Model Comparison Approach

Announcements

- We will first grade Exam 2, then grade HW 3.
- I am going to modify the HW due dates a bit so stay tuned
- Exam 3:
  - Heavily weighted to newer material
  - Older material covered would be things like, define a p-value, confidence intervals, and what types of tests would you use for different research questions. I will NOT make you re-calculate specific tests.
- **Merve is teaching on Thursday 11/9**

Recap

- We've compared a lot of means:
    - sample mean ($\bar{x}$) to population mean (μ); paired or not paired
    - 2 sample means ($\bar{x_1}$ vs. $\bar{x_2}$)

This time

- t-tests through oneway ANOVA
- BUT, we're going to take a different approach...

Model Comparisons

Scenario 1

Gerrymandering

- Depending on the estimate you pick, about 53% of voters in Wisconsin were Democrats in 2016.
- So our best estimate of the percentage of voters that are Democrat in any *district* might be 53%
- Now that 2016 feels like a million years ago, you find that in actuality it was 52% of voters in Wisconsin were Democrats in 2016.
- *Question: Was our population estimate of 53% significantly different from our sample estimate of 52%?*

one-sample t-test

Model Comparisons

In the normal **one-sample t-test**

- $\bar{x} = \mu$
- $H_0 : \bar{x} - \mu = 0$
- $H_A : \bar{x} - \mu \neq 0$

Model Comparisons

Another way of thinking about it -- what does a model of the null hypothesis look like?

- If there is truly no difference between means (the null is true), then the best way to summarize the data is to use the *population mean*. Let's use that as our **estimator**.

- If we use the population mean as our estimator, we can look at error or **residuals**. The actual data points - estimator (a deviation score!). We want to do a good job of predicting. So we want this error to be as SMALL as possible!

- If there *is* a difference between means (the null should be rejected), then the best way of summarizing our data should be to use the sample mean.

Which had the smallest prediction error?

model that uses population mean as the estimator?

model that uses sample mean as the estimator?

Model Comparisons

Let's break it down into **full** and **restricted** models:

- **Restricted Model:** reflects what we are testing *against*.
- **Full Model:** allows us to fully include all information we might have.
- Size of the effect is calculated as the following:

$$\frac{(E_r - E_f)/(df_r - df_f)}{E_f/df_f}$$

- E_r is the error from the restricted model
- E_f is the error from the full model
- df_r is the degrees of freedom from the restricted model
- df_f is the degrees of freedom from the full model

The Data

```r
dems <- data.frame(Dem = c(30, 69, 99, 77, 29, 37, 38, 37))
dems
```

```
##   Dem
## 1  30
## 2  69
## 3  99
## 4  77
## 5  29
## 6  37
## 7  38
## 8  37
```

The Restricted Model

Step 1: Get the deviation scores. In the restricted model, we are subtracting from our population estimate of 53%. That is, 53% is our PREDICTION of the **restricted** model

```
dems$deviationScores <- dems$Dem - 53
dems
```

```
##   Dem deviationScores
## 1  30             -23
## 2  69              16
## 3  99              46
## 4  77              24
## 5  29             -24
## 6  37             -16
## 7  38             -15
## 8  37             -16
```

The Restricted Model

Step 2: Square the deviation scores

```
dems$dev2 <- dems$deviationScores ^2
dems
```

```
##   Dem deviationScores dev2
## 1  30             -23  529
## 2  69              16  256
## 3  99              46 2116
## 4  77              24  576
## 5  29             -24  576
## 6  37             -16  256
## 7  38             -15  225
## 8  37             -16  256
```

The Restricted Model

Step 3: Get the sum of the square deviation scores. This is our **ERROR** term. It is the squared errors.

```
Er <- sum(dems$dev2)
dems
```

```
##   Dem deviationScores dev2
## 1  30             -23  529
## 2  69              16  256
## 3  99              46 2116
## 4  77              24  576
## 5  29             -24  576
## 6  37             -16  256
## 7  38             -15  225
## 8  37             -16  256
```

```
Er
```

```
## [1] 4790
```

The Restricted Model

Step 4: Determine the degrees of freedom.

- Degrees of freedom deals with how much information is *free to vary*
- You get a feel for this the more you practice
- In this restricted model, there is nothing that we are "guessing" or estimating. So there is nothing to subtract. $df = n$, **df = 8**

```
dfr <- 8
```

The Full Model

Step 1: Get the deviation scores. In the full model, we are subtracting from our population estimate of 52%. That is, 52% is the PREDICTION of the **full** model.

```
demsFull = dems %>%
  select(Dem)

demsFull$deviationScores <- demsFull$Dem - 52
demsFull
```

```
##   Dem deviationScores
## 1  30             -22
## 2  69              17
## 3  99              47
## 4  77              25
## 5  29             -23
## 6  37             -15
## 7  38             -14
## 8  37             -15
```

The Full Model

Step 2: Square the deviation scores

```
demsFull$dev2 <- demsFull$deviationScores^2
demsFull
```

```
##   Dem deviationScores dev2
## 1  30             -22  484
## 2  69              17  289
## 3  99              47 2209
## 4  77              25  625
## 5  29             -23  529
## 6  37             -15  225
## 7  38             -14  196
## 8  37             -15  225
```

The Full Model

Step 3: Get the sum of the square deviation scores. This is our **ERROR** term. It is the squared errors.

```
Ef <- sum(demsFull$dev2)
demsFull
```

```
##   Dem deviationScores dev2
## 1  30             -22  484
## 2  69              17  289
## 3  99              47 2209
## 4  77              25  625
## 5  29             -23  529
## 6  37             -15  225
## 7  38             -14  196
## 8  37             -15  225
```

```
Ef
```

```
## [1] 4782
```

The Full Model

Step 4: Determine the degrees of freedom.

- Degrees of freedom deals with how much information is *free to vary*
- You get a feel for this the more you practice
- In this full model, we are guessing/estimating our sample mean of 52. $df = n - 1$, **df = 8 - 1 = 7**

```
dff <- 7
```

The Effect

$$\frac{(E_r - E_f)/(df_r - df_f)}{E_f/df_f}$$

```
effect <- ((Er - Ef) / (dfr - dff)) / (Ef/dff)
effect
```

```
## [1] 0.01171058
```

This is our F-statistic. $t^2 = F$. So to get our t-statistic, let's take the square root of our `effect`.

```
tstat <- sqrt(effect)
round(x = tstat, digits = 3)
```

```
## [1] 0.108
```

Model Comparison Approach

Is 0.108 more extreme than our critical value?

- For an $\alpha = .05$ in a two-tailed test with $df = 7$, the critical t value is 2.3650.

Conclusion: No, it's not more extreme than the critical value. The weighted error term from the restricted model is smaller than the weighted error term from the full model -- 53 is a better estimator than 52. The means are not statistically significantly different

Model Comparison Approach

We just did a one-sample t-test! Let's verify our results:

```r
t.test(x = dems$Dem, mu = 53)
```

```
##
##      One Sample t-test
##
## data:  dems$Dem
## t = -0.10822, df = 7, p-value = 0.9169
## alternative hypothesis: true mean is not equal to 53
## 95 percent confidence interval:
##  30.14892 73.85108
## sample estimates:
## mean of x
##        52
```

Scenario 2

- What if now we want to compare the difference in means (of % Democrats) between the 2010 election?
- *Question: are the means of % Democrats significantly different between 2010 and 2016?*

The Data

```
dems <- data.frame(Dem = c(30, 69, 99, 77, 29, 37, 38, 37,
                           30, 62, 50, 69, 27, 29, 44, 45),
                   Year = c(rep("2016", times = 8),
                            rep("2010", times = 8)))
dems$Year <- factor(dems$Year)

dems
```

```
##    Dem Year
## 1   30 2016
## 2   69 2016
## 3   99 2016
## 4   77 2016
## 5   29 2016
## 6   37 2016
## 7   38 2016
## 8   37 2016
## 9   30 2010
## 10  62 2010
## 11  50 2010
## 12  69 2010
## 13  27 2010
## 14  29 2010
## 15  44 2010
## 16  45 2010
```

The Hypotheses

- $H_0 : \bar{x}_{2010} - \bar{x}_{2016} = 0$
- $H_A : \bar{x}_{2010} - \bar{x}_{2016} \neq 0$
- Restricted Model: the best way of minimizing errors is to use the overall **grand mean**
- Full Model: the best way of minimizing errors is to use the **group-specific mean**

The Means

Let's get the grand mean to use in our Restricted model and the means of each group (% Dem in 2010 vs. % Dem in 2016):

```
grandMean <- mean(dems$Dem)

groupMeans <- dems %>%
  group_by(Year) %>%
  summarize(means = mean(Dem))

grandMean
```

```
## [1] 48.25
```

```
groupMeans
```

```
## # A tibble: 2 × 2
##   Year  means
##   <fct> <dbl>
## 1 2010   44.5
## 2 2016   52
```

The Restricted Model

```
dems$Mean <- rep(grandMean, times = nrow(dems))
dems
```

```
##    Dem Year  Mean
## 1   30 2016 48.25
## 2   69 2016 48.25
## 3   99 2016 48.25
## 4   77 2016 48.25
## 5   29 2016 48.25
## 6   37 2016 48.25
## 7   38 2016 48.25
## 8   37 2016 48.25
## 9   30 2010 48.25
## 10  62 2010 48.25
## 11  50 2010 48.25
## 12  69 2010 48.25
## 13  27 2010 48.25
## 14  29 2010 48.25
## 15  44 2010 48.25
## 16  45 2010 48.25
```

The Restricted Model

Step 1: Deviation Scores

```
dems$deviationScores <- dems$Dem - dems$Mean
dems
```

```
##    Dem Year  Mean deviationScores
## 1   30 2016 48.25          -18.25
## 2   69 2016 48.25           20.75
## 3   99 2016 48.25           50.75
## 4   77 2016 48.25           28.75
## 5   29 2016 48.25          -19.25
## 6   37 2016 48.25          -11.25
## 7   38 2016 48.25          -10.25
## 8   37 2016 48.25          -11.25
## 9   30 2010 48.25          -18.25
## 10  62 2010 48.25           13.75
## 11  50 2010 48.25            1.75
## 12  69 2010 48.25           20.75
## 13  27 2010 48.25          -21.25
## 14  29 2010 48.25          -19.25
## 15  44 2010 48.25           -4.25
## 16  45 2010 48.25           -3.25
```

The Restricted Model

Step 2: Square Deviation Scores

```
dems$dev2 <- dems$deviationScores ^2
dems
```

```
##     Dem Year  Mean deviationScores      dev2
## 1    30 2016 48.25          -18.25  333.0625
## 2    69 2016 48.25           20.75  430.5625
## 3    99 2016 48.25           50.75 2575.5625
## 4    77 2016 48.25           28.75  826.5625
## 5    29 2016 48.25          -19.25  370.5625
## 6    37 2016 48.25          -11.25  126.5625
## 7    38 2016 48.25          -10.25  105.0625
## 8    37 2016 48.25          -11.25  126.5625
## 9    30 2010 48.25          -18.25  333.0625
## 10   62 2010 48.25           13.75  189.0625
## 11   50 2010 48.25            1.75    3.0625
## 12   69 2010 48.25           20.75  430.5625
## 13   27 2010 48.25          -21.25  451.5625
## 14   29 2010 48.25          -19.25  370.5625
## 15   44 2010 48.25           -4.25   18.0625
## 16   45 2010 48.25           -3.25   10.5625
```

The Restricted Model

Step 3: Sum of Squares -- our **ERROR** term

```
Er <- sum(dems$dev2)
dems
```

```
##    Dem Year  Mean deviationScores      dev2
## 1   30 2016 48.25          -18.25  333.0625
## 2   69 2016 48.25           20.75  430.5625
## 3   99 2016 48.25           50.75 2575.5625
## 4   77 2016 48.25           28.75  826.5625
## 5   29 2016 48.25          -19.25  370.5625
## 6   37 2016 48.25          -11.25  126.5625
## 7   38 2016 48.25          -10.25  105.0625
## 8   37 2016 48.25          -11.25  126.5625
## 9   30 2010 48.25          -18.25  333.0625
## 10  62 2010 48.25           13.75  189.0625
## 11  50 2010 48.25            1.75    3.0625
## 12  69 2010 48.25           20.75  430.5625
## 13  27 2010 48.25          -21.25  451.5625
## 14  29 2010 48.25          -19.25  370.5625
## 15  44 2010 48.25           -4.25   18.0625
## 16  45 2010 48.25           -3.25   10.5625
```

```
Er
```

```
## [1] 6701
```

The Restricted Model

Step 4: Determine the degrees of freedom.

- Degrees of freedom deals with how much information is *free to vary*
- You get a feel for this the more you practice
- In this restricted model, we are guessing/estimating our grand mean of 48.25. $df = n - 1$, **df = 16 - 1 = 15**

```
dfr <- 15
```

The Full Model

```
dems <- data.frame(Dem = c(30, 69, 99, 77, 29, 37, 38, 37,
                           30, 62, 50, 69, 27, 29, 44, 45),
                   Year = c(rep("2016", times = 8),
                            rep("2010", times = 8)))
dems$Year <- factor(dems$Year)

dems$Mean <- c(rep(groupMeans$means[2], times = 8),
               rep(groupMeans$means[1], times = 8))
dems
```

```
##    Dem Year Mean
## 1   30 2016 52.0
## 2   69 2016 52.0
## 3   99 2016 52.0
## 4   77 2016 52.0
## 5   29 2016 52.0
## 6   37 2016 52.0
## 7   38 2016 52.0
## 8   37 2016 52.0
## 9   30 2010 44.5
## 10  62 2010 44.5
## 11  50 2010 44.5
## 12  69 2010 44.5
## 13  27 2010 44.5
## 14  29 2010 44.5
## 15  44 2010 44.5
```

The Full Model

Step 1: Deviation Scores

```
dems$deviationScores <- dems$Dem - dems$Mean
dems
```

```
##     Dem Year Mean deviationScores
## 1    30 2016 52.0           -22.0
## 2    69 2016 52.0            17.0
## 3    99 2016 52.0            47.0
## 4    77 2016 52.0            25.0
## 5    29 2016 52.0           -23.0
## 6    37 2016 52.0           -15.0
## 7    38 2016 52.0           -14.0
## 8    37 2016 52.0           -15.0
## 9    30 2010 44.5           -14.5
## 10   62 2010 44.5            17.5
## 11   50 2010 44.5             5.5
## 12   69 2010 44.5            24.5
## 13   27 2010 44.5           -17.5
## 14   29 2010 44.5           -15.5
## 15   44 2010 44.5            -0.5
## 16   45 2010 44.5             0.5
```

The Full Model

Step 2: Square Deviation Scores

```r
dems$dev2 <- dems$deviationScores ^2
dems
```

```
##    Dem Year Mean deviationScores    dev2
## 1   30 2016 52.0           -22.0  484.00
## 2   69 2016 52.0            17.0  289.00
## 3   99 2016 52.0            47.0 2209.00
## 4   77 2016 52.0            25.0  625.00
## 5   29 2016 52.0           -23.0  529.00
## 6   37 2016 52.0           -15.0  225.00
## 7   38 2016 52.0           -14.0  196.00
## 8   37 2016 52.0           -15.0  225.00
## 9   30 2010 44.5           -14.5  210.25
## 10  62 2010 44.5            17.5  306.25
## 11  50 2010 44.5             5.5   30.25
## 12  69 2010 44.5            24.5  600.25
## 13  27 2010 44.5           -17.5  306.25
## 14  29 2010 44.5           -15.5  240.25
## 15  44 2010 44.5            -0.5    0.25
## 16  45 2010 44.5             0.5    0.25
```

The Full Model

Step 3: Sum of Squares -- our **ERROR** term

```
Ef <- sum(dems$dev2)
dems
```

```
##    Dem Year Mean deviationScores    dev2
## 1   30 2016 52.0           -22.0  484.00
## 2   69 2016 52.0            17.0  289.00
## 3   99 2016 52.0            47.0 2209.00
## 4   77 2016 52.0            25.0  625.00
## 5   29 2016 52.0           -23.0  529.00
## 6   37 2016 52.0           -15.0  225.00
## 7   38 2016 52.0           -14.0  196.00
## 8   37 2016 52.0           -15.0  225.00
## 9   30 2010 44.5           -14.5  210.25
## 10  62 2010 44.5            17.5  306.25
## 11  50 2010 44.5             5.5   30.25
## 12  69 2010 44.5            24.5  600.25
## 13  27 2010 44.5           -17.5  306.25
## 14  29 2010 44.5           -15.5  240.25
## 15  44 2010 44.5            -0.5    0.25
## 16  45 2010 44.5             0.5    0.25
```

```
Ef
```

```
## [1] 6476
```

The Full Model

Step 4: Determine the degrees of freedom.

- Degrees of freedom deals with how much information is *free to vary*
- You get a feel for this the more you practice
- In this full model, we are guessing/estimating our 2 means (mean for 2010 and mean for 2016). $df = n - 2$, **df = 16 - 2 = 14**

```
dff <- 14
```

The Effect

$$\frac{(E_r - E_f)/(df_r - df_f)}{E_f/df_f}$$

```
effect <- ((Er - Ef) / (dfr - dff)) / (Ef/dff)
effect
```

```
## [1] 0.4864114
```

This is our F-statistic. Remember that $t^2 = F$. So to get our t-statistic, let's take the square root of our `effect`.

```
tstat <- sqrt(effect)
round(x = tstat, digits = 3)
```

```
## [1] 0.697
```

Model Comparison Approach

Is 0.697 more extreme than our critical value?

- For an $\alpha = .05$ in a two-tailed test with $df = 14$, the critical t value is 2.145.

Conclusion: No, it's not more extreme than the critical value. The weighted error term for the restricted is smaller than the weighted error term from the full. The grand mean was a better estimator than using individual group means. Therefore, the means are not statistically significantly different.

Model Comparison Approach

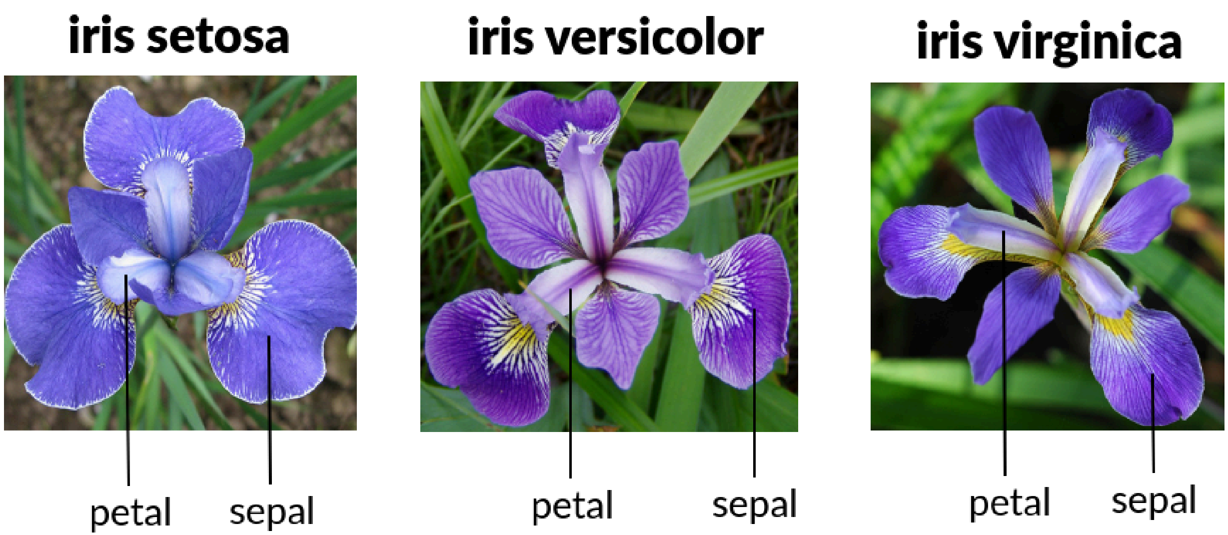
We just did an independent-samples t-test! Let's verify our results:

```
t.test(dems$Dem ~ dems$Year)
```

```
##
##      Welch Two Sample t-test
##
## data:  dems$Dem by dems$Year
## t = -0.69743, df = 11.406, p-value = 0.4995
## alternative hypothesis: true difference in means between group 2010 and group 20
## 95 percent confidence interval:
##  -31.06632  16.06632
## sample estimates:
## mean in group 2010 mean in group 2016
##               44.5               52.0
```

Scenario 3

- We have a dataset that looks at the lengths and widths of petals & sepals of the iris flower. It includes 3 different species of irises.
- *Question: are the sepal lengths different amongst the 3 species of irises?*



The Data

```
##    Sepal.Length Species
## 1           5.1  setosa
## 2           4.9  setosa
## 3           4.7  setosa
## 4           4.6  setosa
## 5           5.0  setosa
## 6           5.4  setosa
## 7           4.6  setosa
## 8           5.0  setosa
## 9           4.4  setosa
## 10          4.9  setosa
```

```
##    Sepal.Length    Species
## 51          7.0 versicolor
## 52          6.4 versicolor
## 53          6.9 versicolor
## 54          5.5 versicolor
## 55          6.5 versicolor
## 56          5.7 versicolor
## 57          6.3 versicolor
## 58          4.9 versicolor
## 59          6.6 versicolor
## 60          5.2 versicolor
```

```
##     Sepal.Length   Species
## 101          6.3 virginica
## 102          5.8 virginica
## 103          7.1 virginica
## 104          6.3 virginica
## 105          6.5 virginica
## 106          7.6 virginica
## 107          4.9 virginica
## 108          7.3 virginica
## 109          6.7 virginica
## 110          7.2 virginica
```

Reorder the data for teaching purposes

```
iris_sorted = iris %>%
  group_by(Sepal.Length, Species) %>%
  arrange(Sepal.Length)
```

The Hypotheses

- $H_0 : \bar{x}_{setosa} = \bar{x}_{versicolor} = \bar{x}_{virginica}$
- $H_A : \bar{x}_{setosa} \neq \bar{x}_{versicolor} \neq \bar{x}_{virginica}$
- Restricted Model: the best way of minimizing errors is to use the overall **grand mean**
- Full Model: the best way of minimizing errors is to use the **group-specific means**

The Means

Let's get the grand mean to use in our Restricted model and the means of each group:

```
grandMean <- mean(iris$Sepal.Length)

groupMeans <- iris %>%
  group_by(Species) %>%
  summarize(means = mean(Sepal.Length))

grandMean
```

```
## [1] 5.843333
```

```
groupMeans
```

```
## # A tibble: 3 × 2
##   Species    means
##   <fct>      <dbl>
## 1 setosa      5.01
## 2 versicolor  5.94
## 3 virginica   6.59
```

The Restricted Model

```
##   Sepal.Length Species     Mean
## 1          5.1  setosa 5.843333
## 2          4.9  setosa 5.843333
## 3          4.7  setosa 5.843333
## 4          4.6  setosa 5.843333
```

```
##    Sepal.Length    Species     Mean
## 51          7.0 versicolor 5.843333
## 52          6.4 versicolor 5.843333
## 53          6.9 versicolor 5.843333
## 54          5.5 versicolor 5.843333
```

```
##     Sepal.Length   Species     Mean
## 101          6.3 virginica 5.843333
## 102          5.8 virginica 5.843333
## 103          7.1 virginica 5.843333
## 104          6.3 virginica 5.843333
```

The Restricted Model

Step 1: Deviation Scores

```
##   Sepal.Length Species     Mean deviationScores
## 1          5.1  setosa 5.843333      -0.7433333
## 2          4.9  setosa 5.843333      -0.9433333
## 3          4.7  setosa 5.843333      -1.1433333
## 4          4.6  setosa 5.843333      -1.2433333

##    Sepal.Length    Species     Mean deviationScores
## 51          7.0 versicolor 5.843333       1.1566667
## 52          6.4 versicolor 5.843333       0.5566667
## 53          6.9 versicolor 5.843333       1.0566667
## 54          5.5 versicolor 5.843333      -0.3433333

##     Sepal.Length   Species     Mean deviationScores
## 101          6.3 virginica 5.843333      0.45666667
## 102          5.8 virginica 5.843333     -0.04333333
## 103          7.1 virginica 5.843333      1.25666667
## 104          6.3 virginica 5.843333      0.45666667
```

The Restricted Model

Step 2: Square Deviation Scores

```
##   Sepal.Length Species     Mean deviationScores      dev2
## 1          5.1  setosa 5.843333      -0.7433333 0.5525444
## 2          4.9  setosa 5.843333      -0.9433333 0.8898778
## 3          4.7  setosa 5.843333      -1.1433333 1.3072111
## 4          4.6  setosa 5.843333      -1.2433333 1.5458778
```

```
##    Sepal.Length    Species     Mean deviationScores      dev2
## 51          7.0 versicolor 5.843333       1.1566667 1.3378778
## 52          6.4 versicolor 5.843333       0.5566667 0.3098778
## 53          6.9 versicolor 5.843333       1.0566667 1.1165444
## 54          5.5 versicolor 5.843333      -0.3433333 0.1178778
```

```
##     Sepal.Length   Species     Mean deviationScores        dev2
## 101          6.3 virginica 5.843333      0.45666667 0.208544444
## 102          5.8 virginica 5.843333     -0.04333333 0.001877778
## 103          7.1 virginica 5.843333      1.25666667 1.579211111
## 104          6.3 virginica 5.843333      0.45666667 0.208544444
```

The Restricted Model

Step 3: Sum of Squares -- our **ERROR** term

```
Er <- sum(restricted$dev2)
Er
```

```
## [1] 102.1683
```

The Restricted Model

Step 4: Determine the degrees of freedom.

- Degrees of freedom deals with how much information is *free to vary*
- You get a feel for this the more you practice
- In this restricted model, we are guessing/estimating our grand mean of 5.843. $df = n - 1$, **df = 150 - 1 = 149**

```
dfr <- 149
```

The Full Model

```
##   Sepal.Length Species  Mean
## 1          5.1  setosa 5.006
## 2          4.9  setosa 5.006
## 3          4.7  setosa 5.006
## 4          4.6  setosa 5.006

##    Sepal.Length    Species  Mean
## 51          7.0 versicolor 5.936
## 52          6.4 versicolor 5.936
## 53          6.9 versicolor 5.936
## 54          5.5 versicolor 5.936

##     Sepal.Length   Species  Mean
## 101          6.3 virginica 6.588
## 102          5.8 virginica 6.588
## 103          7.1 virginica 6.588
## 104          6.3 virginica 6.588
```

The Full Model

Step 1: Deviation Scores

```
##   Sepal.Length Species  Mean deviationScores
## 1          5.1  setosa 5.006           0.094
## 2          4.9  setosa 5.006          -0.106
## 3          4.7  setosa 5.006          -0.306
## 4          4.6  setosa 5.006          -0.406
```

```
##    Sepal.Length    Species  Mean deviationScores
## 51          7.0 versicolor 5.936           1.064
## 52          6.4 versicolor 5.936           0.464
## 53          6.9 versicolor 5.936           0.964
## 54          5.5 versicolor 5.936          -0.436
```

```
##     Sepal.Length   Species  Mean deviationScores
## 101          6.3 virginica 6.588          -0.288
## 102          5.8 virginica 6.588          -0.788
## 103          7.1 virginica 6.588           0.512
## 104          6.3 virginica 6.588          -0.288
```

The Full Model

Step 2: Square Deviation Scores

```
##   Sepal.Length Species  Mean deviationScores     dev2
## 1          5.1  setosa 5.006           0.094 0.008836
## 2          4.9  setosa 5.006          -0.106 0.011236
## 3          4.7  setosa 5.006          -0.306 0.093636
## 4          4.6  setosa 5.006          -0.406 0.164836

##    Sepal.Length    Species  Mean deviationScores     dev2
## 51          7.0 versicolor 5.936           1.064 1.132096
## 52          6.4 versicolor 5.936           0.464 0.215296
## 53          6.9 versicolor 5.936           0.964 0.929296
## 54          5.5 versicolor 5.936          -0.436 0.190096

##     Sepal.Length   Species  Mean deviationScores     dev2
## 101          6.3 virginica 6.588          -0.288 0.082944
## 102          5.8 virginica 6.588          -0.788 0.620944
## 103          7.1 virginica 6.588           0.512 0.262144
## 104          6.3 virginica 6.588          -0.288 0.082944
```

The Full Model

Step 3: Sum of Squares -- our **ERROR** term

```
Ef <- sum(full$dev2)
Ef
```

```
## [1] 38.9562
```

The Full Model

Step 4: Determine the degrees of freedom.

- Degrees of freedom deals with how much information is *free to vary*
- You get a feel for this the more you practice
- In this full model, we are guessing/estimating our 3 means (mean for each species). $df = n - 3$, **df = 150 - 3 = 147**

```
dff <- 147
```

The Effect

$$\frac{(E_r - E_f)/(df_r - df_f)}{E_f/df_f}$$

```
effect <- ((Er - Ef) / (dfr - dff)) / (Ef/dff)
effect
```

```
## [1] 119.2645
```

This is our F-statistic. This is an ANOVA, so we can stick with the F-statistic. (back to this in a sec)

```
round(x = effect, digits = 3)
```

```
## [1] 119.265
```

Model Comparison Approach

Is 119.265 more extreme than our critical value?

- The critical value for this test is 3.058

Conclusion: The weighted error term of the restricted is larger than the weighted error term of the full. Using each group's mean was a better estimator, compared to the overall grand mean. Therefore, the means are statistically significantly different.

Model Comparison Approach

We just did a oneway ANOVA! Let's verify our results:

```
summary(aov(Sepal.Length ~ Species, data = iris))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Species       2  63.21  31.606   119.3 <2e-16 ***
## Residuals   147  38.96   0.265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Linking Distributions

- ANOVA is a comparison of means. But we are interested in variance. That is, we are trying to figure out the source of the variance (more on this next time)
- When we want to make inferences about sample variability, we need to know the sampling distribution.
- We can make use of the χ^2 distribution here. It has a single parameter, ν, related to the sample size, N. There is an important relationship between the normal distribution and the χ^2 distribution:
    - The sum of squared standard normal variables will have a χ^2 distribution.

Linking Distributions

- The F-distribution is a ratio of two χ^2 variables -- 2 different variances
- However, it's weighted by degrees of freedom. But there needs to be a df for the numerator AND df for the denominator
- So the F-distribution has 2 parameters: 2 different degrees of freedom. As we've talked about it thus far, think of this as the df for the full and restricted models. Next time, we'll translate into ANOVA terminology

$$F_{\nu_1 \nu_2} = \frac{\frac{\chi^2_{\nu_1}}{\nu_1}}{\frac{\chi^2_{\nu_2}}{\nu_2}}$$

The F-distribution is one-sided (like the χ^2). You only care about the upper tail. You can't have a negative variance. If this is the ratio of 2 variances (ish), it's not going to be negative (F-statistic ≥ 0)

The t, χ^2, and F distributions all depend on the normal distribution assumption. The normal distribution is the "parent" population.

As sample size increases, t and χ^2 converge on the normal. F converges on : $\frac{\chi^2_{\nu_1}}{\nu_1}$ with the numerator depending on the normal

But keep in mind

These probability distributions have a key assumption:

- the sample is a random selection from the population

If the sample is not a random selection, the rules of probability don't apply.

Utility

Why do this painful process?

A model is what **YOU** define. It's how you think the world works. The restricted model is really just an embodiment of the null hypothesis! The full model is the embodiment of the alternative hypothesis!

Minimizing error terms is how we evaluate multitudes of models!

> *All models are wrong, but some are useful* - George Box

Plus, model comparison frameworks come up more formally in some advanced types of statistics.

Next time

Translate all of this into classic, textbook ANOVA terminology