Lecture 15 - ANOVA

Sta102 / BME102

Colin Rundel

October 26, 2015

Wolf River

- 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.
- Given that these compounds are denser than water, researchers believe that their molecules are likely to be deposited in sediment, and are thus more likely to be found in higher concentrations near the bottom of the river.

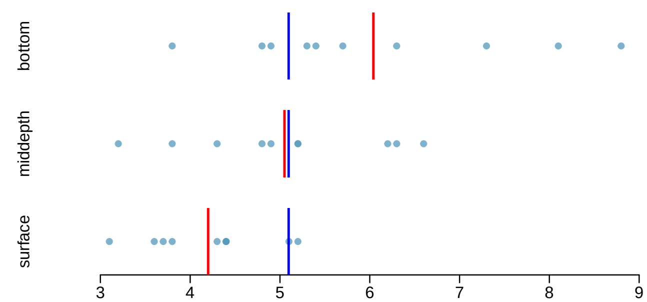
Wolf River - Data

Aldrin concentration (ng / L) 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	5.10	1.37

Research question

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

- To compare means of 2 groups we use a T statistic (this is the distribution of the sampling distribution).
- To compare means of 3 or more groups we use a new test called *ANOVA* (analysis of variance) and a new test statistic / sampling distribution - F.

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,

$$\mu_1 = \mu_2 = \cdots = \mu_k,$$

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

H_A : At least one pair of means are different.

Note - this hypothesis test does not tell us if all the means are different or only if one pair is different, more on how to do that later.

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.
   - Particularly important when the sample sizes are small.

   How do we check for normality?

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

   How can we check this condition?

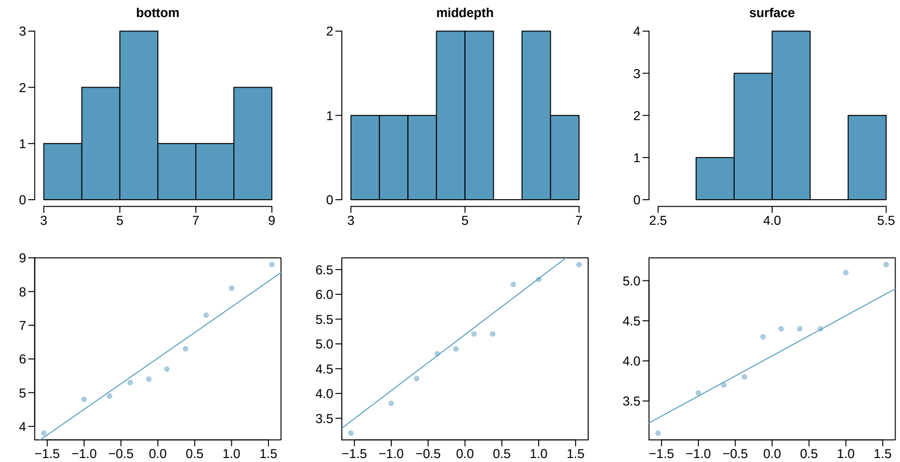
(1) Independence

Does this condition appear to be satisfied for the Wolf River data?

In this study the we have no reason to believe that the aldrin concentration measurements are dependent of each other.

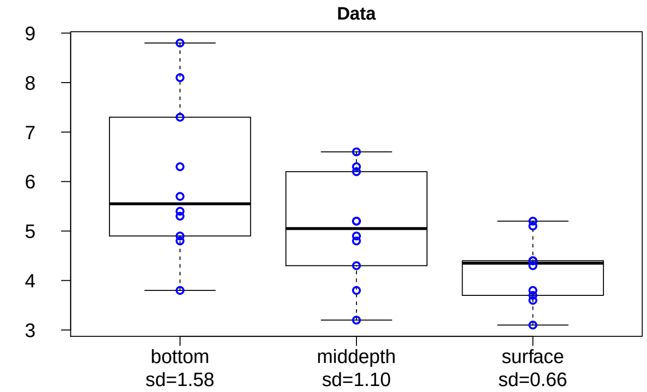
(2) Approximately normal

Does this condition appear to be satisfied?



(3) Constant variance

Does this condition appear to be satisfied?



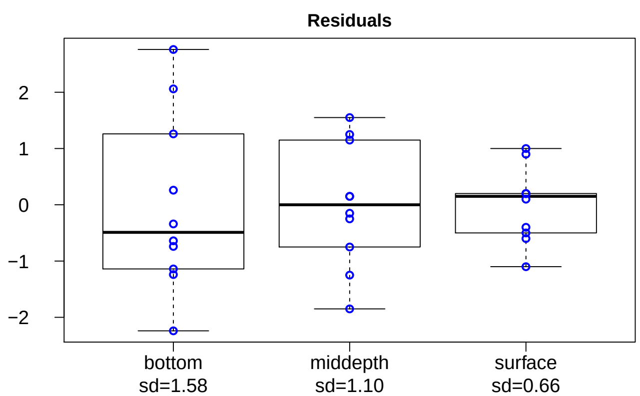
In this case it is somewhat hard to tell since the means are different.

(3) Constant variance - Residuals

One of the ways to think about each data point is as follows:

$$y_{ij} = \mu_i + \epsilon_{ij}$$

where ϵ_{ij} is called the residual ($\epsilon_{ij} = y_{ij} - \mu_i$).



t test vs. ANOVA - Purpose

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.

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

t test vs. ANOVA - Method

t test

Compute a test statistic (a ratio).

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

ANOVA

Compute a test statistic (a ratio).

$$F = \frac{\text{variability btw. groups}}{\text{variability w/in groups}}$$

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

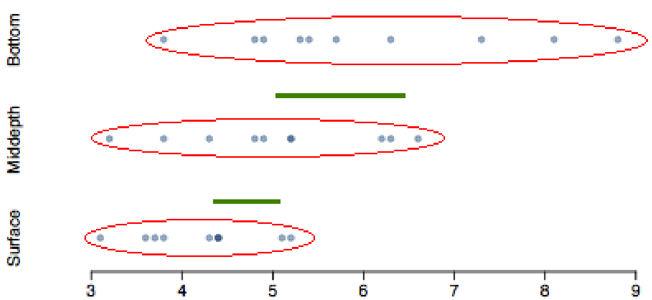
t test vs. ANOVA

- With only two groups t-test and ANOVA are equivalent, but only if we use a *pooled variance* (this assumes the two groups are the same, *not* just that the means are equal) in the denominator of the test statistic.
- With more than two groups, ANOVA compares the sample means to an overall *grand mean*.

Test statistic

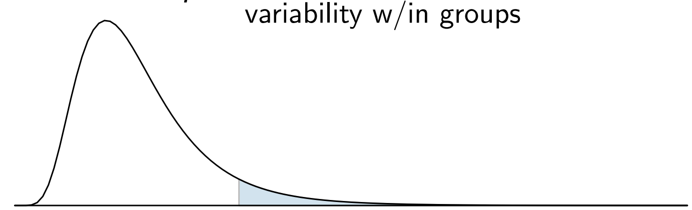
Does there appear to be a lot of variability within groups? How about between groups?

$$F = \frac{\text{variability btw. groups}}{\text{variability w/in groups}}$$



F distribution and p-value

$$F = \frac{\text{variability btw. groups}}{\text{variability w/in groups}}$$



- 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.

Types of Variability

For ANOVA we think of our variability (uncertainty) in terms of three separate quantities:

- Total variability - all of the variability in the data, ignoring any explanatory variable(s). (You can think of this as being analogous to the sample variance of all the data)
- Group variability - variability between the group means and the grand mean.
- Error variability - the sum of the variability within each group. (You can think of this as being analogous to the sum of sample variances for each group or the sum of the variances of the residuals)

Sum of squares and Variability

Mathematically, we can think of the unnormalized measures of variability as follows:

- Total variability - Sum of Squares Total
$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\mu_\cdot)^2 = Var(Y_{ij})$$
- Group variability - Sums of Squares Group
$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mu_i-\mu_\cdot)^2 = \sum_{i}^{k} n_i(\mu_i-\mu_\cdot)^2$$
- Error variability - Sum of Squares Error
$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\mu_i)^2 = \sum_{i=1}^{k} Var(Y_{i\cdot})$$

Partitioning Sums of Squares

With a little bit of careful algebra we can show that:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\mu_\cdot)^2 = \sum_{i}^{k} n_i(\mu_i-\mu_\cdot)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\mu_i)^2$$

$$\text{Total Variability} = \text{Group Variability (w/in)} + \text{Error Variability (btw)}$$
$$\text{Sum of Squares Total} = \text{Sum of Squares Group} + \text{Sum of Squares Error}$$

ANOVA Output

The results of an ANOVA is usually summarized in a tabular form that includes these measures of uncertainty as well as the calculation of the F test statistic.

		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			

ANOVA output - 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 n_i 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

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

ANOVA output (cont.) - 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.

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

ANOVA output (cont.) - 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:

$$SSE = SST - SSG$$

$$SSE = 54.29 - 16.96 = 37.33$$

ANOVA output

		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: $df_E = df_T - df_G = n - k$

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

ANOVA output (cont.) - MS

		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

Mean square values are calculated as sum of squares divided by the degrees of freedom.

$$\begin{aligned} MSG &= 16.96/2 = 8.48 \\ MSE &= 37.33/27 = 1.38 \end{aligned}$$

ANOVA output (cont.) - 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.

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

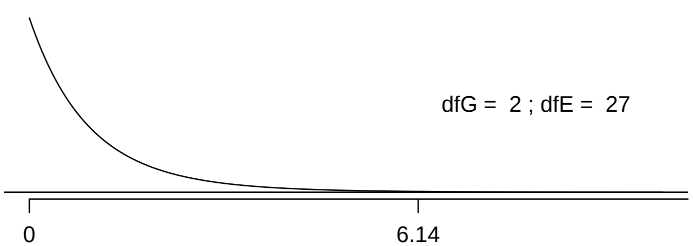
$$F = \frac{8.48}{1.38} = 6.14$$

ANOVA output (cont.) - P-value

		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

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

Conclusion

- If p-value is small (less than α), reject H_0. The data provide convincing evidence that at least one pair of means differ (but we say specifically which pair).
- 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).

What is the conclusion of the hypothesis test for Wolf river?

- The data provide convincing evidence that the average aldrin concentration is different for at least one pair.

Which means differ?

- We've 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* or *multiple testing*.
- If there are k groups, then there are $K = \binom{k}{2} = \frac{k(k-1)}{2}$ possible pairs.
- One common approach is the *Bonferroni correction* that uses a *stringent* significance level for each test:

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

  where K is the number of comparisons being considered.

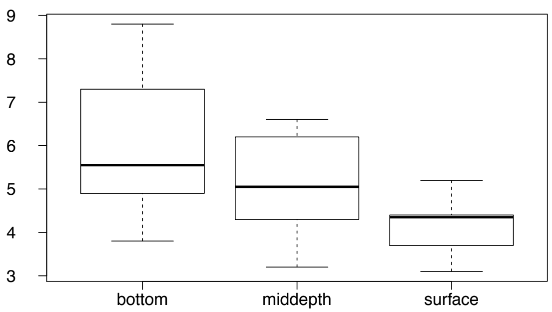
Determining the modified α

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 or two sample t tests for determining which pairs of groups have significantly different means?

$$\alpha^* = 0.05/3 = 0.0167$$

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.)

For an ANOVA we make have an assumption that all the groups have equal variance, this is not a part of a normal t-test. When performing a posthoc test we should maintain this assumption and use a pooled estimate of variability and the appropriate degrees of freedom associated with this estimate for our t distribution.

- Replace within-group sample standard deviations with MSE, which is s^2_{pooled}
- Use the error degrees of freedom $(n-k)$ for t-distributions

Difference in two means - ANOVA posthoc test

$$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \approx \sqrt{\frac{MSE}{n_1} + \frac{MSE}{n_2}}$$

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

	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			

	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

$$T_{df_E} = \frac{(\bar{x}_b - \bar{x}_m) - 0}{\sqrt{\frac{MSE}{n_b} + \frac{MSE}{n_m}}}$$

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

$$0.05 < p - value < 0.10 \qquad \text{(two-sided)}$$

$$\alpha^\star = 0.05/3 = 0.0167$$

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.

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

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

$$T_{27} = \frac{(6.04 - 4.02)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{2.02}{0.53} = 3.81$$

$$p - value = P(T_{27} > 3.81 \text{ or } T_{27} < -3.81)$$

$$< 0.01$$

$$\alpha^\star = 0.05/3 = 0.0167$$

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

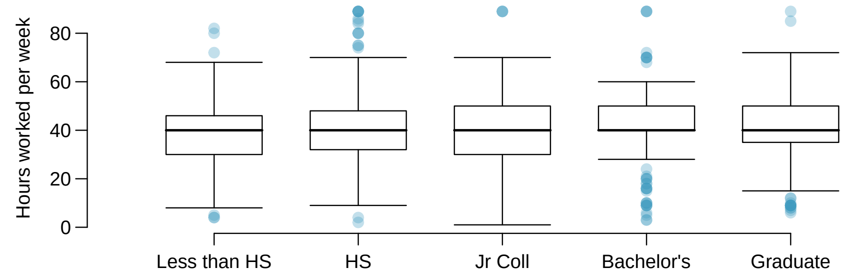
GSS - Hours worked vs Education

Previously we have seen data from the General Social Survey in order to compare the average number of hours worked per week by US residents with and without a college degree. However, this analysis didn't take advantage of the original data which contained more accurate information on educational attainment (less than high school, high school, junior college, Bachelor's, and graduate school).

Using ANOVA, we can consider educational attainment levels for all 1,172 respondents at once instead of re-categorizing them into two groups. On the following slide are the distributions of hours worked by educational attainment and relevant summary statistics that will be helpful in carrying out this analysis.

GSS - Hours worked vs Education (data)

	Educational attainment					
	Less than HS	HS	Jr Coll	Bachelor's	Graduate	Total
Mean	38.67	39.6	41.39	42.55	40.85	40.45
SD	15.81	14.97	18.1	13.62	15.51	15.17
n	121	546	97	253	155	1,172



GSS - Hours worked vs Education (ANOVA table)

Given what we know, fill in the unknowns in the ANOVA table below.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
degree	*4*	*2006.16*	501.54	*2.189*	0.0682
Residuals	*1167*	267382	*229.12*		
Total	*1171*	*269388.16*			

	Educational attainment					
	Less than HS	HS	Jr Coll	Bachelor's	Graduate	Total
Mean	38.67	39.6	41.39	42.55	40.85	40.45
SD	15.81	14.97	18.1	13.62	15.51	15.17
n	121	546	97	253	155	1,172