Inference for Comparing Many Means

Dr. Ab Mosca (they/them)

Slides based off slides courtesy of OpenIntro and John McGreedy of Johns Hopkins University

- 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

- 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

- 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

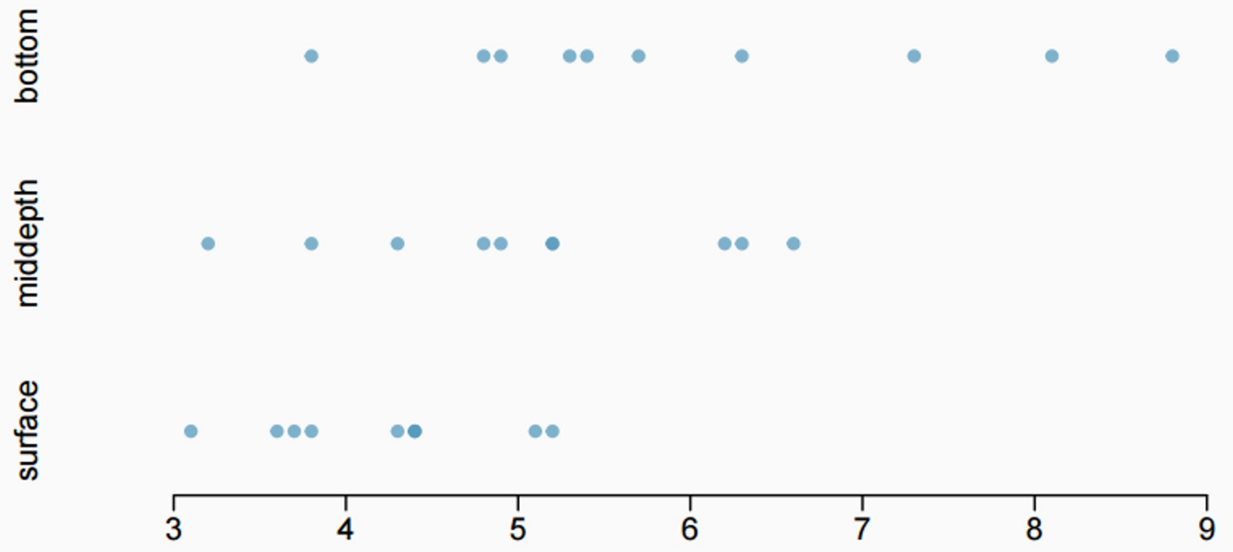
Data

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

Exploratory analysis

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



	n	mean	sd
bottom	10	6.04	1.58
middepth	10	5.05	1.10
surface	10	4.20	0.66
overall	30	5.1 0	1.37

Research question

Is there a difference between the mean aldrin concentrations among the three levels?

Research question

Is there a difference between the mean aldrin concentrations among the three levels?

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

Research question

Is there a difference between the mean aldrin concentrations among the three levels?

- 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

ANOVA

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

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

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

(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

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

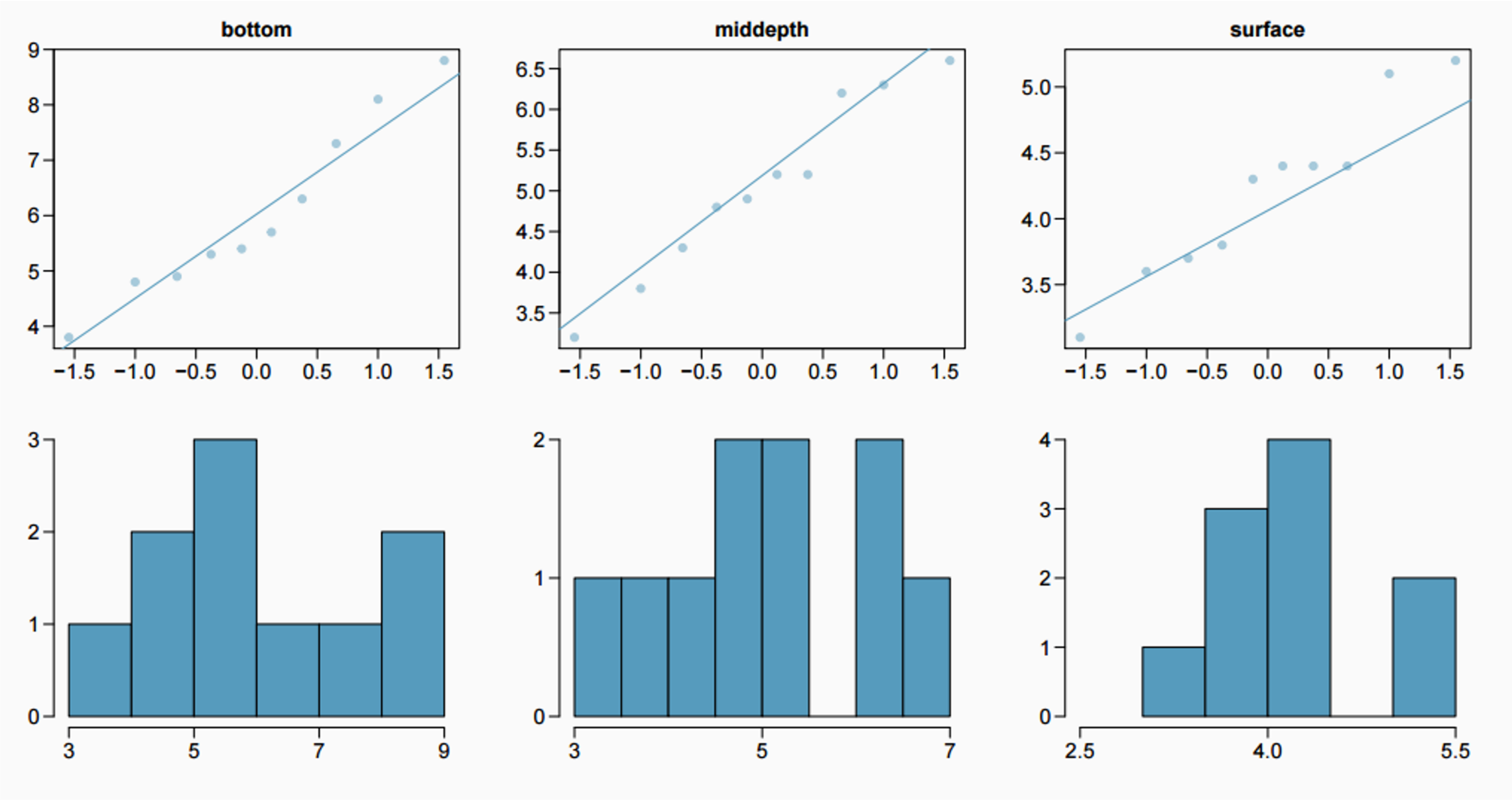
1. The observations should be independent within and between groups
    - If the data are a simple random sample from less than 10% of the population, this condition is satisfied
    - Carefully consider whether the data may be independent (e.g. no pairing)
    - Always important, but sometimes difficult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

(2)approximately normal

Does this condition appear to be satisfied?



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? – Look at distributions, consider data

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?

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

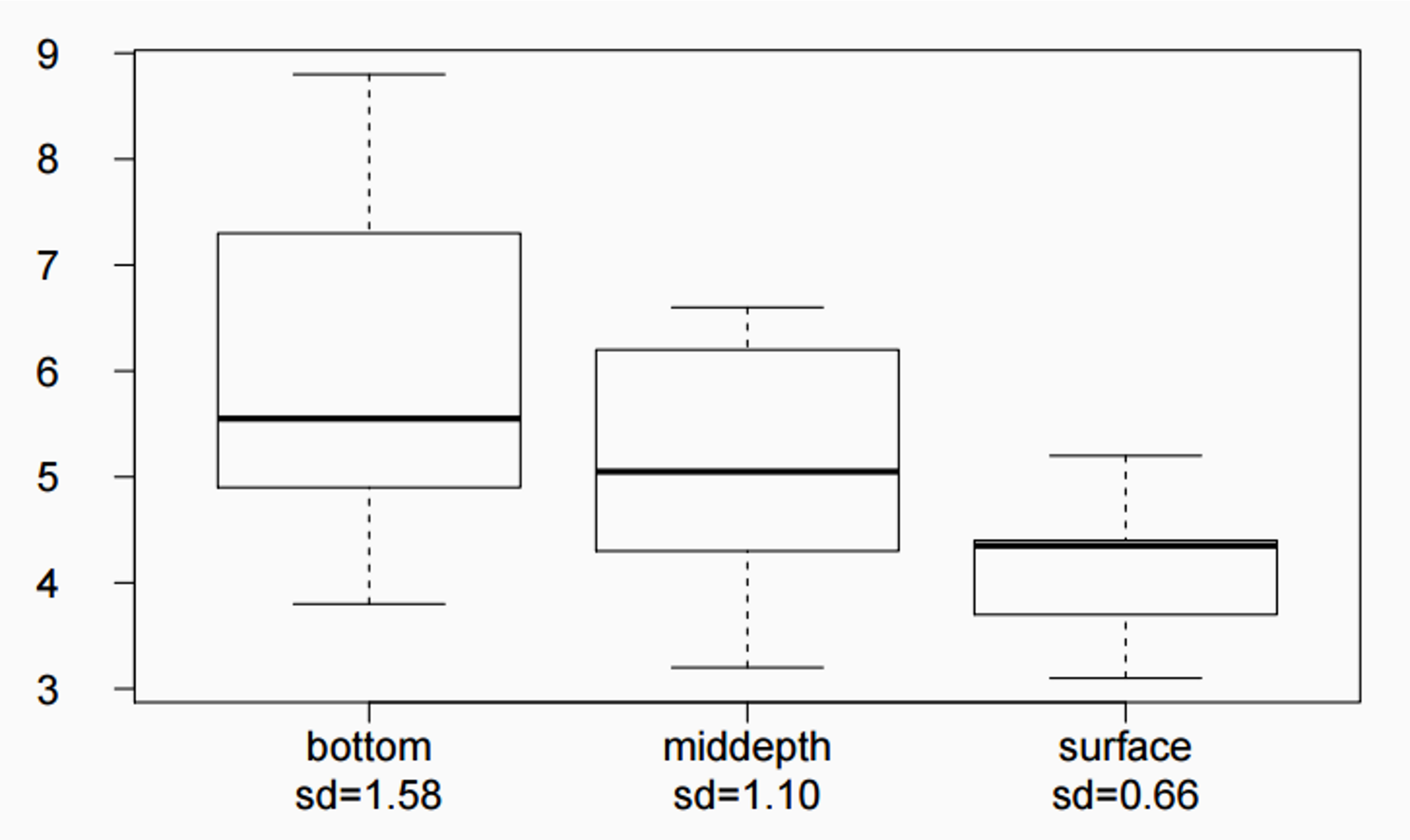
How do we check for normality? – Look at distributions, consider data

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? – Look at variance in each group

(3)constant variance

Does this condition appear to be satisfied?



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

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

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

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 (for surface, middle, bottom measurements)

A. $H_0 : \mu_B = \mu_M = \mu_S$

$H_A : \mu_B \neq \mu_M \neq \mu_S$

B. $H_0 : \mu_B \neq \mu_M \neq \mu_S$

$H_A : \mu_B = \mu_M = \mu_S$

C. $H_0 : \mu_B = \mu_M = \mu_S$

H_A : At least one mean is different

D. $H_0 : \mu_B = \mu_M = \mu_S = 0$

H_A : At least one mean is different

E. $H_0 : \mu_B = \mu_M = \mu_S$

$H_A : \mu_B > \mu_M > \mu_S$

Hypotheses (for surface, middle, bottom measurements)

A. $H_0 : \mu_B = \mu_M = \mu_S$

$H_A : \mu_B \neq \mu_M \neq \mu_S$

B. $H_0 : \mu_B \neq \mu_M \neq \mu_S$

$H_A : \mu_B = \mu_M = \mu_S$

C. $H_0 : \mu_B = \mu_M = \mu_S$

H_A : *At least one mean is different*

D. $H_0 : \mu_B = \mu_M = \mu_S = 0$

H_A : At least one mean is different

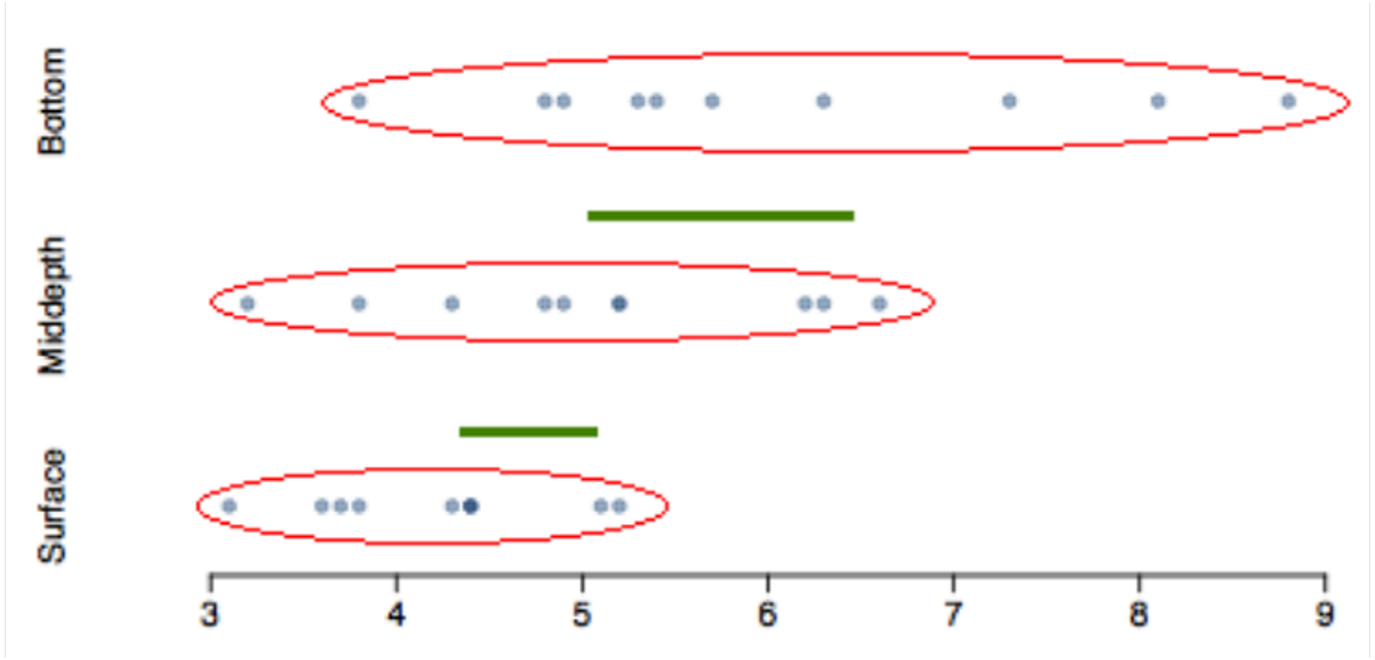
E. $H_0 : \mu_B = \mu_M = \mu_S$

$H_A : \mu_B > \mu_M > \mu_S$

Test statistic

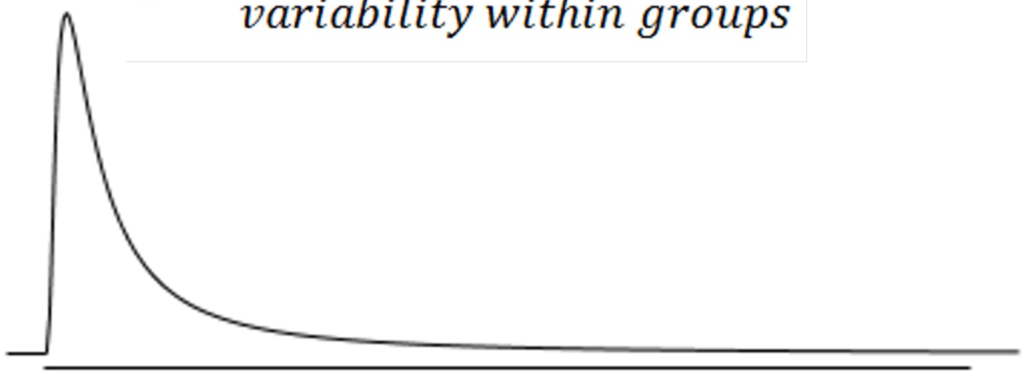
Does there appear to be a lot of variability within groups? How about between groups?

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



F distribution and p-value

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



- 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

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(*G*roup)	depth	2	16.96	8.48	6.13	0.0063
(*E*rror)	Residuals	27	37.33	1.38		
	*T*otal	29	54.29			

Degrees of freedom associated with ANOVA

- groups: $df_G = k - 1$, where k is the number of groups
- total: $df_T = n - 1$, where n is the total sample size
- error: $df_E = df_T - df_G$

		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			

Degrees of freedom associated with ANOVA

- groups: $df_G = k - 1$, where k is the number of groups
- total: $df_T = n - 1$, where n is the total sample size
- error: $df_E = df_T - df_G$

Calculate the various 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			

Degrees of freedom associated with ANOVA

- groups: $df_G = k - 1$, where k is the number of groups
- total: $df_T = n - 1$, where n is the total sample size
- error: $df_E = df_T - df_G$

Calculate the various degrees of freedom

- $df_G = k - 1 = 3 - 1 = 2$
- $df_T = n - 1 = 30 - 1 = 29$
- $df_E = 29 - 2 = 27$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(*G*roup)	depth	2	16.96	8.48	6.13	0.0063
(*E*rror)	Residuals	27	37.33	1.38		
	*T*otal	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 n_i 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

Calculate 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 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

Calculate SST

		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:

$$SSE = SST - SSG$$

Calculate SSE

		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:

$$SSE = SST - SSG$$

Calculate SSE

$$SSE = 54.29 - 16.96 = 37.33$$

		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 square error

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

Calculate Mean square error for Group and Error

		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

Calculate Mean square error for Group and Error

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

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(*G*roup)	depth	2	16.96	8.48	*6.14*	0.0063
(*E*rror)	Residuals	27	37.33	1.38		
	*T*otal	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

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

Calculate *F*

		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

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

Calculate *F*

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

		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			

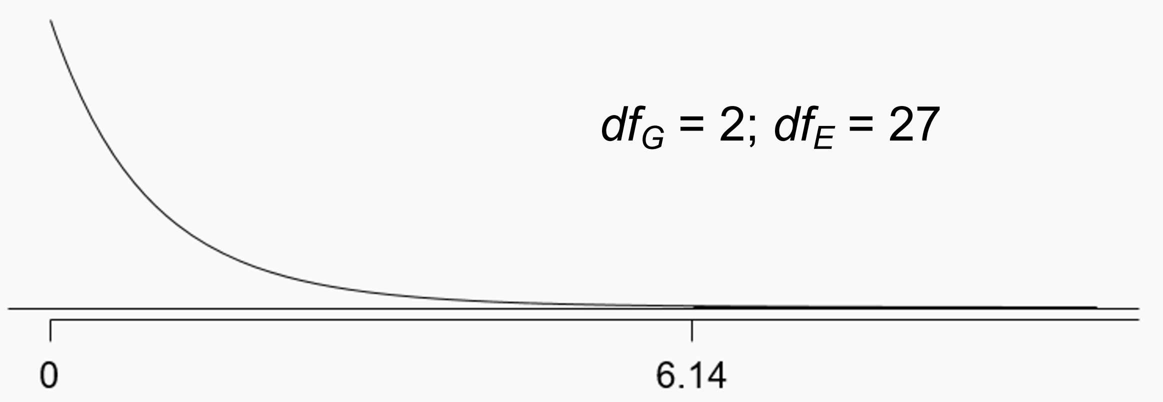
p-value

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

		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			

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

		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			

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)

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 Type 1 Error rate increases
- This issue is resolved by using a modified significance level

Multiple comparisons

- The scenario of testing many pairs of groups is called multiple comparisons

Multiple comparisons

- The scenario of testing many pairs of groups is called multiple comparisons
- The **Bonferroni correction** suggests that a more stringent significance level is more appropriate for these tests:

$$\alpha^* = \alpha / K$$

where K is the number of comparisons being considered

Multiple comparisons

- The scenario of testing many pairs of groups is called multiple comparisons
- The **Bonferroni correction** suggests that a more stringent significance level is more appropriate for these tests:

$$\alpha^* = \alpha / K$$

where K is the number of comparisons being considered

- If there are k groups, then usually all possible pairs are compared and $K = \frac{k(k-1)}{2}$

Determining the modified α

In the aldrin data set depth has 3 levels: bottom, mid-depth, and surface. If α = 0.05, what should be the modified significance level for two sample t tests for determining which pairs of groups have significantly different means?

$$K = \frac{k(k-1)}{2}$$

A. α* = 0.05
B. α* = 0.05/2 = 0.025
C. α* = 0.05/3 = 0.0167
D. α* = 0.05/6 = 0.0083

Determining the modified α

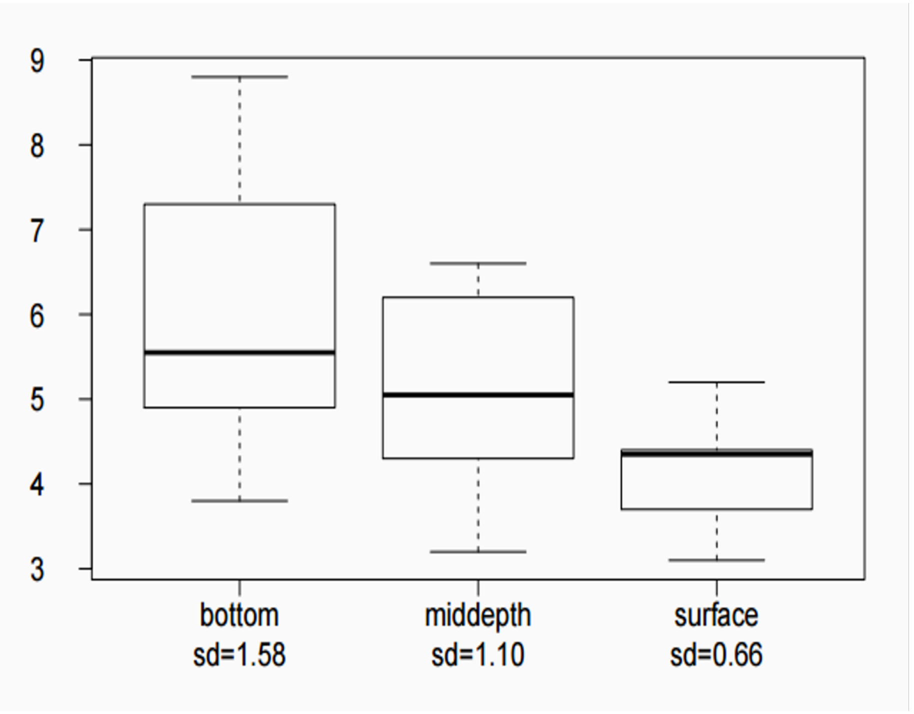
In the aldrin data set depth has 3 levels: bottom, mid-depth, and surface. If $\alpha = 0.05$, what should be the modified significance level for two sample t tests for determining which pairs of groups have significantly different means?

$$K = \frac{k(k-1)}{2}$$

A. $\alpha^* = 0.05$
B. $\alpha^* = 0.05/2 = 0.025$
C. *$\alpha^* = 0.05/3 = 0.0167$*
D. $\alpha^* = 0.05/6 = 0.0083$

Which means differ?

Based on the box plots below, which means would you expect to be significantly different?



A. bottom & surface

B. bottom & mid-depth

C. mid-depth & surface

D. bottom & mid-depth;
   mid-depth & surface

E. bottom & mid-depth;
   bottom & surface;
   mid-depth & surface

Which means differ? (cont.)

If the ANOVA assumption of equal variability across groups is satisfied, we can use the data from all groups to estimate variability:

- Estimate any within-group standard deviation with $\sqrt{MSE}$, which is s_{pooled}
- Use the error degrees of freedom, $n - k$, for t-distributions

Difference in two means: after ANOVA

$$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{MSE}{n_1} + \frac{MSE}{n_2}}$$

Is there a difference between the average aldrin concentration at the bottom and at mid depth?

	n	mean	sd
bottom	10	6.04	1.58
middepth	10	5.05	1.10
surface	10	4.2	0.66
overall	30	5.1	1.37

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

$$T_{df_E} = \frac{\left(\bar{x}_{bottom} - \bar{x}_{middepth}\right)}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{middepth}}}}$$

Is there a difference between the average aldrin concentration at the bottom and at mid depth?

	n	mean	sd
bottom	10	*6.04*	1.58
middepth	10	*5.05*	1.10
surface	10	4.2	0.66
overall	30	5.1	1.37

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

$$T_{df_E} = \frac{(\bar{x}_{bottom} - \bar{x}_{middepth})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{middepth}}}}$$

$$T_{27} = \frac{(6.04 - 5.05)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{0.99}{0.53} = 1.87$$

Is there a difference between the average aldrin concentration at the bottom and at mid depth?

	n	mean	sd
bottom	10	6.04	1.58
middepth	10	5.05	1.10
surface	10	4.2	0.66
overall	30	5.1	1.37

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

$$T_{df_E} = \frac{(\bar{x}_{bottom} - \bar{x}_{middepth})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{middepth}}}}$$

$$T_{27} = \frac{(6.04 - 5.05)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{0.99}{0.53} = 1.87$$

$$0.05 < p - value < 0.10 \quad (two - sided)$$

Is there a difference between the average aldrin concentration at the bottom and at mid depth?

	n	mean	sd
bottom	10	6.04	1.58
middepth	10	5.05	1.10
surface	10	4.2	0.66
overall	30	5.1	1.37

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

$$T_{df_E} = \frac{(\bar{x}_{bottom} - \bar{x}_{middepth})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{middepth}}}}$$

$$T_{27} = \frac{(6.04 - 5.05)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{0.99}{0.53} = 1.87$$

$$0.05 < p - value < 0.10 \quad (two - sided)$$

$$\alpha^* = \frac{0.05}{3} = 0.0167$$

Is there a difference between the average aldrin concentration at the bottom and at mid depth?

	n	mean	sd
bottom	10	6.04	1.58
middepth	10	5.05	1.10
surface	10	4.2	0.66
overall	30	5.1	1.37

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

$$T_{df_E} = \frac{(\bar{x}_{bottom} - \bar{x}_{middepth})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{middepth}}}}$$

$$T_{27} = \frac{(6.04 - 5.05)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{0.99}{0.53} = 1.87$$

$$0.05 < p - value < 0.10 \quad (two - sided)$$

$$\alpha^* = \frac{0.05}{3} = 0.0167$$

Fail to reject H_0, the data do not provide convincing evidence of a difference between the average aldrin concentrations at bottom and mid depth

Pairwise comparisons

Is there a difference between the average aldrin concentration at the bottom and at surface?

	n	mean	sd
bottom	10	*6.04*	1.58
middepth	10	*5.05*	1.10
surface	10	4.2	0.66
overall	30	5.1	1.37

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

$$T_{df_E} = \frac{(\bar{x}_{bottom} - \bar{x}_{surface})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{surface}}}}$$

Pairwise comparisons

Is there a difference between the average aldrin concentration at the bottom and at surface?

	n	mean	sd
bottom	10	*6.04*	1.58
middepth	10	*5.05*	1.10
surface	10	4.2	0.66
overall	30	5.1	1.37

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

$$T_{df_E} = \frac{(\bar{x}_{bottom} - \bar{x}_{surface})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{surface}}}}$$

$$T_{27} = \frac{(6.04 - 4.2)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{1.84}{0.53} = 3.47$$

Pairwise comparisons

Is there a difference between the average aldrin concentration at the bottom and at surface?

	n	mean	sd
bottom	10	6.04	1.58
middepth	10	5.05	1.10
surface	10	4.2	0.66
overall	30	5.1	1.37

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

$$T_{df_E} = \frac{(\bar{x}_{bottom} - \bar{x}_{surface})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{surface}}}}$$

$$T_{27} = \frac{(6.04 - 4.2)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{1.84}{0.53} = 3.47$$

$$p - value < 0.01 \quad (two - sided)$$

Pairwise comparisons

Is there a difference between the average aldrin concentration at the bottom and at surface?

	n	mean	sd
bottom	10	*6.04*	1.58
middepth	10	*5.05*	1.10
surface	10	4.2	0.66
overall	30	5.1	1.37

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

$$T_{df_E} = \frac{(\bar{x}_{bottom} - \bar{x}_{surface})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{surface}}}}$$

$$T_{27} = \frac{(6.04 - 4.2)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{1.84}{0.53} = 3.47$$

$$p - value < 0.01 \quad (two - sided) \qquad \alpha^* = \frac{0.05}{3} = 0.0167$$

Pairwise comparisons

Is there a difference between the average aldrin concentration at the bottom and at surface?

	n	mean	sd
bottom	10	6.04	1.58
middepth	10	5.05	1.10
surface	10	4.2	0.66
overall	30	5.1	1.37

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

Reject H_0, the data provide convincing evidence of a difference between the average aldrin concentrations at bottom and surface

$$T_{df_E} = \frac{(\bar{x}_{bottom} - \bar{x}_{surface})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{surface}}}}$$

$$T_{27} = \frac{(6.04 - 4.2)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{1.84}{0.53} = 3.47$$

$$p - value < 0.01 \quad (two - sided) \qquad \alpha^* = \frac{0.05}{3} = 0.0167$$