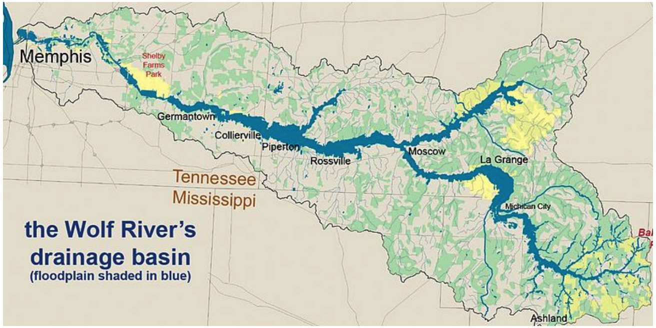
MATH 1052H - S62 - Lecture 06

More than Two Means

Some Context

In our previous lectures, we've discussed single sample means, paired sample means (two means), and independent sample means (two means). We hinted that something we might be interested in was more-than-two samples. In this lecture, we will start the discussion of that idea.

Case Study: Aldrin in the 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.

Case Study: Aldrin - Study Methods

- 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 than near mid-depth.

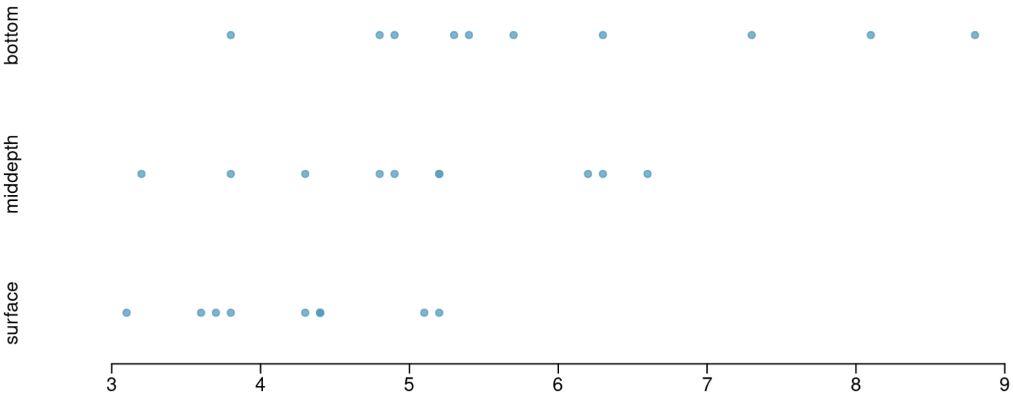
Data

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

```
##    aldrin    depth
## 1     3.8   bottom
## 2     4.8   bottom
## 3     4.9   bottom
## 4     5.3   bottom
## 10    8.8   bottom
## 11    3.2 middepth
## 12    3.8 middepth
## 13    4.3 middepth
## 20    6.6 middepth
## 21    3.1  surface
## 22    3.6  surface
## 23    3.7  surface
```

Exploratory analysis

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



```
##              n mean   sd
## 1   bottom 10 6.04 1.58
## 2 middepth 10 5.05 1.10
## 3  surface 10 4.20 0.66
```

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.

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 mean is different than others.

Conditions

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

How do we check for normality?

- 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.

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

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).

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

z/t test vs. ANOVA

For both:

- 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.

In general:

- 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 (back to our Case Study)

What are the correct hypotheses for testing for a difference between the mean aldrin concentrations among the three levels?

- $H_0 : \mu_B = \mu_M = \mu_S$ versus $H_A : \mu_B \neq \mu_M \neq \mu_S$
- $H_0 : \mu_B \neq \mu_M \neq \mu_S$ versus $H_A : \mu_B = \mu_M = \mu_S$
- $H_0 : \mu_B = \mu_M = \mu_S$ versus H_A : At least one mean is different.
- $H_0 : \mu_B = \mu_M = \mu_S = 0$ versus H_A : At least one mean is different.
- $H_0 : \mu_B = \mu_M = \mu_S$ versus $H_A : \mu_B > \mu_M > \mu_S$

Hypotheses

What are the correct hypotheses for testing for a difference between the mean aldrin concentrations among the three levels?

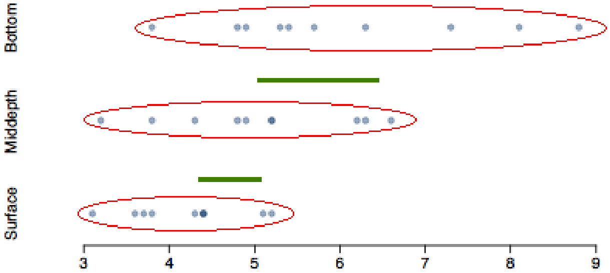
- $H_0 : \mu_B = \mu_M = \mu_S$ versus $H_A : \mu_B \neq \mu_M \neq \mu_S$
- $H_0 : \mu_B \neq \mu_M \neq \mu_S$ versus $H_A : \mu_B = \mu_M = \mu_S$
- **$H_0 : \mu_B = \mu_M = \mu_S$ versus H_A : At least one mean is different.**
- $H_0 : \mu_B = \mu_M = \mu_S = 0$ versus H_A : At least one mean is different.
- $H_0 : \mu_B = \mu_M = \mu_S$ versus $H_A : \mu_B > \mu_M > \mu_S$

ANOVA and the F test

Test statistic

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

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



Test statistic (cont.)

$$F = \frac{\text{variability between groups}}{\text{variability within groups}} = \frac{MSG}{MSE}$$

- **MSG** is Mean Square between Groups

  $$df_G = k - 1$$

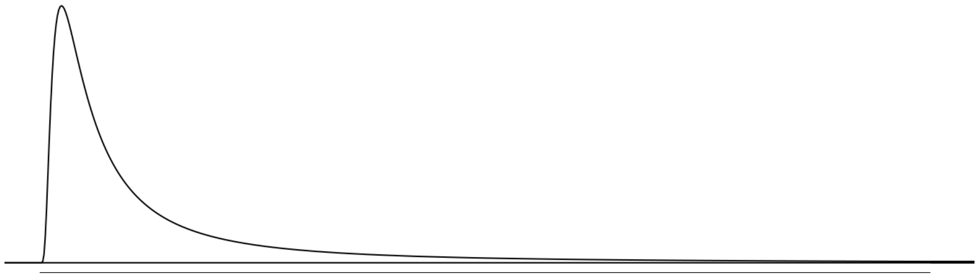
  where k is number of groups
- **MSE** is Mean Square Error - variability in residuals

  $$df_E = n - k$$

  where n is number of observations.

F distribution and p-value

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



- 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.

ANOVA output, deconstructed

```r
# our data is stored as a data.frame called aldrin
our_model <- lm(aldrin ~ depth, data = aldrin)
summary( aov( our_model ) )
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## depth        2  16.96   8.480   6.134 0.00637 **
## Residuals   27  37.33   1.383
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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)
## depth        2  16.96   8.480   6.134 0.00637 **
## Residuals   27  37.33   1.383
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
##             Df Sum Sq Mean Sq F value  Pr(>F)   
## depth        2  16.96   8.480   6.134 0.00637 **
## Residuals   27  37.33   1.383                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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.

I will never make you compute this by hand!

```
##             Df Sum Sq Mean Sq F value  Pr(>F)   
## depth        2  16.96   8.480   6.134 0.00637 **
## Residuals   27  37.33   1.383                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sum of squares total, SST

Measures the variability between groups

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

where x_i represent each observation in the dataset.

Again, not something we do by hand!

```
##             Df Sum Sq Mean Sq F value  Pr(>F)   
## depth        2  16.96   8.480   6.134 0.00637 **
## Residuals   27  37.33   1.383                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sum of squares error, SSE

Measures the variability within groups:

$$SSE = SST - SSG$$

(might make you do this one by hand, because it’s easy!)

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## depth         2  16.96   8.480   6.134 0.00637 **
## Residuals    27  37.33   1.383
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Mean square error

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

(also easy, also may be asked!)

$$MSG = 16.96/2 = 8.480$$

$$MSE = 37.33/27 = 1.383$$

```
##             Df Sum Sq Mean Sq F value  Pr(>F)   
## depth        2  16.96   8.480   6.134 0.00637 **
## Residuals   27  37.33   1.383                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test statistic, F value

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

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

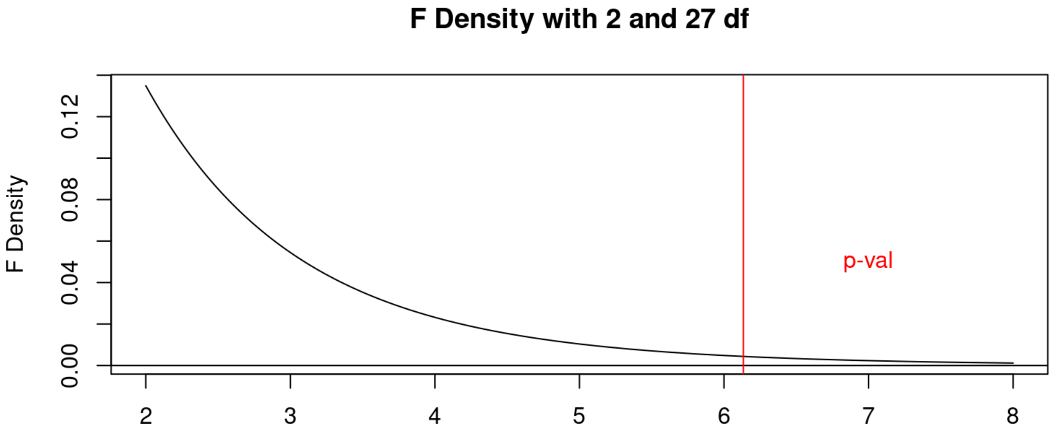
$$F = \frac{8.480}{1.383} = 6.134$$

```
##             Df Sum Sq Mean Sq F value  Pr(>F)   
## depth        2  16.96   8.480   6.134 0.00637 **
## Residuals   27  37.33   1.383                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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)   
## depth        2  16.96   8.480   6.134 0.00637 **
## Residuals   27  37.33   1.383                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



```
##             Df Sum Sq Mean Sq F value  Pr(>F)   
## depth        2  16.96   8.480   6.134 0.00637 **
## Residuals   27  37.33   1.383                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
pf(q = 6.134, df1 = 2, df2 = 27, lower.tail = FALSE)
```

```
## [1] 0.006366302
```

Conclusion - in context

What is the conclusion of the hypothesis test?

The data provide convincing evidence that the average aldrin concentration

- is different for all groups.
- on the surface is lower than the other levels.
- is different for at least one group.
- 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

- is different for all groups.
- on the surface is lower than the other levels.
- **is different for at least one group.**
- 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).

Checking conditions

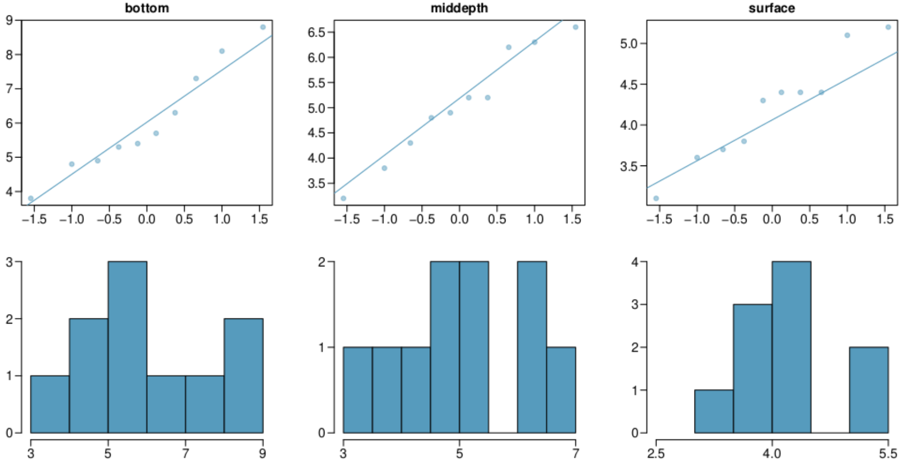
(1) independence

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.

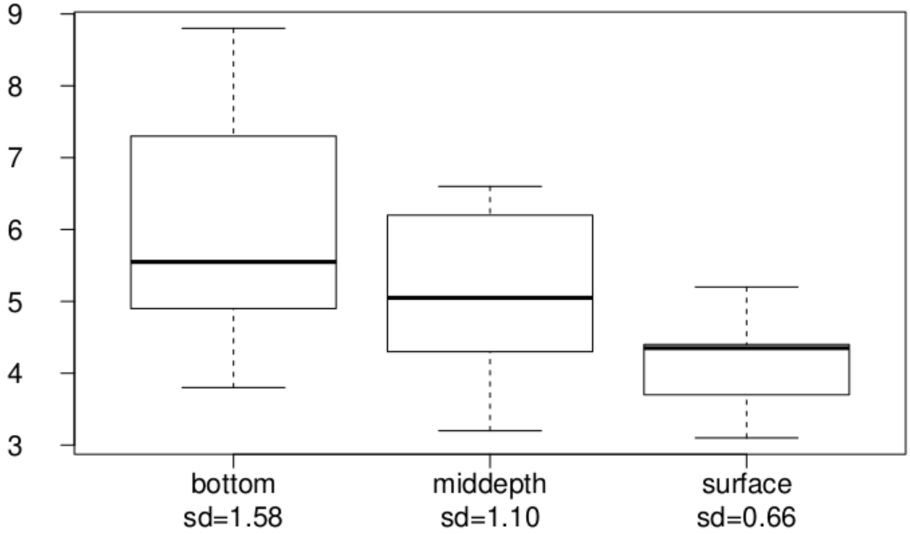
(2) approximately normal

Does this condition appear to be satisfied?



(3) constant variance

Does this condition appear to be satisfied?



Multiple comparisons & Type 1 error rate

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?

Pitfalls

- 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:

$$\alpha^\star = \alpha/K$$

  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 $\alpha = 0.05$, what should be the modified significance level for two sample t tests for determining which pairs of groups have significantly different means?

- $\alpha^* = 0.05$
- $\alpha^* = 0.05/2 = 0.025$
- $\alpha^* = 0.05/3 = 0.0167$
- $\alpha^* = 0.05/6 = 0.0083$

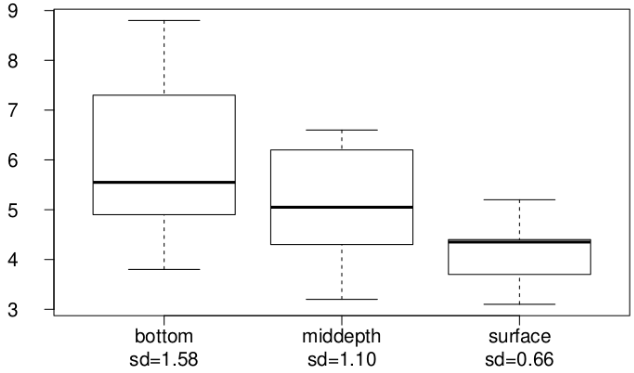
Determining the modified α

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?

- $\alpha^* = 0.05$
- $\alpha^* = 0.05/2 = 0.025$
- $\alpha^* = 0.05/3 = 0.0167$
- $\alpha^* = 0.05/6 = 0.0083$

Which means differ?

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



- bottom & surface
- bottom & mid-depth
- mid-depth & surface
- bottom & mid-depth; mid-depth & surface
- 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?

- From the original summary table, mean of bottom is 6.04, mean of middepth is 5.05
- From the ANOVA table, the Residuals have df 27 and Mean Sq 1.38

$$T_{df_E} = \frac{(\bar{x}_{bottom} - \bar{x}_{middepth})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{middepth}}}} \qquad T_{27} = \frac{(6.04 - 5.05)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{0.99}{0.53} = 1.87$$

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

Finish

Find the p-value:

```r
pt(q = 1.87, df = 27, lower.tail = FALSE) * 2
```

```
## [1] 0.0723635
```

Then since the test has $\alpha^\star = 0.0167$: **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: Bottom and Surface

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

$$T_{df_E} = \frac{(\bar{x}_{bottom} - \bar{x}_{surface})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{surface}}}} \qquad T_{27} = \frac{(6.04 - 4.02)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{2.02}{0.53} = 3.81$$

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

```
pt(q = 3.81, df = 27, lower.tail = FALSE)
```

```
## [1] 0.0003650144
```

Conclusion: **reject H_0**, the data provide convincing evidence of a difference between the average aldrin concentrations at bottom and surface.

Conclusion, Notes

What we have described in this lecture is known in the literature as “one-way ANOVA”. There is only one factor variable across the data, and any data source you use will be stored in a data.frame with two columns only.

This data is the natural output of **experiments** (remember Chapter 1 of our text from 1051H?), and we will discuss more complicated examples of this later in the term. This topic plus linear regression are the two most useful and common models used in science, so pay attention to them - you’ll definitely encounter them again!