Two’s Company, Three’s . . . ANOVA!

Kathleen M. Saul

May 10, 2011

Thinking back to two sample tests

- To evaluate the difference between two independent samples, we use a t-statistic and t-test:
- When the population variances equal,
  - $t = \dfrac{(xbar_1 - xbar_2) - (mu_1 - mu_2)}{sqrt\{sp^2 * [(1/n_1) + (1/n_2)]\}}$
    - From our null hypothesis, $(mu_1 - mu_2) = 0$
    - The numerator then relects the difference between the means of the two samples and denominator contains variance terms capturing the variation within each sample.

To compare three or more independent samples

- Example: Comparing sediment load in Mack Creek vs. that in Lookout Creek vs. that in McRae Creek in HJ Andrews Experimental Forest
- Data would take the form of three columns of independent measurements

ANOVA = Analysis of Variance

- Ho: $mu_1 = mu_2 = . . .= mu_i$
- Ha: At least two means differ
- ANOVA analyzes sample means and variances to determine if the population means differ.
  - Requires the response variable be normally distributed with equal population variances.
- For a One Way ANOVA, we have data of the form:

Group 1	Group 2	Group 3	. . .	Group K
x11	x21	x31		xk1
x12	x22	x32		xk2
. . .				. . .
$x1n_1$	$x2n_2$	$x3n_3$		xkn_l

(K columns and n_l rows)

- SSE = Variation of each observation around the group mean
  - $SSE = \Sigma_k \Sigma_{ni} (x_{kni} – xbar_{ni})^2$
- SSG = Variation of the group means around the overall mean
  - $SSG = \Sigma_{ni} (xbar_{ni} – xbarbar)^2$
- SST = Variation of each observation around the overall mean
  - $SST = \Sigma_k \Sigma_{ni} (x_{kni} – xbarbar)^2$

- If $N = n_1 + n_2 + . . . + n_k$
- $MSE = SSE / (N – k)$
- $MST = SST / (N – 1)$
  - Where $(N – K)$ and $(N – 1)$ represent the associated degrees of freedom.

- The statistic of interest is F

  $$F = \frac{\text{variance between samples}}{\text{variance within samples}}$$

  F = MST / MSE

  with numerator df = (k – 1) and

  denominator df = (N – k)

Consider my fictitious data on the HJ Andrews creek sediment loads:

Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
MackCreek	9	105	11.66666667	13.75		
LookoutCreek	9	177	19.66666667	14.5		
McRaeCreek	9	190	21.11111111	11.86111111		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	465.8518519	2	232.9259259	17.42105263	2.11971E-05	3.402826105
Within Groups	320.8888889	24	13.37037037			
Total	786.7407407	26				

- If the p-value associated with F is less than alpha, we reject the null that all the population means are equal.
- But how do we know which groups differ???
  - Look at the pairwise differences in means:
    - Examine the Least Significant Difference (LSD)
    - LSD = $t_{(alpha/2)}$ * sqrt {MSE ($1/n_a + 1/n_b$)}

      where MSE is the variation within groups
    - mu_a and mu_b differ significantly if

      $|xbar_a – xbar_b| > LSD$

- But, we want the probability of a Type I error (alpha) to be no more than alpha.
- To correct this we must partition alpha for each of the pairs so that the total equals alpha.
- Let C = {k * (k – 1)} / 2
  - Where k is the number of pairwise combinations
- Then, newalpha = alpha / C
- Use $t_{(newalpha / 2)}$ to determine the LSD with df = (N – k)

Multiple Comparisons			
			LSD
Treatment	Treatment	Difference	Alpha = 0.0008333
MackCreek	*LookoutCreek*	-8.00	6.58
	McRaeCreek	-9.44	6.58
LookoutCreek	*McRaeCreek*	-1.44	6.58

We would conclude that Mack Creek and Lookout Creek have significantly different sediment loads, as do Mack Creek and McRae Creek. Lookout Creek and McRae Creek do not differ significantly.

Two Way ANOVA: Randomized Block Design

Consider data of the form:

		Treatment		
Group	A	B	C	D
1				
2				
3				
4				
5				
. . .				
n				

The response variable is expected to be normally distributed, and population variances are assumed equal.

- Ho: $mu_A = mu_B = mu_C$
- Ha: At least two means differ
- Results in output of the form:

ANOVA							
Source of Variation							
		SS	df	MS	F	p-value	F Crit
Rows							
Columns							
Error							
Total							

- The F statistic for the rows (Group) indicates whether there are statistically significant differences between the groups.
- The F statistic for the columns (Treatment) tells if the means of the treatments statistically differ.
- Follow up by testing pairwise to determine which pairs differ.

Two Factor ANOVA: Factorial Experiment

- For data that take the form:

		Factor A: Fertilizer		
		Brand X	Brand Y	None
Factor B: Light Condition				
Full Sun				
Shade				

  - All possible combinations of levels of factors are considered.
  - Assumes samples are independent.

- Need to perform multiple F tests
- First:
  - Ho: No difference between the means of the a levels of factor A
  - Ha: At least two means differ
- Next:
  - Ho: No difference between the means of the b levels of factor B
  - Ha: At least two means differ
- Finally:
  - Factors A and B do not interact to affect mean responses
  - Factors A and B do interact to affect mean responses

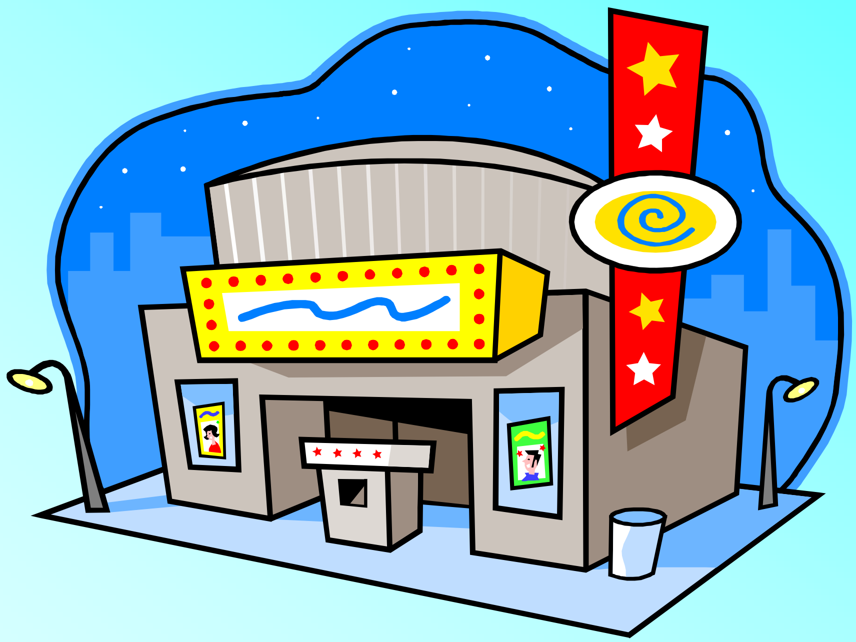
- Results in an output of the form:

ANOVA							
Source of Variation							
		SS	df	MS	F	p-value	F Crit
Factor A							
Factor B							
Interaction							
Within							
Total							

- Use the associated F statistics and p-values to test the three sets of hypotheses.

Coming Attractions:

R commands and examples of output.



UNIVERSITY OF DELAWARE

FIGHTIN' BLUE HENS