

Pedro Peres-Neto, PhD @com_ecology · 3h

One thing that many find challenging when they start writing is building a good narrative structure. But exercising building narratives (in my opinion) is a big component in how researchers practice creativity & intuition. Bottom line, AI language models won't exercise for you 🤖

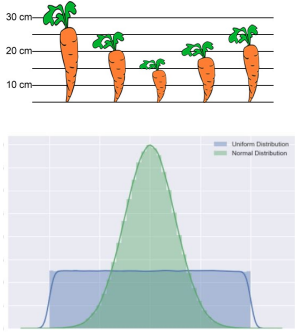
Federico Riva @riva_ecology · Jan 18

I truly believe that integrating ChatGPT in manuscript writing is a major mistake for science. Writing clearly requires clear ideas, and letting an algorithm fill the spaces based on what other people have written before us is the quickest way to stagnate in normal science.

[Show this thread](#)

1

The role of normality in biology - We often work with continuous variables that are assumed to be "normally" distributed



2

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

Confidence intervals and statistical hypothesis testing are frameworks based on sampling theory.

Here, sampling theory relates to repeated sample to model (derive) the expectations (probabilities of sample values) under sampling variation for statistical populations.

Repeated sample is used to derive the sampling distribution used in confidence intervals and statistical hypothesis testing (lecture 3).

BUT: Repeated sampling is only possible making certain assumptions about the statistical population.

3

Why sampling properties of estimators are important?

The mean of all possible sample means (i.e., sampling distribution) ALWAYS equals the population mean regardless of the original distribution of the population. As such, the sample mean is an unbiased (“honest”) estimator of the true population; i.e., in average the arithmetic mean equals the true population mean value (parameter).

4

PROPERTY OF THE MEAN AS AN ESTIMATOR: The mean of all possible sample means (i.e., sampling distribution) ALWAYS equals the population mean regardless of the original distribution of the population – the case of a tiny uniform distribution

1,2,3,4,5; population mean=3.0

All possible 15 samples (with replacement) and their means for n=2:

(1,1) = 1.0	(1,2) = 1.5	(2,3) = 2.5	(3,4) = 3.5	(4,5) = 4.5
(2,2) = 2.0	(1,3) = 2.0	(2,4) = 3.0	(3,5) = 4.0	
(3,3) = 3.0	(1,4) = 2.5	(2,5) = 3.5		
(4,4) = 4.0	(1,5) = 3.0			
(5,5) = 5.0				

Notice that permutations, i.e., (1,2) = (2,1) are not shown but should be considered

The mean of all sample means is always equal to the population mean

$$\frac{(1.0 + 2.0 + 3.0 + 4.0 + 5.0 + 1.5 + 2.0 + 2.5 + 3.0 + 2.5 + 3.0 + 3.5 + 3.5 + 4.0 + 4.5)}{15} = 3.0$$

6 sample means smaller than the true population value [in red]
 6 sample means greater than the true population value [in green]
 3 sample means equal to the true population value [in black]

5

Why sampling properties of estimators are important?

Even though the mean of all possible variances is equal to the variance of normally distributed populations (and also for many non-normally distributed populations, i.e., **robust against normality**), the sampling properties of confidence intervals and statistical hypothesis testing may not hold when populations are not normally distributed.

For instance, a 95% confidence interval may end up being in reality smaller (e.g., 93%) or larger (e.g., 97%) if the population is quite different from normal. And statistical hypothesis testing may have type I errors that are not equal to alpha (as is the case normally or closely to normally distributed populations).

We covered these issues in BIOL322 and will revisit them later on in the course in respect to advanced methods.

6

Why sampling properties of estimators are important?

Again, the sample variance is often a robust estimator for the true population variance for non-normally distributed populations. In other words, the mean of all sample standard variance will be often very close to the true population variance for non-normally distributed populations.

However, given that we don't know when this is the case, commonly, statistical procedures based on the standard deviation (e.g., **t-test**, **ANOVA**, **regression**) "assume" normality.

Normality is needed to make sure that estimates (from samples; e.g., t value, F value) can be properly contrasted with the sampling distribution that was assumed to be true (theoretical) and that P-values are then properly estimated.

7

Despite these very detailed characteristics, 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).

The normal distribution is a model that needed to be used to build sampling distributions.

One way to be normal, but infinite ways to be any other type of distribution; that said, the normal distribution approximates many biological distributions!

And remember that sample means and variances (key statistical estimators) are robust against normality so it works well for populations that are slightly "non-normal" (i.e., approximately normal).

8

Why is the mean an unbiased estimator?

Because the mean of all possible possible sample means equals the population mean (parameter) only when the population is normally distributed.

Because the mean of all possible possible sample means equals the population mean (parameter) regardless whether the population is normally distributed or not.

Because the mean of all possible possible sample means does not equal the population mean (parameter).

9

Why is the mean always an unbiased estimator?

Because the mean of all possible possible sample means equals the population mean (parameter) only when the population is normally distributed.

Because the mean of all possible possible sample means equals the population mean (parameter) regardless whether the population is normally distributed or not.

Because the mean of all possible possible sample means does not equal the population mean (parameter).

10

Why is the variance not always an unbiased estimator?

Because the variance of all possible possible sample variance equals the population variance (parameter) when the population is normally distributed.

We can't guarantee this property for highly non-normal distributions.

11


The road of statistics: avoid bias when estimating population parameters from sample values - the role of degrees of freedom!




12

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

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

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


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


13

Let's use a computational approach to understand the performance of these two estimators for the population variance:

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

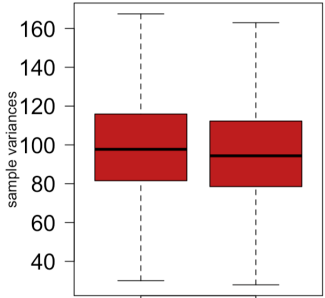
```

samples <- replicate(1000000, rnorm(n=30, mean=350, sd=10))
var.based.popMean <- function(x, mu) {sum((x-mu)^2/(length(x)))}
var.based.n <- function(x){sum((x-mean(x))^2)/(length(x))}
sample.var.based.Pop <- apply(X=samples, MARGIN=2, FUN=var.based.popMean, mu=350)
sample.var.n.instead <- apply(X=samples, MARGIN=2, FUN=var.based.n)
boxplot(sample.var.based.Pop, sample.var.n.instead,
        outline=FALSE, col="firebrick",
        cex.axis=1.5, las=1, ylab="sample variances")
    
```

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

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

14



```

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

The mean of s^2 for the estimator based on the population mean μ divided by n was unbiased (i.e., pretty much the population σ^2 ; would had been exactly $\sigma^2 = 100$ with infinite sampling); whereas the estimator based on the sample Y divided by n was biased.

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

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

15

Note the asymmetry of the sampling distribution of variances; hence the median is not exactly equal to the mean. But the variance is unbiased when based on μ 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
    
```

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

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

16

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

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

👍

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

👎

17

There is a correction factor for the sample bias in s^2 called Bessel's correction (but seems that Gauss 1823 came up with it first; <https://mathworld.wolfram.com/BesselsCorrection.html>)

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

👍

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

👎

18

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")
    
```

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

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

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

19

```

> 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!

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

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

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

20

```

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

Note though that:

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

is slightly more precise than:

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

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

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

21

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?

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

👎

22

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

Proof of Bessel's Correction

Bessel's correction is the division of the sample variance by $n - 1$ rather than n . Each day, though, a math proof that this correction results in an unbiased estimator of the population variance.

Example: If 100 people are asked, "How many children do you have?" and the responses are $0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100$, then the sample variance s^2 is calculated as follows: $s^2 = \frac{1}{100} \sum_{i=1}^{100} (x_i - \bar{x})^2$.

When the formula above is used, the sample variance s^2 is calculated as follows: $s^2 = \frac{1}{100} \sum_{i=1}^{100} (x_i - \bar{x})^2$.

The proof involves a straightforward calculation to show that the estimator in Equation 3 is unbiased. The proof involves a straightforward calculation to show that the estimator in Equation 3 is unbiased. The proof involves a straightforward calculation to show that the estimator in Equation 3 is unbiased.

Let's begin:

Proof

Let's prove that the following estimator for the population variance is unbiased:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \tag{1}$$

We'll have the expectation of the estimator and substitute:

$$E\left[\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2\right] = \frac{1}{n-1} \sum_{i=1}^n E[(x_i - \bar{x})^2]$$

$$= \frac{1}{n-1} \sum_{i=1}^n \left[E[x_i^2] - 2\bar{x}E[x_i] + \bar{x}^2 \right]$$

$$= \frac{1}{n-1} \left[\sum_{i=1}^n E[x_i^2] - 2\bar{x} \sum_{i=1}^n E[x_i] + n\bar{x}^2 \right]$$

$$= \frac{1}{n-1} \left[\sum_{i=1}^n E[x_i^2] - 2\bar{x} \sum_{i=1}^n \mu + n\bar{x}^2 \right]$$

$$= \frac{1}{n-1} \left[\sum_{i=1}^n E[x_i^2] - 2n\bar{x}\mu + n\bar{x}^2 \right]$$

$$= \frac{1}{n-1} \left[\sum_{i=1}^n E[x_i^2] - 2n\bar{x}\mu + n\bar{x}^2 \right]$$

Note that this is valid because:

$$E\left[\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2\right] = \frac{1}{n-1} \sum_{i=1}^n E[(x_i - \bar{x})^2]$$

with eq 1 holds because the variance of x_i is:

Note that since x_i is an i.i.d. random variable, any of the x_i , $i = 1, 2, \dots, n$, has the same variance. Furthermore, recall that for any random variable Z :

$$\text{Var}(Z) = E(Z^2) - (E(Z))^2 \quad \text{and} \quad E(Z^2) = \text{Var}(Z) + (E(Z))^2$$

So we can write:

$$E[x_i^2] = \text{Var}(x_i) + (E[x_i])^2$$

$$= \sigma^2 + \mu^2$$

$$E[x_i] = \text{Var}(x_i) + (E[x_i])^2$$

$$= \sigma^2 + \mu^2$$

Thus, we have:

$$\text{Var}(x_i) = \sigma^2 = E[x_i^2] - (E[x_i])^2$$

$$= \sigma^2 = E[x_i^2] - \mu^2$$

Finally, let's put everything together:

$$E\left[\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2\right] = \frac{1}{n-1} \sum_{i=1}^n (E[x_i^2] - \mu^2)$$

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

$$= \frac{1}{n-1} \sum_{i=1}^n \sigma^2$$

$$= \frac{1}{n-1} n \sigma^2$$

$$= \sigma^2$$

What we have shown is that our estimator is off by a constant, $(1 - \frac{1}{n}) = \frac{n-1}{n}$. If we want an unbiased estimator, we should multiply both sides of Equation 3 by the inverse of the constant:

$$E\left[\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2\right] = \frac{1}{n-1} \sum_{i=1}^n (\sigma^2 + \mu^2 - \mu^2) = \sigma^2$$

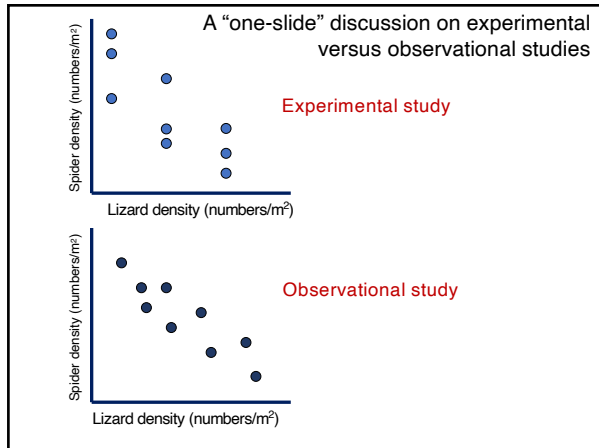
And this new estimator is exactly what we wanted to prove. Bessel's correction results in an unbiased estimator for the population variance.

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

23



24



25

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

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

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

ANOVA

26

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

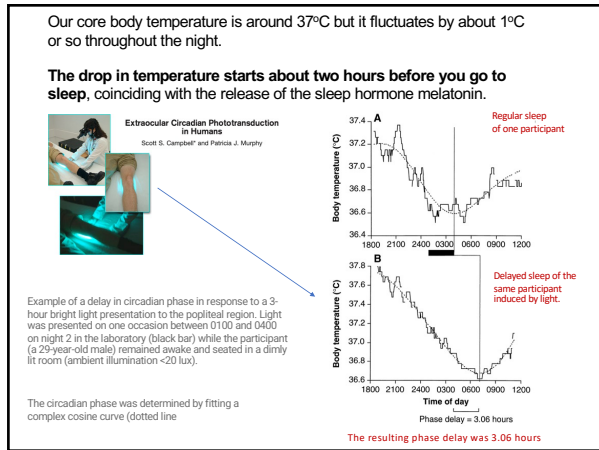
Scott S. Campbell* and Patricia J. Murphy

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 knees). A systematic relation was found between the timing of the light pulses 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.

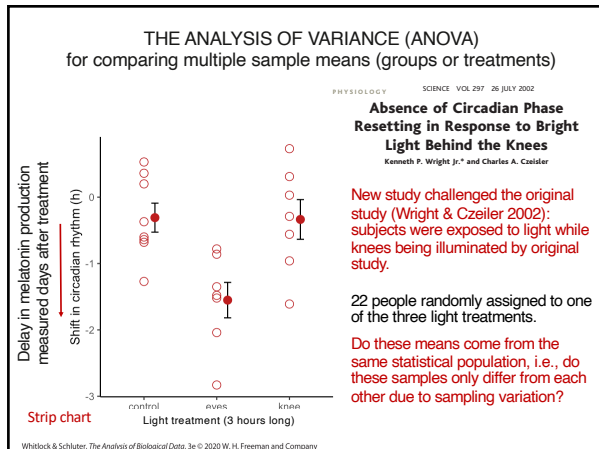
SCIENCE • VOL. 279 • 16 JANUARY 1998

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

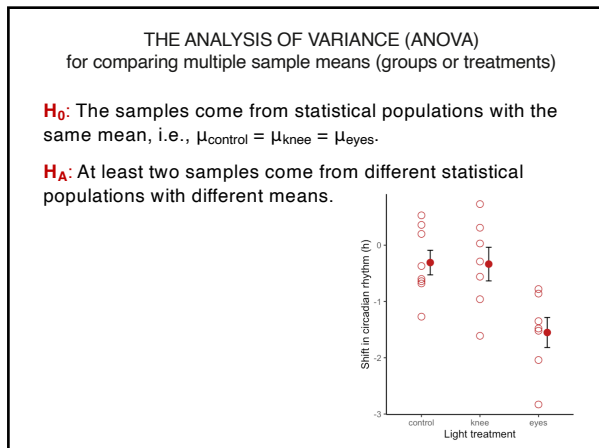
27



28



29



30

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

H₀: 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.

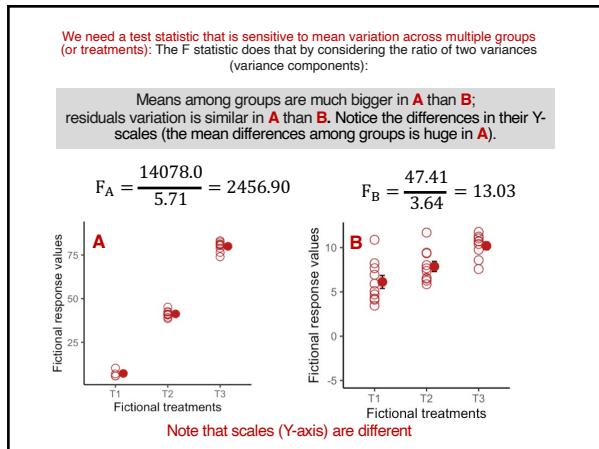
Which is to say:

H₀: Differences in means among groups are due to **sampling error from the same population.**

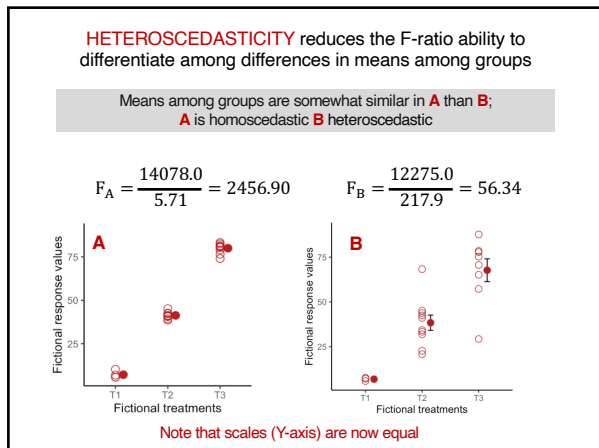
H_A: Differences in means among groups are NOT due to **sampling error from the same population.**

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.

31



32



33

Verbal representation of equations

Let's talk ANOVA "jargon"

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

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

34

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!

$F = \frac{s_b^2}{s_w^2} = \frac{\text{Degrees of freedom of } MS_{\text{groups}}}{g-1}$

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

35

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!

$F = \frac{s_b^2}{s_w^2} = \frac{\text{Degrees of freedom of } MS_{\text{groups}}}{g-1}$

MS_{errors} (w=within groups) Error Mean Square $\sum_{i=1}^g (n_i - 1) s_i^2$ Variance of each group s_i^2 Big "N": sum of all sample sizes across groups $\sum_{i=1}^g (n_i - 1) = (N-g)$ Number of groups g Degrees of freedom of MS_{groups} Sample size of each group n_i

36

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

Let's suppose two groups for simplicity!

group 1

1 2 3 4 5

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

37

g₁: 1 2 3 4 5 $\bar{X}_1 = 3.0$ $\bar{X}_2 = 12.0$
 g₂: 10 11 12 13 14 $s_1^2 = 2.5$ $s_2^2 = 2.5$

$\bar{X} = (1+2+3+4+5+10+11+12+13+14)/10 = 7.5$
 $MS_{groups} = \text{variance among group means (due to "treatment")}$
 $= (5 \times (3.0 - 7.5)^2 + 5 \times (12.0 - 7.5)^2) / (2-1) =$
 $202.5 / (2-1) = 202.5$

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

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

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

Mean of each group Total mean
 Variance of each group
 Big "N": sum of all sample sizes across groups
 $\sum_{i=1}^g (n_i - 1) \rightarrow (N-g)$

38

g₁: 1 2 3 4 5
 g₂: 10 11 12 13 14

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

Mean of each group Total mean
 Variance of each group
 Big "N": sum of all sample sizes across groups
 $\sum_{i=1}^g (n_i - 1) \rightarrow (N-g)$

$MS_{error} \quad \bar{X}_1 = 3.0 \quad \bar{X}_2 = 12.0$
 $s_1^2 = 2.5 \quad s_2^2 = 2.5$

$MS_{error} = \text{variance within groups (residuals)}$

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

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

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

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

39

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

$$MS_{groups} = \frac{(5 \times (3.0 - 7.5)^2 + 5 \times (12.0 - 7.5)^2)}{(2-1)} = \frac{202.5}{(2-1)} = 202.5$$

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

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

$$MS_{error} = \text{variance within groups (residuals)}$$

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

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

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

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

40



41

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

42

“The knees who say night”

Shift in circadian rhythm (h)

control noise eyes
Light treatment

Statistical Conclusion?

H₀: 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
---

```

43

“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		

44

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** involve subtractions from means so that they would be biased if not divided by adjustments (degrees of freedom) to produce **mean square deviations**.

45

“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?

46

How can we conceptualize the construction of the F distribution?

The statistical “machinery”:

- 1) Assume that 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.

47

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**.

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.

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

(8,7,7) observations

Probability density

F values

control eyes knee

Light treatment

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

48

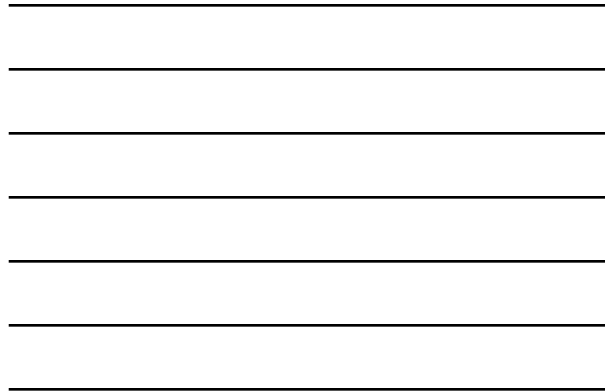
The F (sampling) distribution assuming that H_0 is true
 H_0 : Differences in means among groups are due to **sampling error from the same population.**

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

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.

49



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

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

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 ($N-g$)

50



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

```
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
```

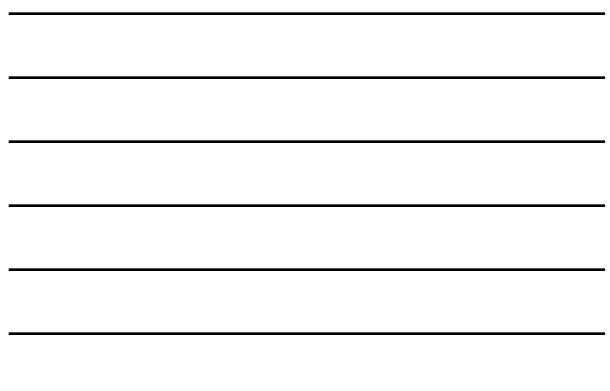
Degrees of freedom Observed F-value (observed test statistic) P-value

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).

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.

52

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.*

53

“The knees who say night”

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

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!

54



55

Analysis of Variance
(more than one factor)
Multifactorial - ANOVA

Part I - Introduction

56

Some types of ANOVA designs:

Single-factor ANOVA (Intro stats)
Factorial designs (crossed) – today
Mixed models

57

Research question (my own fictional example; real examples will be seen in the next lecture and tutorials):

Do exercise and diet affect weight loss?

How would you set a study to test this question?

Why fictional? The context of the problem itself seems to be easier to understand than more "biological" applications!

58

Study - Individuals are followed regarding their weight loss after 1 month of exercise (or not) and diet (or not).

Factorial ANOVA - always involves one continuous variable (i.e., response variable = weight loss) and two or more categorical (factors) variables (exercise and diet).

Factors: exercise and diet (two-factorial).

In this example, **exercise** and **diet** are factors with two levels or groups (Yes/No).

Response variable: weight loss.

59

Data structure in a csv file

Diet: yes no
 ↓ ↓ ↓ ↓
 Exercise: Y N Y N
 replicates (5 individuals) ↓ ↓ ↓ ↓

Weight loss: start weight - end weight (in pounds)

A	B	C
Diet	Exercise	WeightLoss
yes	yes	5.8
yes	yes	5.3
yes	yes	5.7
yes	yes	6.1
yes	yes	5.1
no	yes	6.2
no	yes	5.4
no	yes	6.3
no	yes	4.5
no	yes	4.2
yes	no	6.9
yes	no	8.1
yes	no	8.2
yes	no	8.8
yes	no	8.6
no	no	7.1
no	no	8.1
no	no	7.6
no	no	7.4
no	no	7.8

60

Do exercise and diet affect weight loss?

Let's elaborate on this question further:

Main effects

- Are the differences in weight loss only due to exercise alone?
- Are the differences in weight loss only due to diet alone?

Interaction

- Does the effect of diet on weight loss depend on exercise? In other words, are the differences in weight loss attributable to some combinations of exercise and diet? (e.g., the biggest weight loss compared to any other combination of diet and exercise was observed when individuals both dieted and exercised).

61

Treatments

Main effects:

Diet - two treatments (yes/no).

Exercise - two treatments (yes/no).

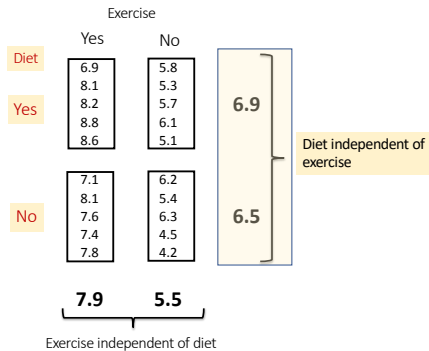
Possible sources of statistical interactions:

Combination of diet and exercise treatments - four pairwise combinations of means:

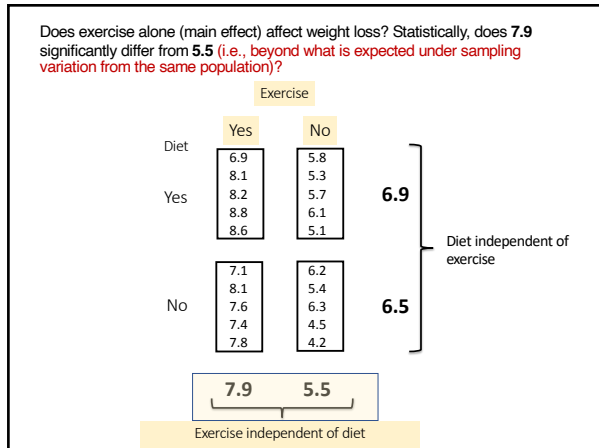
- 1) No exercise but diet.
- 2) Exercise but no diet.
- 3) No exercise and no diet.
- 4) Exercise and diet.

62

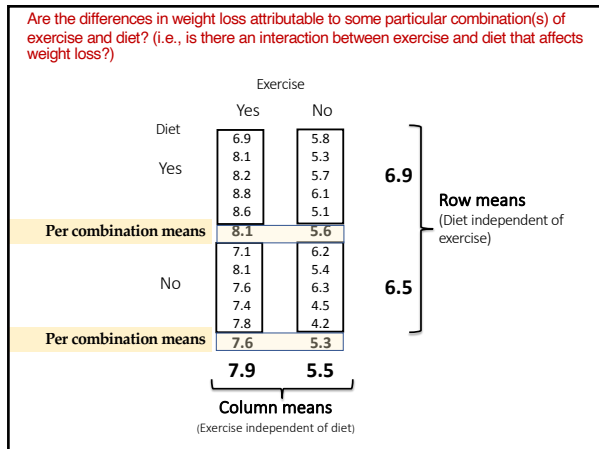
Does diet alone (main effect) affect weight loss? Statistically, does **6.9** significantly differ from **6.5** (i.e., beyond what is expected under sampling variation from the same population)?



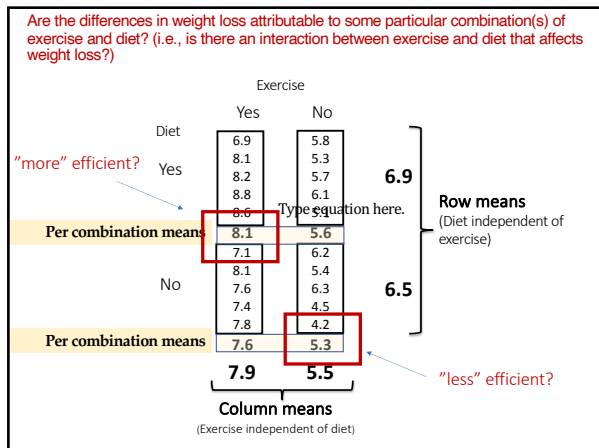
63



64



65



66

Stating the 3 possible sets of statistical hypotheses in a two-factorial design:

Does *dieting* affect weight loss? DIET (main effect 1)

H₀: There is no difference between diet treatments in mean weight loss (in the population).

H_A: There is a difference between diet treatments in mean weight loss (in the population).

67

Stating the 3 possible sets of statistical hypotheses in a two-factorial design:

Does *exercising* affect weight loss? EXERCISE (main effect 2)

H₀: There is no difference between exercise treatments in mean weight loss (in the population).

H_A: There is a difference between exercise treatments in mean weight loss (in the population).

68

Stating the 3 possible sets of statistical hypotheses in a two-factorial design:

Are the differences in weight loss attributable to some combinations of exercise and diet? (interaction effect)

H₀: The effect of diet on weight loss does not depend on exercise in the population (*or vice versa*).

H_A: The effect of diet on weight loss depends on exercise in the population (*or vice versa*).

69

Type of effects in this study:

Fixed: The levels in a factor are specifically chosen by the researcher (diet and exercise)

Note: The typical ANOVA design (simple or factorial) is conducted assuming a fixed design (we will see other designs later in the course).

70

ANOVA Table

Source of variation	Df	SS	Mean SS	F value	Prob
Diet	1	0.800	0.800	1.8089	0.1974
Exercise	1	28.800	28.800	65.1215	<0.0000001
Diet x Exercise	1	0.072	0.072	0.1628	0.6919
residuals	16	7.076	0.442		

H₀: There is no difference between diet treatments in mean weight loss.
 H_a: There is a difference between diet treatments in mean weight loss.

H₀: There is no difference between exercise treatments in mean weight loss.
H_a: There is a difference between exercise treatments in mean weight loss.

H₀: The effect of diet on weight loss does not depend on exercise (or vice versa).
 H_a: The effect of diet on weight loss depends on exercise (or vice versa).

Research conclusion: Only exercise affects weight loss!

71

ANOVA Table (R) versus publication quality

Response: PopA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Diet	1	0.800	0.8000	1.8089	0.1974
Exercise	1	28.800	28.8000	65.1215	4.954e-07 ***
Diet:Exercise	1	0.072	0.0720	0.1628	0.6919
Residuals	16	7.076	0.4422		

↓

Source of variation	Df	SS	Mean SS	F value	Prob
Diet	1	0.800	0.800	1.8089	0.1974
Exercise	1	28.800	28.800	65.1215	<0.0000001
Diet x Exercise	1	0.072	0.072	0.1628	0.6919
residuals	16	7.076	0.442		

Conclusion: There is a difference between exercise treatments in mean weight loss (in the population).

72

ANOVA Table (details on degrees of freedom)

Source of variation	Df	SS	Mean SS	F value	Prob
Diet	1	0.800	0.800	1.8089	0.1974
Exercise	1	28.800	28.800	65.1215	<0.0000001
Diet x Exercise	1	0.072	0.072	0.1628	0.6919
residuals	16	7.076	0.442		

df (diet) = number of levels (k) - 1 = 2 - 1 = 1

df (exercise) = number of levels (m) - 1 = 2 - 1 = 1

df (Interaction) = (m - 1).(k - 1) = (2 - 1).(2 - 1) = 1

df (residuals) = (N - m - k) = (20 - 2 - 2) = 16

N = total number of observations across all factors and levels

73

- Next lecture:**
- 1) Real examples of two-way ANOVA designs.
 - 2) Plotting and understanding significant interaction terms.
 - 3) How to test for assumptions (one-way and multi-factorial ANOVA).
 - 4) Identifying which pairs of means significantly differ to find the meaningful interactions (e.g., mean of weight loss with no exercise *versus* mean of weight loss with diet).

74