# Section 7.5
# Comparing means with ANOVA

Stats 7 Summer Session II 2022

# Motivating scenario

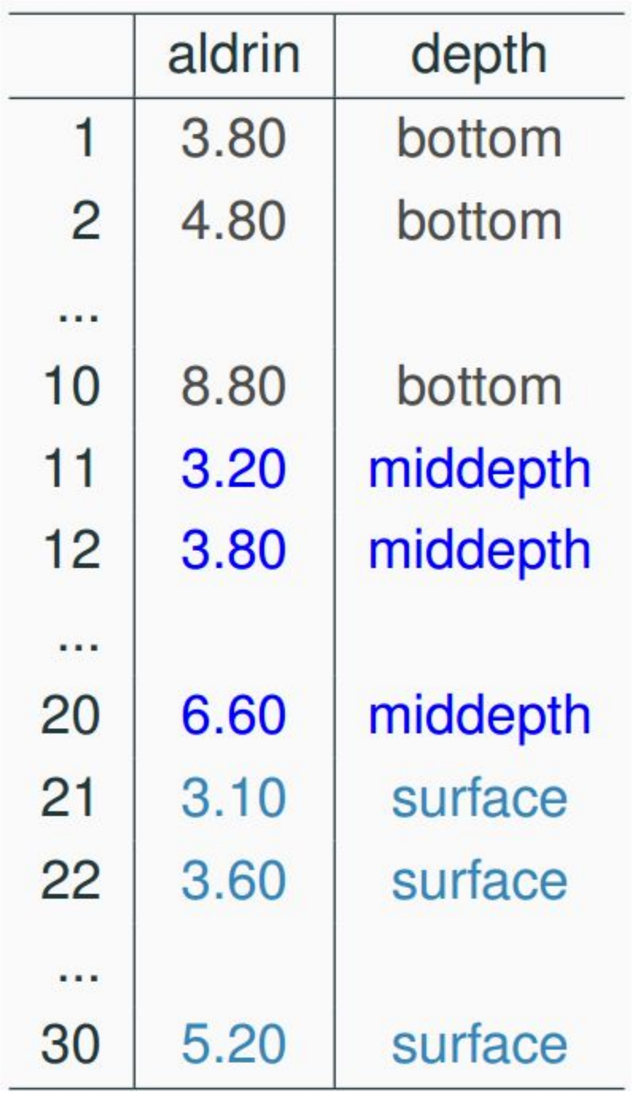
- Sometimes we want to compare means across many groups. We might initially think to do pairwise comparisons.
    - For example, if there were three groups, we might be tempted to compare the first mean with the second, then with the third, and then finally compare the second and third means for a total of three comparisons.
    - However, this strategy can be treacherous. If we have many groups and do many comparisons, it is likely that we will eventually find a difference just by chance, even if there is no difference in the populations.
- Instead, we should apply a holistic test to check whether there is evidence that at least one pair groups are in fact different, and this is where ANOVA saves the day.

Memphis
Shelby Farms Park
Germantown
Collierville
Piperton
Rossville
Moscow
La Grange
Michigan City
Ba
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
- 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

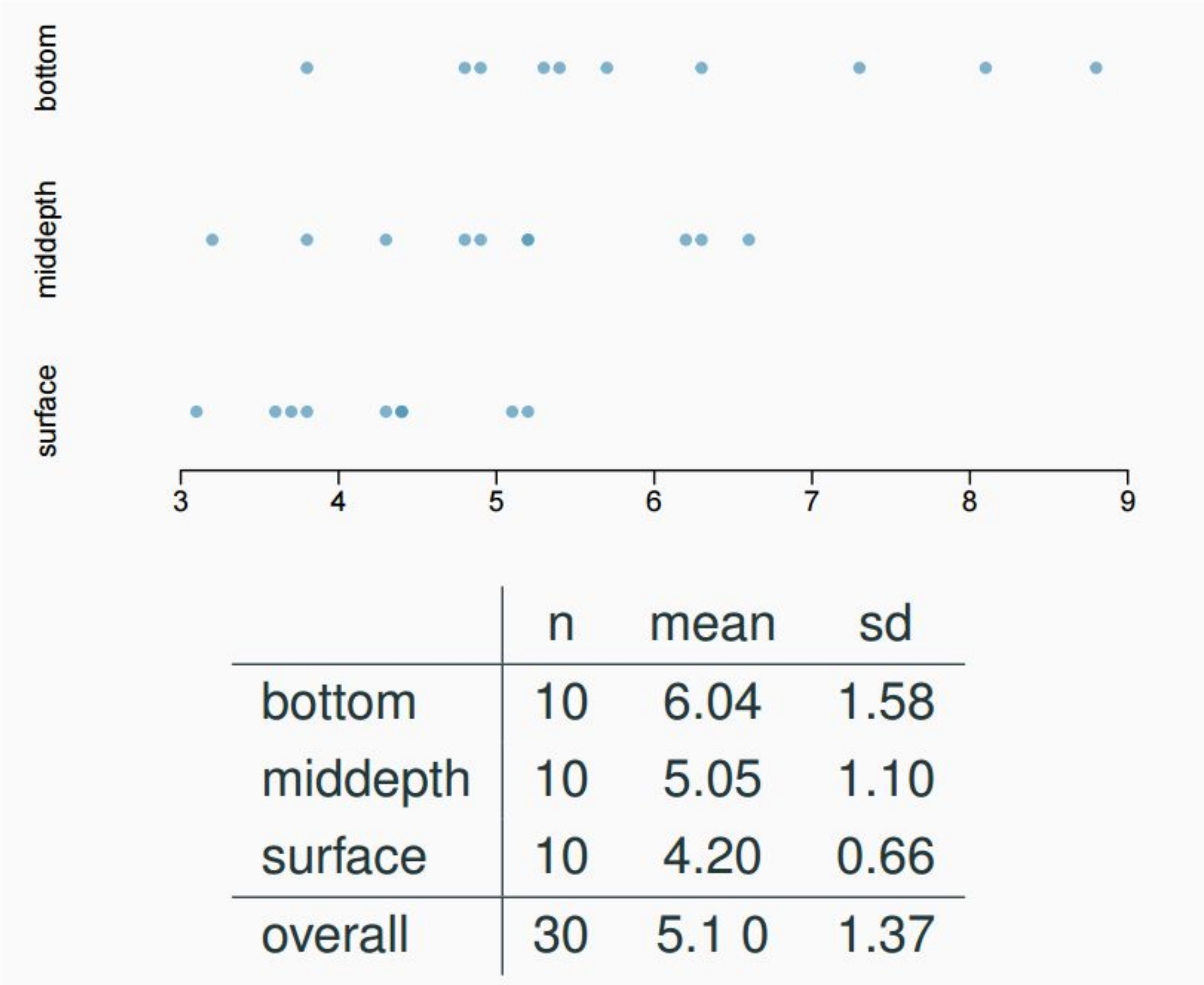
Aldrin concentration (nanograms per liter) at three levels of depth



| | 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?

- To compare means of 2 groups we use a *Z (if we know σ)* or a *T* statistic *(if σ is unknown)*
- 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

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

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

where $\mu_i$ 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

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?

# *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)

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

- The p-value for a F-statistic is always the area to the right of the test statistic since the F value is always positive (similar to chi square test statistics)
- 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
    - Using the pooled standard variance for standard error of a difference in 2 means would only be done if we had reason to believe the standard deviations of the two populations were similar
- With more than two groups, ANOVA compares the sample means to an overall grand mean

# Hypotheses

*There are only one set of hypotheses for ANOVA, some difference or none.*

A. $H_0 : \mu_{\mathrm{B}} = \mu_{\mathrm{M}} = \mu_{\mathrm{S}}$

$H_A : \mu_{\mathrm{B}} \neq \mu_{\mathrm{M}} \neq \mu_{\mathrm{S}}$

B. $H_0 : \mu_{\mathrm{B}} \neq \mu_{\mathrm{M}} \neq \mu_{\mathrm{S}}$

$H_A : \mu_{\mathrm{B}} = \mu_{\mathrm{M}} = \mu_{\mathrm{S}}$

C. $H_0 : \mu_{\mathrm{B}} = \mu_{\mathrm{M}} = \mu_{\mathrm{S}}$

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

D. $H_0 : \mu_{\mathrm{B}} = \mu_{\mathrm{M}} = \mu_{\mathrm{S}} = 0$

$H_A$ : At least one mean is different

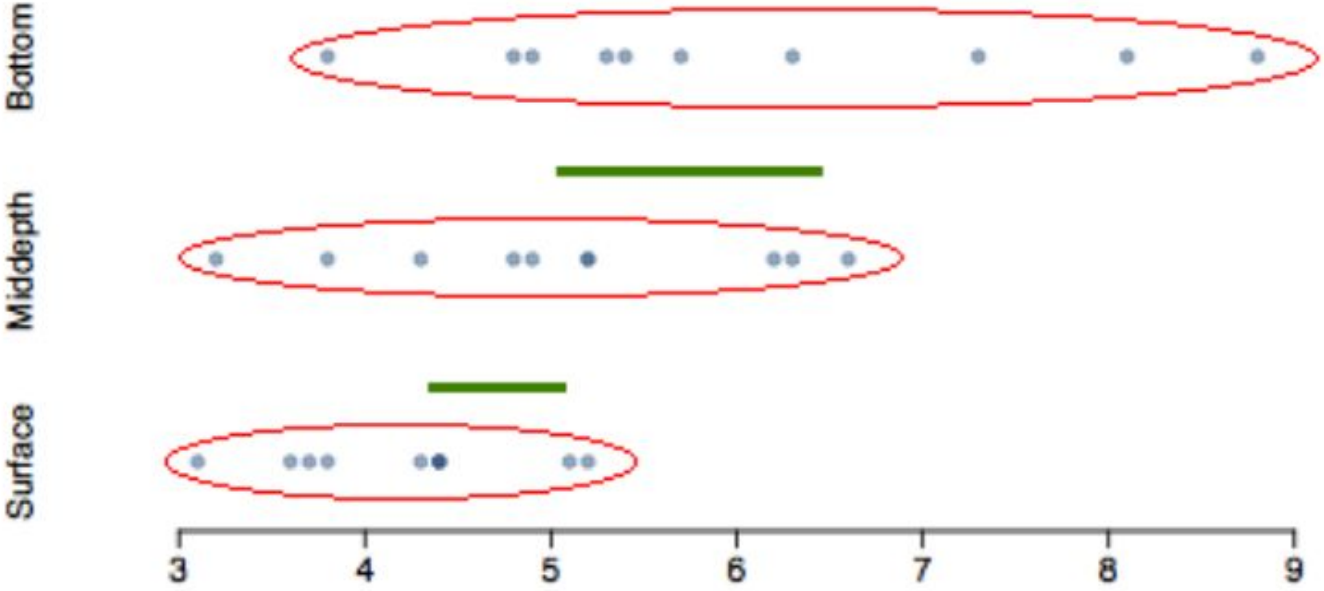
[✓ E. $H_0 : \mu_{\mathrm{B}} = \mu_{\mathrm{M}} = \mu_{\mathrm{S}}$

$H_A$ : At least one mean is different]

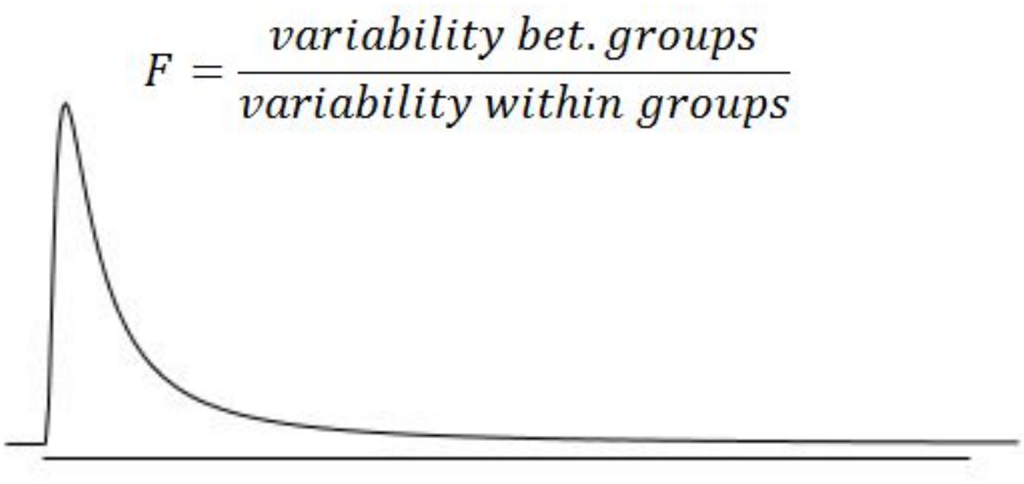
# Test statistic

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

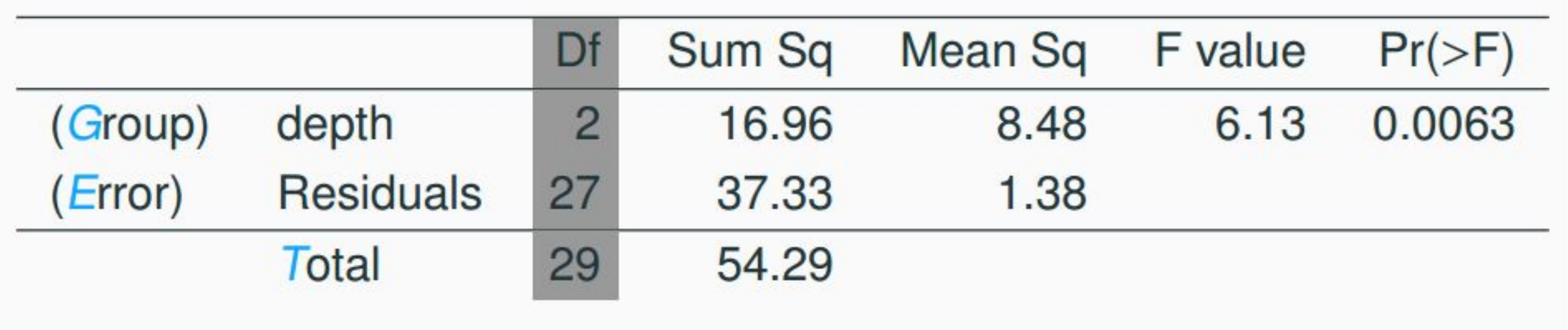


$$F = \frac{variability\ bet.\ groups}{variability\ within\ groups}$$

- 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

We will later discuss how to use software to do our calculations and get the table below.

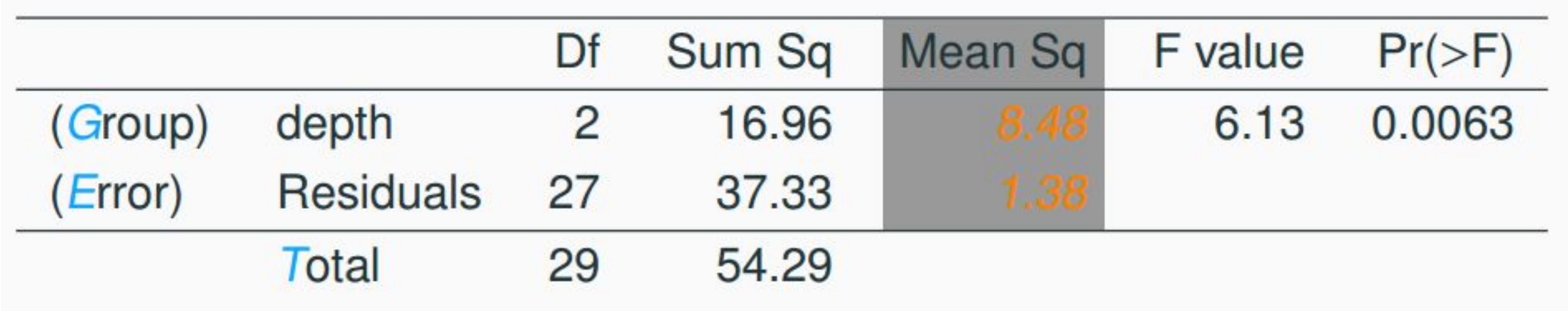
For now we will focus on understanding the table.



| | | 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: $df_E = df_T - df_G$

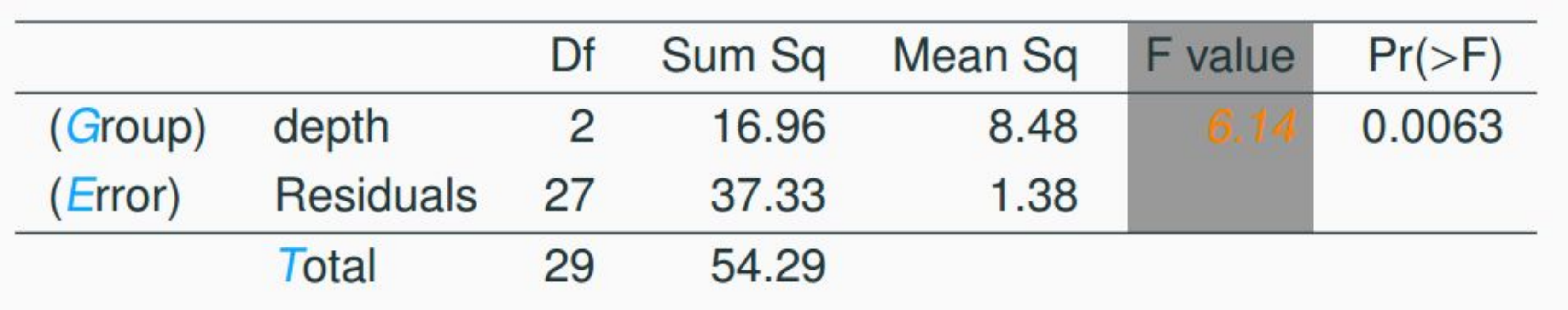


| | | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|---|
| (Group) | depth | 2 | 16.96 | 8.48 | 6.13 | 0.0063 |
| (Error) | Residuals | 27 | 37.33 | 1.38 | | |
| | Total | 29 | 54.29 | | | |

## Mean square error

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

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



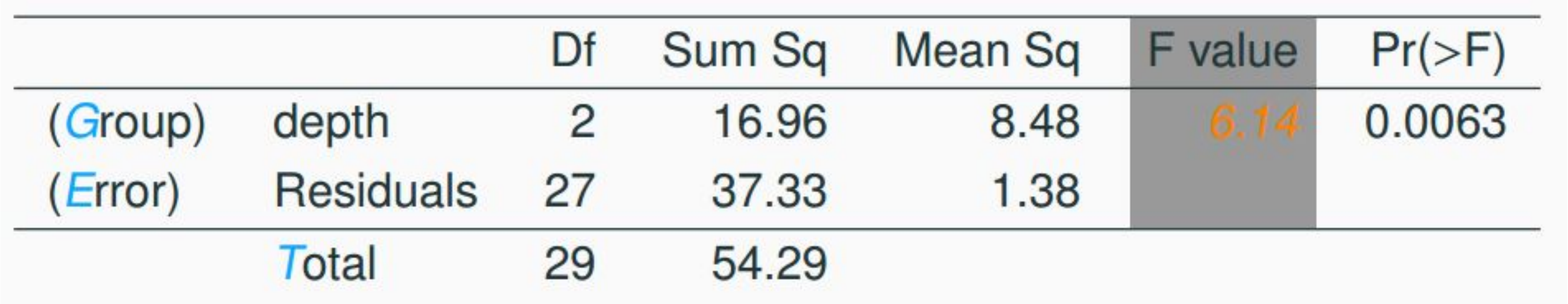
|  |  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|---|
| (Group) | depth | 2 | 16.96 | 8.48 | 6.14 | 0.0063 |
| (Error) | Residuals | 27 | 37.33 | 1.38 |  |  |
|  | Total | 29 | 54.29 |  |  |  |

## Test statistic, *F* value

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

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

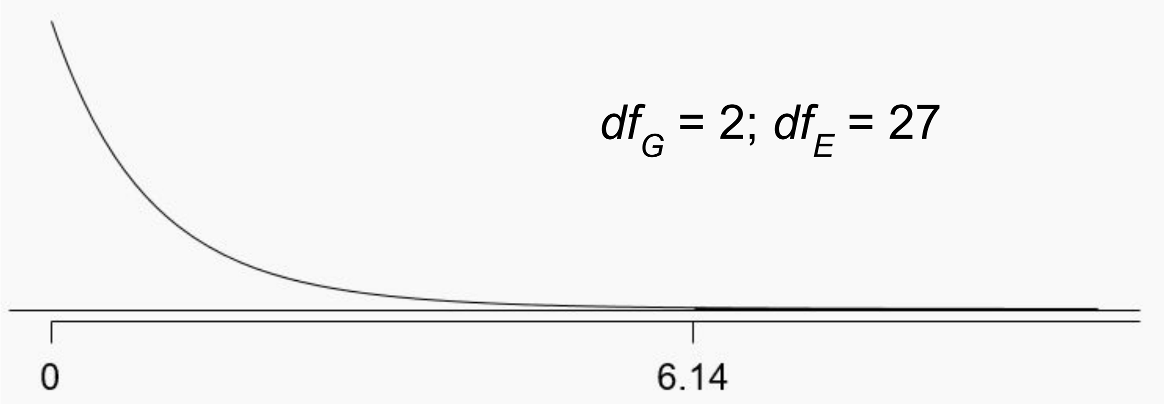
$$F = \frac{8.48}{1.38} = 6.14$$



| | | 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 = 0.0063

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

- 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)

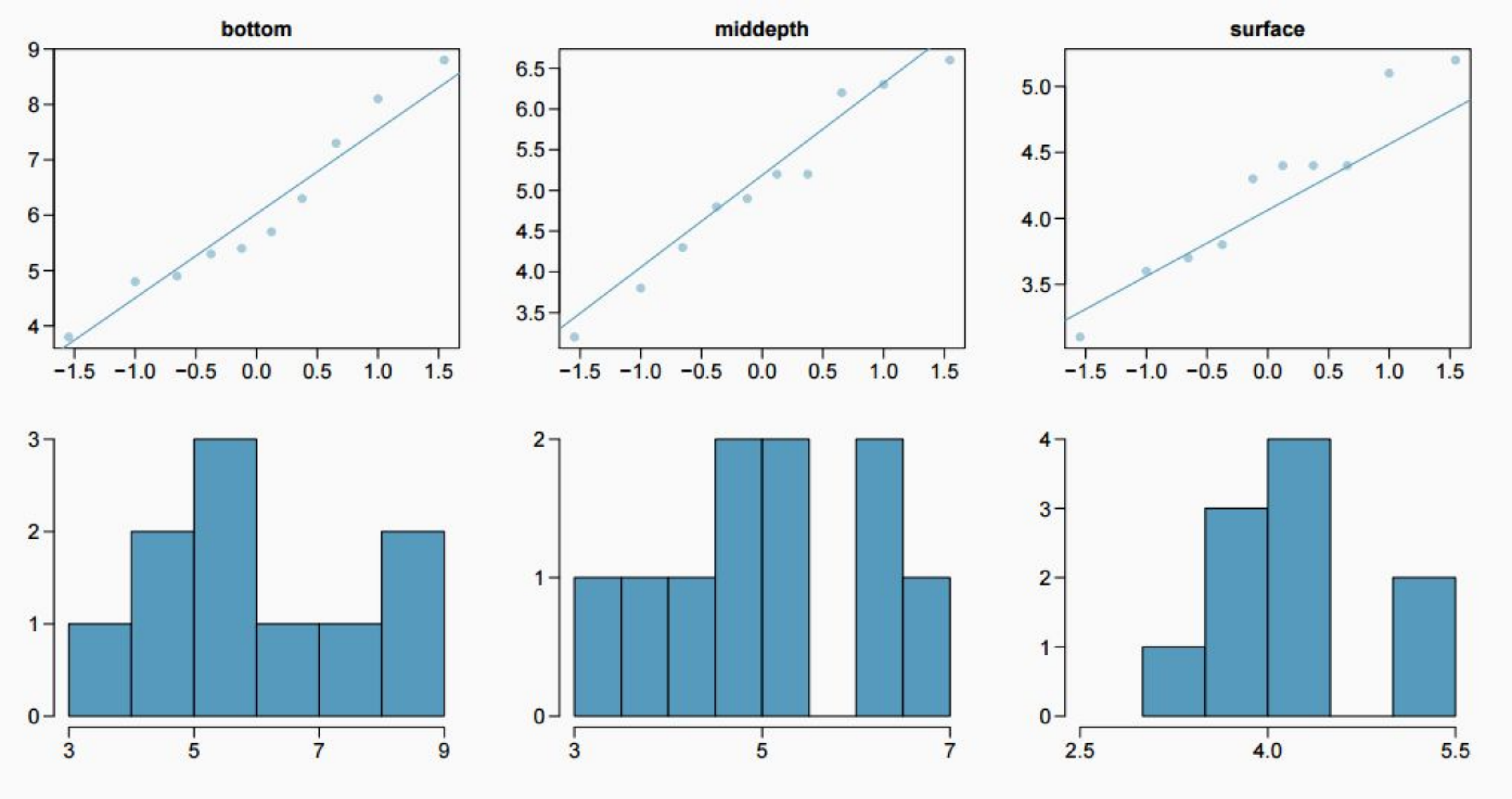
# (1) independence

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*

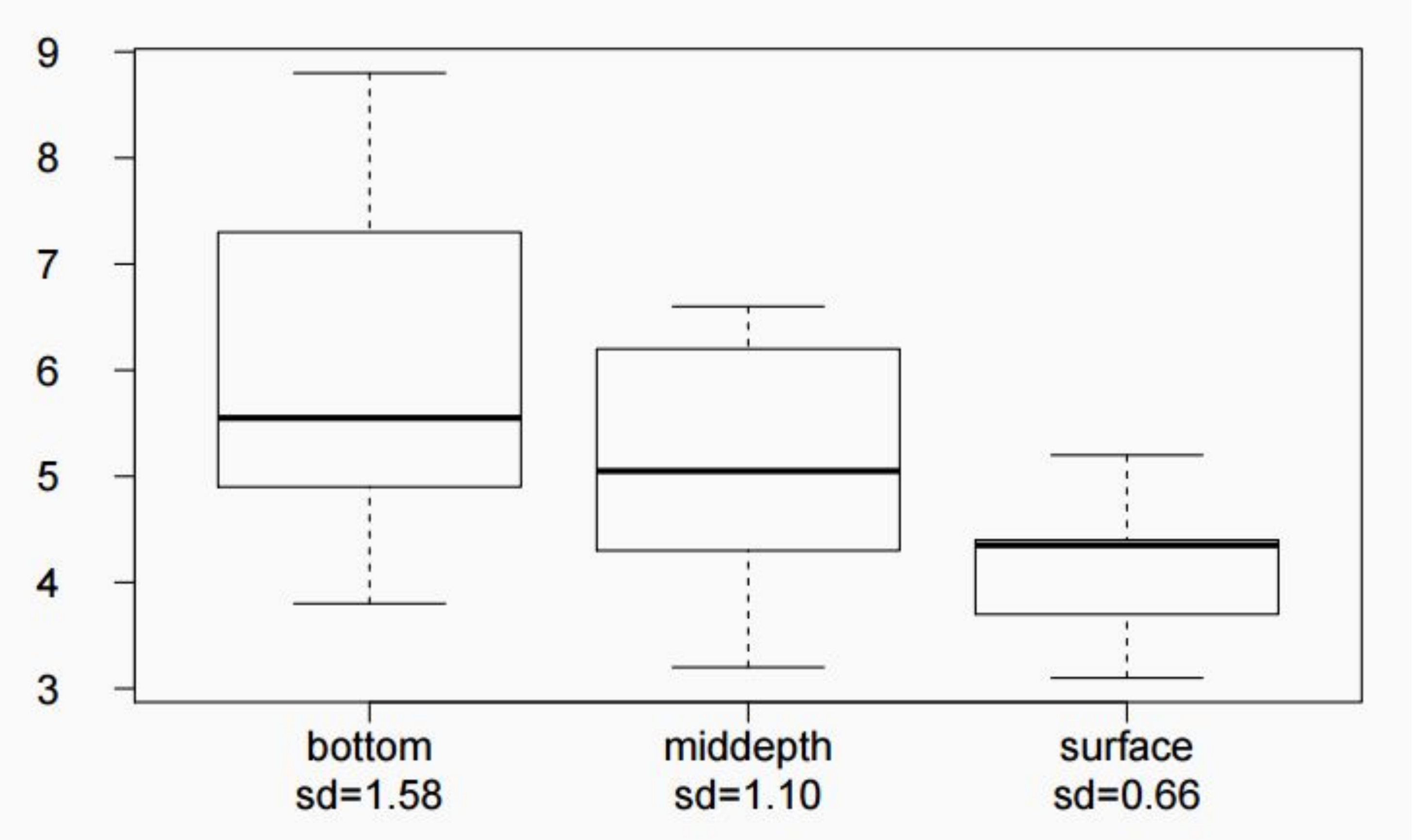
# (2) approximately normal

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?”
- 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
- 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 $\alpha$

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?

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? (cont.)

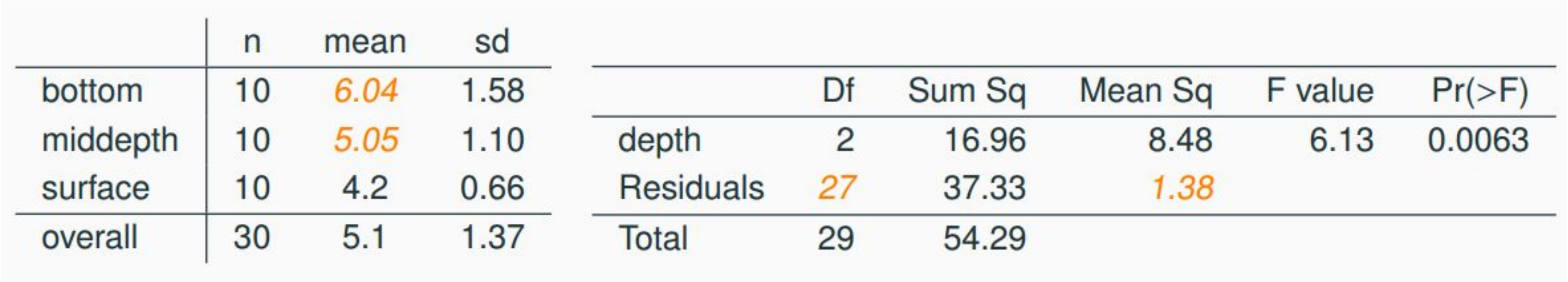
If the ANOVA assumption of equal variability across groups is satisfied, we can use the data from all groups to estimate variability:

- 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 | | | |

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

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

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

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

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?

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

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

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

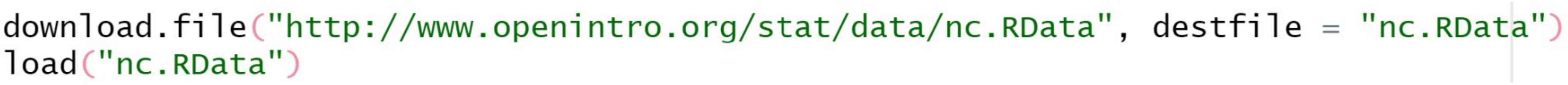
$$\alpha^* = \frac{0.05}{3} = 0.0167$$

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

# ANOVA in R

In 2004, the state of North Carolina released a large data set containing information on births recorded in this state. This data set is useful to researchers studying the relation between habits and practices of expectant mothers and the birth of their children. We will work with a random sample of observations from this data set.

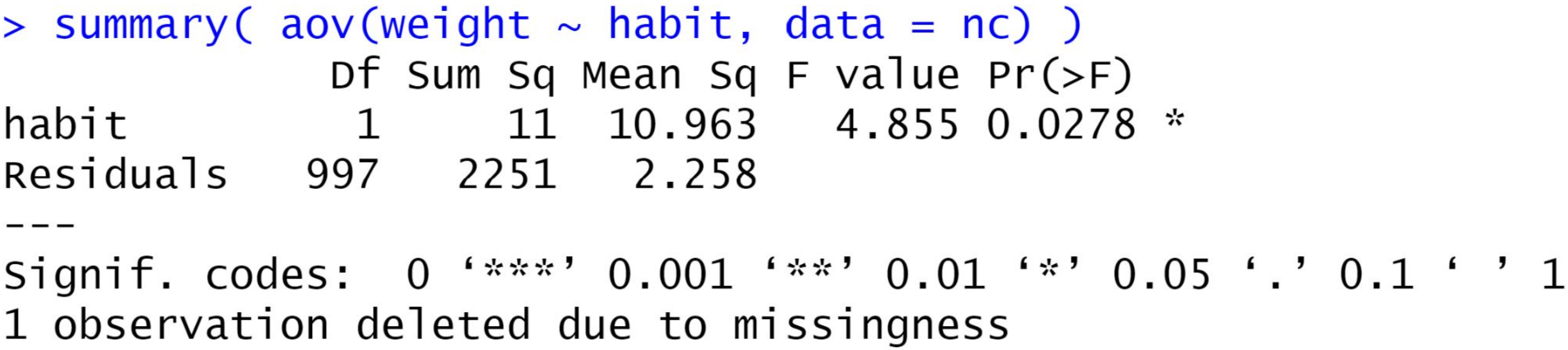
To read in this data run the two lines of code below



```
download.file("http://www.openintro.org/stat/data/nc.RData", destfile = "nc.RData")
load("nc.RData")
```

# ANOVA in R

Consider the possible relationship between a mother’s smoking habit, (categorical variable called “habit”), and the weight of her baby, (numerical variable called “weight”).



```
> summary( aov(weight ~ habit, data = nc) )
             Df Sum Sq Mean Sq F value Pr(>F)  
habit         1     11  10.963   4.855 0.0278 *
Residuals   997   2251   2.258                 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
1 observation deleted due to missingness
```

What is the value of the test statistic?

$F = 4.855$

What is the p-value?

P-value = 0.0278

What would our conclusion be?

There is strong evidence (p-value = 0.0278 < $\alpha$ = 0.05) of a difference in mean birth weight for mothers with different smoking habits.