F tests for ANOVA (Sleuth3 Section 5.3)

Iris Flowers Example

We have measurements of the characteristics of iris flowers, from each of three different species. To make it possible to do some calculations by hand in a bit, I have subset to just 5 flowers from each species.

The ANOVA Model

- We have I groups ($I = 3$ for iris example)
- Total sample size n ($n = 15$ for iris example with reduced sample size)
- Observations in group i follow a Normal(μ_i, σ^2) distribution
  - (Potentially) different mean for each group
  - Same variance across all groups
- All observations are independent of each other: knowing that one is above its group mean doesn't tell you whether or not another is above its group mean.

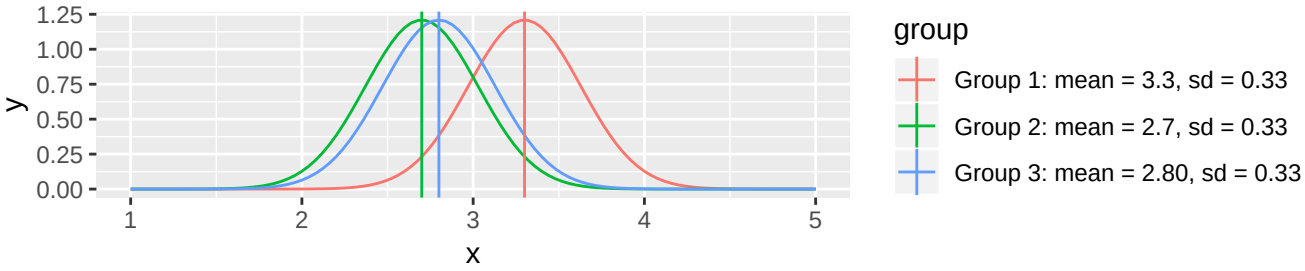
Parameters:

μ_1 = Average sepal width among all setosa flowers (in the region where the flowers in the sample were found?)

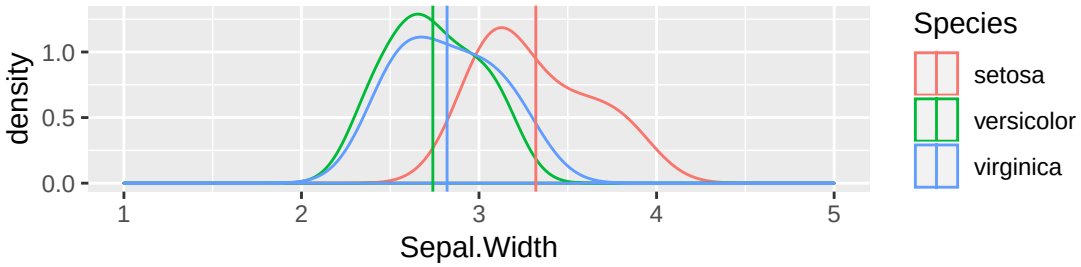
μ_2 = Average sepal width among all versicolor flowers (in the region where the flowers in the sample were found?)

μ_3 = Average sepal width among all virginica flowers (in the region where the flowers in the sample were found?)

Theoretical model:



Compare to the plot for our iris data:



Some questions we might ask:

- Overall, is there a difference in mean sepal widths for the three species? (F test)
  - $H_0 : \mu_1 = \mu_2 = \mu_3$
- Is the mean for setosa flowers different from the mean for versicolor flowers? (t test)
  - $H_0 : \mu_1 = \mu_2$, or $\mu_1 - \mu_2 = 0$
- Is the mean for the setosa flower, found in the arctic, different from the mean for non-arctic flowers? (t test)
  - $H_0 : \frac{1}{2}(\mu_2 + \mu_3) = \mu_1$, or $\frac{1}{2}(\mu_2 + \mu_3) - \mu_1 = 0$

F Test Concepts

Notation:

- i: which group? ($i = 1$, 2, or 3 for iris flowers since there are $I = 3$ species)
- j: which observational unit within its group? (if $i = 2$ and $j = 3$, we're talking about the 3rd versicolor flower)
- Y_{ij}: response variable value for unit j in group i
- $\bar{Y}_i$: sample mean for group i

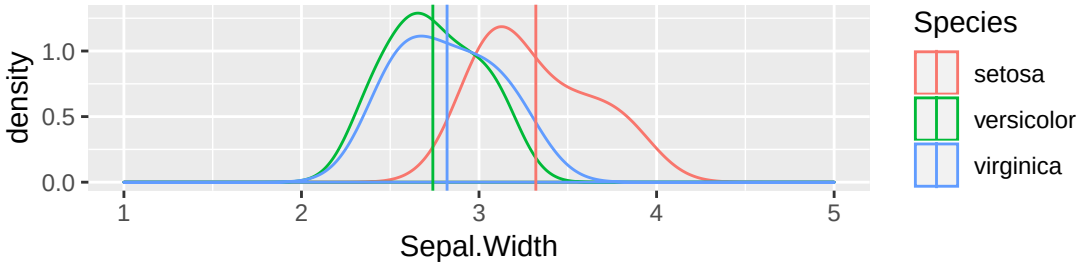
Test set up

Suppose we are conducting a test of $H_0 : \mu_1 = \mu_2 = \mu_3$ vs. H_A : at least one of the means differs from the others

We frame this as a comparison of two models.

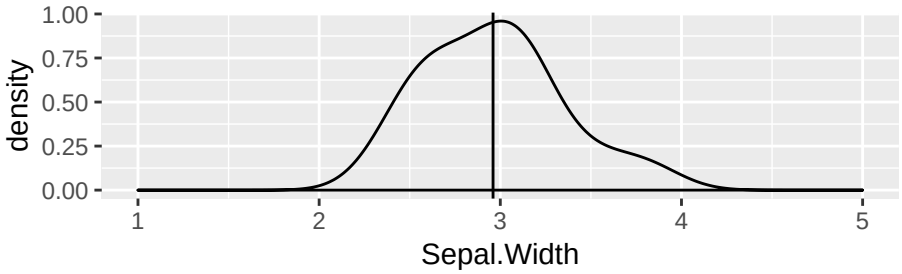
1. Full Model, separate means for all groups (corresponds to H_A)

3 mean parameters: μ_1, μ_2, μ_3



2. Reduced Model, one mean common to all observations (corresponds to H_0)

1 mean parameter: μ



We can measure the usefulness of a model by the size of its residuals

- Imagine if we knew an iris flower was from the setosa species, and we wanted to guess its sepal width.
  - We could guess the group mean for setosa flowers, $\bar{Y}_1$
- **Residual**: difference between observed value for response variable and fitted value for response variable.

$$res_{ij} = Y_{ij} - \bar{Y}_i$$

- In general:

$$\text{Better Model} \Leftrightarrow \text{Better Guesses} \Leftrightarrow \text{Smaller Residuals}$$

- The Full Model will have smaller residuals (on average) than the Reduced Model
- Question the F test answers: are the residuals from the full model smaller than the residuals from the reduced model by a statistically significant margin?

Measuring the size of residuals from a model

- Residual Sum of Squares: Square the residuals and add them up

$$\sum_i \sum_j (res_{ij})^2 = \sum_i \sum_j (Y_{ij} - \bar{Y}_i)^2$$

- Mean Squared Residual:

$$\frac{\text{Residual Sum of Squares}}{\text{Degrees of Freedom}}$$

Extra Sum of Squares

$$\text{Extra Sum of Squares} = \text{Residual Sum of Squares, Reduced Model} - \text{Residual Sum of Squares, Full Model}$$

- Always positive because
  - Reduced Model is more limited than Full Model
  - Reduced Model has larger residuals than Full Model
- If Extra Sum of Squares is really big, the Full Model is much better than the Reduced Model
- Extra Sum of Squares has degrees of freedom equal to the difference in degrees of freedom for the full model and the reduced model. In this case:

$$df = (n-1) - (n-I) = I - 1$$

F Statistic

- "How big is the improvement in Residual Sum of Squares from using the Full Model instead of the Reduced Model"?
  - Size of improvement is measured relative to the size of residuals in the full model

$$\begin{aligned} F &= \frac{\text{(Extra Sum of Squares)/(Extra Degrees of Freedom)}}{\text{(Residual Sum of Squares, Full Model)/(Degrees of Freedom, Full Model)}} \\ &= \frac{\text{(Extra Sum of Squares)}/(I-1)}{\text{(Residual Sum of Squares, Full Model)}/(n-I)} \end{aligned}$$

- If $H_0 : \mu_1 = \mu_2 = \mu_3$ is **true**, then...
  - Full Model **isn't better** than Reduced Model
  - Residual Sum of Squares, Full Model is **similar to** Residual Sum of Squares, Reduced Model
  - Extra Sum of Squares is **small**
  - F Statistic is **small**
- If $H_O : \mu_1 = \mu_2 = \mu_3$ is **not true**, then...
  - Full Model **is better** than Reduced Model
  - Residual Sum of Squares, Full Model is **smaller than** Residual Sum of Squares, Reduced Model
  - Extra Sum of Squares is **large**
  - F Statistic is **large**
- **A large value of F statistic is evidence against** H_0
- We have to keep track of two degrees of freedom: $I - 1$, $n - I$.

R Code

```
iris_fit <- lm(Sepal.Width ~ Species, data = iris)
anova(iris_fit)
```

```
## Analysis of Variance Table
##
## Response: Sepal.Width
##           Df Sum Sq Mean Sq F value Pr(>F)
## Species    2  0.988 0.49400  5.6565 0.0186 *
## Residuals 12  1.048 0.08733
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What is the conclusion of the test?

Verifying the ANOVA table calculations

The ANOVA table is intended to help organize calculations for the F test. Let's work through where all the numbers in the table come from.

Source	Sum of Squares	Degrees of Freedom	Mean Square	F	p-value
Extra					
Full Model					
Reduced Model					

1. Calculate Residual Sum of Squares for the Full Model (Use the table on the next page. I promise I'll never make you do this by hand again.)
   a. Find the group mean for each of the three groups
   b. Find the residual for each observation for the Full Model (observed value minus group mean)
   c. Find the squared residual for each observation
   d. Find the sum of squared residuals for the full model.
   e. Enter the result in the "Full Model Sum of Squares" cell in the ANOVA table.
2. Calculate Residual Sum of Squares for the Reduced Model (I will also never make you do this by hand again.)
   a. Find the mean of all 15 observations
   b. Find the residual for each observation for the Reduced Model
   c. Find the squared residual for each observation
   d. Find the sum of squared residuals for the reduced model.
   e. Enter the result in the "Reduced Model Sum of Squares" cell in the ANOVA table.
3. Calculate the Extra Sum of Squares and enter it in the ANOVA table
4. Enter the degrees of freedom for the Full Model in the ANOVA table (n - I)
5. Enter the degrees of freedom for the Reduced Model in the ANOVA table (n - 1)
6. Enter the degrees of freedom for the Extra Sum of Squares in the ANOVA table (I - 1)
7. Find the Mean Square for the Full Model (Sum of Squares divided by its degrees of freedom)
8. Find the Mean Square for the Extra Sum of Squares (Sum of Squares divided by its degrees of freedom)
9. Find the F statistic (Mean Square for Extra divided by Mean Square for Full Model)
10. Find the p-value. Your R code is `pf(5.6565, df1 = 2, df2 = 12, lower.tail = FALSE)`

		Full Model			Reduced Model		
Y_{ij}	Species	Group Mean	Residual	Squared Residual	Grand Mean	Residual	Squared Residual
3.1	setosa						
3.0	setosa						
3.5	setosa						
3.2	setosa						
3.8	setosa						
2.7	versicolor						
2.4	versicolor						
2.6	versicolor						
2.9	versicolor						
3.1	versicolor						
3.2	virginica						
2.5	virginica						
2.8	virginica						
2.6	virginica						
3.0	virginica						
Total							