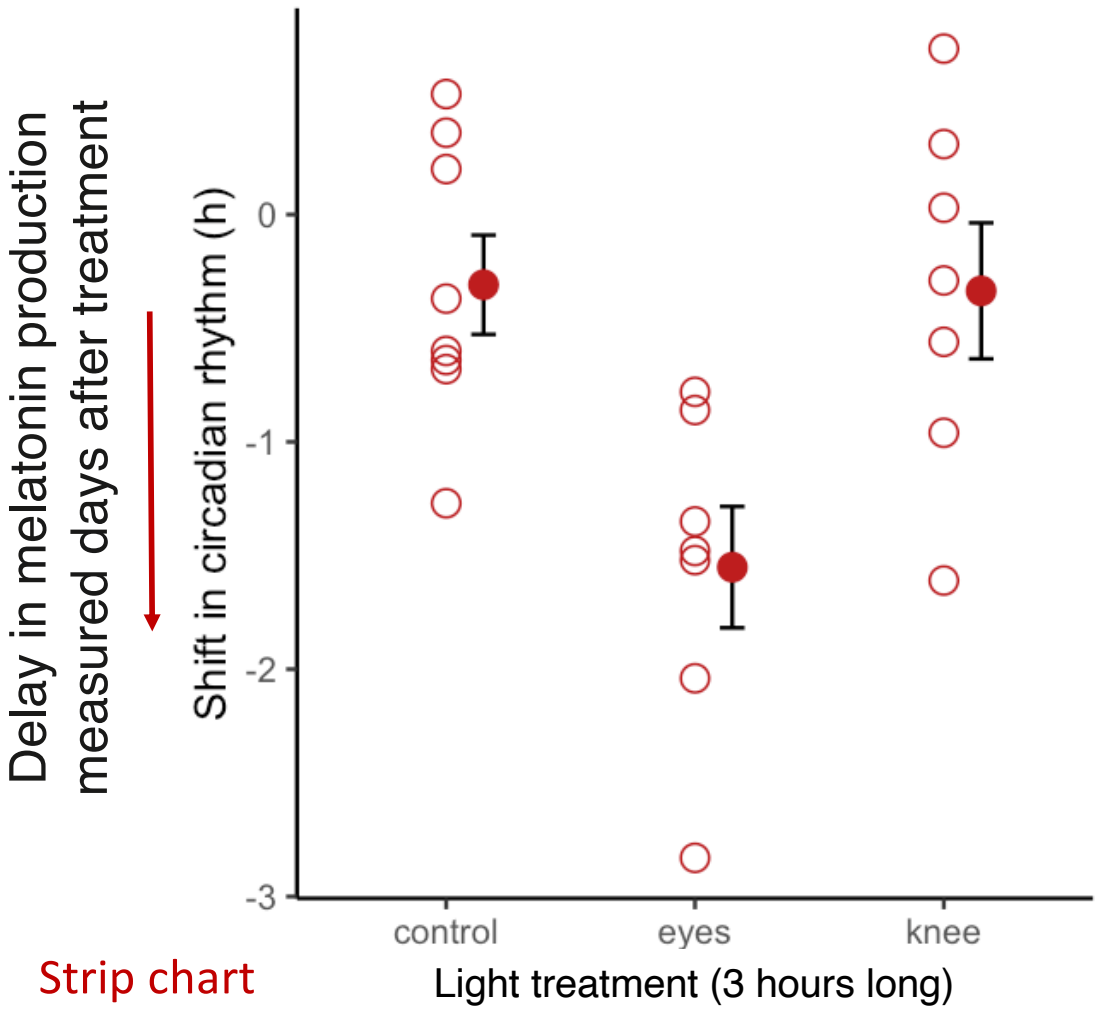
Delayed sleep of the same participant induced by light.

Example of a delay in circadian phase in response to a 3-hour bright light presentation to the popliteal region. Light was presented on one occasion between 0100 and 0400 on night 2 in the laboratory (black bar) while the participant (a 29-year-old male) remained awake and seated in a dimly lit room (ambient illumination <20 lux).

The circadian phase was determined by fitting a complex cosine curve (dotted line

The resulting phase delay was 3.06 hours

THE ANALYSIS OF VARIANCE (ANOVA)
for comparing multiple sample means (groups or treatments)



Strip chart

PHYSIOLOGY

SCIENCE VOL 297 26 JULY 2002

Absence of Circadian Phase Resetting in Response to Bright Light Behind the Knees

Kenneth P. Wright Jr.* and Charles A. Czeisler

New study challenged the original study (Wright & Czeiler 2002): subjects were exposed to light while knees being illuminated by original study.

22 people randomly assigned to one of the three light treatments.

Do these means come from the same statistical population, i.e., do these samples only differ from each other due to sampling variation?

THE ANALYSIS OF VARIANCE (ANOVA)
for comparing multiple sample means (groups or treatments)

H_0: The samples come from statistical populations with the same mean, i.e., $\mu_{control} = \mu_{knee} = \mu_{eyes}$.

H_A: At least two samples come from different statistical populations with different means.

THE ANALYSIS OF VARIANCE (ANOVA)
for comparing multiple sample means among groups (also called treatments or treatment levels)

H_0: The samples come from statistical populations with the same mean, i.e., $\mu_{control} = \mu_{knee} = \mu_{eyes}$.

H_A: At least two samples come from different statistical populations with different means.

Which is to say:

H_0: Differences in means among groups are due to **sampling error (sampling variation) from the same population**.

H_A: Differences in means among groups are NOT due to **sampling error (sampling variation) from the same population**.

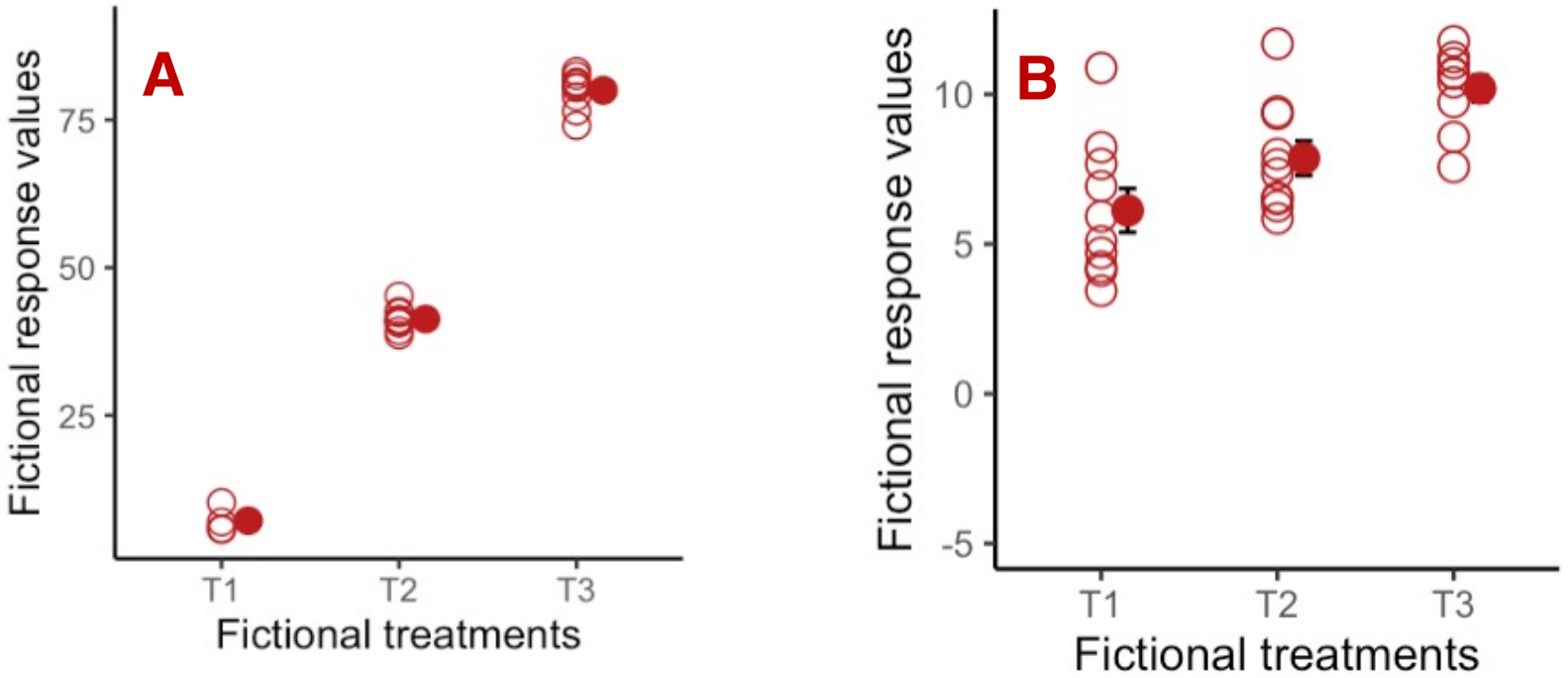
Remember: ***Sampling error*** is due to sampling variation, i.e., samples that come from the same statistical population may differ in their means just due to chance alone.

We need a test statistic that is sensitive to mean variation across multiple groups (or treatments): The F statistic does that by considering the ratio of two variances (variance components):

Means among groups are much bigger in **A** than **B**;
residuals variation is similar across treatments in both **A** than **B.** Notice the differences in their Y-scales (the mean differences among groups is huge in **A**).

$$F_A = \frac{14078.0}{5.71} = 2456.90$$

$$F_B = \frac{47.41}{3.64} = 13.03$$

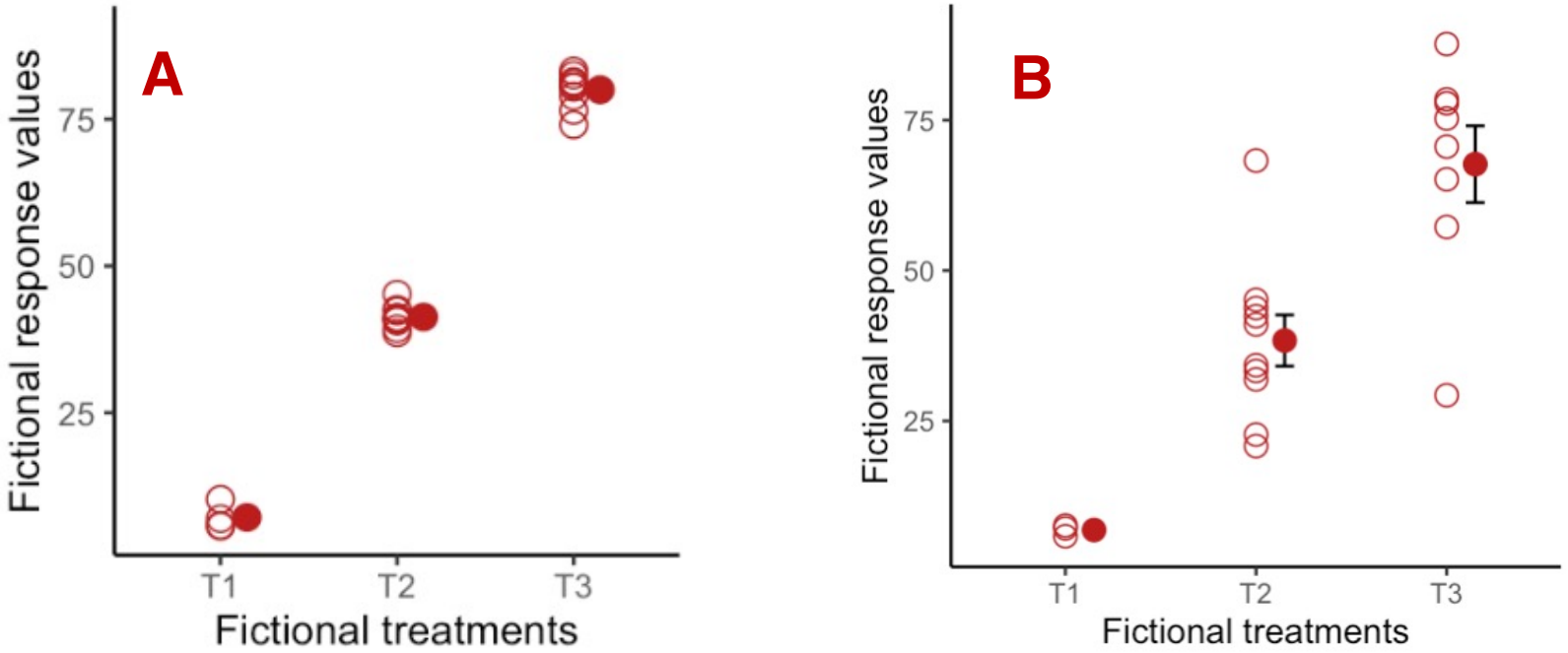


Note that scales (Y-axis) are different

HETEROSCEDASTICITY reduces the F-ratio ability to differentiate among differences in means among groups

Means among groups are somewhat similar in **A** than **B**;
A is homoscedastic **B** heteroscedastic

$$F_A = \frac{14078.0}{5.71} = 2456.90 \qquad F_B = \frac{12275.0}{217.9} = 56.34$$



Note that scales (Y-axis) are now equal

Verbal representation of equations

Let's talk ANOVA "jargon"

$$F = \frac{\text{variance among group means (due to "treatment")}}{\text{variance within groups (caller error or residual variation not explained by the mean within groups)}}$$

$$F = \frac{\text{Group Mean Square}}{\text{Error Mean Square}} = \frac{MS_{groups}}{MS_{error}}$$

The **F-statistic compares signal to noise** by contrasting **between-group variation** with **within-group variation**.

The F statistic measures the variance among groups but accounting for the variance within groups

Group Mean Square

MS_{groups}

(b=between or among)

Mean of each group

Total mean!

$$F = \frac{s_b^2}{s_w^2} = \frac{\sum_{i=1}^{g} n_i(\bar{X}_i - \bar{\bar{X}})^2}{g-1}$$

Degrees of freedom of MS_{groups}

MS_{errors}

(w=within groups)

Error Mean Square

The F statistic in the ANOVA context is so important that is more than worth knowing how it works!

The F statistic measures the variance among groups but accounting for the variance within groups

The F statistic in the ANOVA context is so important that is more than worth knowing how it works!

Group Mean Square

MS_{groups}

(b=between or among)

Mean of each group

Total mean!

$$F = \frac{s_b^2}{s_w^2} = \frac{\dfrac{\sum_{i=1}^{g} n_i(\overline{X}_i - \overline{\overline{X}})^2}{g-1}}{\dfrac{\sum_{i=1}^{g}(n_i - 1)s_i^2}{\sum_{i=1}^{g}(n_i - 1)}}$$

Degrees of freedom of MS_{groups}

Variance of each group

Big "N"; sum of all sample sizes across groups

Number of groups

$= (N-g)$

Degrees of freedom of MS_{groups}

Sample size of each group

MS_{errors}

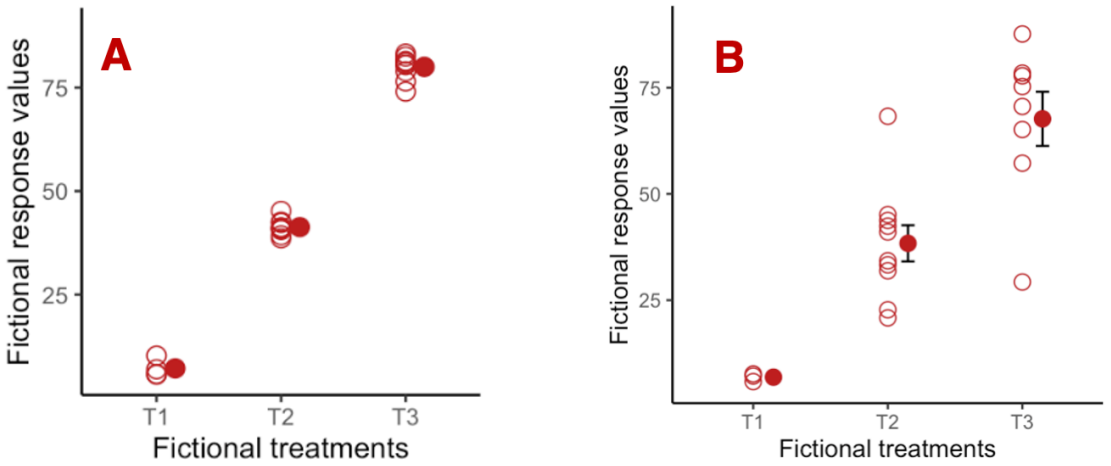
(w=within groups)

Error Mean Square

The **F-statistic compares signal to noise** by contrasting **between-group variation** with **within-group variation**. When variances are equal across groups (**homoscedasticity**), the within-group variance provides a stable and efficient estimate of background noise, allowing true differences among means to stand out clearly.

Heteroscedasticity makes some groups so noisy that the test treats all differences among means as less reliable.

$$F_A = \frac{14078.0}{5.71} = 2456.90 \qquad F_B = \frac{12275.0}{217.9} = 56.34$$



Note that scales (Y-axis) are now equal

A small example: worth doing it “by hand”!

Let’s suppose two groups for simplicity!

group 1

1	2	3	4	5

$\overline{X}_1 = 3.0$

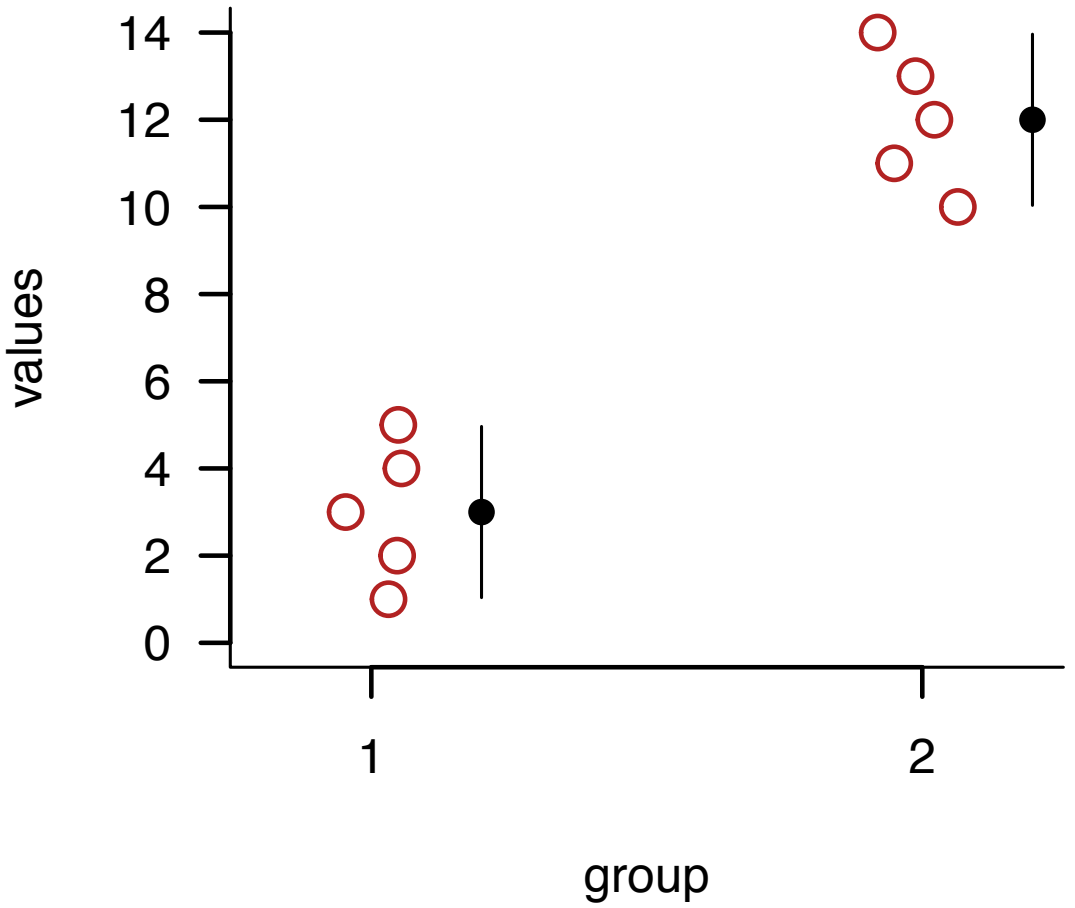
$s_1^2 = 2.5$

group 2

10	11	12	13	14

$\overline{X}_2 = 12.0$

$s_2^2 = 2.5$



g_1: 1 2 3 4 5

g_2: 10 11 12 13 14

$\overline{X}_1 = 3.0 \quad \overline{X}_2 = 12.0$

$s_1^2 = 2.5 \quad s_2^2 = 2.5$

$\overline{\overline{X}}$ = (1+2+3+4+5+10+11+12+13+14)/10 = **7.5**

MS_{groups}= variance among group means (due to "treatment")

= (5x(**3.0** - 7.5)^2 + 5x(**12.0** - 7.5)^2)/(2-1) =

202.5/(2-1) = **202.5**

$df(MS_{groups}) = g - 1$

$$F = \frac{\mathbf{202.5}}{???} = ???$$

Mean of each group — Total mean!

$$F = \frac{s_b^2}{s_w^2} = \frac{\dfrac{\sum_{i=1}^{g} n_i(\overline{X}_i - \overline{\overline{X}})^2}{g-1}}{\dfrac{\sum_{i=1}^{g}(n_i - 1)s_i^2}{\sum_{i=1}^{g}(n_i - 1)}}$$

MS_{groups}

Variance of each group

Big "N"; sum of all sample sizes across groups

$\sum_{i=1}^{g}(n_i - 1) \rightarrow = (N-g)$

g_1: 1 2 3 4 5

g_2: 10 11 12 13 14

Mean of each group Total mean!

$$F = \frac{s_b^2}{s_w^2} = \frac{\dfrac{\sum_{i=1}^{g} n_i(\overline{X}_i - \overline{\overline{X}})^2}{g-1}}{\dfrac{\sum_{i=1}^{g}(n_i - 1)s_i^2}{\sum_{i=1}^{g}(n_i - 1)}}$$

Variance of each group

Big "N"; sum of all sample sizes across groups

$\sum_{i=1}^{g}(n_i - 1) \rightarrow$ = (N-g)

MS_{error}

$\overline{X}_1 = 3.0 \quad \overline{X}_2 = 12.0$

$s_1^2 = 2.5 \quad s_2^2 = 2.5$

MS_{error} = variance within groups (residuals)

$MSE_1 = (1-3.0)^2+(2-3.0)^2 +(3-3.0)^2 +(4-3.0)^2 +(5-3.0)^2 =$ **10**

$MSE_2 = (10-12.0)^2 +(11-12.0)^2 +(12-12.0)^2 +(13-12.0)^2 +(14-12.0)^2 =$ **10**

$MS_{error} = (MSE_1 + MSE_2)/(N-g)=(10+10) / (10-2) = 20/8 =$ **2.5**

$df(MS_{error}) = N-g = 10 - 2 = 8$

$\overline{\overline{X}}$ = (1+2+3+4+5+10+11+12+13+14)/10 = **7.5**

MS_{groups}=

= (5x(**3.0** - 7.5)^2 + 5x(**12.0** - 7.5)^2)/(2-1) =

202.5/(2-1) = **202.5**

$df(MS_{groups}) = g - 1 = 2-1$

$$F = \frac{\mathbf{202.5}}{\mathbf{2.5}} = 81$$

MS_{error} = variance within groups (residuals)

MSE_1 = (1-3.0)^2+(2-3.0)^2 +(3-3.0)^2 +(4-3.0)^2 +(5-3.0)^2 = **10**

MSE_2= (10-12.0)^2 +(11-12.0)^2 +(12-12.0)^2 +(13-12.0)^2 +(14-12.0)^2 = **10**

$MS_{error} = (MSE_1 + MSE_2)/(N-g)=(10+10) / (10-2) = 20/8 =$ **2.5**

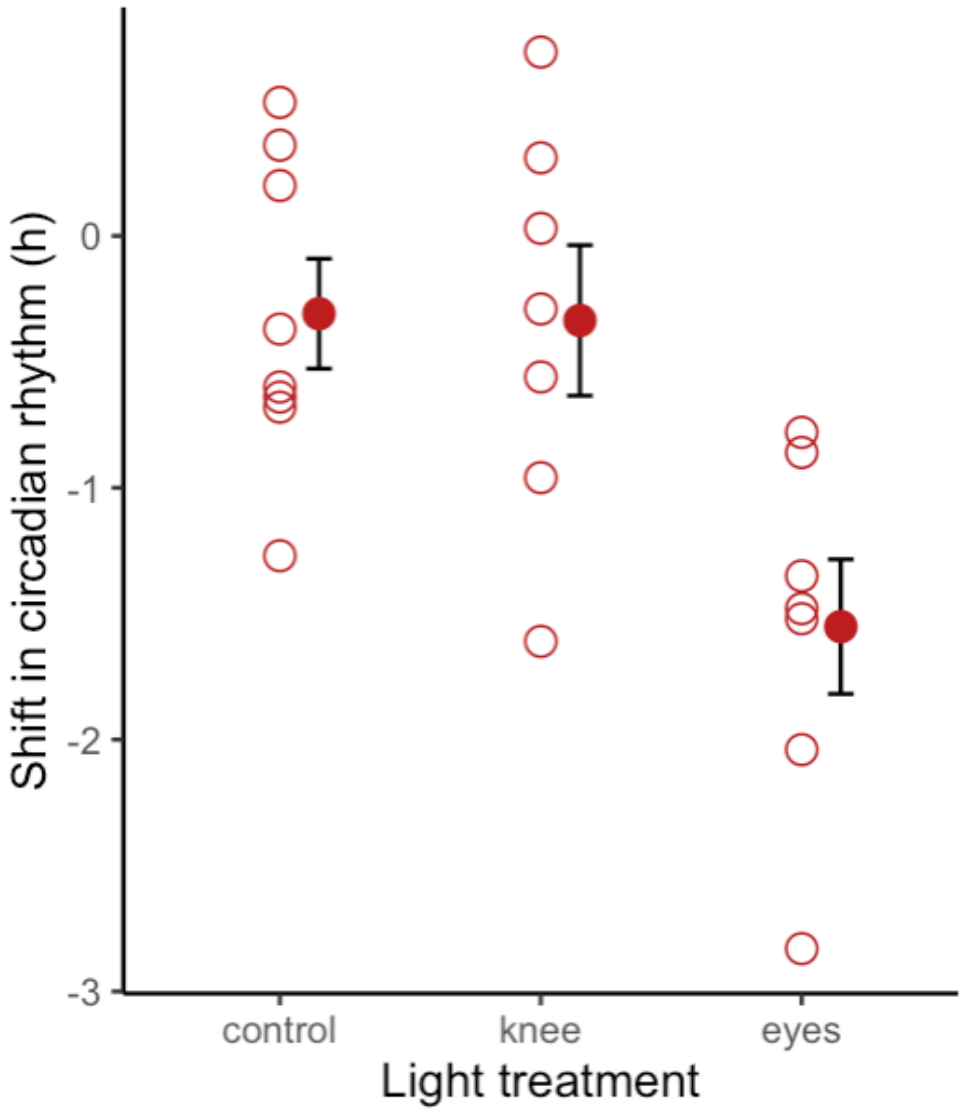
$df(MS_{error}) = N-g = 10 – 2 = 8$

Let’s take a break - 1 minute

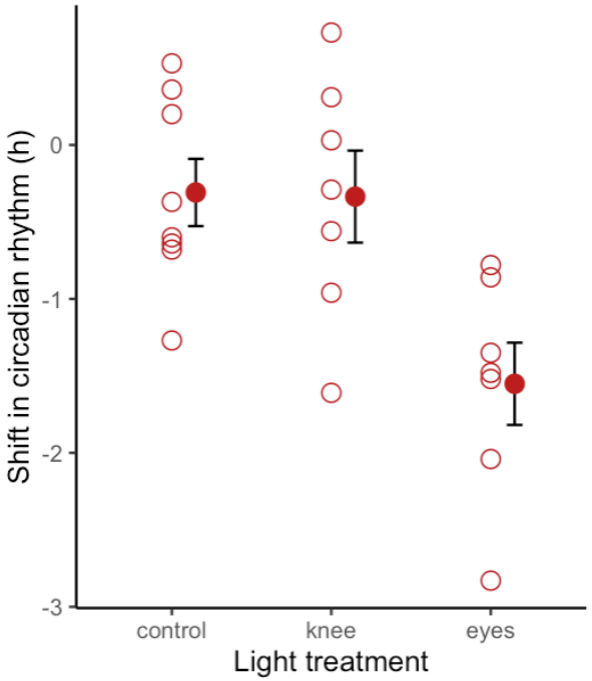
LET's go back to the "The knees who say night"

A	B
treatment	shift
control	0.53
control	0.36
control	0.2
control	-0.37
control	-0.6
control	-0.64
control	-0.68
control	-1.27
knee	0.73
knee	0.31
knee	0.03
knee	-0.29
knee	-0.56
knee	-0.96
knee	-1.61
eyes	-0.78
eyes	-0.86
eyes	-1.35
eyes	-1.48
eyes	-1.52
eyes	-2.04
eyes	-2.83

data in a csv file



“The knees who say night”



Statistical Conclusion?

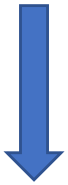
H_0: The samples come from the same population.

H_A: At least two samples come from different populations.

```
summary(aov(shift ~ treatment, data=circadian))
            Df Sum Sq Mean Sq F value  Pr(>F)
treatment    2  7.224   3.612   7.289 0.00447 **
Residuals   19  9.415   0.496
---
```

“The knees who say night”

```
summary(aov(shift ~ treatment, data=circadian))
            Df Sum Sq Mean Sq F value  Pr(>F)
treatment    2  7.224   3.612   7.289 0.00447 **
Residuals   19  9.415   0.496
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



ANOVA Table – reporting quality

Source of variation	Sum of squares	df	Mean square	F	P
Between	7.224	2	3.612	7.289	0.00447
Within	9.415	19	0.496		

Remembering the role of degrees of freedom

Source of variation	Sum of squares	df	Mean square	F	P
Between	7.224	2	3.612	7.289	0.00447
Within	9.415	19	0.496		

Remember that the calculations of sum of squares (as in variance) involve subtractions from means so that they would be biased if not divided by adjustments (degrees of freedom) to produce mean square deviations.

“The knees who say night”

ANOVA Table

Source of variation	Sum of squares	df	Mean square	F	P
Between	7.224	2	3.612	7.289	0.00447
Within	9.415	19	0.496		

H_0: The samples come from the same population.

H_A: At least two samples come from different populations.

Reject H_0

How does the ANOVA significance test work?

How can we conceptualize the construction of the F distribution?

The statistical “machinery”:

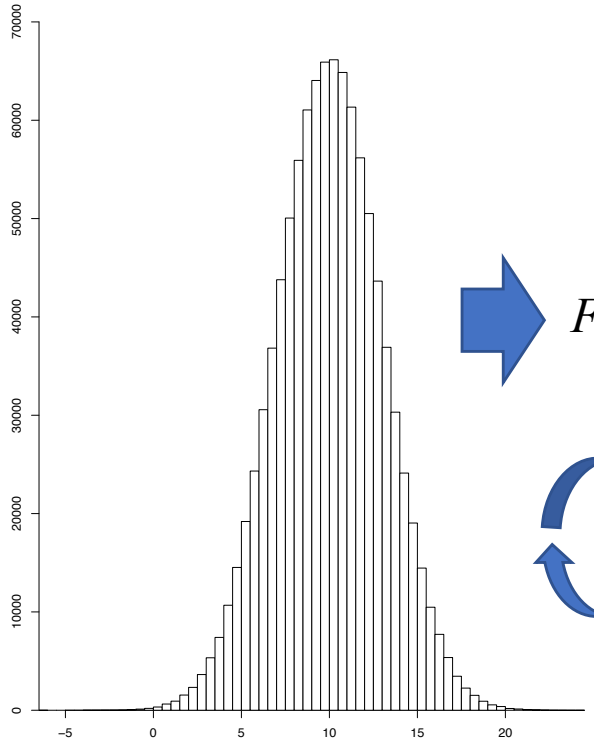
1) Assume that $\mathbf{H_0}$ is true (i.e., samples come from the same population; i.e., population having the same mean and same variance).

2) Sample from the population the appropriate number of groups (samples) respecting the sample size of each group.

3) Repeat step 2 a large (or infinite) number of times and each time calculate the F statistic.

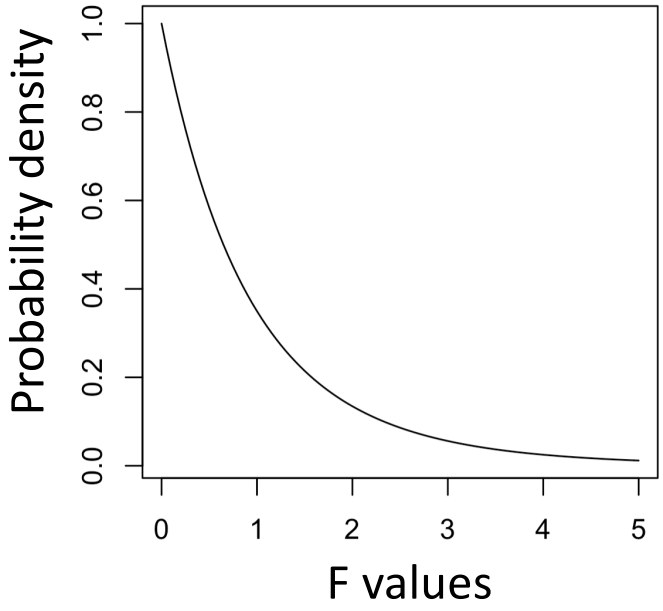
The F (sampling) distribution assuming that H_0 is true

H_0: Differences in means among groups are due to **sampling error from the same population**.

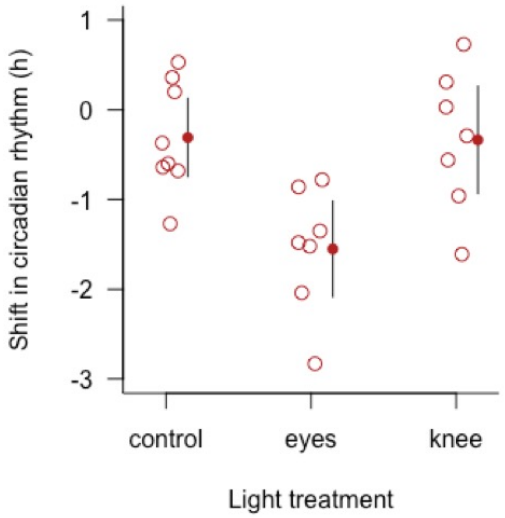


$$F = \frac{s_b^2}{s_w^2} = \frac{\dfrac{\sum_{i=1}^{g} n_i(\bar{X}_i - \bar{\bar{X}})^2}{g-1}}{\dfrac{\sum_{i=1}^{g}(n_i - 1)s_i^2}{\sum_{i=1}^{g}(n_i - 1)}}$$

(8,7,7) observations



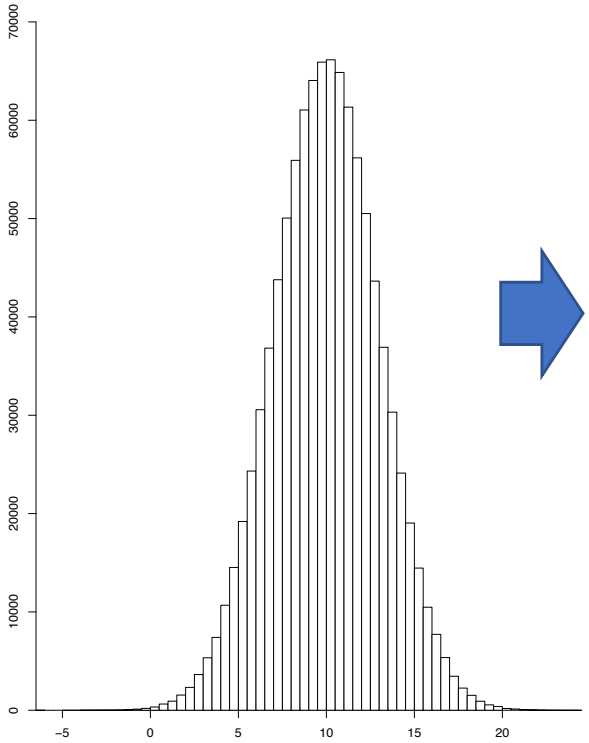
Sample from the same (normally distributed) population (i.e., assume that ***H_0 is true***), respecting the original number of groups and their sample sizes.



Control: 8 observations
Eyes: 7 observations
Knee: 7 observations

The F (sampling) distribution assuming that H_0 is true

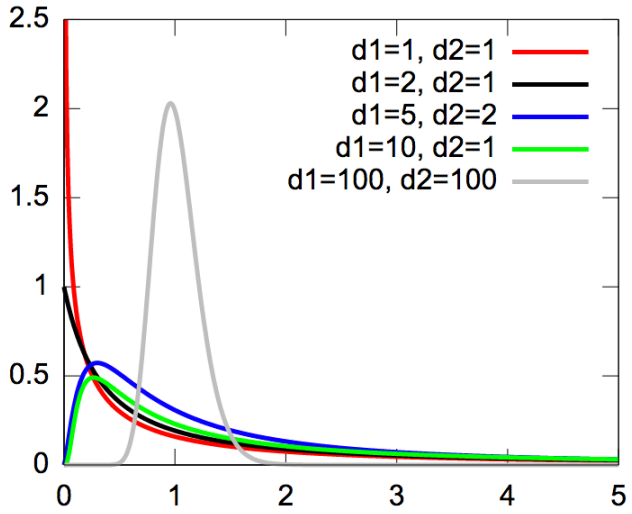
H_0: Differences in means among groups are due to **sampling error from the same population**.



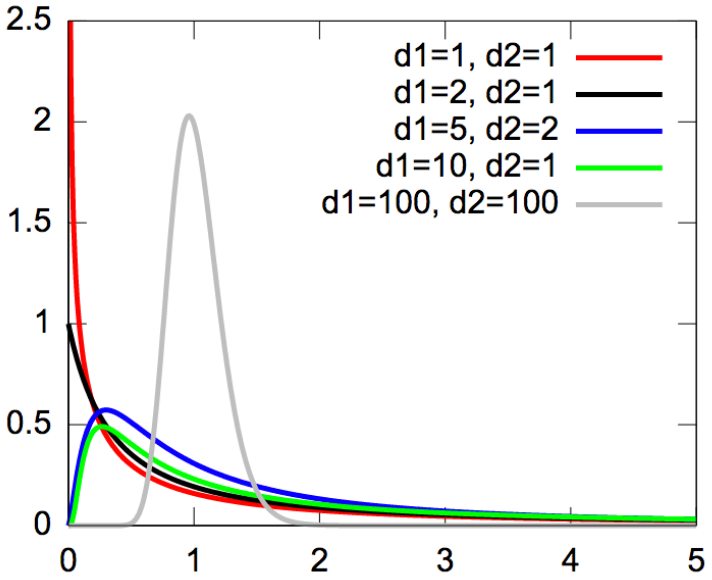
$$F = \frac{s_b^2}{s_w^2} = \frac{\dfrac{\sum_{i=1}^{g} n_i(\bar{X}_i - \bar{\bar{X}})^2}{g-1}}{\dfrac{\sum_{i=1}^{g}(n_i - 1)s_i^2}{\sum_{i=1}^{g}(n_i - 1)}}$$

Different number of groups and different number of observations per group generate different shapes for the F distribution.

Sample from the same (normally distributed) population (i.e., ***H_0 is true***), respecting the original number of groups and their sample sizes.



The F distribution assuming that H_0 is true (i.e., the sampling distribution of the test statistic F when H_0 is true).



Mean of each group

Total mean!

$$F = \frac{s_b^2}{s_w^2} = \frac{\dfrac{\sum_{i=1}^{g} n_i(\bar{X}_i - \bar{\bar{X}})^2}{g-1}}{\dfrac{\sum_{i=1}^{g}(n_i - 1)s_i^2}{\sum_{i=1}^{g}(n_i - 1)}}$$

df_1

Variance of each group

Big "N"; sum of all sample sizes across groups

$\sum_{i=1}^{g}(n_i - 1) \rightarrow = (N-g)$

df_2

The numerator degrees of freedom is based on the number of groups (g-1) and the denominator degrees of freedom depends on the total number of observations (N-g)

```
summary(aov(shift ~ treatment, data=circadian))
            Df Sum Sq Mean Sq F value  Pr(>F)
treatment    2  7.224   3.612   7.289 0.00447 **
Residuals   19  9.415   0.496
```

Degrees of freedom

Observed F-value (observed test statistic)

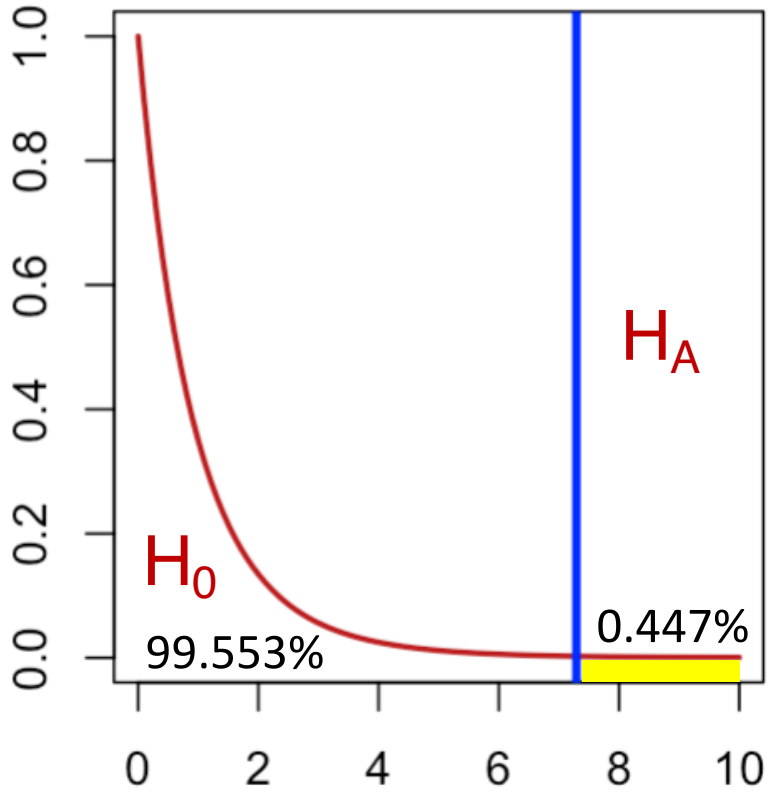
P-value

ANOVA is a one-sided (one-tail) test by design

H_0: The samples come from statistical populations with the same mean, i.e., $\mu_{control} = \mu_{knee} = \mu_{eyes}$.

H_A: At least two samples come from different statistical populations with different means.

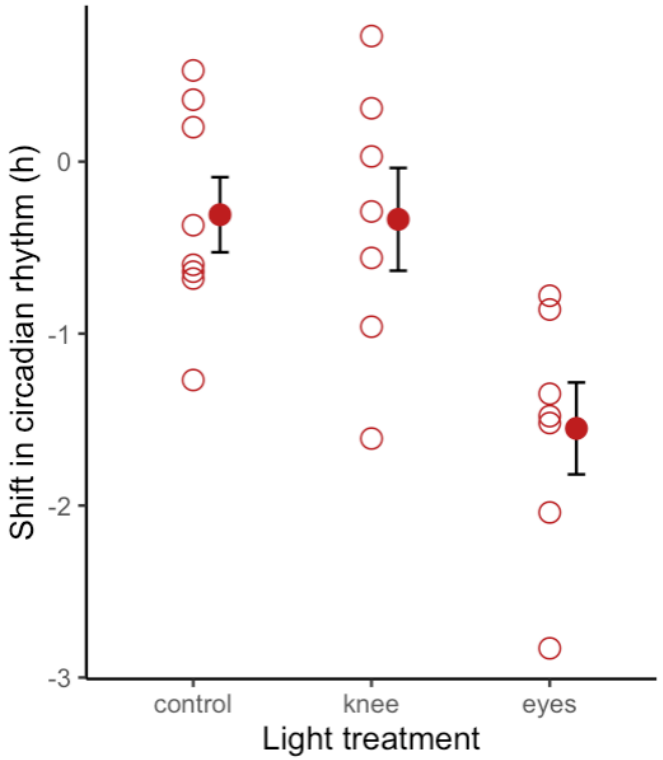
The probability of rejection of **H_0** (P-value) is estimated as the number of F-values in the null distribution equal or greater than the observed F-value (i.e., one tailed-test).



THE ANALYSIS OF VARIANCE (ANOVA)
for comparing multiple sample means (groups or treatments)

H_0: The samples come from statistical populations with the same mean, i.e., $\mu_{control} = \mu_{knee} = \mu_{eyes}$.

H_A: At least two samples come from different statistical populations with different means.



Statistical conclusion: Light treatment influences shifts in circadian rhythm.

Research conclusion: Light treatment influences shifts in circadian rhythm.

ANOVA

Assumptions are the same as for the independent two sample t-test:

- Each of the observations is a random sample from its population (whether they are the same or different populations).

- The variable (e.g., shift in circadian rhythm) is normally distributed in each (treatment) population. More on that in another lecture.

- The variances are equal among all populations from which the treatments were sampled (otherwise the F values change in ways that may not measure difference among means). More on that in another lecture.

"The knees who say night"

H_0: $\mu_{control} = \mu_{knee} = \mu_{eyes}$

H_A: at least one population mean (μ) is different from another population mean or other population means.

Conclusion?

Significant, but how?

How do we know which group means differ from one another?

Why not simply not contrast all pairs of means using a two-sample mean t-test?

Control vs. knee; control vs. eyes; knee vs. eyes?

More later in the course!