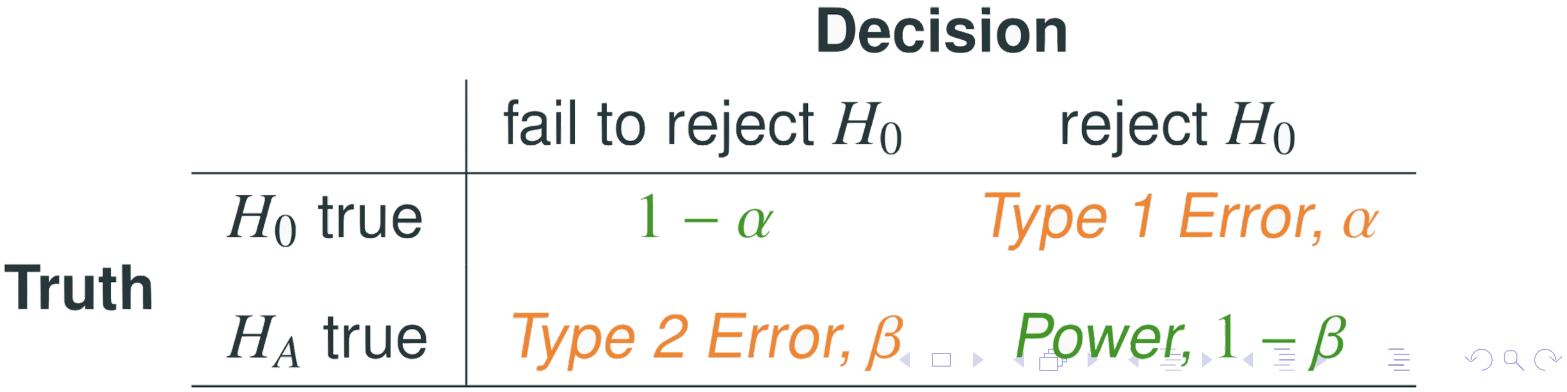
Multiple Hypothesis Testing

Material: Extra material from Introduction to Statistical Learning can be found on the webpage.

Type I error

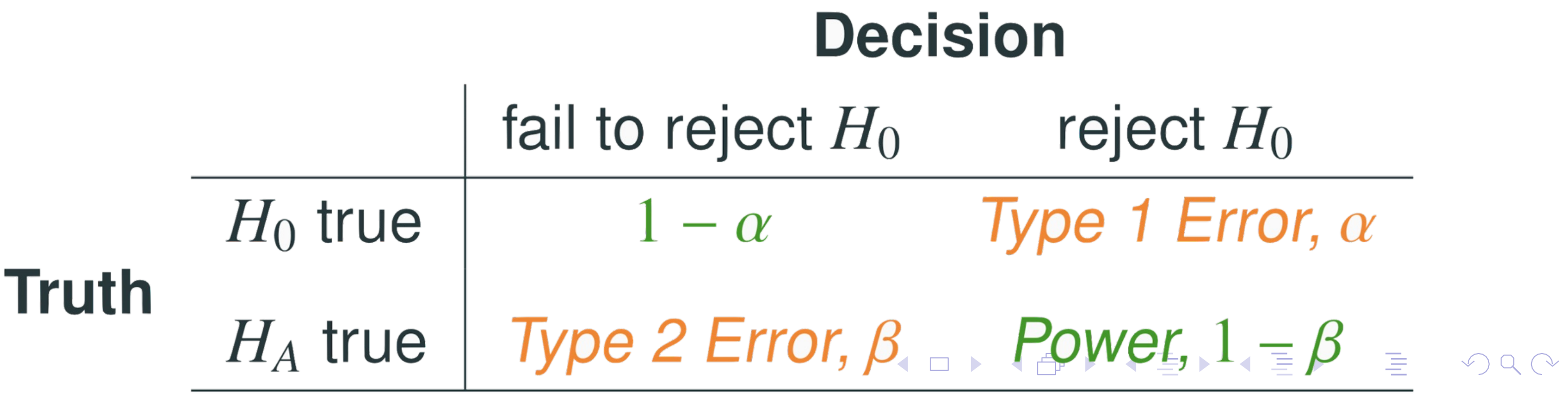
- As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a significance level of 0.05, α = 0.05.
- Type I error rate:
  - P (Type I error) = P (Reject $H_0 | H_0$ is true) ≤ α
- Increasing α increases the Type I error rate.
- When we select α we control for the tolerance we have for type I errors.



		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1 - \alpha$	Type 1 Error, α
	H_A true	Type 2 Error, β	Power, $1 - \beta$

Type II error

- If the alternative hypothesis is actually true, what is the chance that we make a Type II Error, i.e. we fail to reject the null hypothesis even when we should reject it?
- The answer is not obvious, but
  - If the true population average is very close to the null hypothesis value, it will be difficult to detect a difference (and reject H_0).
  - If the true population average is very different from the null hypothesis value, it will be easier to detect a difference.
- The probability of correctly rejecting the null is the **power** of the test.



		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1-\alpha$	*Type 1 Error,* α
	H_A true	*Type 2 Error,* β	*Power,* $1-\beta$

Multiple testing

- Now assume we want to test multiple hypotheses

$$H_{01}, \ldots, H_{0m}$$

- If we reject all null hypotheses for which the p-value falls below 0.05, then how many Type I errors will we make?

A thought experiment

- Suppose that we flip a fair coin ten times, and we wish to test H_0: the coin is fair.

- We'll probably get approximately the same number of heads and tails.

- The p-value probably won't be small. We do not reject H_0.

- But what if we flip 1,024 fair coins ten times each?

Multiple hypotheses testing

- Suppose we test $H_{01}, \ldots, H_{0m}$, all of which are true, and reject any null hypothesis with a p-value below 0.05.
- Then we expect to falsely reject approximately $0.05 \times m$ null hypotheses.
- If $m = 10{,}000$, then we expect to falsely reject 500 null hypotheses by chance!
- That's a lot of Type I errors, i.e. false discoveries/false positives!
- Example: Genome-wide association studies.

Family-wise error rate

The probability of making at least one type 1 error

	Fail to reject H_0	Reject H_0	
H_0 true	U	V	m_0
H_1 true	W	S	$m - m_0$
	$m - R$	R	m

$$FWER = \mathrm{P}(\mathrm{V} \geq 1) =$$

$$1 - P\ (V = 0) =$$

$$1 - P(do\ not\ falsely\ reject\ any\ null\ hypothesis) =$$

$$1 - P(\bigcap_{j=1}^{m} do\ not\ falsely\ reject\ H_{0j})$$

Family-wise error rate

	Fail to reject H_0	Reject H_0	
H_0 true	U	V	m_0
H_1 true	W	S	$m - m_0$
	$m - R$	R	m

$$FWER = \mathrm{P}(\mathrm{V} \geq 1) = 1 - P\,(V = 0) =$$
$$1 - P(\cap_{j=1}^{m} do\,not\,falsely\,reject\,H_{0j})$$

If the tests are independent and all H_{0j} are true

$$FWER = 1 - \prod(P(do\,not\,falsely\,reject\,H_{oj}) = 1 - (1 - a)^m$$

If $m = 3, a = 0.05, FWER = 0.143$
If $m = 10, a = 0.05, FWER = 0.402$

Multiple hypotheses testing

$$FWER = P\ (falsely\,reject\,at\,least\,one\,hypothesis) =$$

$$P\left(\cup_{j=1}^{m} A_j\right) \leq \sum_{j=1}^{m} P\left(A_j\right)$$

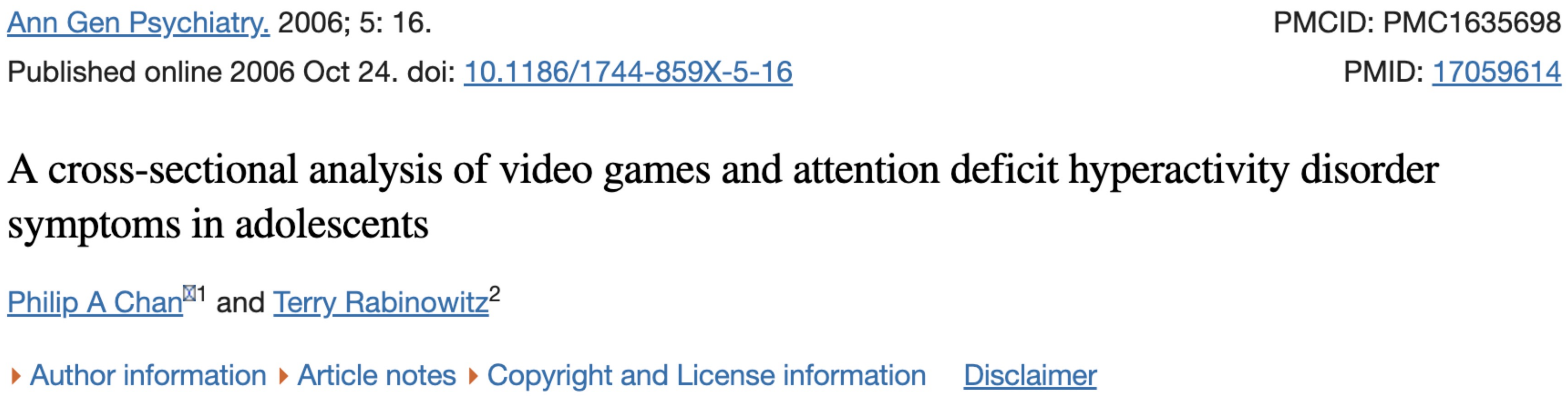
where A_j is the event that we falsely reject the $j-th$ null hypothesis. If we only reject hypotheses when the p-value is less than α/m, then

$$\text{FWER} \leq \sum_{j=1}^{m} P(A_j) \leq \sum_{j=1}^{m} \frac{\alpha}{m} = \alpha$$

because $P\ (A_j\ ) \leq\ \alpha/m$

This is the Bonferroni Correction: to control FWER at level α, reject any null hypothesis with p-value below α/m

Example: Video Games and ADHD



Ann Gen Psychiatry. 2006; 5: 16.

PMCID: PMC1635698

Published online 2006 Oct 24. doi: 10.1186/1744-859X-5-16

PMID: 17059614

A cross-sectional analysis of video games and attention deficit hyperactivity disorder symptoms in adolescents

Philip A Chan✉[1] and Terry Rabinowitz[2]

▸ Author information ▸ Article notes ▸ Copyright and License information Disclaimer

Measuring the effect of four types of videogames/media usage on 5 outcomes related to ADHD

Internet
TV
Console Video Games
Internet Video Games

Young’s Addiction Scale
Conner’s Scale: Oppositional
Conner’s Scale: Inattention
Conner’s Scale: Hyperactivity
Conner’s Scale: ADHD

Example: Video Games and ADHD

vs	Internet	TV	VG-C	VG-I
Young’s Addiction Scale	0.804	**0.040**	< **0.001**	**<0.001**
Conner’s Scale: Oppositional	0.096	0.397	0.917	0.826
Conner’s Scale: Inattention	0.289	0.311	**0.001**	**<0.001**
Conner’s Scale: Hyperactivity	0.901	0.397	0.800	0.142
Conner’s Scale: ADHD	0.115	0.343	**0.018**	**0.020**

- If we reject H_{0j} if the p-value is less than $\alpha = 0.05$, we will conclude that TV, VG-C, VG-I significantly affect YAS, VG-C and VG-I significantly affect Inattention and ADHD.

- However, we have tested multiple hypotheses, so the FWER is greater than 0.05 .

- Assuming that all null hypotheses are true, what is the FWER?

Example: Video Games and ADHD

vs	Internet	TV	VG-C	VG-I
Young's Addiction Scale	0.804	0.040	**< 0.001**	**<0.001**
Conner's Scale: Oppositional	0.096	0.397	0.917	0.826
Conner's Scale: Inattention	0.289	0.311	**0.001**	**<0.001**
Conner's Scale: Hyperactivity	0.901	0.397	0.800	0.142
Conner's Scale: ADHD	0.115	0.343	0.018	0.020

- Using the Bonferroni correction we will reject p-values less than $\alpha/20 = 0.0025$.

- If we reject H_{0j} if the p-value is less than 0.0025, we will conclude that VG-C, VG-I significantly affect YAS and inattention

- If you had performed 1000 tests, the p-value for controlling FWER at level α would be: 5×10^{-5}

Holm’s method for controling FWER

- Compute p-values, p1, . . . , pm for the m null hypotheses

$$H_{01}, \ldots, H_{0m}.$$

- Order the m p-values so that $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$.
- Define

$$P_L = min_j\, P_{(j)} > \frac{\alpha}{m+1-j}$$

- Reject all null hypotheses H_{0j} for which $p_j < p_{(L)}$

- Holm’s method controls the FWER at level α.

Bonferroni vs Holm

Consider m = 5

p-values

$p_1 = 0.006$, $p_2 = 0.918$, $p_3 = 0.012$, $p_4 = 0.601$, $p_5 = 0.756$.

Then

$p_{(1)} = 0.006$, $p_{(2)} = 0.012$, $p_{(3)} = 0.601$, $p_{(4)} = 0.756$, $p_{(5)} = 0.918$.

- Bonferroni?
- Bonferroni - Holm?

Bonferroni vs Holm

- Bonferroni is simple ... reject any null hypothesis with a p-value below α/m.
- Holm is slightly more complicated, but it will lead to more rejections while controlling FWER!!
- Holm is a better choice

The False Discovery Rate

	Fail to reject H_0	Reject H_0	
H_0 true	U	V	m_0
H_1 true	W	S	$m - m_0$
	$m - R$	R	m

- The FWER rate focuses on controlling $P\ (V\ >\ 1)$, i.e., the probability of falsely rejecting any null hypothesis.
- This is a tough ask when m is large! It will cause us to be super conservative (i.e. to very rarely reject).
- Instead, we can control the false discovery rate:
  - FDR $= E(V/R)$

The False Discovery Rate

$$\mathrm{FDR} = E\left(\frac{V}{R}\right) = E\left(\frac{\text{number of false rejections}}{\text{total number of rejections}}\right)$$

- A scientist conducts a hypothesis test on each of m = 20, 000 drug candidates.
- She wants to identify a smaller set of promising candidates to investigate further.
- She wants reassurance that this smaller set is really “promising”, i.e. not too many falsely rejected H_0’s.
- FWER controls $P(at\ least\ one\ false\ rejection)$.
- FDR controls the fraction of candidates in the smaller set that are really false rejections. This is what she needs!

Benjamini-Hochberg procedure for controlling FDR

1. Specify q, the level at which to control the FDR.
2. Compute p-values $p_1, \ldots, p_m$ for the null hypotheses $H_{01}, \ldots, H_{0m}$.
3. Order the p-values so that $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$.
4. Define $L = max_j : p_{(j)} < qj/m$.
5. Reject all null hypotheses H_{0j} for which $p_{(j)} \leq p_{(L)}$.

Then, FDR $\leq q$.

FWER vs FDR

Consider m = 5

p-values

$p_1 = 0.006$, $p_2 = 0.918$, $p_3 = 0.012$, $p_4 = 0.601$, $p_5 = 0.756$.

Then

$p_{(1)} = 0.006$, $p_{(2)} = 0.012$, $p_{(3)} = 0.601$, $p_{(4)} = 0.756$, $p_{(5)} = 0.918$.

- Bonferroni?
- Bonferroni-Holm?
- Benjamini-Hochberg?

FWER vs FDR

Consider $m = 4$

p-values

$$p_1 = 0.01, p_2 = 0.04, p_3 = 0.03\,, p_4 = 0.005$$

- Bonferroni?
- Bonferroni-Holm?
- Benjamini-Hochberg?

Comparing means with ANOVA

Material: DeGroot and Schervish 9.7, 11.6
OpenStatistics Chapter 7.3

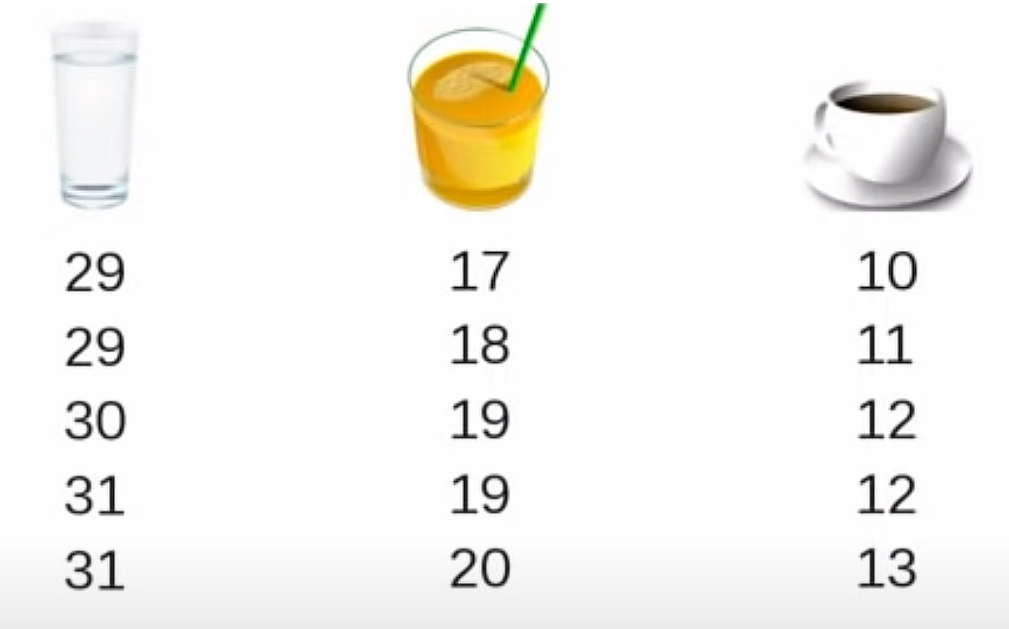
Slides adopted from Openintro.org

Research question:

You want to test if drinking different beverages affects your reaction time.

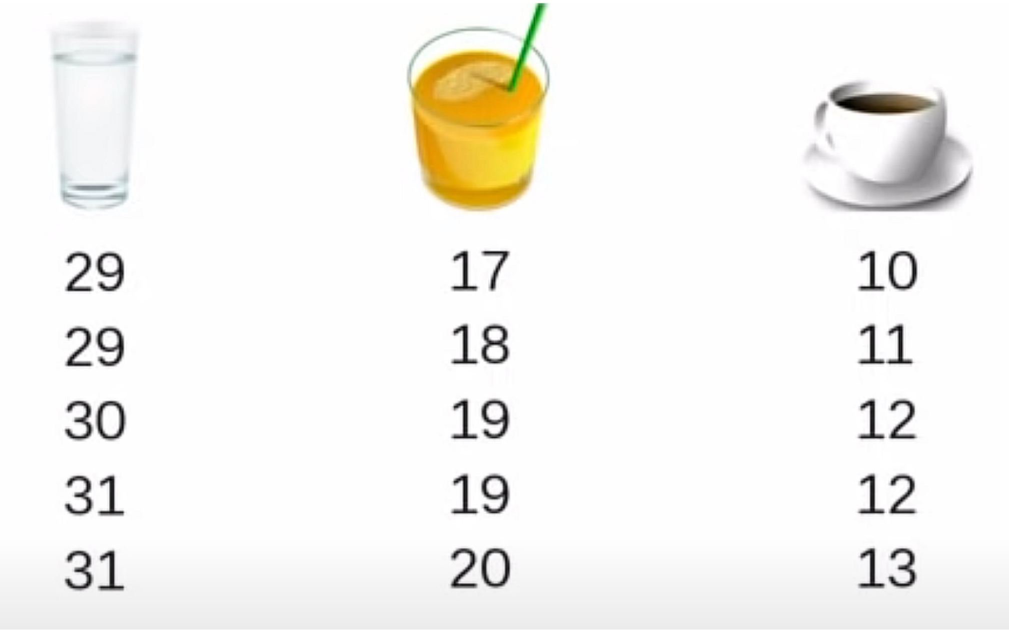
You give split your subjects in 3 groups.
You give each group water, tea, and coffee, respectively
You measure their reaction time.

Scenario 1:



29	17	10
29	18	11
30	19	12
31	19	12
31	20	13

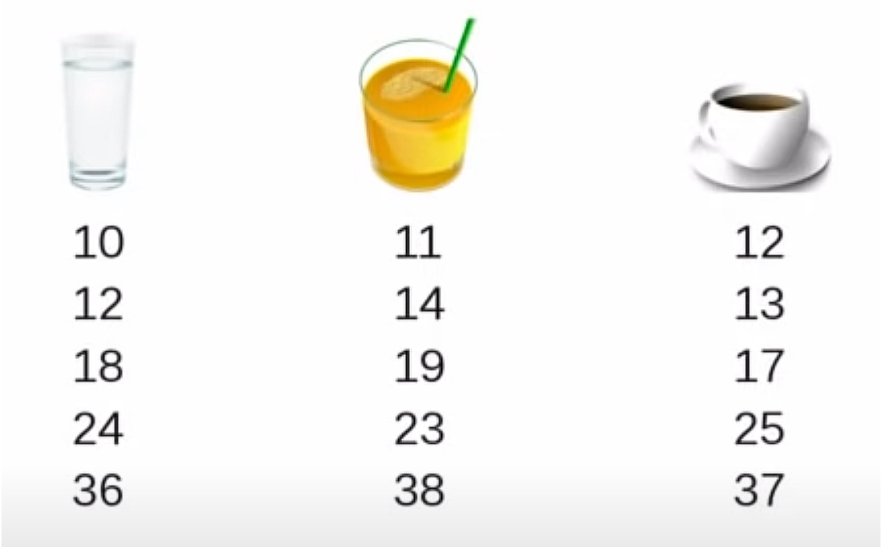
Scenario 1:



29	17	10
29	18	11
30	19	12
31	19	12
31	20	13

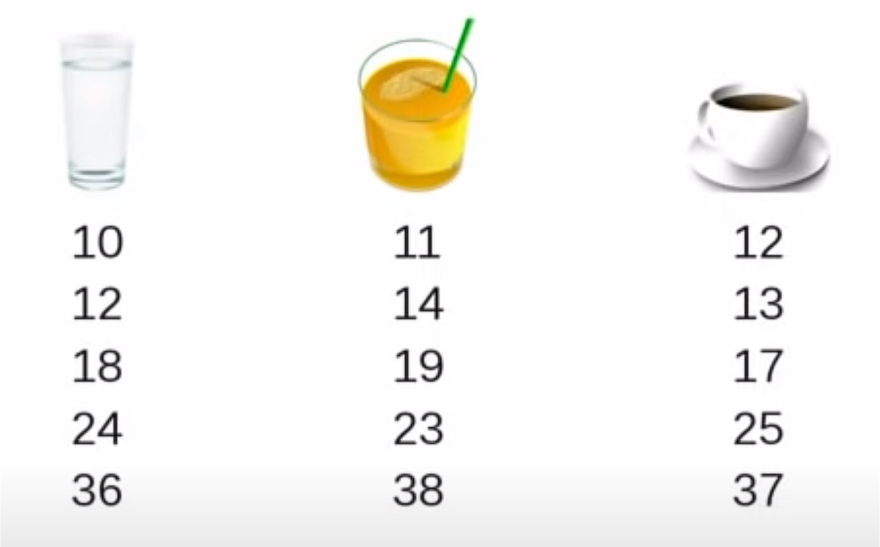
You have little variablity within each group, but different groups look different.

Scenario 2:



10	11	12
12	14	13
18	19	17
24	23	25
36	38	37

Scenario 2:



10	11	12
12	14	13
18	19	17
24	23	25
36	38	37

You have lots of variablity within each group, but different groups look the same.

ANOVA

Figure out how much of the total variance comes from:

a) The variance between the groups

b) The variance within the groups

Calculate the ratio:

$$F = \frac{\text{variance between groups}}{\text{variance within groups}}$$

Research question

Is there a difference between the mean response time among the three beverages?

Research question

Is there a difference between the mean response time among the three beverages?

- To compare means of 2 groups we use a Z or a T statistic

Research question

Is there a difference between the mean response time among the three beverages?

- To compare means of 2 groups we use a Z or a T statistic
- To compare means of 3+ groups we use a new test called *ANOVA* and a new statistic called F

The F distributions

Definition: Let Y and W be independent random variables such that

- Y has the χ^2 distribution with m degrees of freedom and
- W has the χ^2 distribution with n degrees of freedom,

where m and n are positive integers.

Then the random variable $X = \dfrac{Y/m}{W/n}$ follows an F-distribution with m and n degrees of freedom.

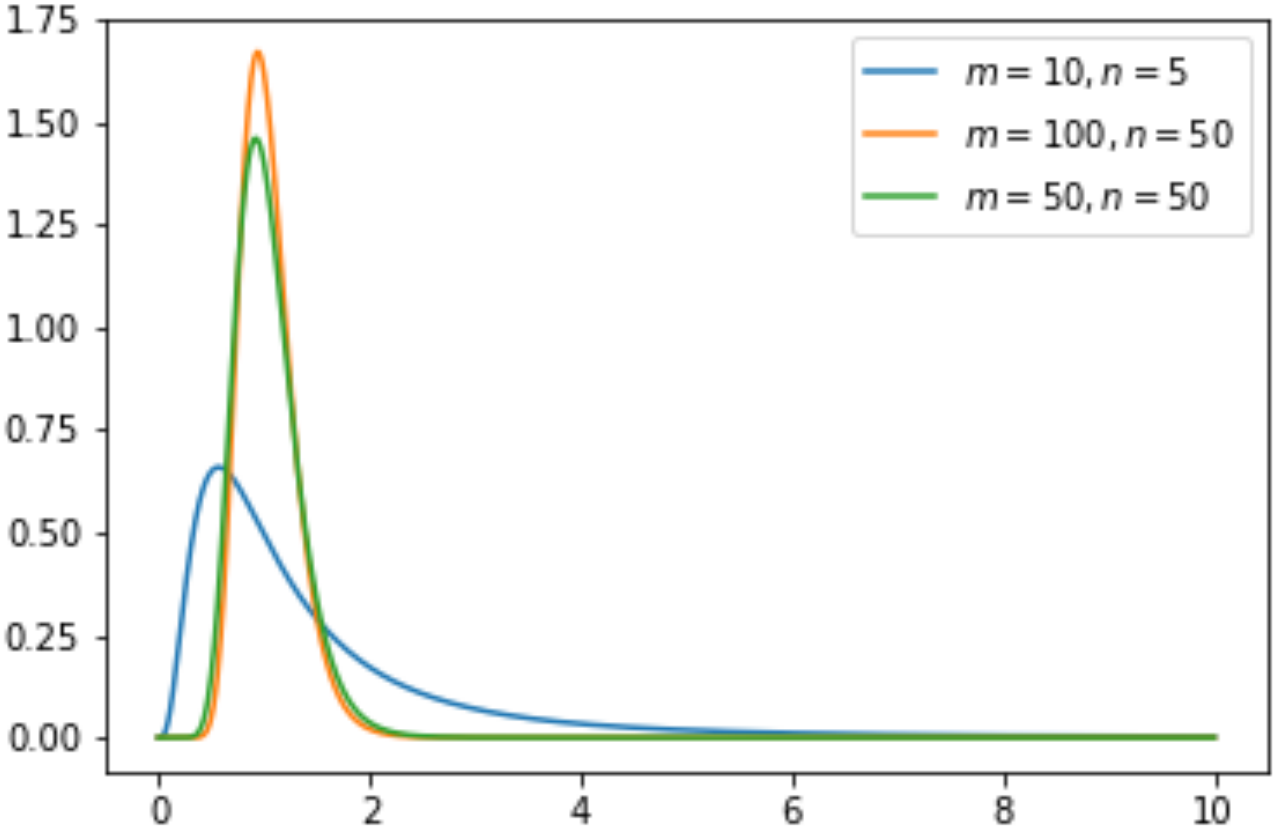
The F distributions

pdf:

$$f(x) = \frac{\Gamma\left[\frac{1}{2}(m+n)\right] m^{\frac{m}{2}} n^{\frac{n}{2}}}{\Gamma\left(\frac{1}{2}m\right)\Gamma\left(\frac{1}{2}n\right)} \times \frac{x^{\frac{m}{2}-1}}{(mx+n)^{\frac{m+n}{2}}}, \qquad x > 0$$

Python scipy.stats.f
Quantile using f.ppf

The F distributions



Comparing variances of two normals

Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$

Then $\frac{S_x^2}{\sigma_x^2} \sim \chi_{n-1}^2$, where $S_x^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$

Comparing variances of two normals

Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$

Then $\frac{S_x^2}{\sigma_x^2} \sim \chi_{n-1}^2$, where $S_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2$

Let $\mathrm{V}^* = \frac{S_x^2/[(m-1)\sigma_1^2]}{S_y^2/[(n-1)\sigma_2^2]}$

Then $V \sim F$ distribution with $m-1, n-1$ degrees of freedom.

If $\sigma_1^2 = \sigma_2^2$, then $V = \frac{S_x^2/(m-1)}{S_y^2/(n-1)}$ also follows the same distribution

ANOVA

ANOVA is used to assess whether the mean of the outcome variable is different for different levels of a categorical variable

z/*t* test vs. ANOVA - Purpose

z/*t* test

Compare means from two groups to see whether they are so far apart that the observed difference cannot reasonably be attributed to sampling variability

$$H_0 : \mu_1 = \mu_2$$

ANOVA

Compare the means from two or more groups to see whether they are so far apart that the observed differences cannot all reasonably be attributed to sampling variability

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

ANOVA

ANOVA is used to assess whether the mean of the outcome variable is different for different levels of a categorical variable

H_0: The mean outcome is the same across all categories,

$$\mu_1 = \mu_2 = \ldots = \mu_k,$$

where μ_i represents the mean of the outcome for observations in category *i*

ANOVA

ANOVA is used to assess whether the mean of the outcome variable is different for different levels of a categorical variable

H_0 : The mean outcome is the same across all categories,

$$\mu_1 = \mu_2 = \ldots = \mu_k,$$

where μ_i represents the mean of the outcome for observations in category *i*

H_A : At least one mean is different than others

z/*t* test vs. ANOVA - Method

z/*t* test

Compute a test statistic (a ratio)

$$z/t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE(\bar{x}_1 - \bar{x}_2)}$$

ANOVA

Compute a test statistic (a ratio)

$$F = \frac{variability\ bet.groups}{variability\ within\ groups}$$

z/*t* test vs. ANOVA - Method

z/*t* test

Compute a test statistic (a ratio)

$$z/t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE(\bar{x}_1 - \bar{x}_2)}$$

ANOVA

Compute a test statistic (a ratio)

$$F = \frac{variability\ bet.groups}{variability\ within\ groups}$$

- Large test statistics lead to small p-values
- If the p-value is small enough H_0 is rejected, we conclude that the population means are not equal

z/*t* test vs. ANOVA

- With only two groups t-test and ANOVA are equivalent, but only if we use a pooled standard variance in the denominator of the test statistic
- With more than two groups, ANOVA compares the sample means to an overall grand mean

Hypotheses

A. $H_0 : \mu_{\mathrm{W}} = \mu_{\mathrm{T}} = \mu_{\mathrm{C}}$

$H_A : \mu_{\mathrm{W}} \neq \mu_{\mathrm{T}} \neq \mu_{\mathrm{C}}$

B. $H_0 : \mu_{\mathrm{W}} \neq \mu_{\mathrm{T}} \neq \mu_{\mathrm{C}}$

$H_A : \mu_{\mathrm{W}} = \mu_{\mathrm{T}} = \mu_{\mathrm{C}}$

C. $H_0 : \mu_{\mathrm{W}} = \mu_{\mathrm{T}} = \mu_{\mathrm{C}}$

H_A : At least one mean is different

A. $H_0 : \mu_{\mathrm{W}} = \mu_{\mathrm{T}} = \mu_{\mathrm{C}} = 0$

H_A : At least one mean is different

E. $H_0 : \mu_{\mathrm{W}} = \mu_{\mathrm{T}} = \mu_{\mathrm{C}}$

$H_A : \mu_{\mathrm{B}} > \mu_{\mathrm{M}} > \mu_{\mathrm{C}}$

Hypotheses

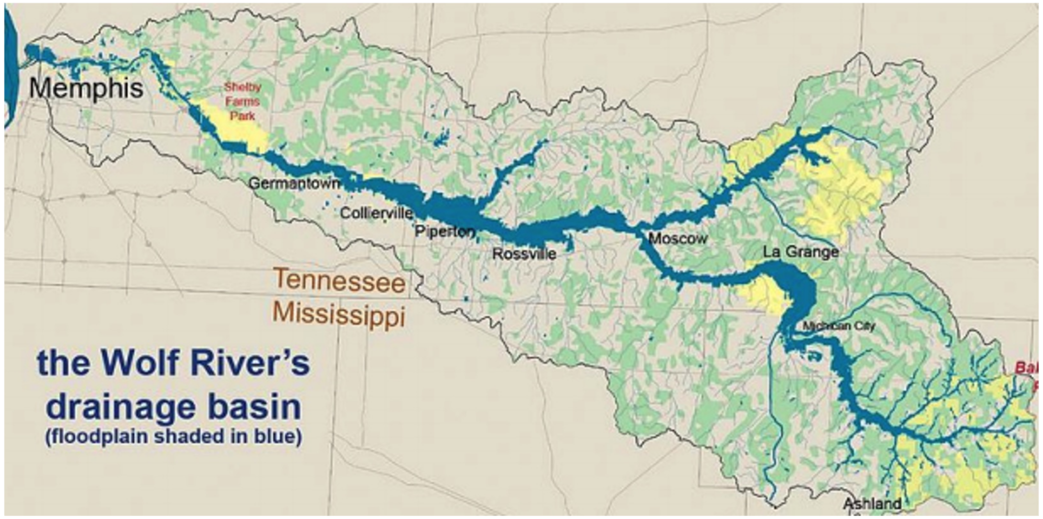
A. $H_0 : \mu_{\mathrm{W}} = \mu_{\mathrm{T}} = \mu_{\mathrm{C}}$
$H_A : \mu_{\mathrm{W}} \neq \mu_{\mathrm{T}} \neq \mu_{\mathrm{C}}$

B. $H_0 : \mu_{\mathrm{W}} \neq \mu_{\mathrm{T}} \neq \mu_{\mathrm{C}}$
$H_A : \mu_{\mathrm{W}} = \mu_{\mathrm{T}} = \mu_{\mathrm{C}}$

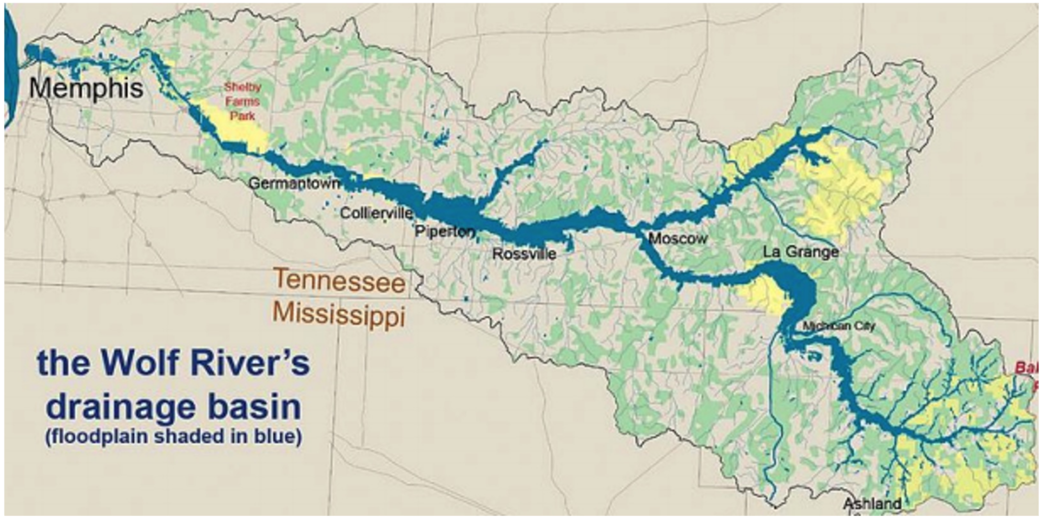
C. $H_0 : \mu_W = \mu_T = \mu_C$
H_A : *At least one mean is different*

A. $H_0 : \mu_{\mathrm{W}} = \mu_{\mathrm{T}} = \mu_{\mathrm{C}} = 0$
H_A : At least one mean is different

E. $H_0 : \mu_{\mathrm{W}} = \mu_{\mathrm{T}} = \mu_{\mathrm{C}}$
$H_A : \mu_{\mathrm{B}} > \mu_{\mathrm{M}} > \mu_{\mathrm{C}}$



- The Wolf River in Tennessee flows past an abandoned site once used by the pesticide industry for dumping wastes, including chlordane (pesticide), aldrin, and dieldrin (both insecticides)



- The Wolf River in Tennessee flows past an abandoned site once used by the pesticide industry for dumping wastes, including chlordane (pesticide), aldrin, and dieldrin (both insecticides)
- These highly toxic organic compounds can cause various cancers and birth defects

Memphis
Shelby Farms Park
Germantown
Collierville
Piperton
Rossville
Moscow
La Grange
Michigan City
Ashland
Tennessee
Mississippi
the Wolf River's drainage basin
(floodplain shaded in blue)

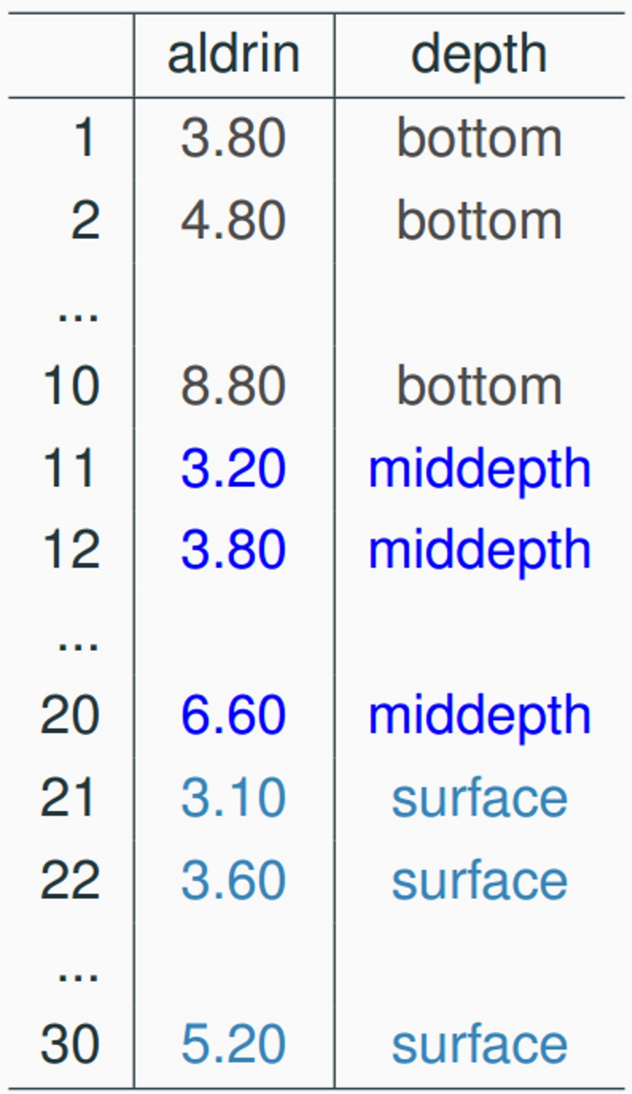
- The Wolf River in Tennessee flows past an abandoned site once used by the pesticide industry for dumping wastes, including chlordane (pesticide), aldrin, and dieldrin (both insecticides)
- These highly toxic organic compounds can cause various cancers and birth defects
- The standard methods to test whether these substances are present in a river is to take samples at six-tenths depth

Memphis
Shelby Farms Park
Germantown
Collierville
Piperton
Rossville
Moscow
La Grange
Tennessee
Mississippi
Michigan City
Ashland
the Wolf River's drainage basin
(floodplain shaded in blue)

- The Wolf River in Tennessee flows past an abandoned site once used by the pesticide industry for dumping wastes, including chlordane (pesticide), aldrin, and dieldrin (both insecticides)
- These highly toxic organic compounds can cause various cancers and birth defects
- The standard methods to test whether these substances are present in a river is to take samples at six-tenths depth
- But since these compounds are denser than water and their molecules tend to stick to particles of sediment, they are more likely to be found in higher concentrations near the bottom

Data

Aldrin concentration (nanograms per liter) at three levels of depth

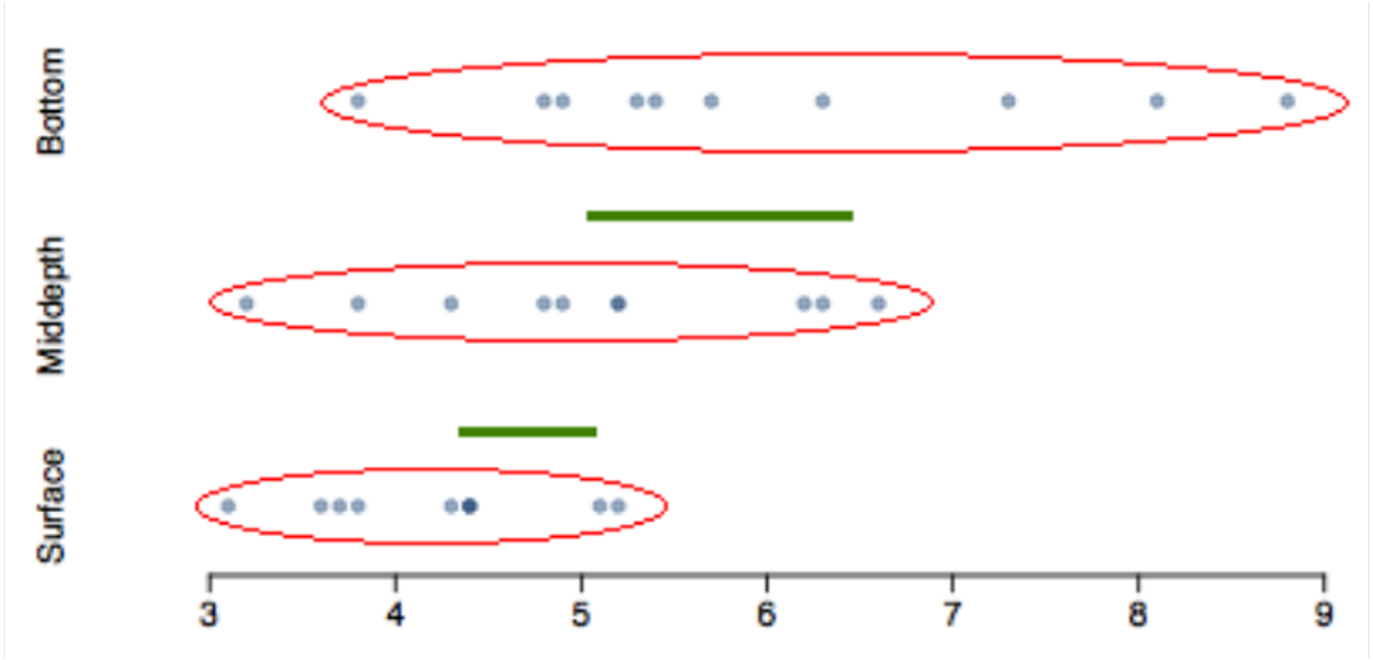


	aldrin	depth
1	3.80	bottom
2	4.80	bottom
...		
10	8.80	bottom
11	3.20	middepth
12	3.80	middepth
...		
20	6.60	middepth
21	3.10	surface
22	3.60	surface
...		
30	5.20	surface

Test statistic

Does there appear to be a lot of variability within groups? How about between groups?

$$F = \frac{variability\ bet.\ groups}{variability\ within\ groups}$$



Measuring variability

Total:

$$SST = \sum_{i} (x_i - \bar{x})^2$$

Between Groups:

$$SSG = \sum_{i=1}^{p} n_i(\bar{x_i} - \bar{x})^2$$

Residual:

$$SSE = \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x_i}\right)^2$$

Measuring variability

Total:

$$SST = \sum_{i} (x_i - \bar{x})^2$$

Between Groups:

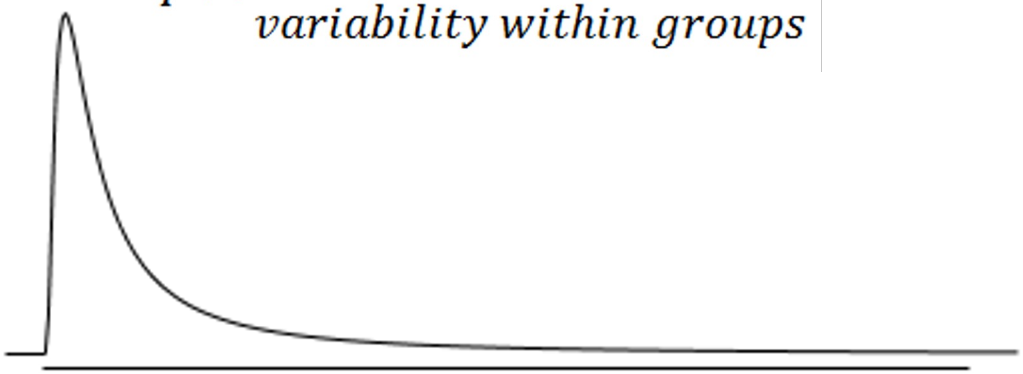
$$SSG = \sum_{i=1}^{p} n_i(\overline{x_i} \; - \bar{x})^2$$

Residual:

$$SSE = \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(x_{ij} - \overline{x_i}\right)^2$$

$$SST = SSG + SSE$$

F distribution and p-value



$$F = \frac{variability\ bet.\,groups}{variability\ within\ groups}$$

- Large values of the F statistic lead to small p-values, which leads to rejecting In order to be able to reject H_0, we need a small p-value, which requires a large F statistic
- In order to obtain a large F statistic, variability between sample means needs to be greater than variability within sample means

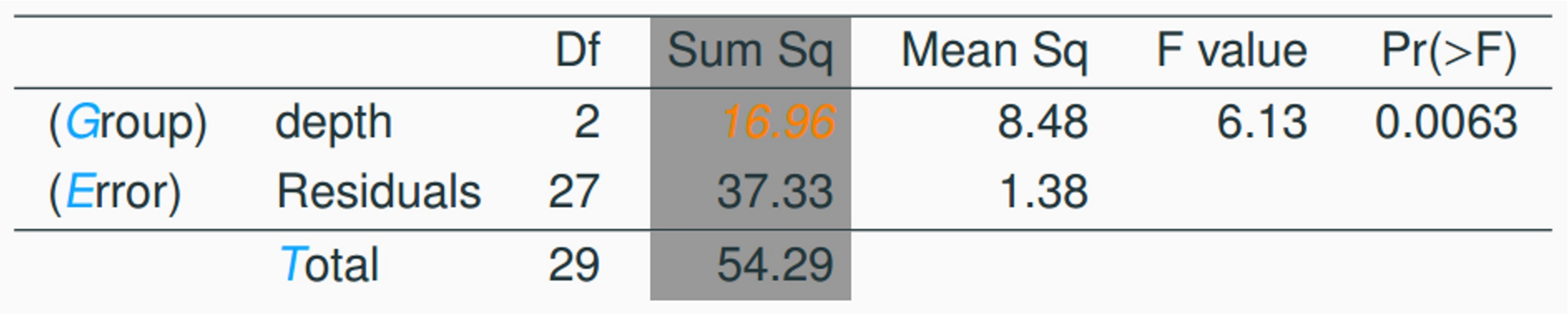
Theorem

Suppose $\mu_1 = \mu_2 = \cdots = \mu_k$

Then

$$F = \frac{SSG/(k-1)}{SSE/(n-k)}$$

has the F distribution with $k-1$ and $n-k$ degrees of freedom



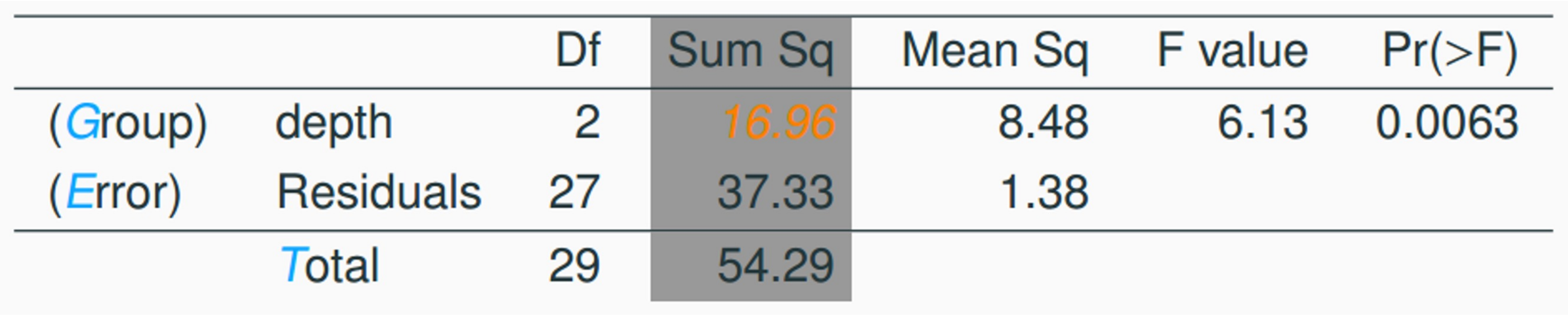
		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares between groups, SSG

Measures the variability between groups

$$SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

where is each group size, $\bar{x}_i$ is the average for each group, $\bar{x}$ is the overall (grand) mean



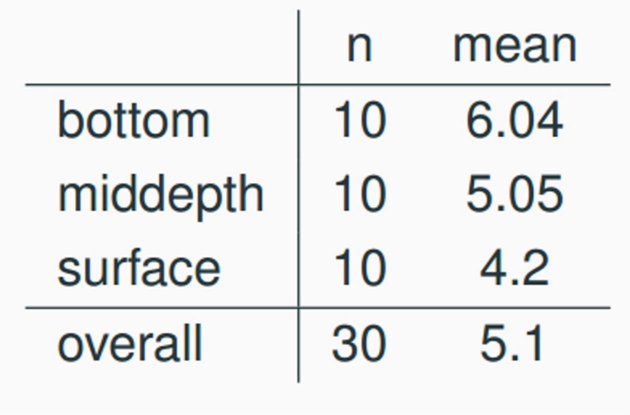
		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares between groups, SSG

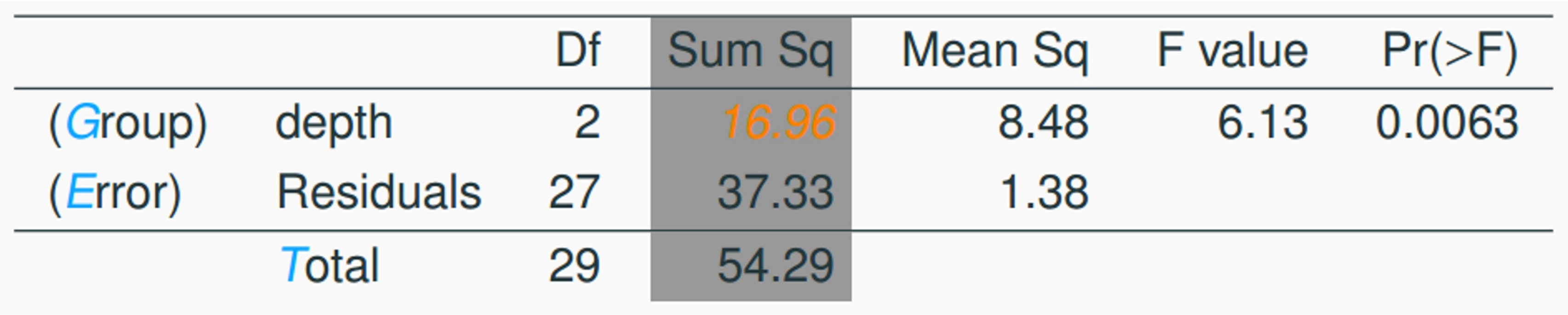
Measures the variability between groups

$$SSG = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2$$

where is each group size, $\bar{x}_i$ is the average for each group, $\bar{x}$ is the overall (grand) mean



	n	mean
bottom	10	6.04
middepth	10	5.05
surface	10	4.2
overall	30	5.1



		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

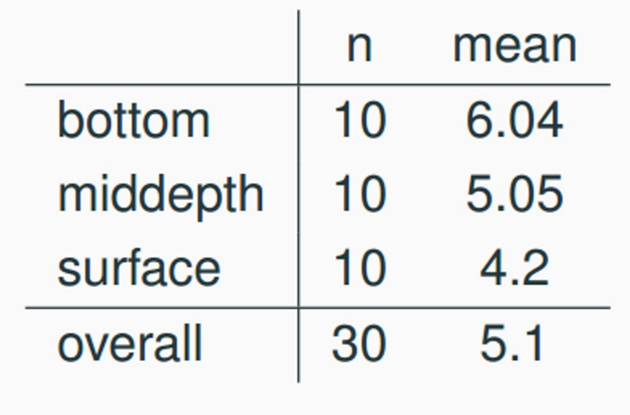
Sum of squares between groups, SSG

Measures the variability between groups

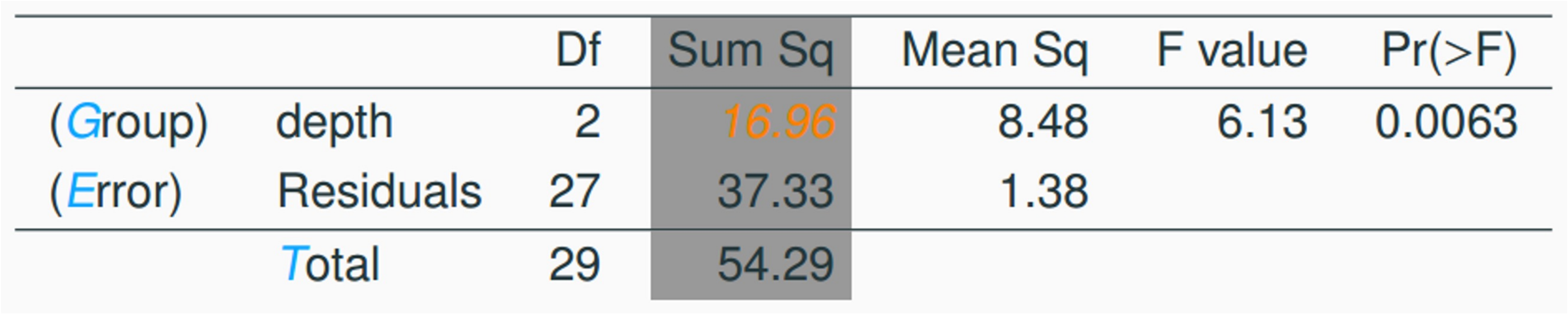
$$SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

where is each group size, $\bar{x}_i$ is the average for each group, $\bar{x}$ is the overall (grand) mean

$$SSG = (10 \times (6.04 - 5.1)^2)$$



	n	mean
bottom	10	6.04
middepth	10	5.05
surface	10	4.2
overall	30	5.1



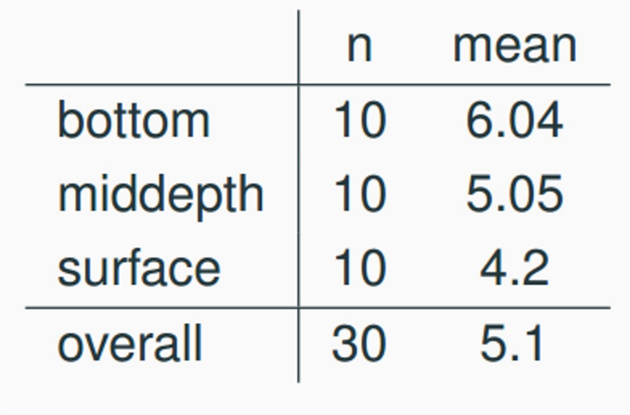
		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares between groups, SSG

Measures the variability between groups

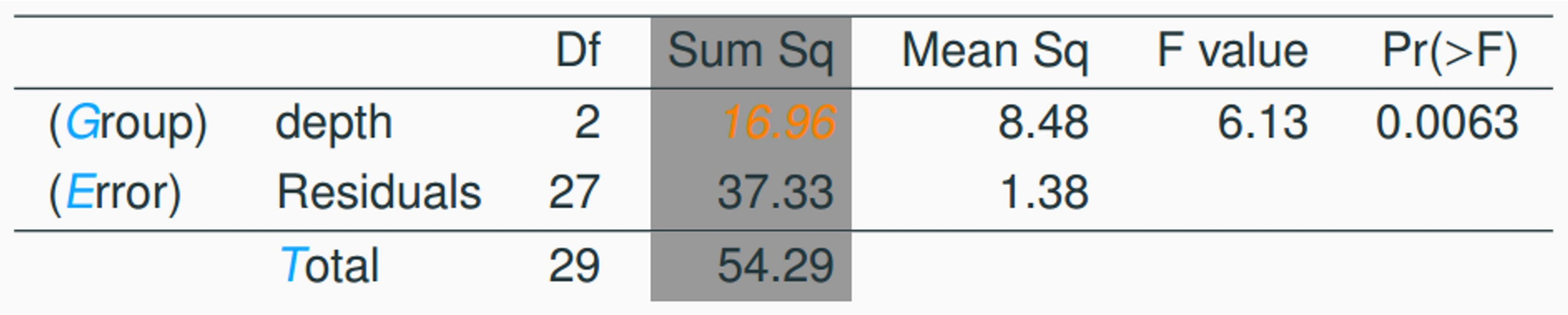
$$SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

where is each group size, $\bar{x}_i$ is the average for each group, $\bar{x}$ is the overall (grand) mean



	n	mean
bottom	10	6.04
middepth	10	5.05
surface	10	4.2
overall	30	5.1

$$SSG = (10 \times (6.04 - 5.1)^2)$$
$$+ (10 \times (5.05 - 5.1)^2)$$



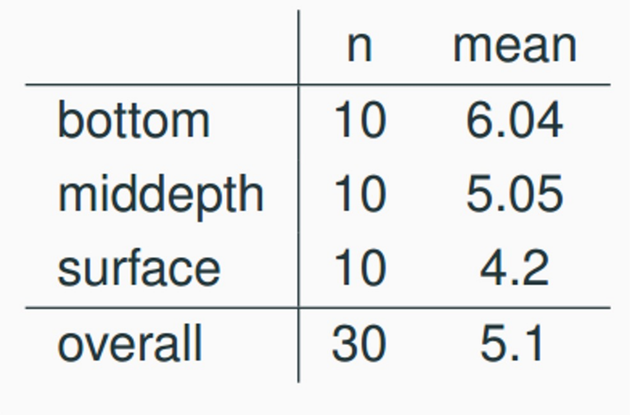
		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares between groups, SSG

Measures the variability between groups

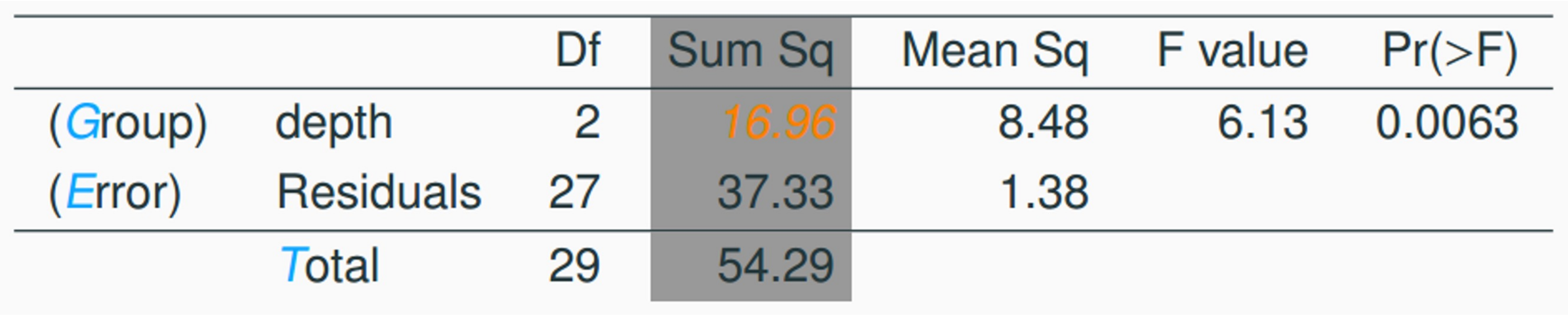
$$SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

where is each group size, $\bar{x}_i$ is the average for each group, $\bar{x}$ is the overall (grand) mean



	n	mean
bottom	10	6.04
middepth	10	5.05
surface	10	4.2
overall	30	5.1

$$SSG = (10 \times (6.04 - 5.1)^2)$$
$$+ (10 \times (5.05 - 5.1)^2)$$
$$+ (10 \times (4.2 - 5.1)^2)$$



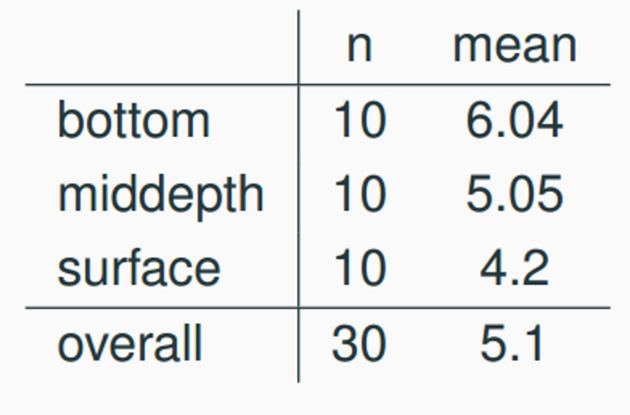
		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares between groups, SSG

Measures the variability between groups

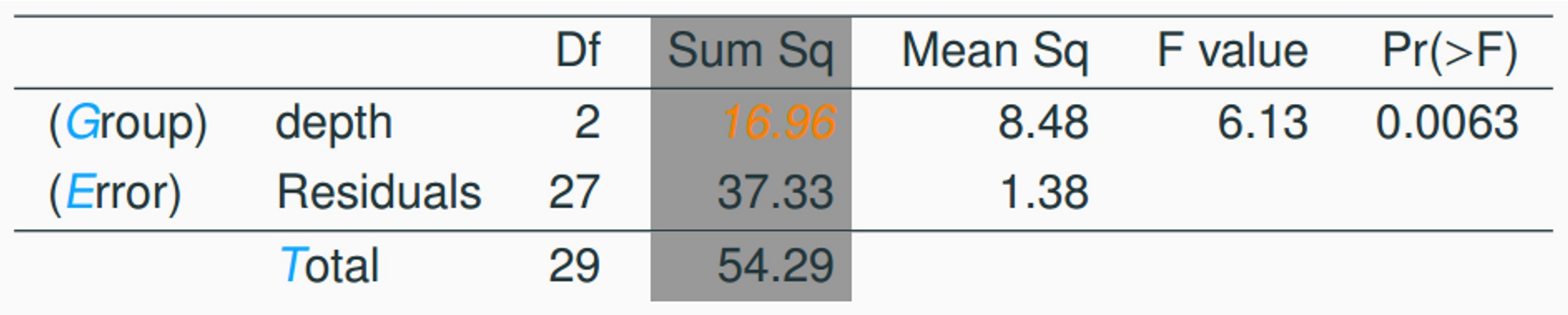
$$SSG = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2$$

where is each group size, $\bar{x}_i$ is the average for each group, $\bar{x}$ is the overall (grand) mean



	n	mean
bottom	10	6.04
middepth	10	5.05
surface	10	4.2
overall	30	5.1

$$\begin{aligned} SSG = {} & (10 \times (6.04 - 5.1)^2) \\ & + (10 \times (5.05 - 5.1)^2) \\ & + (10 \times (4.2 - 5.1)^2) \end{aligned}$$



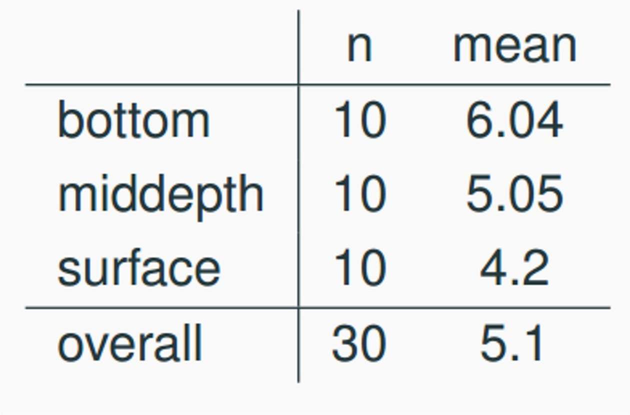
		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares between groups, SSG

Measures the variability between groups

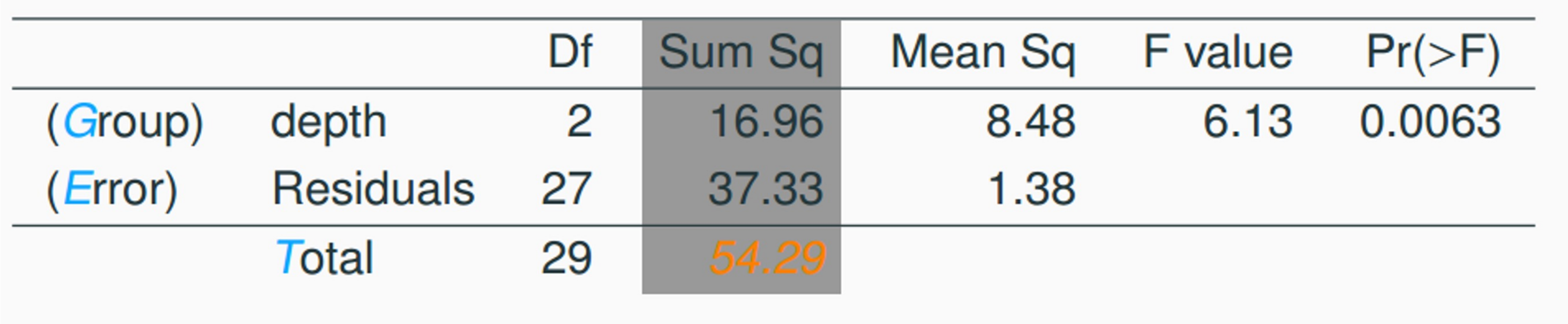
$$SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

where is each group size, $\bar{x}_i$ is the average for each group, $\bar{x}$ is the overall (grand) mean



	n	mean
bottom	10	6.04
middepth	10	5.05
surface	10	4.2
overall	30	5.1

$$\begin{aligned} SSG = &\ (10 \times (6.04 - 5.1)^2) \\ &+ (10 \times (5.05 - 5.1)^2) \\ &+ (10 \times (4.2 - 5.1)^2) \\ &= 16.96 \end{aligned}$$



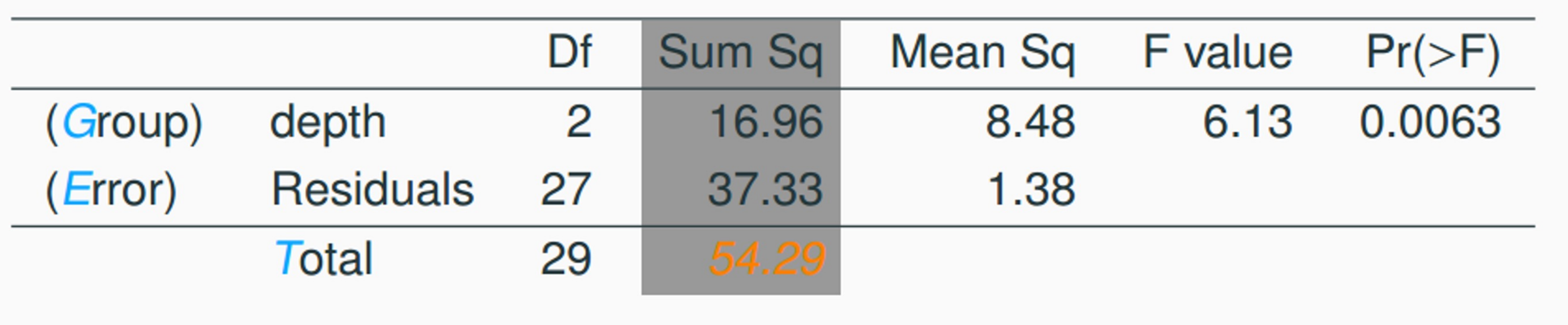
		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares total, SST

Measures the variability between groups

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

where x_i represent each observation in the dataset



		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	*54.29*			

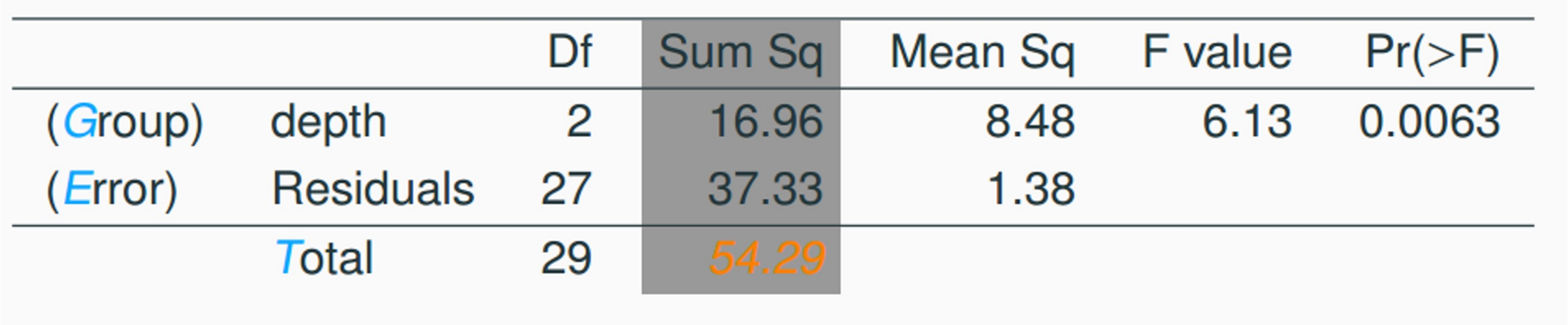
Sum of squares total, SST

Measures the variability between groups

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

where x_i represent each observation in the dataset

$SST = (3.8 - 5.1)^2 + (4.8 - 5.1)^2 + (4.9 - 5.1)^2 + \ldots + (5.2 - 5.1)^2$



		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

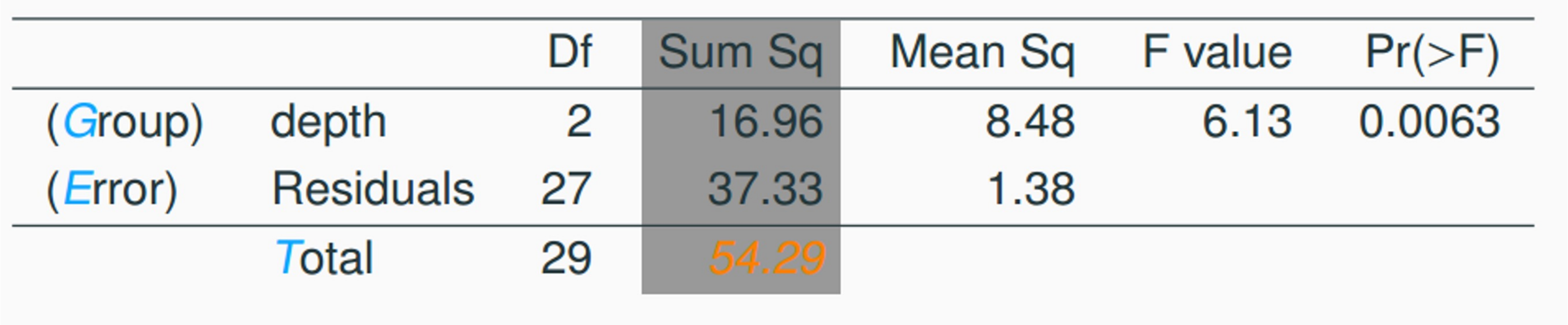
Sum of squares total, SST

Measures the variability between groups

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

where x_i represent each observation in the dataset

$$\begin{aligned} SST &= (3.8 - 5.1)^2 + (4.8 - 5.1)^2 + (4.9 - 5.1)^2 + \ldots + (5.2 - 5.1)^2 \\ &= (-1.3)^2 + (-0.3)^2 + (-0.2)^2 + \ldots + (0.1)^2 \end{aligned}$$



		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

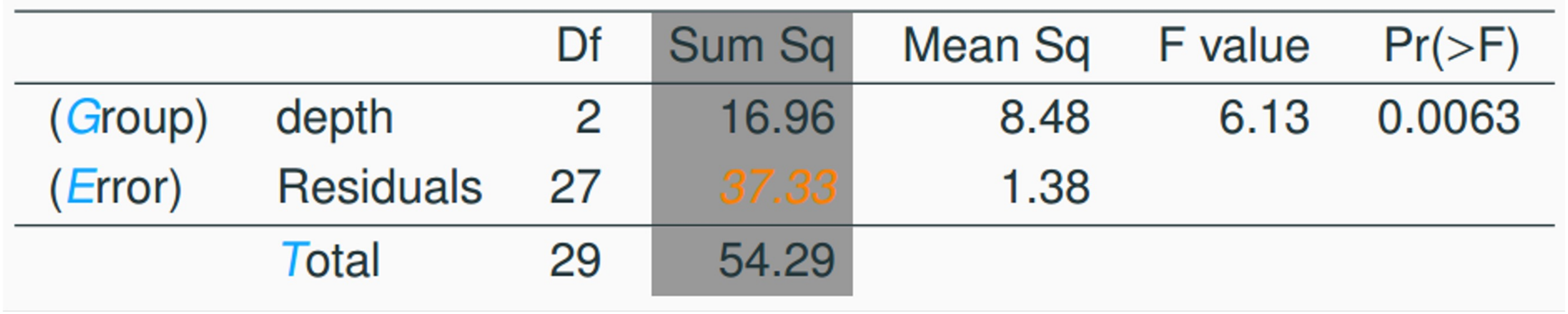
Sum of squares total, SST

Measures the variability between groups

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

where x_i represent each observation in the dataset

$$\begin{aligned} SST &= (3.8 - 5.1)^2 + (4.8 - 5.1)^2 + (4.9 - 5.1)^2 + \ldots + (5.2 - 5.1)^2 \\ &= (-1.3)^2 + (-0.3)^2 + (-0.2)^2 + \ldots + (0.1)^2 \\ &= 1.69 + 0.09 + 0.04 + \ldots + 0.01 \\ &= 54.29 \end{aligned}$$

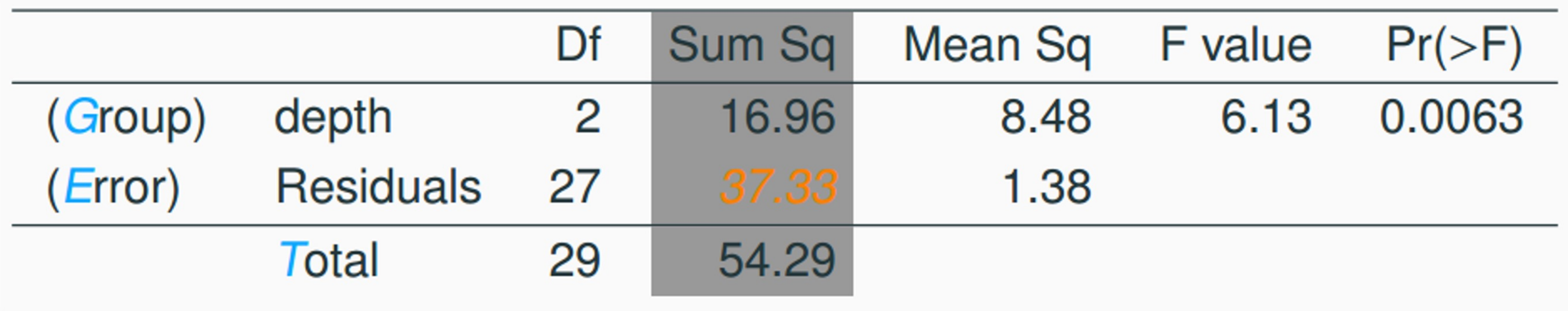


		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares error, SSE

Measures the variability within groups:

$$SSE = SST - SSG$$



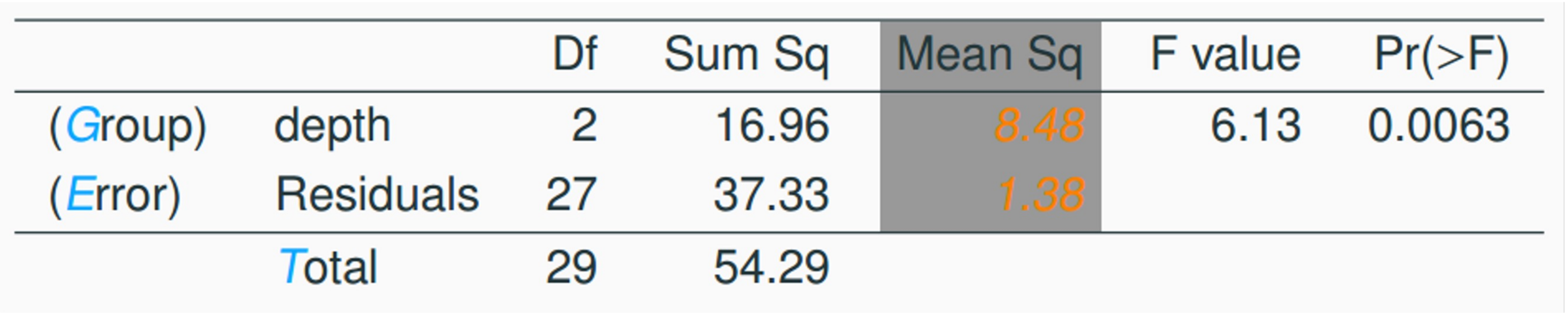
		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares error, SSE

Measures the variability within groups:

$$SSE = SST - SSG$$

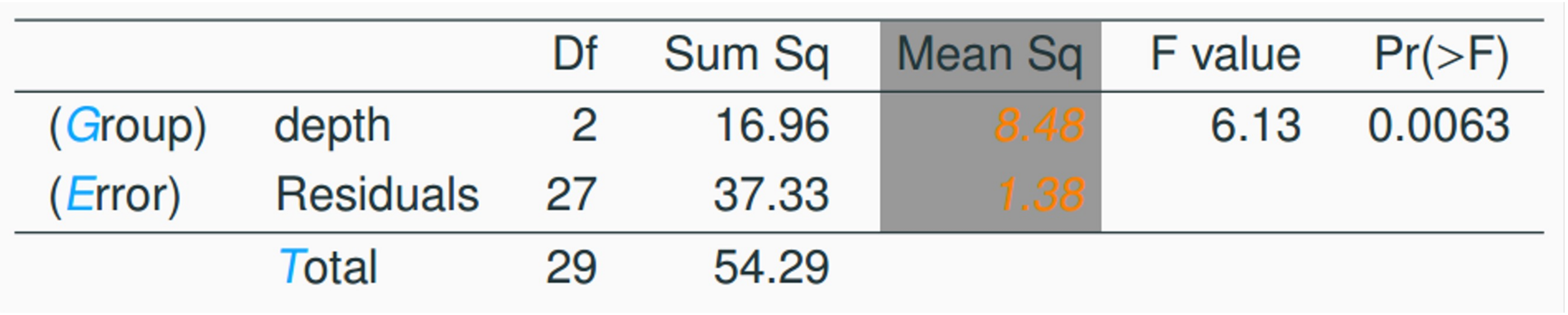
$$SSE = 54.29 - 16.96 = 37.33$$



		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Mean squared error

Mean squared error is calculated as sum of squares divided by the degrees of freedom

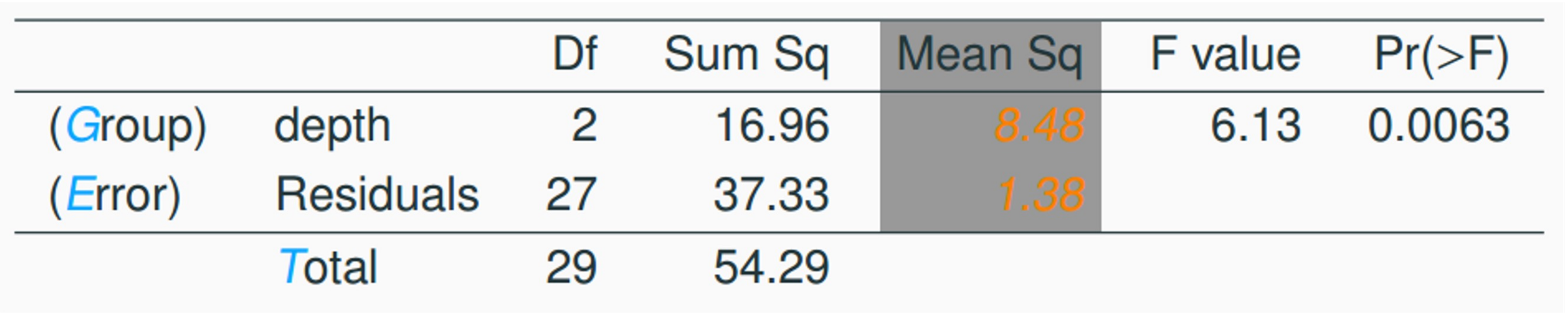


		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Mean squared error

Mean squared error is calculated as sum of squares divided by the degrees of freedom

$$MSG = 16.96/2 = 8.48$$

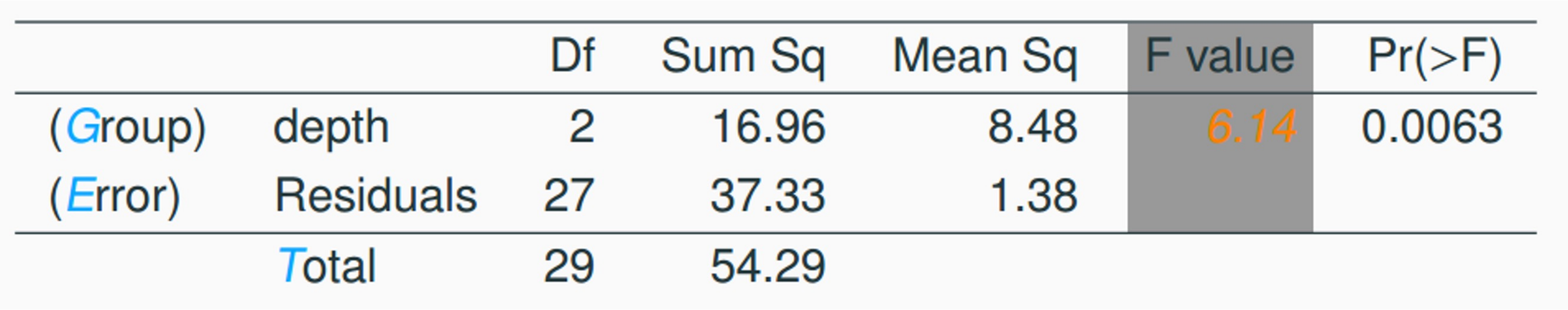


		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Mean square error

Mean square error is calculated as sum of squares divided by the degrees of freedom

$$MSG = 16.96/2 = 8.48$$
$$MSE = 37.33/27 = 1.38$$

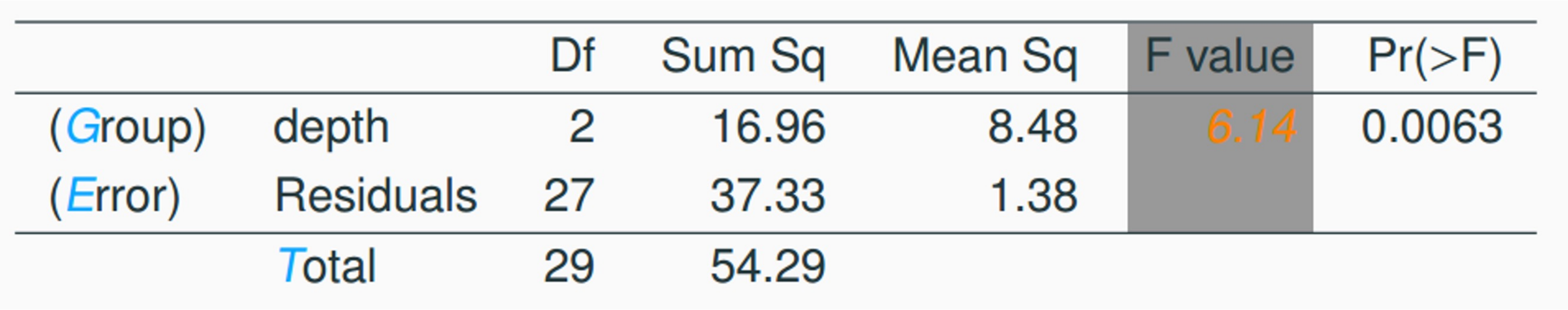


		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.14	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Test statistic, *F* value

As we discussed before, the *F* statistic is the ratio of the between group and within group variability

$$F = \frac{MSG}{MSE}$$

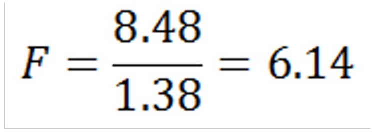


		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.14	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

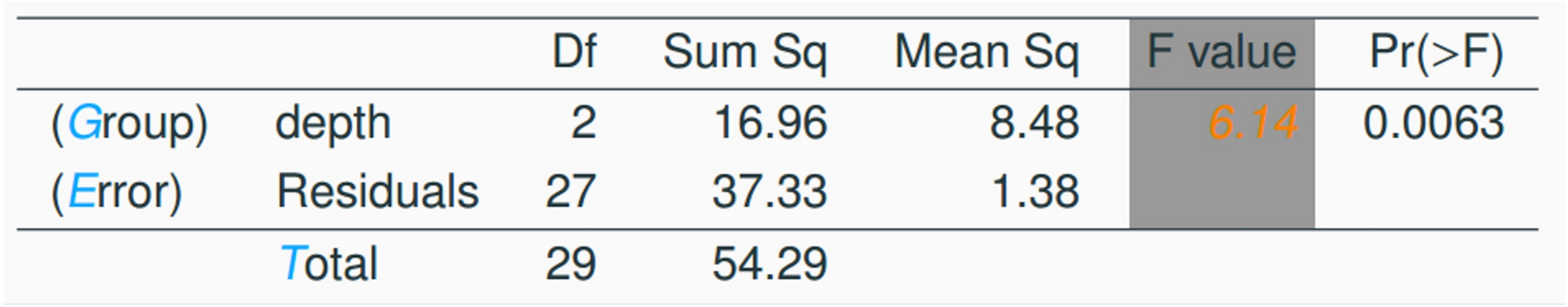
Test statistic, *F* value

As we discussed before, the *F* statistic is the ratio of the between group and within group variability

$$F = \frac{MSG}{MSE}$$



$$F = \frac{8.48}{1.38} = 6.14$$



		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.14	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

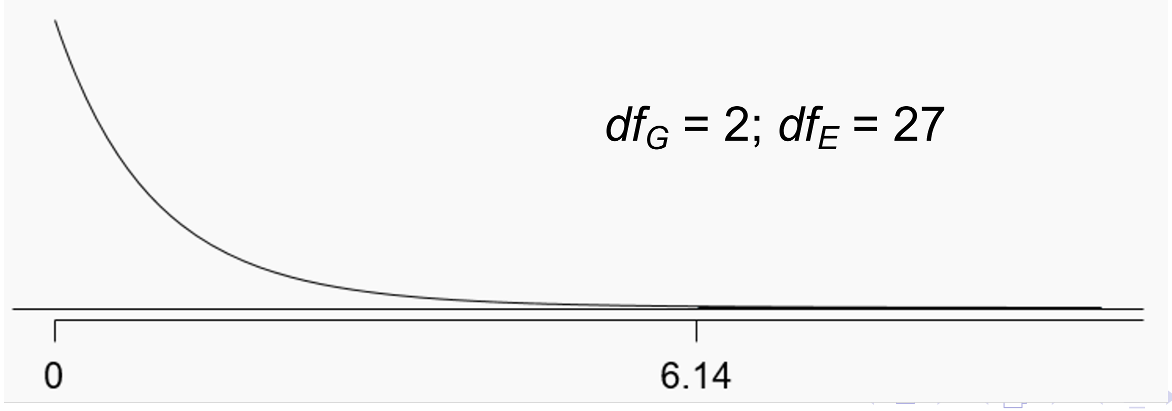
p-value

p-value is the probability of at least as large a ratio between the “between group” and “within group” variability, if in fact the means of all groups are equal. It’s calculated as the area under the F curve, with degrees of freedom df_G and df_E, above the observed F statistic.

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	6.14	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

p-value

p-value is the probability of at least as large a ratio between the "between group" and "within group" variability, if in fact the means of all groups are equal. It's calculated as the area under the *F* curve, with degrees of freedom df_G and df_E, above the observed *F* statistic.



Conclusion - in context

What is the conclusion of the hypothesis test?

The data provide convincing evidence that the average aldrin concentration

A. is different for all groups
B. on the surface is lower than the other levels
C. is different for at least one group
D. is the same for all groups

Conclusion - in context

What is the conclusion of the hypothesis test?

The data provide convincing evidence that the average aldrin concentration

A. is different for all groups
B. on the surface is lower than the other levels
C. is different for at least one group
D. is the same for all groups

Conclusion

- If p-value is small (less than α), reject H_0. The data provide convincing evidence that at least one mean is different from (but we can’t tell which one)

Conclusion

- If p-value is small (less than α), reject H_0. The data provide convincing evidence that at least one mean is different from (but we can’t tell which one)
- If p-value is large, fail to reject H_0. The data do not provide convincing evidence that at least one pair of means are different from each other, the observed differences in sample means are attributable to sampling variability (or chance)

Conditions

1. The observations should be independent within and between groups
   - If the data are a simple random sample from less than 10% of the population, this condition is satisfied
   - Carefully consider whether the data may be independent (e.g. no pairing)
   - Always important, but sometimes difficult to check

Conditions

1. The observations should be independent within and between groups
   - If the data are a simple random sample from less than 10% of the population, this condition is satisfied
   - Carefully consider whether the data may be independent (e.g. no pairing)
   - Always important, but sometimes difficult to check

2. The observations within each group should be nearly normal
   - Especially important when the sample sizes are small

How do we check for normality?

Conditions

1. The observations should be independent within and between groups
   - If the data are a simple random sample from less than 10% of the population, this condition is satisfied
   - Carefully consider whether the data may be independent (e.g. no pairing)
   - Always important, but sometimes difficult to check

2. The observations within each group should be nearly normal
   - Especially important when the sample sizes are small

How do we check for normality?

3. The variability across the groups should be about equal
   - Especially important when the sample sizes differ between groups

How can we check this condition?

(1)independence

Does this condition appear to be satisfied?

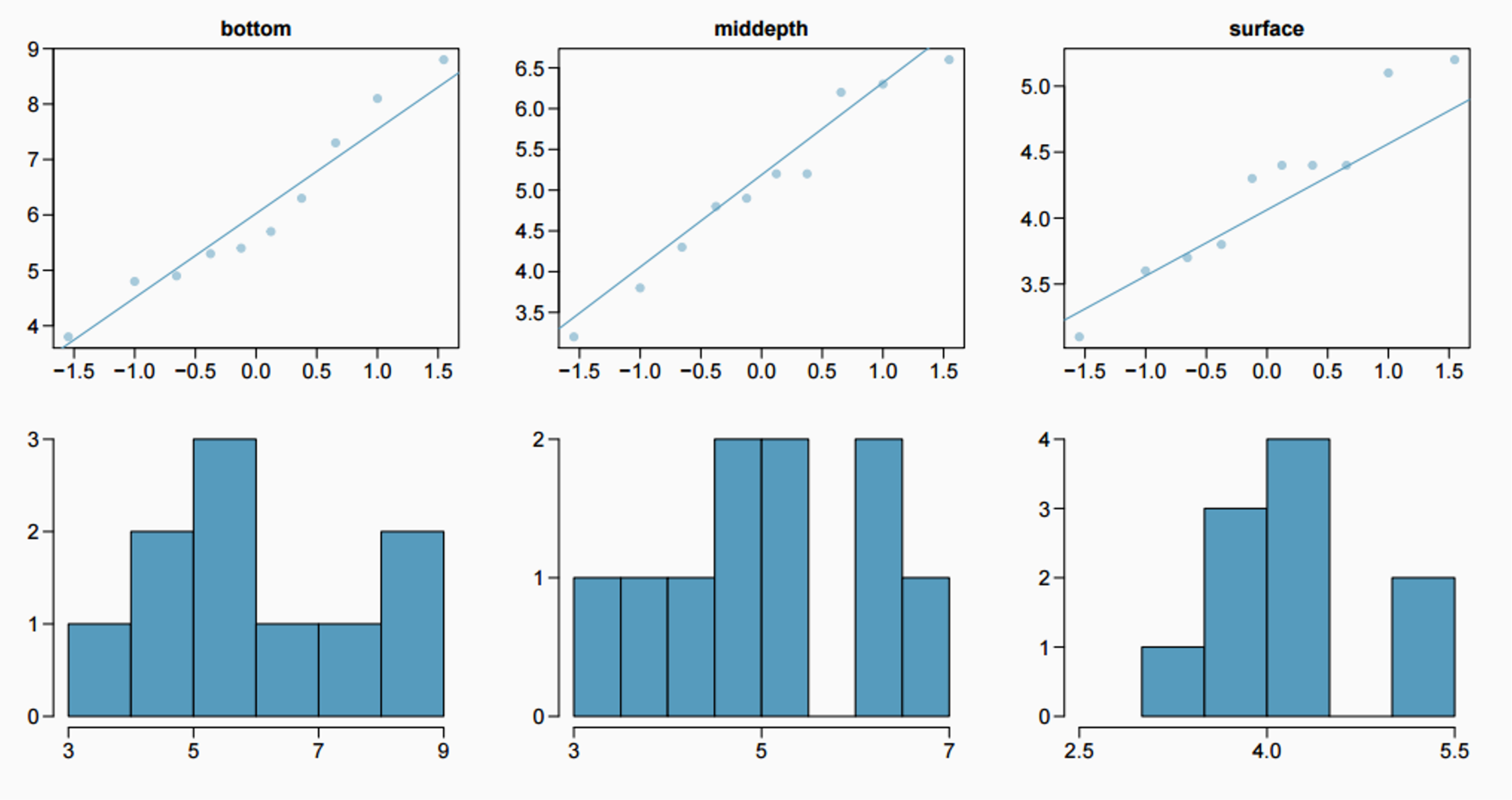
(1)independence

Does this condition appear to be satisfied?

In this study the we have no reason to believe that the aldrin concentration won't be independent of each other

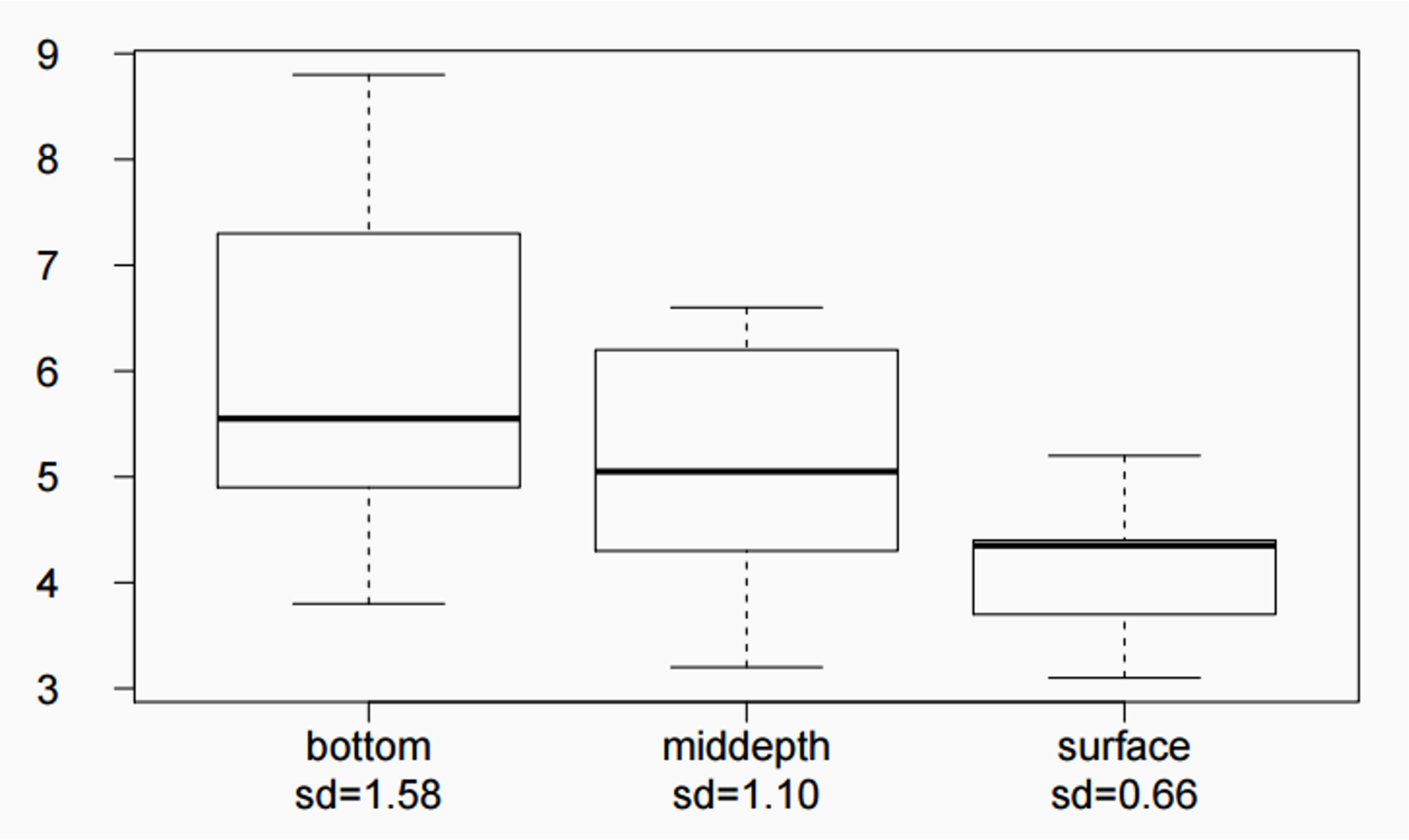
(2)approximately normal

Does this condition appear to be satisfied?



(3)constant variance

Does this condition appear to be satisfied?



Which means differ?

- Earlier we concluded that at least one pair of means differ. The natural question that follows is “which ones?”

Which means differ?

- Earlier we concluded that at least one pair of means differ. The natural question that follows is “which ones?”
- We can do two sample t tests for differences in each possible pair of groups

Which means differ?

- Earlier we concluded that at least one pair of means differ. The natural question that follows is “which ones?”
- We can do two sample t tests for differences in each possible pair of groups

Can you see any pitfalls with this approach?

Which means differ?

- Earlier we concluded that at least one pair of means differ. The natural question that follows is “which ones?”
- We can do two sample t tests for differences in each possible pair of groups

Can you see any pitfalls with this approach?

- When we run too many tests, the Family-wise error rate increases
- We can use: Corrections for multiple comparisons (e.g., Bonferroni)
- Tukey-Kramer tests perform all pairwise comparisons while controlling for FWER at level α

Why not just use pairwise comparisons?

- Controlling for family-wise error rate is conservative

- It may be the case that we end up getting no significant p-values in pairwise comparisons, but a significant ANOVA p-value