

**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*
*variance*

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

*Standard deviation*

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


---

---

---

---

---

---

---

---

1

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

**Sampling Distributions of Means & Variances in Inferential Statistics**

Population Distribution  
 Normal Distribution  
 True Mean ( $\mu$ )  
 True Variance ( $\sigma^2$ )

... and Random Samples

Sampling Distribution of the Mean  
 Sample Mean ( $\bar{x}$ )

Sampling Distribution of the Variance  
 Sample Variance ( $s^2$ )

**Inferential Statistics:**  
 ✓ Confidence Intervals  
 ✓ P-Values  
 ✓ t-Tests  
 ✓ ANOVA  
 ✓ Regression

Confidence Intervals   P-Values   t-Tests   ANOVA   Regression

---

---

---

---

---

---

---

---

2

**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).

**Why is it important to make assumptions about statistical populations of interest?**

**Sampling theory** ▶ Uses repeated sampling to model the expected variability of sample values (probabilities) in a statistical population.

**Repeated sampling** ▶ Helps derive the **sampling distribution** used in confidence intervals and hypothesis tests (Lecture 3).

**Important:** Repeated sampling only works under specific assumptions about the population!

**Statistical inference relies on our assumptions about the population!**

---

---

---

---

---

---

---

---

3



**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

7

---

---

---

---

---

---

---

---

---

---

**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).

8

---

---

---

---

---

---

---

---

---

---

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

### Sampling Distributions of Means & Variances in Inferential Statistics

Population Distribution

Normal Distribution  
True Mean ( $\mu$ )  
True Variance ( $\sigma^2$ )

... and Random Samples

Sampling Distribution of the Mean

Sampling Distribution of the Variance

Inferential Statistics:

- ✓ Confidence Intervals
- ✓ P-Values
- ✓ t-Tests
- ✓ ANOVA
- ✓ Regression

Confidence Intervals, P-Values, t-Tests, ANOVA, Regression

9

---

---

---

---

---

---

---

---

---

---

**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.

---

---

---

---

---

---

---

---

10

**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.

---

---

---

---

---

---

---

---

11

**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).*

---

---

---

---

---

---

---

---

12

**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.

---

---

---

---

---

---

---

---

13

*Degrees of Freedom: Why They Matter for Unbiased Estimation*




---

---

---

---

---

---

---

---

14

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

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


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


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

---

---

---

---

---

---

---

---

15

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

$\mu$  below is the population mean (often unknown)  $\sigma^2=100; \sigma=10$

```
# 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)
```

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

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

---

---

---

---

---

---

---

---

---

---

---

---

16

Complete code (if you're curious)

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

# Variance using the TRUE population mean (mu 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 mu", "Estimated mu (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 (sigma^2 = 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")
```

---

---

---

---

---

---

---

---

---

---

---

---

17

```
> 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  $\mu$  divided by  $n$  was unbiased (i.e., pretty much the population  $\sigma^2$ ; it would have been exactly  $\sigma^2 = 100$  with infinite sampling); whereas the estimator based on the sample  $Y$  divided by  $n$  was biased.

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

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

---

---

---

---

---

---

---

---

---

---

---

---

18

Note the asymmetry of the sampling distribution of variances.

The variance is unbiased when based on  $\mu$  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
    
```

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

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


---

---

---

---

---

---

---

---

---

---

---

---

19

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

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

👍

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

👎

---

---

---

---

---

---

---

---

---

---

---

---

20

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).

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

👍

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

👎

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

---

---

---

---

---

---

---

---

---

---

---

---

21

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")
    
```

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

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

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

---

---

---

---

---

---

---

---

---

---

---

---

22

```

> mean(sample.var.based.Pop)
[1] 99.93689
> mean(sample.standard.var)
[1] 99.93232
> mean(sample.var.n.instead)
[1] 96.66124
    
```

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

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

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

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

---

---

---

---

---

---

---

---

---

---

---

---

23

## 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  $n$ ?

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

👍 But why?

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

👎

---

---

---

---

---

---

---

---

---

---

---

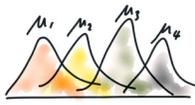
---

24



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)



ANOVA

$\mu_1 = \mu_2 = \mu_3 = \mu_4 ?$

28

---

---

---

---

---

---

---

---

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.genomenetwork.org/articles/08\\_02/bright\\_knees.shtml](http://www.genomenetwork.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 knees). 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 extraocular 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

29

---

---

---

---

---

---

---

---

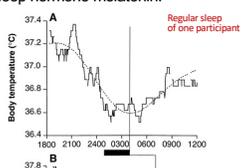
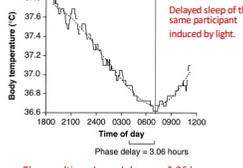
Our core body temperature is around 37°C but it fluctuates by about 1°C or so throughout the night.

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

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

30

---

---

---

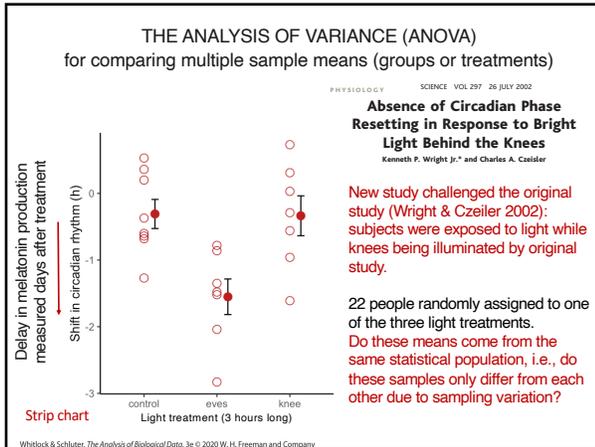
---

---

---

---

---




---

---

---

---

---

---

---

---

---

---

---

---

31

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

**H<sub>0</sub>:** The samples come from statistical populations with the same mean, i.e.,  $\mu_{\text{control}} = \mu_{\text{knee}} = \mu_{\text{eyes}}$ .

**H<sub>A</sub>:** At least two samples come from different statistical populations with different means.

---

---

---

---

---

---

---

---

---

---

---

---

32

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

**H<sub>0</sub>:** The samples come from statistical populations with the same mean, i.e.,  $\mu_{\text{control}} = \mu_{\text{knee}} = \mu_{\text{eyes}}$ .

**H<sub>A</sub>:** At least two samples come from different statistical populations with different means.

**Which is to say:**

**H<sub>0</sub>:** Differences in means among groups are due to **sampling error (sampling variation) from the same population.**

**H<sub>A</sub>:** 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.

---

---

---

---

---

---

---

---

---

---

---

---

33

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

---

---

---

---

---

---

---

---

34

**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

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

$$F_B = \frac{12275.0}{217.9} = 56.34$$

Note that scales (Y-axis) are now equal

---

---

---

---

---

---

---

---

35

**Verbal representation of equations**

Let's talk ANOVA "jargon"

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

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

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

---

---

---

---

---

---

---

---

36

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

Group Mean Square  $MS_{groups}$  (b=between or among)  $\sum_{i=1}^g n_i (\bar{X}_i - \bar{X})^2$

Mean of each group  $\bar{X}_i$

Total mean!  $\bar{X}$

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

$$F = \frac{S_b^2}{S_w^2} = \frac{\sum_{i=1}^g n_i (\bar{X}_i - \bar{X})^2}{g - 1} \div \frac{\sum_{i=1}^g (n_i - 1) s_i^2}{\sum_{i=1}^g (n_i - 1)}$$

Degrees of freedom of  $MS_{groups}$  is  $g - 1$

$MS_{errors}$  (w=within groups) Error Mean Square

37

---

---

---

---

---

---

---

---

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

Group Mean Square  $MS_{groups}$  (b=between or among)  $\sum_{i=1}^g n_i (\bar{X}_i - \bar{X})^2$

Mean of each group  $\bar{X}_i$

Total mean!  $\bar{X}$

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

$$F = \frac{S_b^2}{S_w^2} = \frac{\sum_{i=1}^g n_i (\bar{X}_i - \bar{X})^2}{g - 1} \div \frac{\sum_{i=1}^g (n_i - 1) s_i^2}{\sum_{i=1}^g (n_i - 1)}$$

Degrees of freedom of  $MS_{groups}$  is  $g - 1$

Variance of each group  $s_i^2$

Big "N"; sum of all sample sizes across groups  $\sum_{i=1}^g n_i$

Number of groups  $g$

Degrees of freedom of  $MS_{groups}$  is  $(N - g)$

Sample size of each group  $n_i$

38

---

---

---

---

---

---

---

---

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.

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

Note that scales (Y-axis) are now equal

39

---

---

---

---

---

---

---

---

**A small example: worth doing it "by hand"!**

Let's suppose two groups for simplicity!

**group 1**

1 2 3 4 5

  
 $\bar{X}_1 = 3.0$   
 $s_1^2 = 2.5$

---

---

---

---

---

---

---

---

40

g<sub>1</sub>: 1 2 3 4 5       $\bar{X}_1 = 3.0$      $\bar{X}_2 = 12.0$   
 g<sub>2</sub>: 10 11 12 13 14     $s_1^2 = 2.5$      $s_2^2 = 2.5$

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

MS<sub>groups</sub> = variance among group means (due to "treatment")

$$= (5 \times (3.0 - 7.5)^2 + 5 \times (12.0 - 7.5)^2) / (2-1) =$$

$$202.5 / (2-1) = \mathbf{202.5}$$

df(MS<sub>groups</sub>) = g - 1

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

Mean of each group    Total mean!

$$MS_{groups} = \frac{\sum_{i=1}^g n_i (\bar{X}_i - \bar{X})^2}{g-1}$$

Variance of each group

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

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

---

---

---

---

---

---

---

---

41

g<sub>1</sub>: 1 2 3 4 5  
 g<sub>2</sub>: 10 11 12 13 14

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

MS<sub>error</sub>       $\bar{X}_1 = 3.0$      $\bar{X}_2 = 12.0$   
 $s_1^2 = 2.5$      $s_2^2 = 2.5$

MS<sub>error</sub> = variance within groups (residuals)

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

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

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

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

---

---

---

---

---

---

---

---

42

$\bar{X} = (1+2+3+4+5+10+11+12+13+14)/10 = 7.5$   
 $MS_{groups} =$   
 $= (5 \times (3.0 - 7.5)^2 + 5 \times (12.0 - 7.5)^2) / (2-1) =$   
 $202.5 / (2-1) = 202.5$   
 $df(MS_{groups}) = g - 1 = 2 - 1$

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

$MS_{error} = \text{variance within groups (residuals)}$   
 $MSE_1 = (1-3.0)^2 + (2-3.0)^2 + (3-3.0)^2 + (4-3.0)^2 + (5-3.0)^2 = 10$   
 $MSE_2 = (10-12.0)^2 + (11-12.0)^2 + (12-12.0)^2 + (13-12.0)^2 + (14-12.0)^2 = 10$   
 $MS_{error} = (MSE_1 + MSE_2) / (N-g) = (10+10) / (10-2) = 20/8 = 2.5$   
 $df(MS_{error}) = N-g = 10 - 2 = 8$

---

---

---

---

---

---

---

---

43

Let's take a break - 1 minute




---

---

---

---

---

---

---

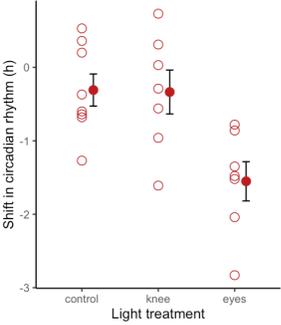
---

44

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




---

---

---

---

---

---

---

---

45

### “The knees who say night”

*Statistical Conclusion?*

**H<sub>0</sub>:** The samples come from the same population.  
**H<sub>A</sub>:** 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
---

```

---

---

---

---

---

---

---

---

46

### “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		

---

---

---

---

---

---

---

---

47

### 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**.

---

---

---

---

---

---

---

---

48

**“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<sub>0</sub>**: The samples come from the same population.

**H<sub>A</sub>**: At least two samples come from different populations.

**Reject H<sub>0</sub>**

How does the ANOVA significance test work?

---

---

---

---

---

---

---

---

---

---

---

---

49

How can we conceptualize the construction of the F distribution?

The statistical “machinery”:

- 1) Assume that **H<sub>0</sub>** 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.

---

---

---

---

---

---

---

---

---

---

---

---

50

**The F (sampling) distribution assuming that H<sub>0</sub> is true**

**H<sub>0</sub>**: Differences in means among groups are due to **sampling error from the same population**.

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

(8,7) observations

Sample from the same (normally distributed) population (i.e., assume that **H<sub>0</sub> is true**), respecting the original number of groups and their sample sizes.

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

---

---

---

---

---

---

---

---

---

---

---

---

51



**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.,  $\mu_{\text{control}} = \mu_{\text{knee}} = \mu_{\text{eyes}}$ .

$H_A$ : 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.

---

---

---

---

---

---

---

---

55

**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.*

---

---

---

---

---

---

---

---

56

**“The knees who say night”**

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

$H_A$ : at least one population mean ( $\mu$ ) 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!**

---

---

---

---

---

---

---

---

57