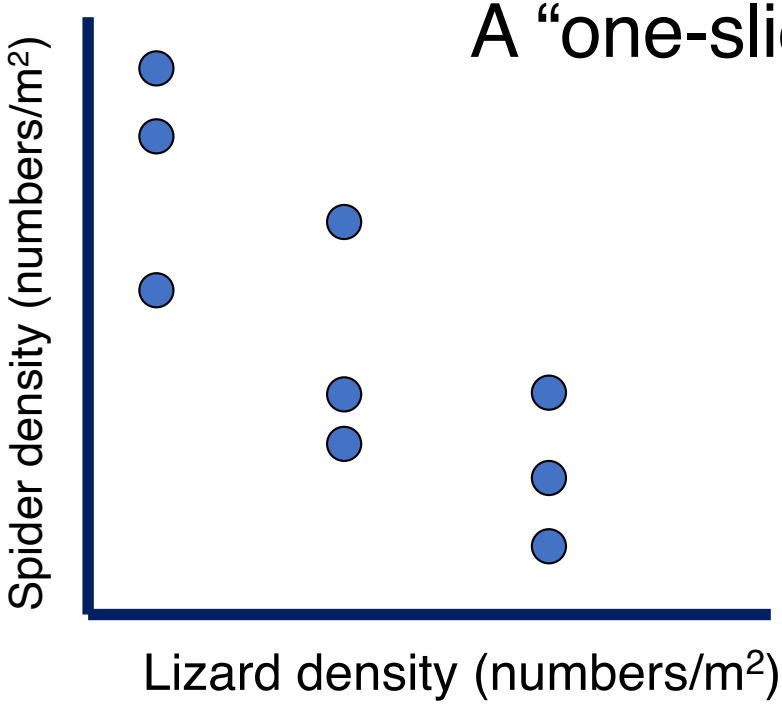
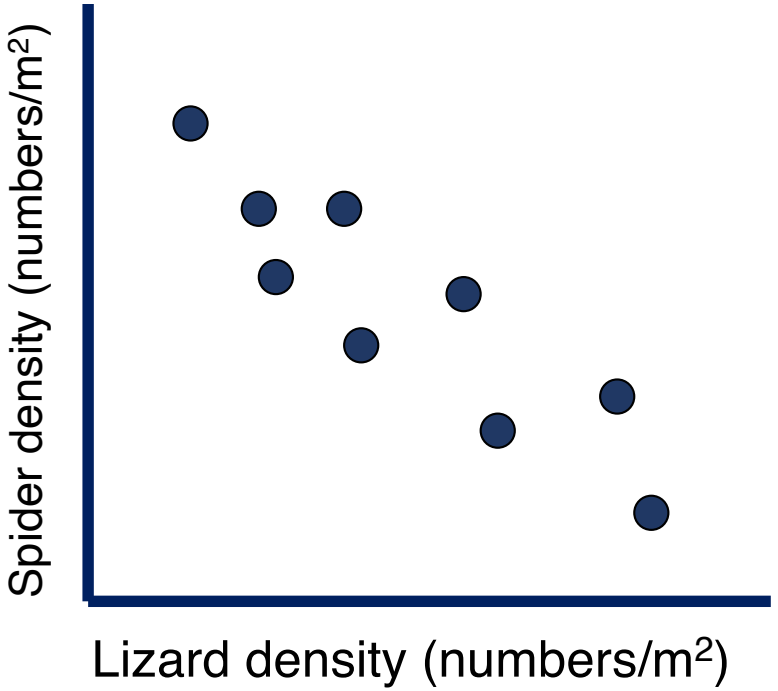
A “one-slide” discussion on experimental versus observational studies



Controlled (Experimental) study; often discrete (categorial) variation in the predictor variable, with no manipulation or assignment of treatments by the researcher.



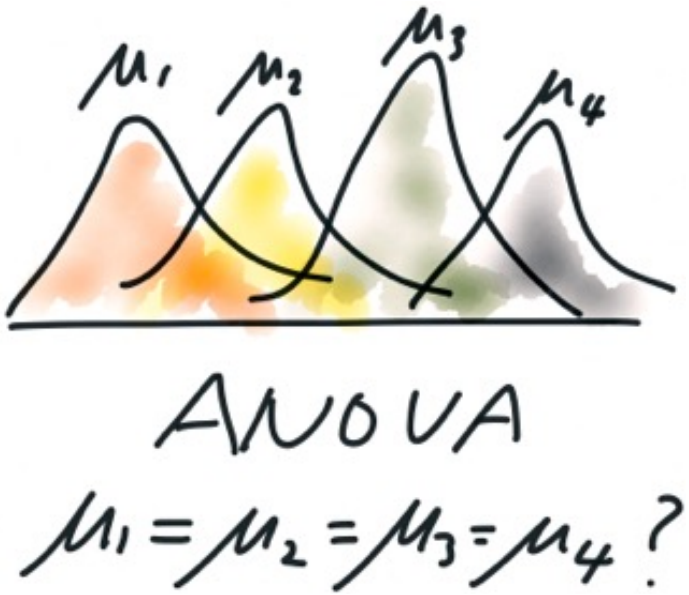
Uncontrolled (Observational) study; often continuous variation in the predictor variable, with no manipulation or assignment of treatments by the researcher.

Classes of statistical designs

Dependent Variable	Independent Variable	
	Continuous	Categorical
Continuous	Regression	t-tests and **ANOVA**
Categorical	Logistic Regression	Tabular

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)

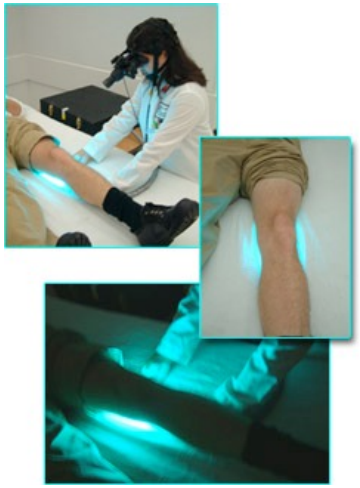


The problem about “The knees who say night”

By Whitlock and Schluter (2009)

A study once suggested that shining light on the back of the knees could reset the human circadian clock and potentially prevent jet lag, sparking excitement among scientists, entrepreneurs, and the public.

The idea challenged conventional understanding, which holds that biological clocks are reset only through light detected by the eyes.



Extraocular Circadian Phototransduction in Humans

Scott S. Campbell* and Patricia J. Murphy

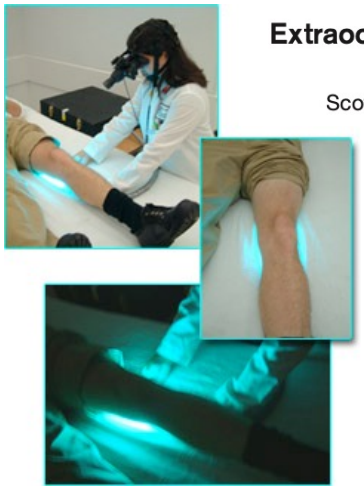
Physiological and behavioral rhythms are governed by an endogenous circadian clock. The response of the human circadian clock to extraocular light exposure was monitored by measurement of body temperature and melatonin concentrations throughout the circadian cycle before and after light pulses presented 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

“Bright light behind the knees is just bright light behind the knees”
http://www.genomenewsnetwork.org/articles/08_02/bright_knees.shtml

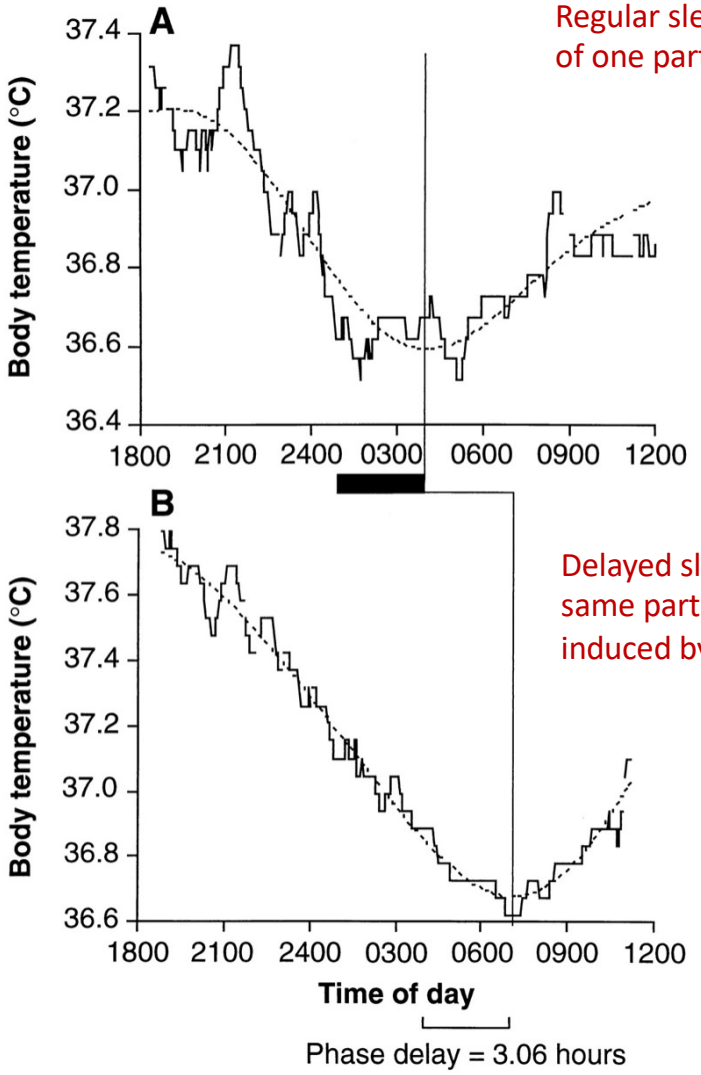
Our core body temperature is around 37ºC but it fluctuates by about 1ºC or so throughout the night.

The drop in temperature starts about two hours before we 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

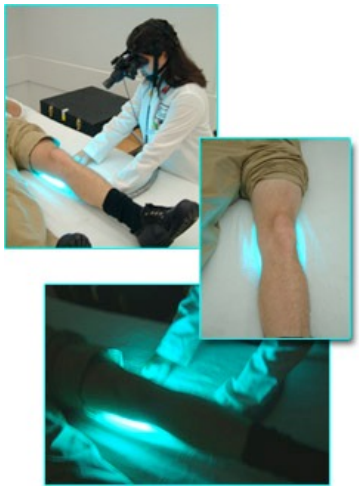
The resulting phase delay was 3.06 hours

“The knees who say night”

A study once suggested that shining light on the back of the knees could reset the human circadian clock and potentially prevent jet lag, sparking excitement among scientists, entrepreneurs, and the public.

The idea challenged conventional understanding, which holds that biological clocks are reset only through light detected by the eyes.

However, later research cast doubt on these findings, raising skepticism about both the mechanism and the potential for commercial light-therapy devices.

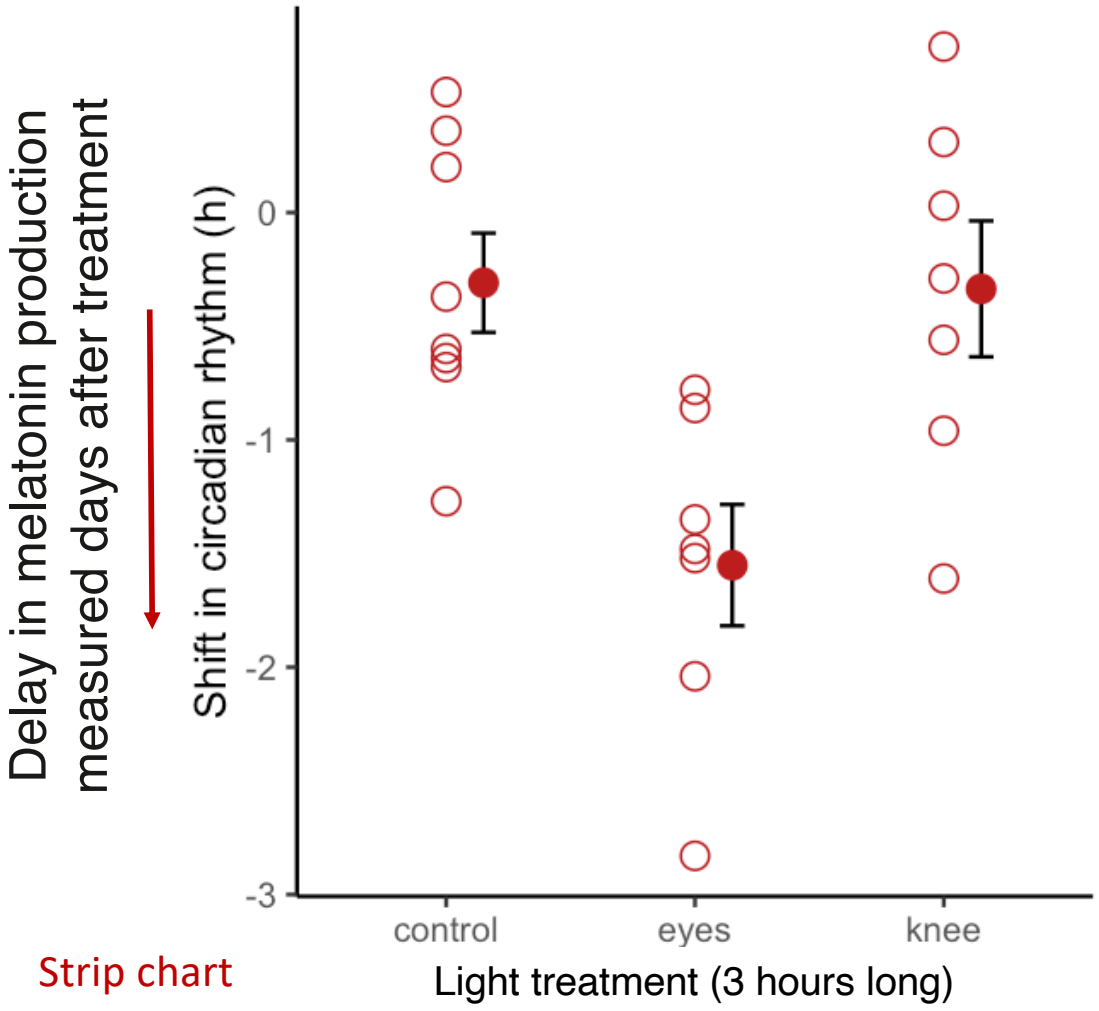


Wright and Czeisler (2002) pointed out that in the original study by Campbell and Murphy (1998), subjects may have been exposed to **light reaching their eyes indirectly**, even though the light was aimed at the back of the knees.

In their replication, they used **tighter experimental controls**, including carefully controlling light exposure to the eyes and improving measurement of circadian phase using melatonin.

Under these stricter conditions, they found **no effect of knee light on circadian rhythms**, suggesting the original results were likely due to **uncontrolled light exposure rather than a real physiological effect**.

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.

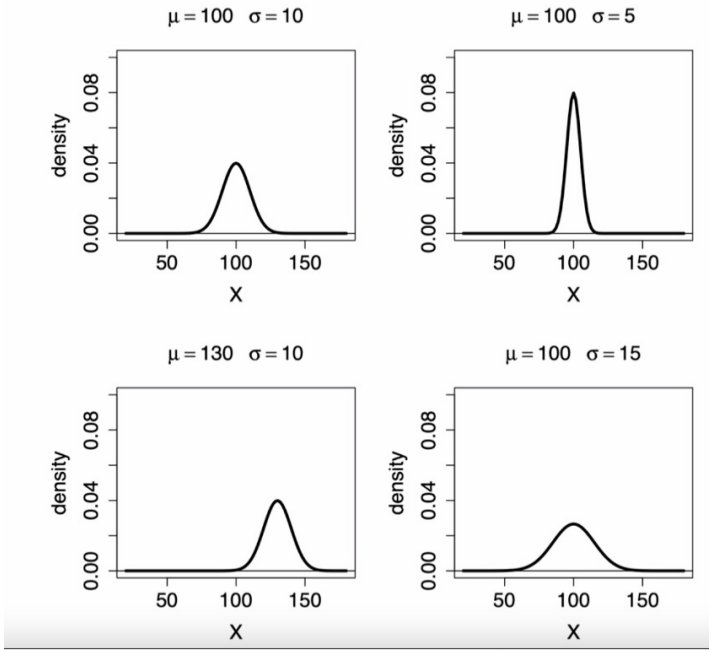
Do these means come from the same statistical population, i.e., do these samples only differ from each other due to sampling variation?

Control = wearable light devices similar but no light.

Do these samples originate from statistical populations with the same mean, or do the observed differences in means among samples arise solely from sampling variation within a single population?

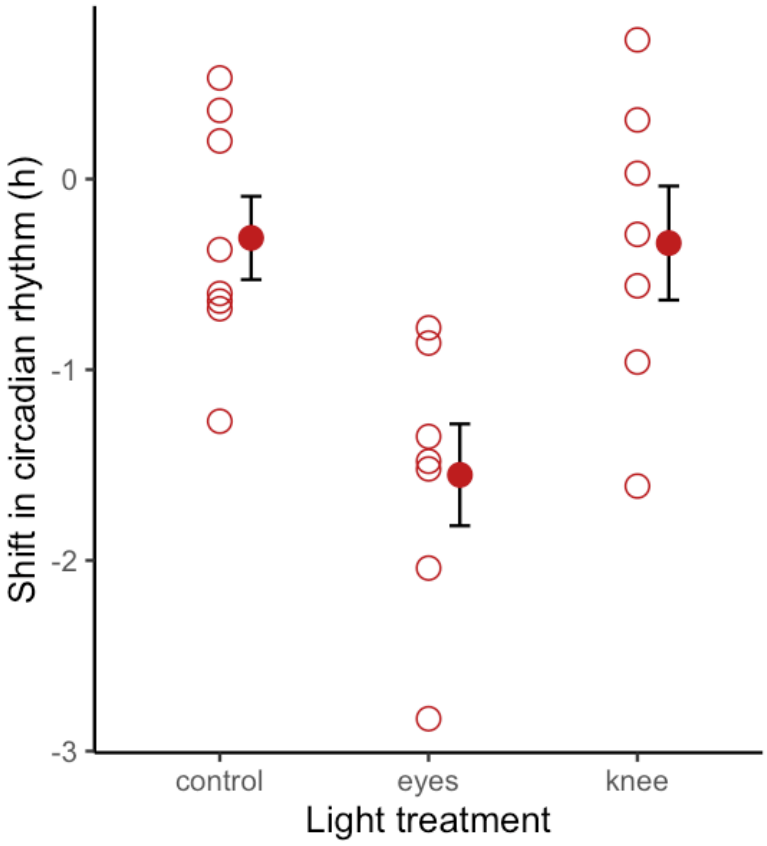
Keep in mind that samples may originate from statistical populations that share the same mean but are not necessarily from the same statistical population.

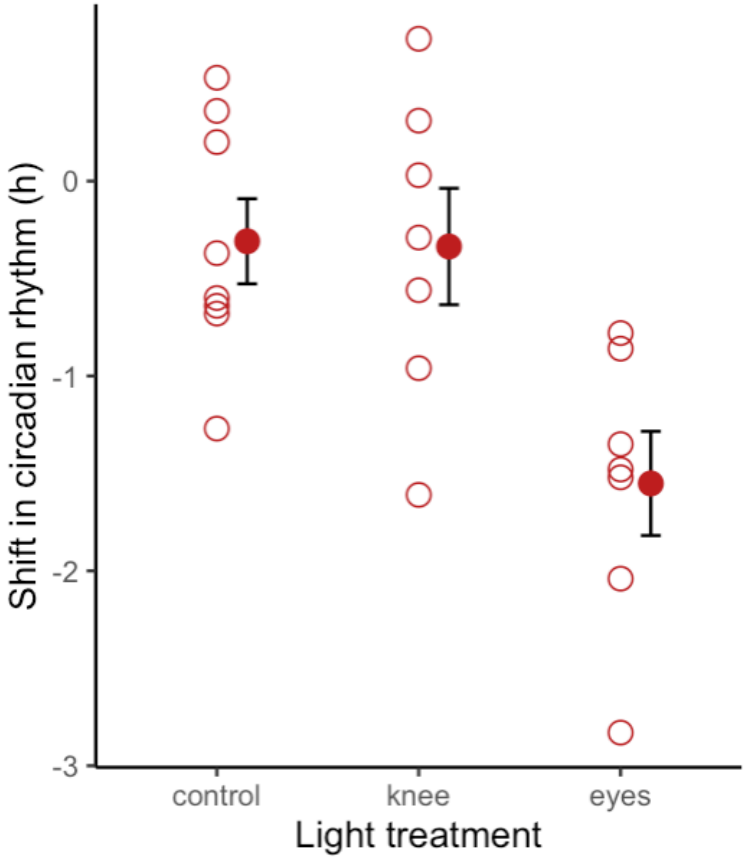
Why is this important? Statistical populations can have identical means but differ in their variances (and other aspects). This indicates that samples could be drawn from distinct populations, yet their means may not differ. This aligns with the null hypothesis (H_0), which assumes no difference in means between groups.

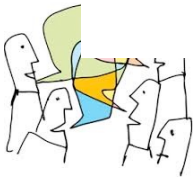


THE ANALYSIS OF VARIANCE (ANOVA)
for comparing multiple sample means (groups or treatments)

Which order of treatments work best?



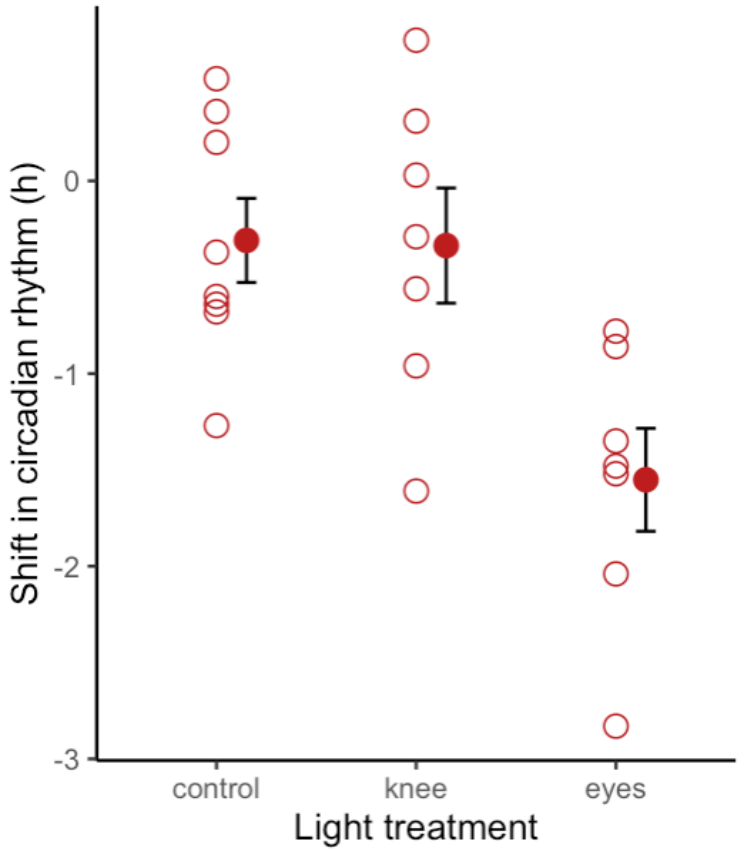




THE ANALYSIS OF VARIANCE (ANOVA)
for comparing multiple sample means (groups or treatments)

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

H_A: At least two samples originate from distinct statistical populations with differing means.



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

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

H_A: At least two samples originate from distinct statistical populations with differing means.

Which is to say:

H_0: Differences in group means are solely due to sampling variation from statistical populations that share the same mean.

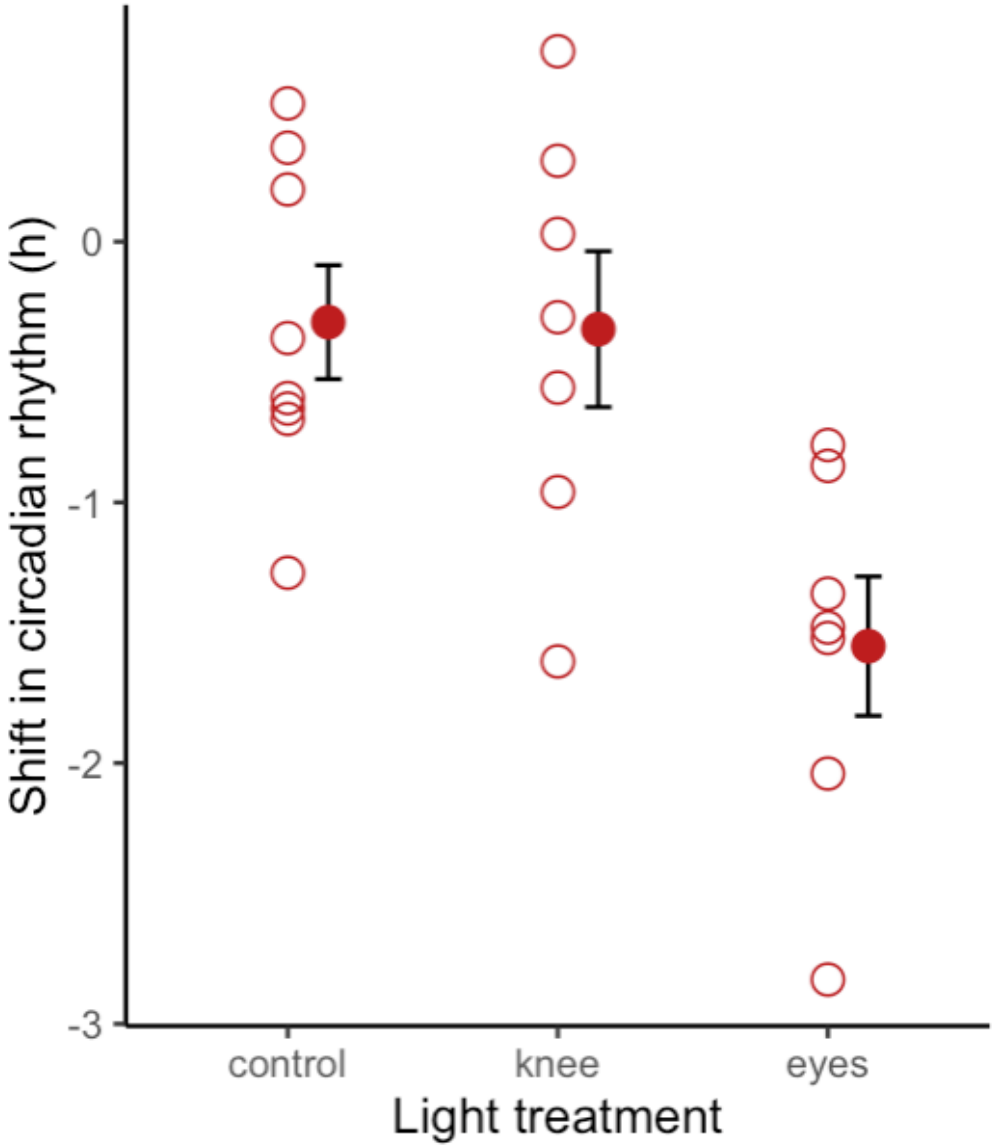
H_A: Differences in group means are not solely due to sampling variation, indicating the populations may differ in their means.

Remember: ***Sampling error*** is due to sampling variation, i.e., samples that come from the statistical populations sharing the same mean.

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

An ANOVA always involves one *continuous* response variable (e.g., shift in circadian rhythm) and **one** *categorical* predictor variable.

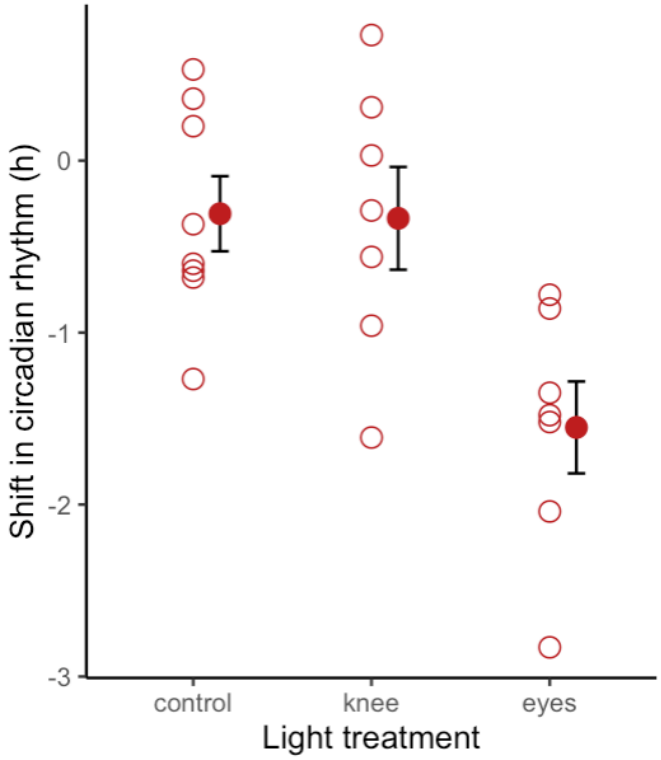
The categorical variable (predictor) is divided into groups which are often referred as treatments or factor levels.



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

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

H_A: At least two samples originate from distinct statistical populations with differing means.



As we are studying one single factor (light), we will use a ***one-way ANOVA***.

If two factors were involved (say light and time of experimentation) that would be a two-way ANOVA (not covered in BIOL322).

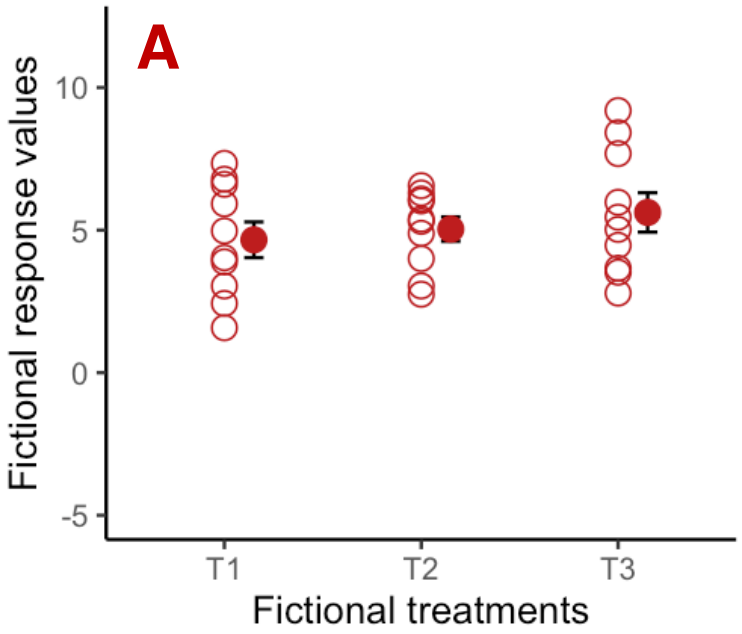
We require a test statistic that can detect variations in means across multiple groups.
The F-statistic achieves this by evaluating the ratio of two variance components.

$$F = \frac{\text{variance among group means (due to “treatment”)}}{\text{variance within groups}}$$

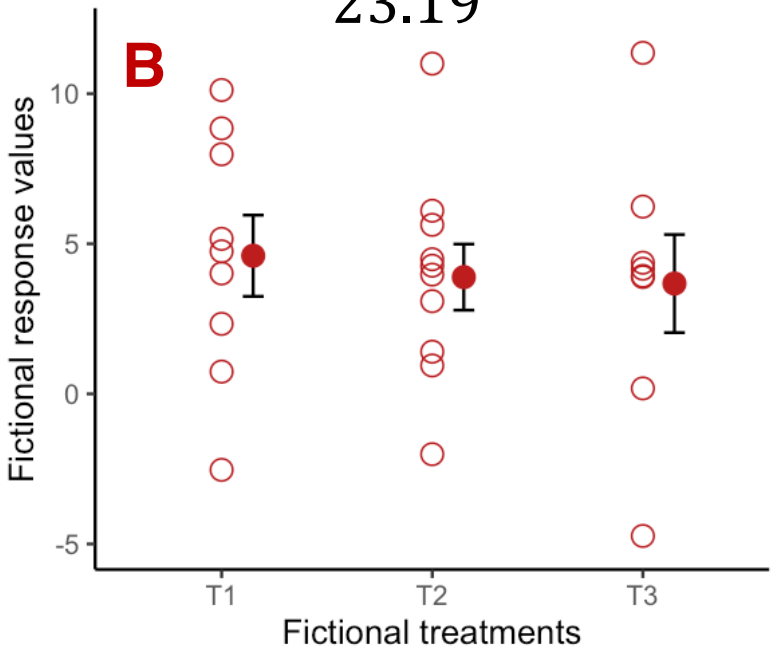
(referred to as error or residual variation, it represents the variation not explained by the differences in means among groups)

Means among groups don’t vary much in both data **A** and **B**,
but residual variation (within groups) is smaller in **A** than **B**.

$$F_A = \frac{2.34}{3.51} = 0.67$$



$$F_B = \frac{11.63}{23.19} = 0.50$$



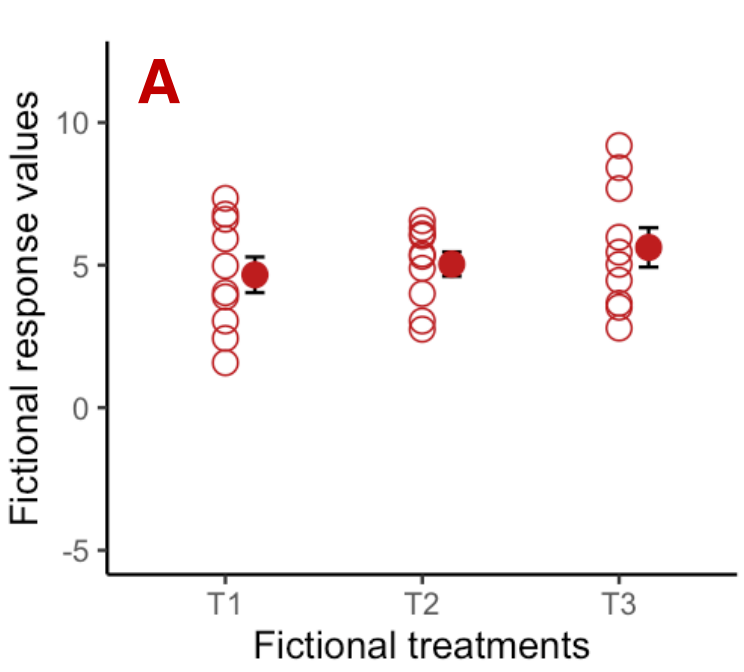
We require a test statistic that can detect variations in means across multiple groups.

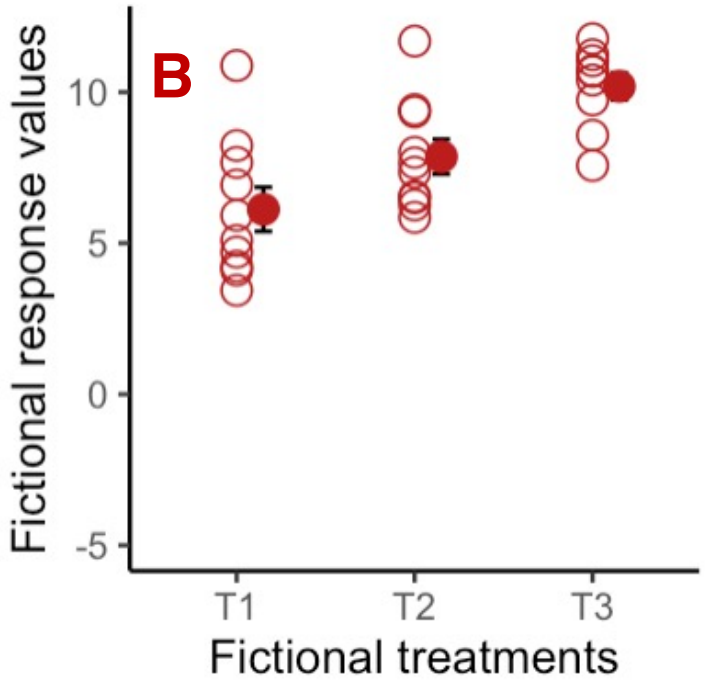
The F-statistic achieves this by evaluating the ratio of two variance components.

Means among groups don't vary in **A** but vary in **B**;
residuals variation is similar in **A** than **B**.

$$F_A = \frac{2.34}{3.51} = 0.67$$

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



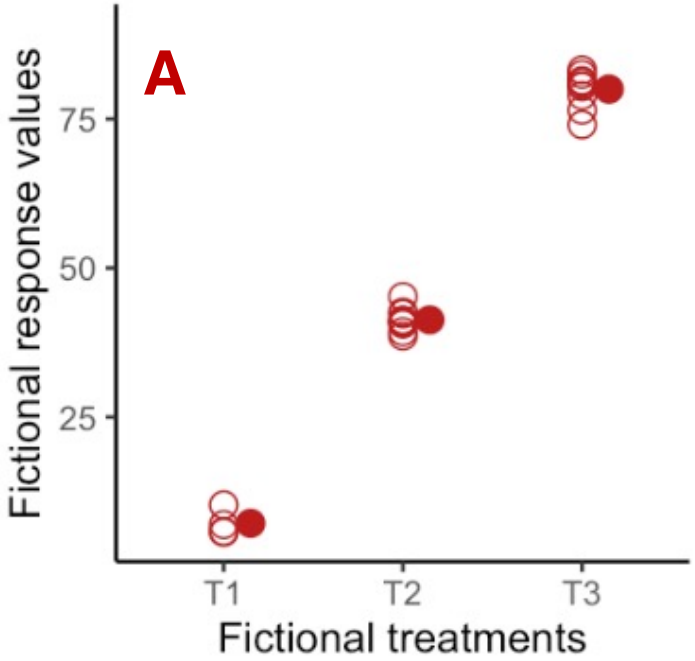


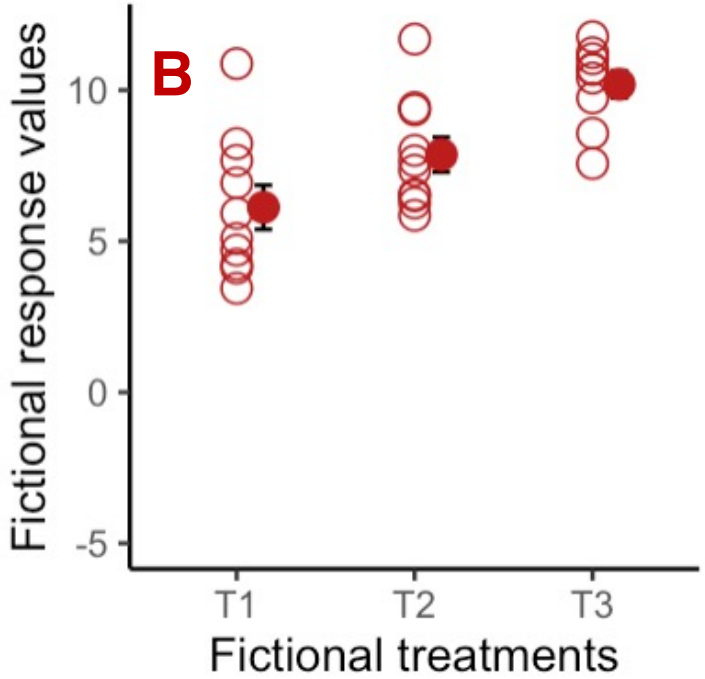
We require a test statistic that can detect variations in means across multiple groups: the F-statistic achieves this by evaluating the ratio of two variance components.

Mean differences among groups are much larger in **A** than **B**; residuals variation is similar in **A** than **B**. Notice the differences in their Y-scales (the mean differences among groups is huge in **A**).

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

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



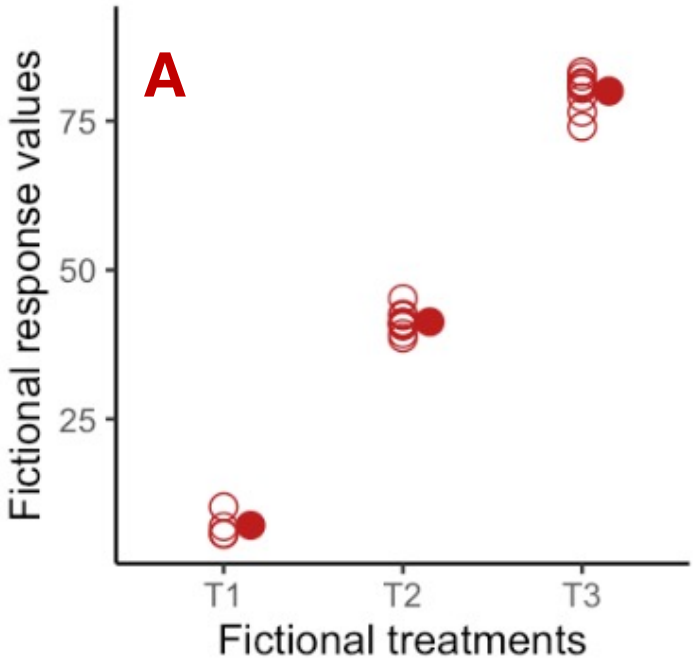


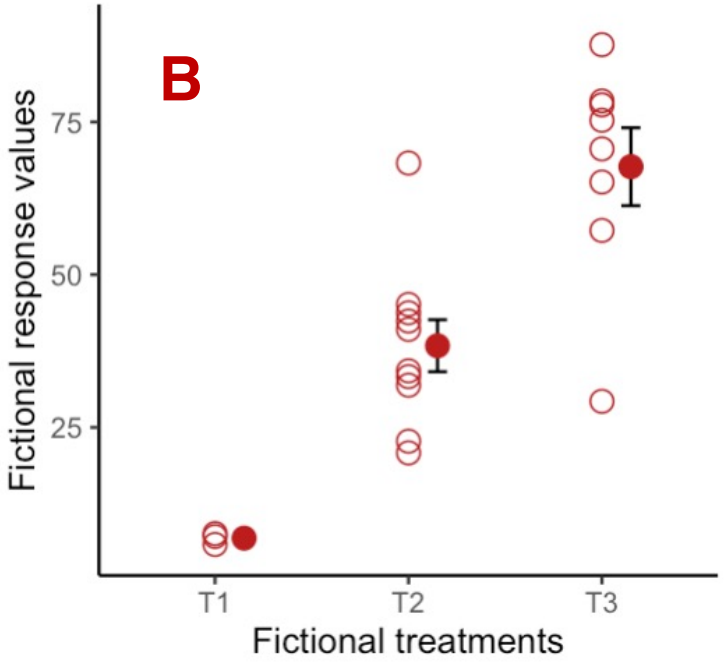
HETEROSCEDASTICITY reduces the F-ratio ability to differentiate among differences in means among groups

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

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

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





We require a test statistic that can detect variations in means across multiple groups: the F-statistic achieves this by evaluating the ratio of two variance components.

Let's talk ANOVA "jargon"

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

You can interpret ANOVA without knowing how it works, but you are less likely to use ANOVA inappropriately if you have some idea of how it works (*Motulsky*)

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):

Some 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 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 evaluates the variance among group means while accounting for the variance within groups.

The F-statistic plays a crucial role in the ANOVA context, making it well worth understanding 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

$= (N-g)$

Number of groups

Degrees of freedom of MS_{groups}

Sample size of each group

MS_{errors}

(w=within groups)

Error Mean Square

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

Let’s assume two groups for simplicity!

group 1

1	2	3	4	5

$\overline{X}_1 = 3.0$

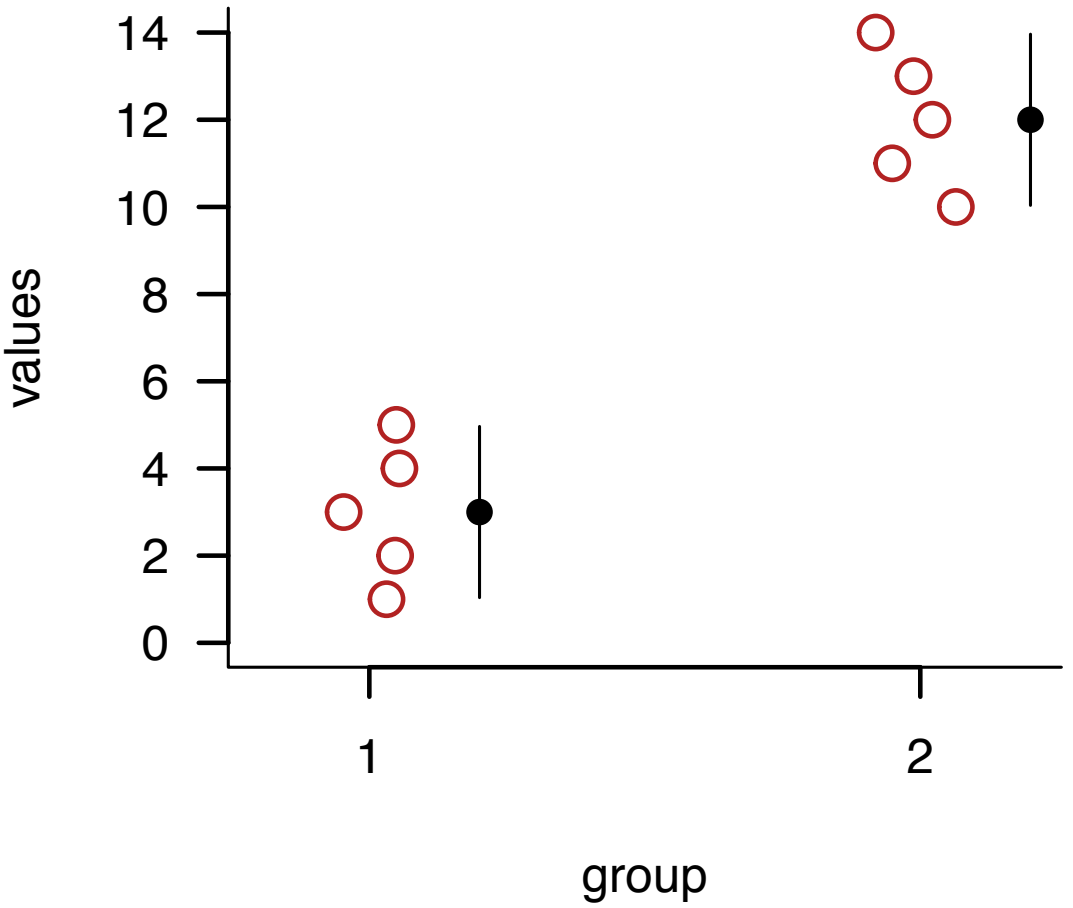
$s_1^2 = 2.5$

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**

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

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

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!

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

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}

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

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

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**

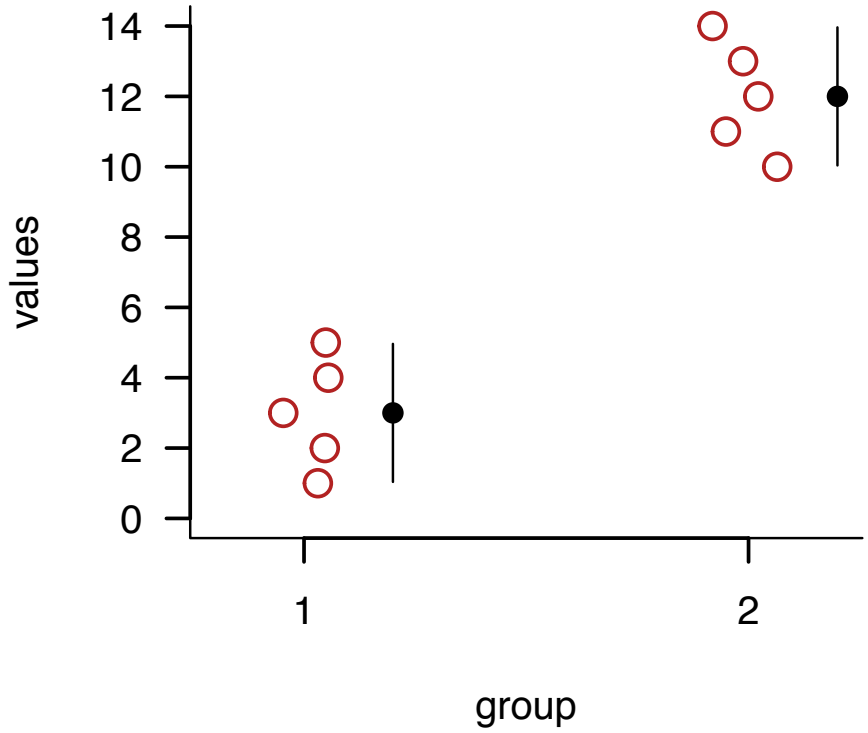
MSE_2= (10-12.0)^2 +(11-12.0)^2 +(12-12.0)^2 +(13-12.0)^2 +(14-12.0)^2 = **10**

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

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

one-way ANOVA in R - step 1: organizing data in a csv file

E	F
values	group
1	1
2	1
3	1
4	1
5	1
10	2
11	2
12	2
13	2
14	2

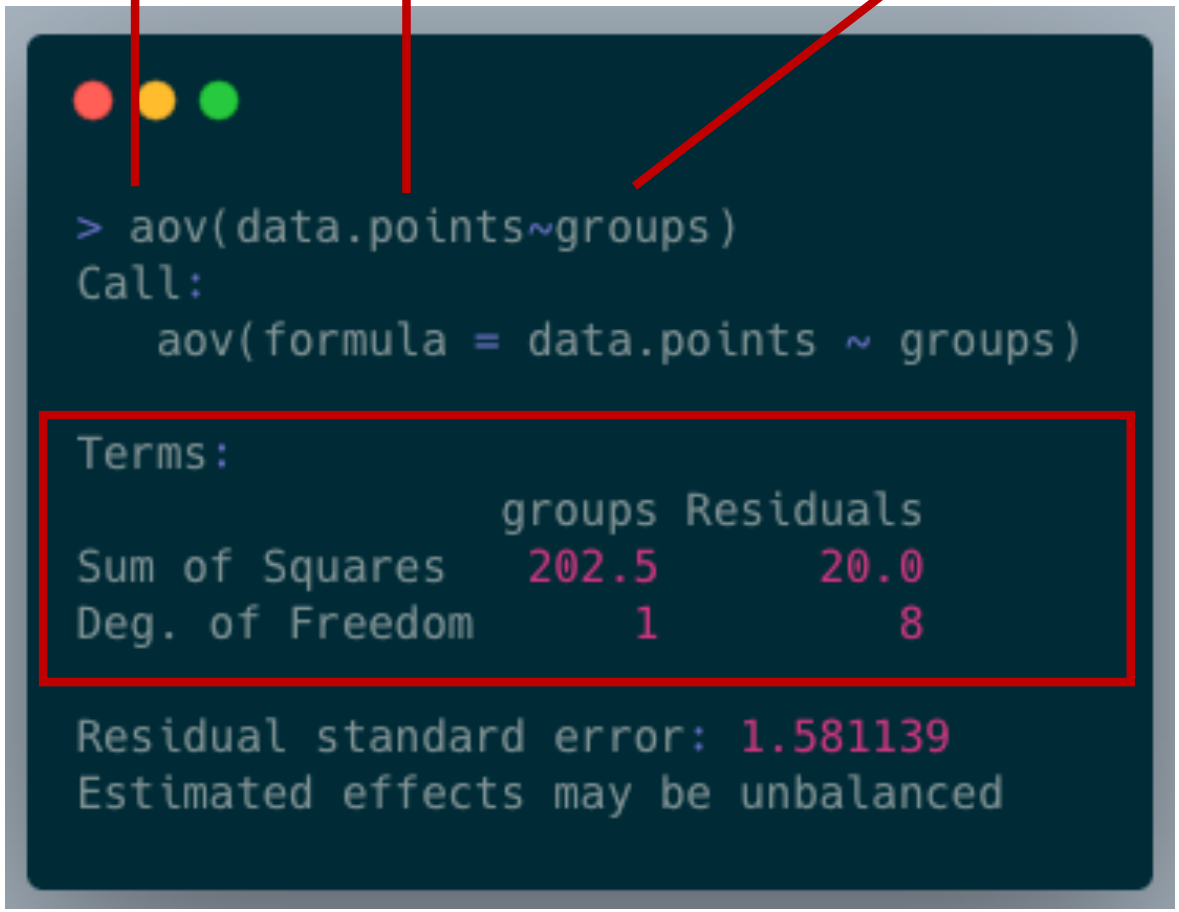


ANOVA in R

Function to
run **A**n **A**nalysis of **V**ariance (aov)

Vector of observations
(1,2,3,4,5,10,11,12,13,14)

Factor identifying group of
the observation
(1,1,1,1,1,2,2,2,2,2)



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

NOTE: ANOVA for two groups is equivalent to the two-sample mean t-test

```
> aov(data.points~groups)
Call:
   aov(formula = data.points ~ groups)

Terms:
                groups Residuals
Sum of Squares   202.5      20.0
Deg. of Freedom      1         8

Residual standard error: 1.581139
Estimated effects may be unbalanced
```

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

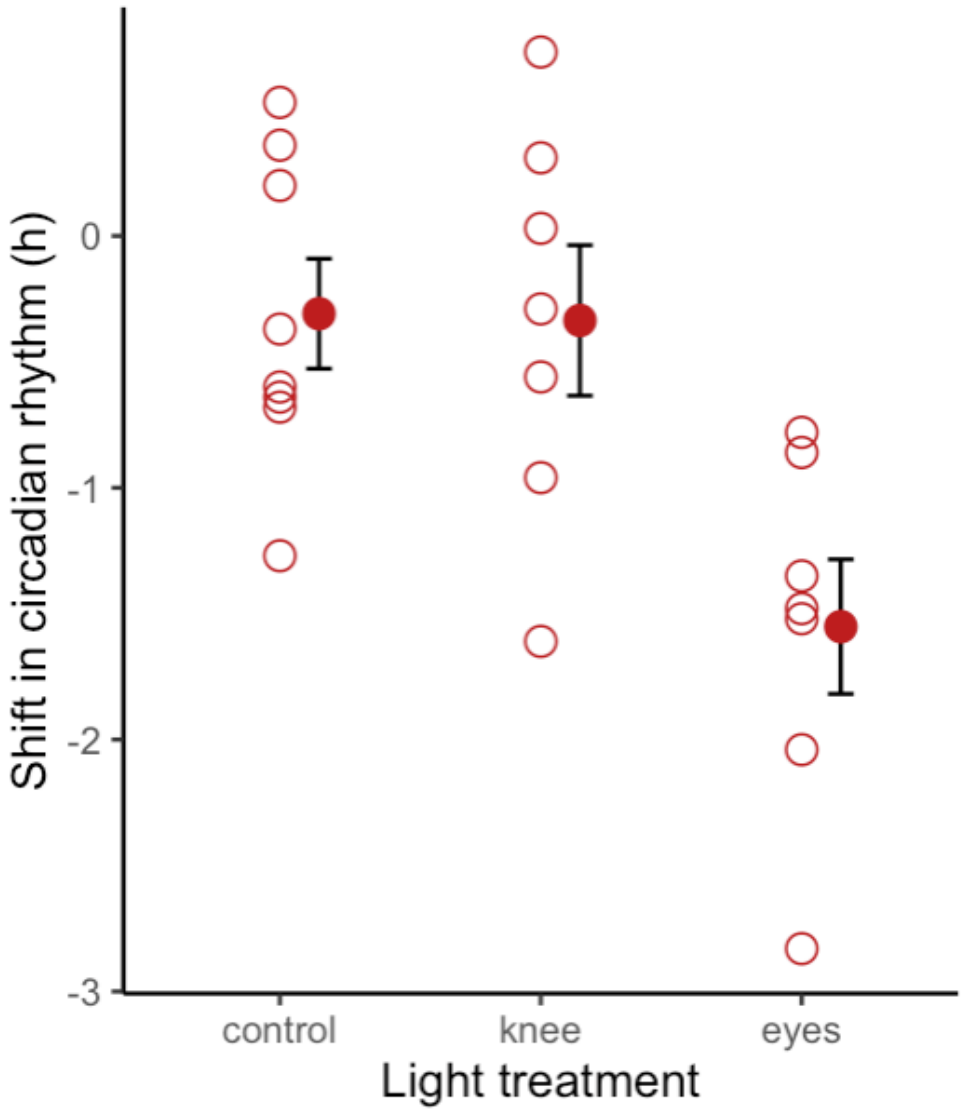
$$t = \sqrt{F} = \sqrt{\frac{\mathbf{202.5}}{\mathbf{2.5}}} = 9$$

```
t.test(data.points~groups,var.equal = TRUE)

    Two Sample t-test

data:  data.points by groups
t = -9, df = 8, p-value = 1.853e-05
```

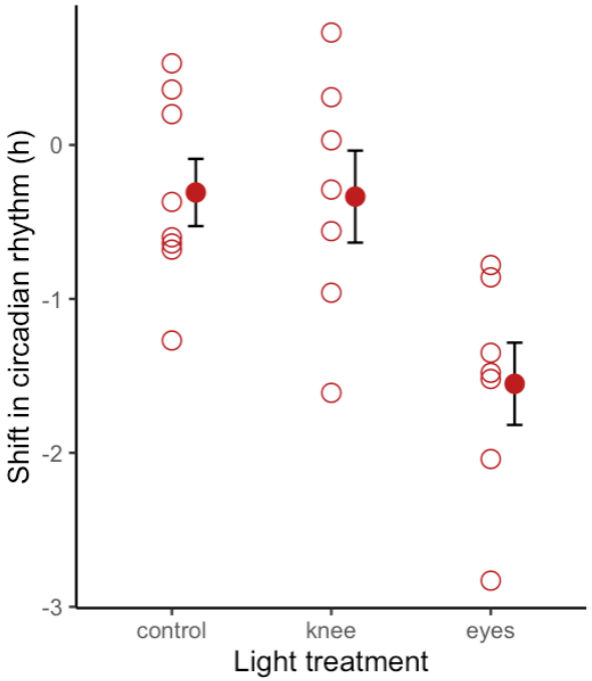
Back to the “The knees who say night”

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?

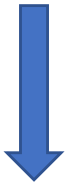
H_0: The samples originate from statistical populations that share the same mean, i.e., $\mu_{control} = \mu_{knee} = \mu_{eyes}$.

H_A: At least two samples originate from distinct statistical populations with differing means.

```
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		

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

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

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

From lecture 11

Computing the sum of squares requires subtracting the sample mean, which introduces bias because the mean is estimated from the same data.

To correct for this, the sum of squares is divided by the appropriate degrees of freedom for the numerator and denominator sum of squares, producing an unbiased estimate of variability.

“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 originate from statistical populations that share the same mean, i.e., $\mu_{control} = \mu_{knee} = \mu_{eyes}$.

H_A: At least two samples originate from distinct statistical populations with differing means.

Reject H_0

How does the ANOVA statistical significance test work?

How to think about the F distribution

The statistical “machinery”:

1) Assume that $\mathbf{H_0}$ is true (i.e., the samples originate from statistical populations that share the same mean and variance).

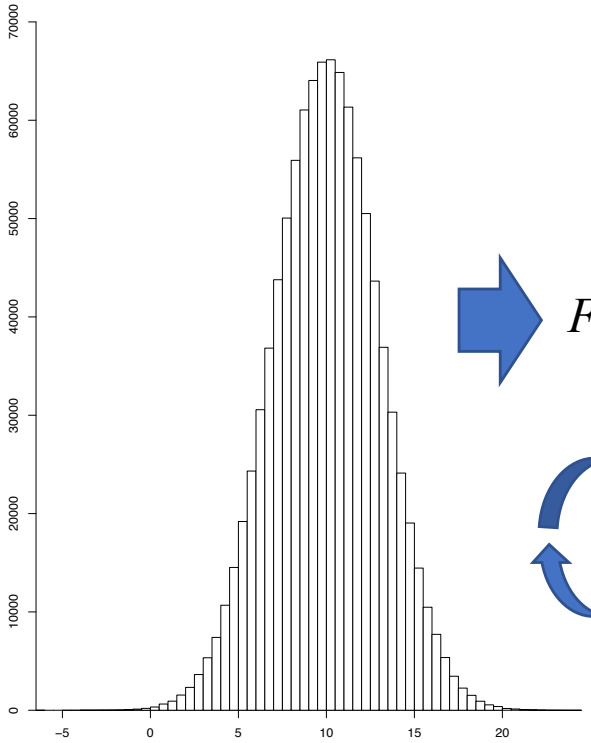
Why assume equal variances? Two populations can share the same mean yet differ in infinitely many ways through their variances. When we also assume normality and equal variances, the populations are effectively identical except for sampling variation. These assumptions simplify the mathematics and allow us to treat the samples as coming from the same statistical population under the null hypothesis.

2) Sample from the statistical 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.

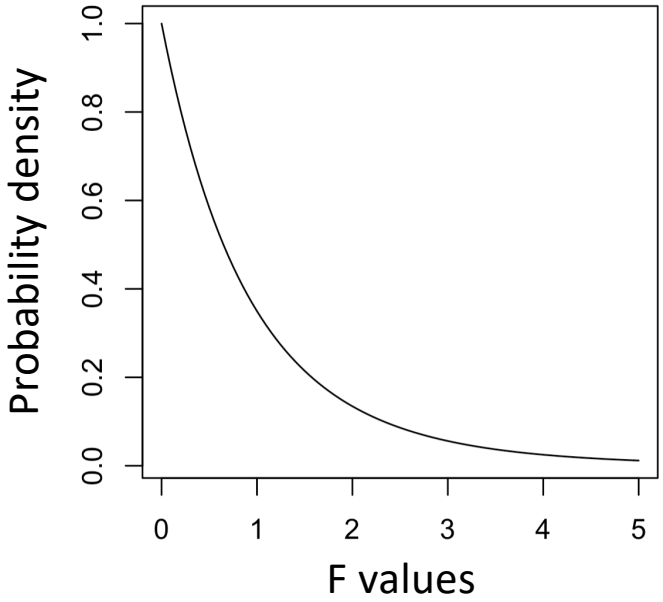
The F (sampling) distribution assuming that H_0 is true

H_0: Differences in group means are solely due to sampling variation from statistical populations that share the same mean.

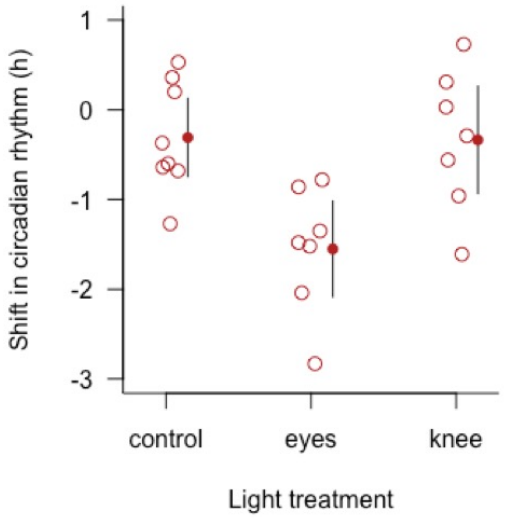


$$F = \frac{s_b^2}{s_w^2} = \frac{\dfrac{\sum_{i=1}^{g} n_i(\bar{X}_i - \bar{\bar{X}})^2}{g-1}}{\dfrac{\sum_{i=1}^{g}(n_i - 1)s_i^2}{\sum_{i=1}^{g}(n_i - 1)}}$$

(8,7,7) observations



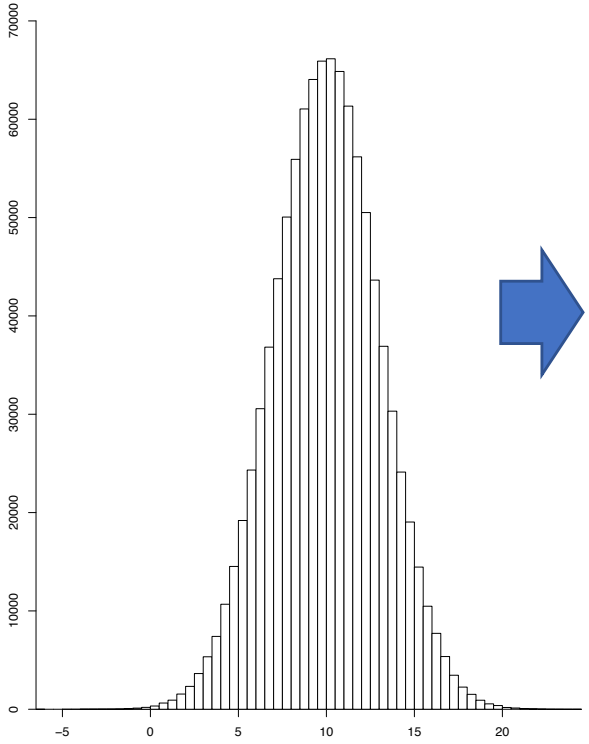
Sample from the same (normally distributed) population (i.e., ***H_0 is true***), respecting the original number of groups and their sample sizes.



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

The F (sampling) distribution assuming that H_0 is true

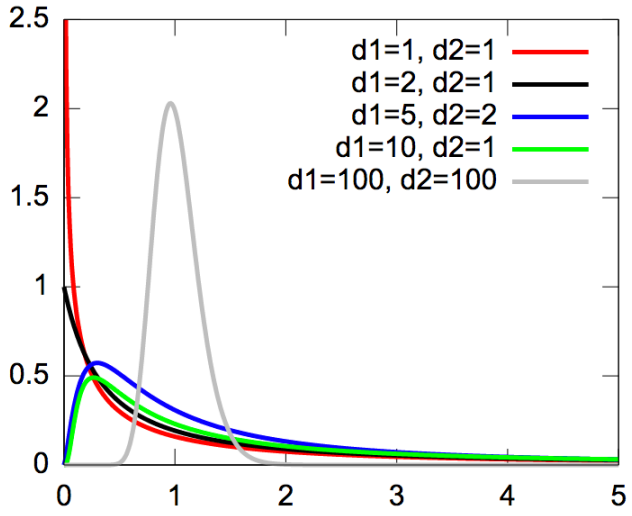
H_0: Differences in group means are solely due to sampling variation from statistical populations that share the same mean.



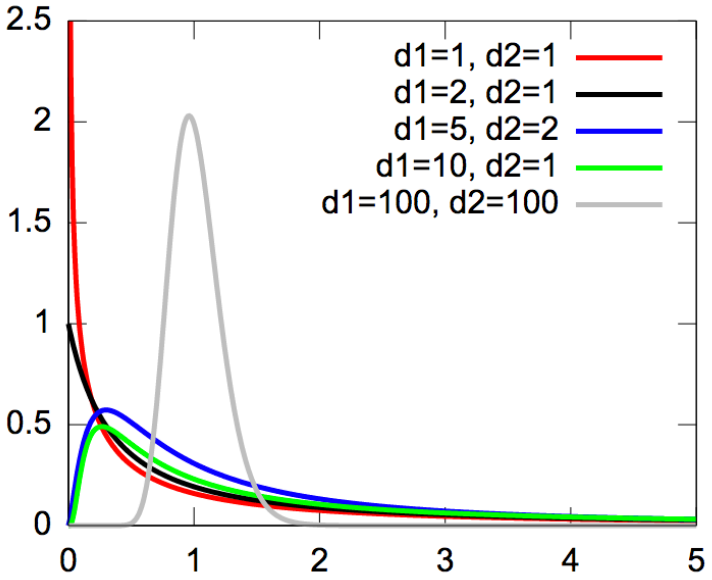
$$F = \frac{s_b^2}{s_w^2} = \frac{\dfrac{\sum_{i=1}^{g} n_i(\bar{X}_i - \bar{\bar{X}})^2}{g-1}}{\dfrac{\sum_{i=1}^{g}(n_i - 1)s_i^2}{\sum_{i=1}^{g}(n_i - 1)}}$$

Varying the number of groups and the number of observations per group results in 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

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

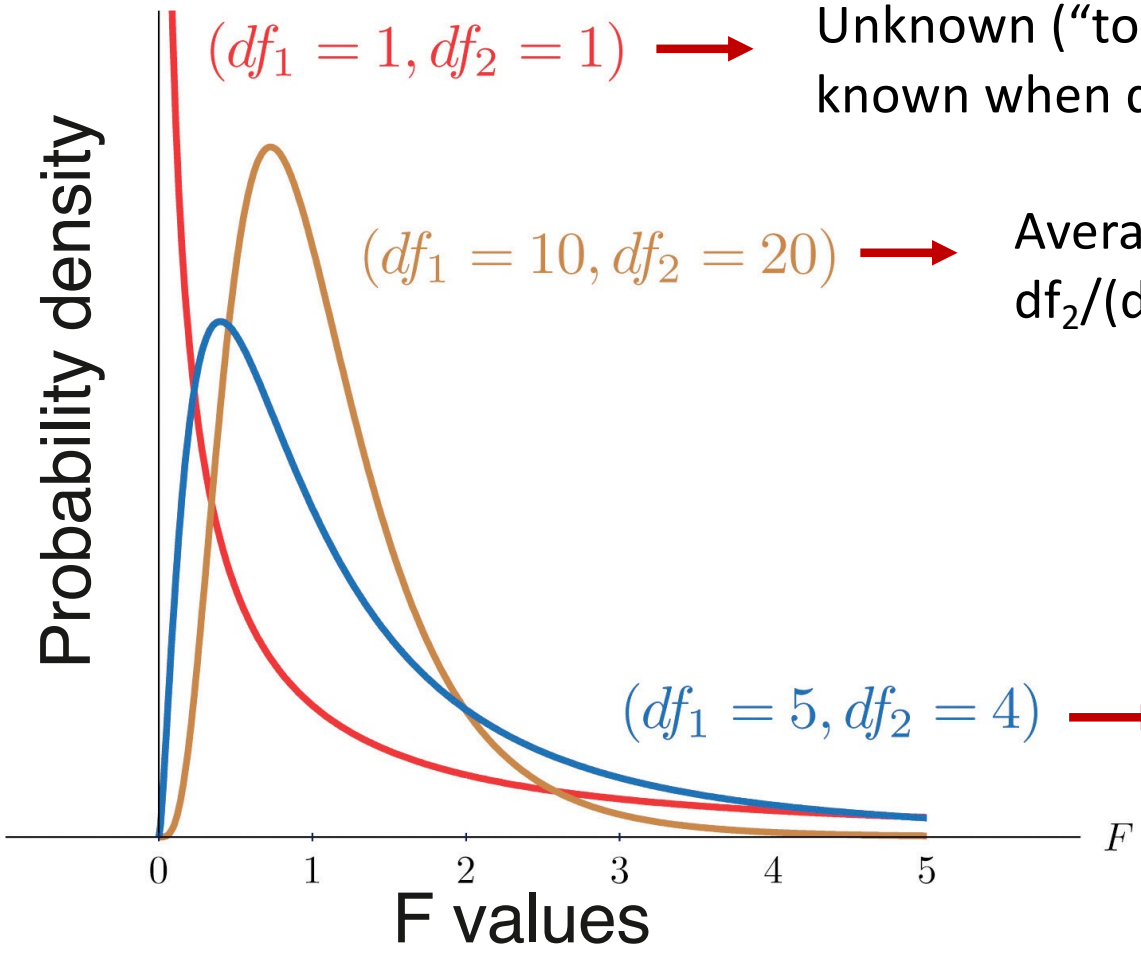
df_2

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

The F (sampling) distribution assuming that H_0 is true

Remember: The average of the t-distribution under **H_0** is always zero;

However, the average of the F-distribution expected under **H_0** depends on its degrees of freedom.



Unknown ("to me");
known when $df_2 > 2$

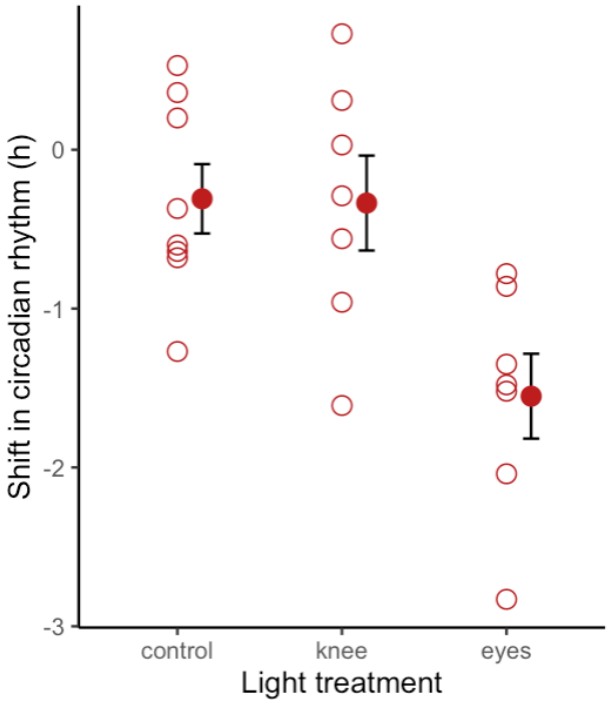
Average F value is
$df_2/(df_2-2) = 20/18 = 1.11$

Average F value is
$df_2/(df_2-2) = 4/2 = 2.00$

The expected mean value for the F distribution under the null hypothesis as a way to also express the null and alternative hypothesis

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

H_A: At least two samples originate from distinct statistical populations with differing means.



Which is to say:

H_0: $F = df_2/(df_2 - 2)$

H_A: $F \neq df_2/(df_2 - 2)$

```
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

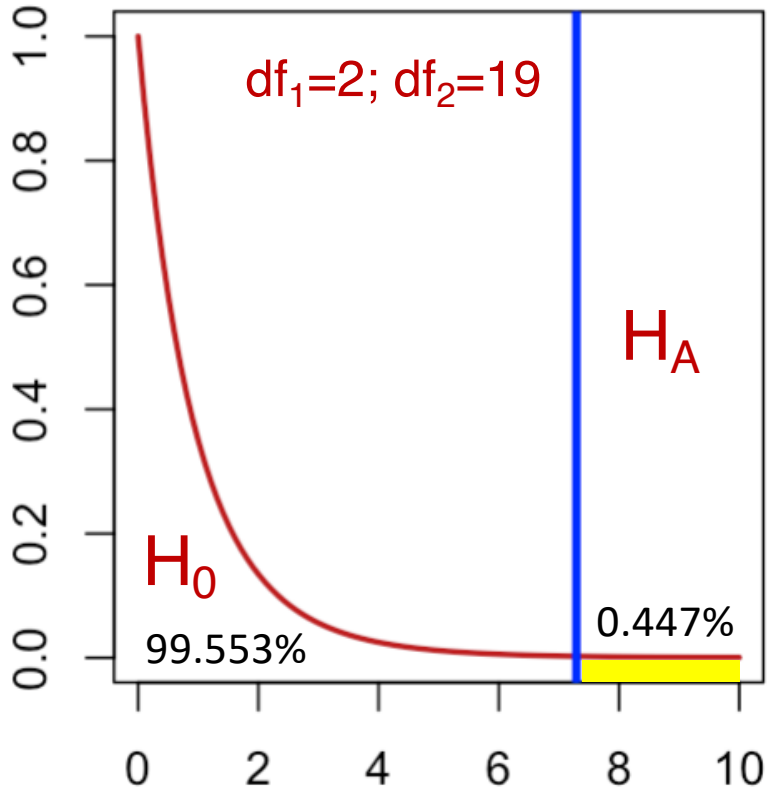
ANOVA is non-directional, but it focuses on large F values because large values indicate that group means differ more than expected by chance. Small F values, in contrast, suggest that the means are very similar.

Thus, although ANOVA tests a non-directional hypothesis, evidence against the null hypothesis appears only in the upper tail of the F-distribution.

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

H_A: At least two samples originate from distinct statistical populations with differing means.

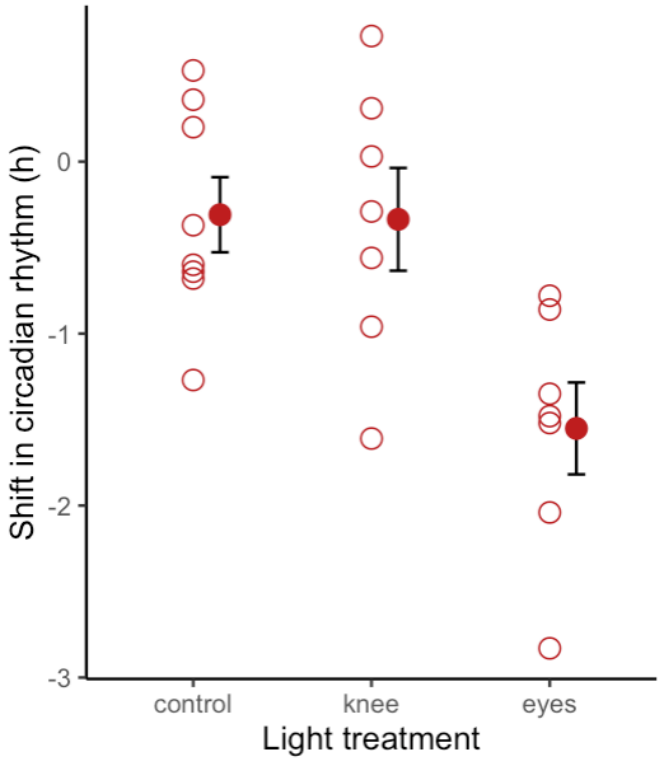
The p-value is estimated as the number of F-values in the null distribution equal or greater than the observed F-value (i.e., one tailed-test).



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

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

H_A: At least two samples originate from distinct statistical populations with differing means.



Research conclusion: light treatment influences shifts in circadian rhythm.

ANOVA assumptions

ANOVA assumptions (similar to the independent two-sample t-test):

Observations are independent random samples from their respective populations.

The response variable (e.g., shift in circadian rhythm) is approximately normally distributed within each treatment group (we will revisit this assumption later).

Variances are equal across groups (homoscedasticity). Unequal variances can distort F-values and lead to unreliable inference about differences among means (discussed in a later lecture).

Differences in variances among populations can be tested using Levene's test. Although the calculation is beyond the scope of BIOL 322, it is important to understand when to use it, what it tests, and how to implement it in R.

```
leveneTest(shift ~ factor(treatment), data=circadian)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  2  0.1586 0.8545
      19
```

P = 0.8545. Based on an alpha = 0.05, we should not reject the null hypothesis that:

$$\sigma^2_{control} = \sigma^2_{knee} = \sigma^2_{eye}$$

Therefore, we can feel confident using a standard ANOVA for these data. When the assumption of equal variances is not met, an alternative such as Welch's ANOVA can be used.

“The knees who say night”

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

H_A: At least two samples originate from distinct statistical populations with differing 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?

Lecture 18