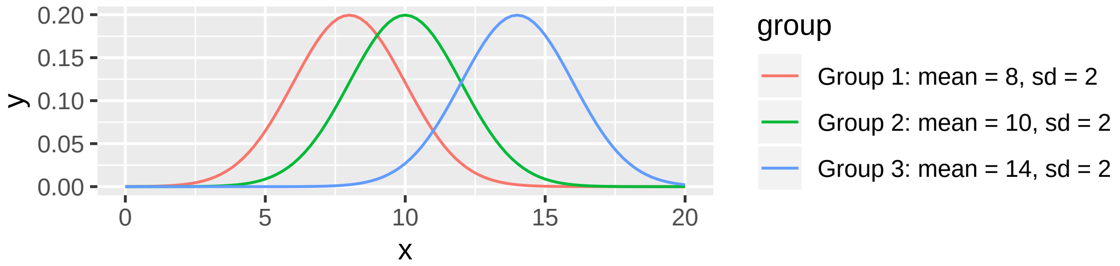
ANOVA: Concepts for t-based and F-based inference

Sleuth3 Sections 6.2 and 5.2

2019-09-20

The ANOVA Model

- Observations in group i follow a Normal(μ_i, σ^2) distribution
  - (Potentially) different mean for each group
  - Same variance across all groups



Notation

- We have I groups ($I = 3$ for iris example)
- Sample size of n_i for group i, total sample size $n = n_1 + n_2 + \cdots + n_I$
- y_{ij}: response variable value for unit j in group i
  - 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)
- $\bar{y}_i$: sample mean for group i

Two Types of Hypotheses:

1. $H_0 : C_1\mu_1 + C_2\mu_2 + \cdots + C_I\mu_I = 0$.
   - Some combination of means is equal to 0
   - Can specify with one = sign
   - Use a t test
2. $H_0 : \mu_1 = \mu_2 = \mu_3$.
   - Some of the group means are actually equal to each other
   - Need multiple = signs to specify
   - Use an F test

t statistic for ANOVA model

	General Set Up	**Single Mean**	**ANOVA (I groups)**
Parameter	A number describing the population we are interested in	μ: population mean (or difference in means with paired data).	γ: linear combination of population means for different groups $\gamma = C_1\mu_1 + C_2\mu_2 + \cdots + C_I\mu_I$
Estimate	An estimate of the parameter based on the data in our sample	$\hat{\mu} = \bar{Y}$: sample mean (or difference in sample means with paired data).	Linear combination of sample means for different groups $\hat{\gamma} = C_1\bar{Y}_1 + C_2\bar{Y}_2 + \cdots C_I\bar{Y}_I$
SD(Estimate)	Measures variability of the estimate across different samples.	$\sigma/\sqrt{n}$	$\sigma\sqrt{\frac{C_1^2}{n_1} + \frac{C_2^2}{n_2} + \cdots + \frac{C_I^2}{n_I}}$
SE(Estimate)	An estimate of SD(Estimate)	$s/\sqrt{n}$	$s_{pooled}\sqrt{\frac{C_1^2}{n_1} + \frac{C_2^2}{n_2} + \cdots + \frac{C_I^2}{n_I}}$
Estimate of σ	How do we estimate the variance of residuals?	Based on squared differences from the overall sample mean $s = \sqrt{\frac{\sum_{j=1}^{n}(y_j - \bar{y})^2}{n-1}}$	Based on squared differences from the group means $s_{pooled} = \sqrt{\frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2}{n-I}}$
t statistic	$t = \frac{\textbf{Estimate} - \textbf{Parameter}}{\textbf{SE(Estimate)}}$	$t = \frac{\bar{Y} - \mu}{s/\sqrt{n}}$	$t = \frac{\hat{\gamma} - \gamma}{s_p\sqrt{\frac{C_1^2}{n_1} + \frac{C_2^2}{n_2} + \cdots + \frac{C_I^2}{n_I}}}$
Degrees of Freedom		$n - 1$	$n - I$
Confidence Interval	**Estimate** $\pm\, t^* SE($**Estimate**$)$	$\bar{Y} \pm t^* SE(\bar{Y})$	$\hat{\gamma} \pm t^* SE(\hat{\gamma})$
P-value	• Calculate the t statistic as above, assuming H_0 is true (plug in the value of the parameter from H_0) • If the null hypothesis were true, what proportion of samples would have a t statistic at least as extreme as the value you just calculated?		

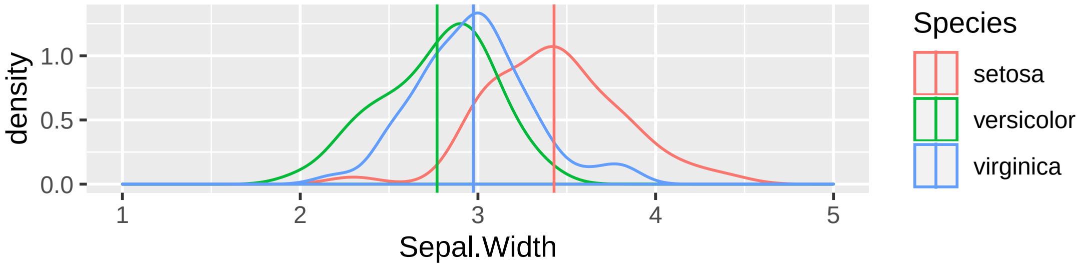
F tests

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.

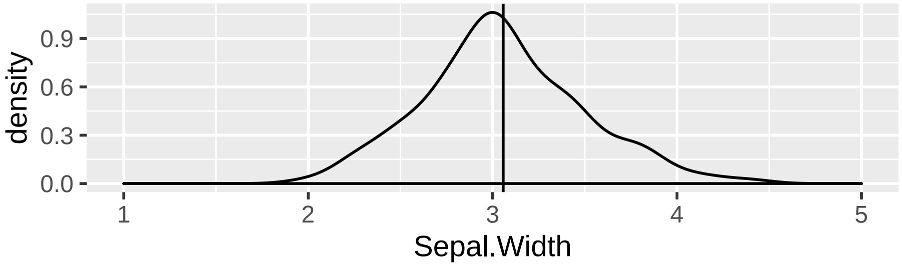
1. Full Model, separate means for all groups (corresponds to H_A)

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



How should we measure the usefulness of a model?

- Suppose we know a flower is from the setosa species, and we want to guess its sepal width. Which guess is better? Why? Can you think of a quantitative way to explain?
  - The group mean for setosa flowers, $\bar{Y}_1$. Location of red line in top plot, about 3.5
  - The overall mean for iris flowers, $\bar{Y}$. Location of black line in lower plot, about 3.

Residuals

- **Residual**: difference between observed value for response variable and fitted value for response variable.

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

- In general: Better Model ⇔ Better Guesses ⇔ Smaller Residuals
- The Full Model will have smaller residuals (on average) than the Reduced Model
- F test answers: are the residuals from the full model enough smaller than the residuals from the reduced model that I think the full model is necessary?

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:

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

Example

Suppose I have just 3 flowers of each species. Below is an example of the calculation of the RSS for the reduced model (one mean) and the full model (separate means for each group).

				Reduced Model			Full Model		
i	j	Species	Sepal Width (Y_{ij})	Mean	Residual	Squared Residual	Mean	Residual	Squared Residual
1	1	setosa	3.9	3.044	0.856	0.733	3.4	0.5	0.25
1	2	setosa	3.1	3.044	0.056	0.003	3.4	-0.3	0.09
1	3	setosa	3.2	3.044	0.156	0.024	3.4	-0.2	0.04
2	1	versicolor	2.4	3.044	-0.644	0.415	2.467	-0.067	0.004
2	2	versicolor	2.6	3.044	-0.444	0.197	2.467	0.133	0.018
2	3	versicolor	2.4	3.044	-0.644	0.415	2.467	-0.067	0.004
3	1	virginica	3.3	3.044	0.256	0.066	3.267	0.033	0.001
3	2	virginica	2.7	3.044	-0.344	0.118	3.267	-0.567	0.321
3	3	virginica	3.8	3.044	0.756	0.572	3.267	0.533	0.284
Total						2.543			1.012

Extra Sum of Squares

$$\begin{aligned}\text{Extra Sum of Squares} &= \text{Residual Sum of Squares, Reduced Model} - \text{Residual Sum of Squares, Full Model}\\ &= 2.543 - 1.012\\ &= 1.531\end{aligned}$$

- 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
- You can calculate the degrees of freedom for the Extra Sum of Squares in either of two ways:
  - difference in degrees of freedom for the full model and the reduced model: $(n-1)-(n-I) = I - 1 = 3 - 1 = 2$
  - difference in number of parameters for the mean between full and reduced model: $3 - 1 = 2$

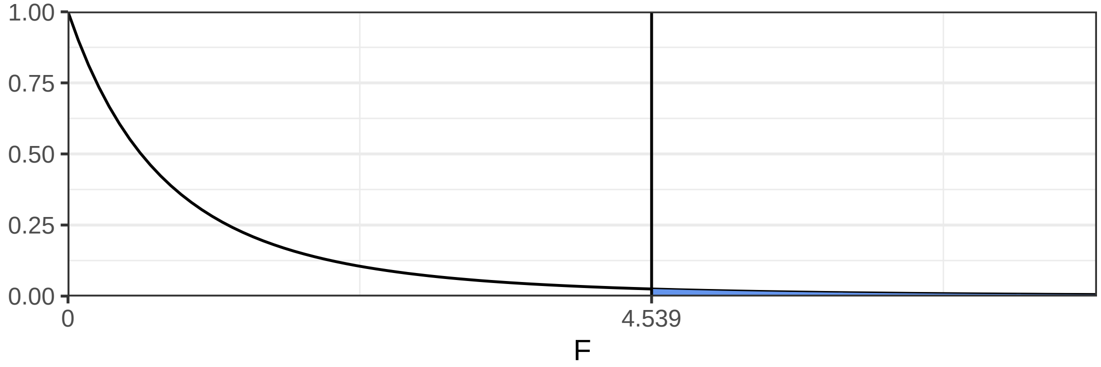
F Statistic

- "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

$$\begin{aligned}F &= \frac{\text{(Extra Sum of Squares)/(Extra Degrees of Freedom)}}{\text{(Residual Sum of Squares, Full Model)/(Degrees of Freedom, Full Model)}}\\ &= \frac{1.531/(3-1)}{1.012/(9-3)}\\ &= 4.539\end{aligned}$$

- 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

- For finding p-values we are interested in the probability of getting an F statistic at least as large as the F statistic we got from our sample, if H_0 is true.



- We have to keep track of two degrees of freedom