Basic Analysis of Variance and the General Linear Model

Psy 420

Andrew Ainsworth

Why is it called analysis of variance anyway?

- If we are interested in group mean differences, why are we looking at variance?
- t-test only one place to look for variability
- More groups, more places to look
- Variance of group means around a central tendency (grand mean – ignoring group membership) really tells us, on average, how much each group is different from the central tendency (and each other)

Why is it called analysis of variance anyway?

- Average mean variability around GM needs to be compared to average variability of scores around each group mean
- Variability in any distribution can be broken down into conceptual parts:

  total variability = (variability of each group mean around the grand mean) + (variability of each person’s score around their group mean)

General Linear Model (GLM)

- The basis for most inferential statistics (e.g. 420, 520, 524, etc.)
- Simple form of the GLM

$$\text{score} = \text{grand mean} + \text{independent variable} + \text{error}$$

$$Y = \mu + \alpha + \varepsilon$$

General Linear Model (GLM)

- The basic idea is that everyone in the population has the same score (the grand mean) that is changed by the effects of an independent variable (A) plus just random noise (error)
- Some levels of A raise scores from the GM, other levels lower scores from the GM and yet others have no effect.

General Linear Model (GLM)

- Error is the “noise” caused by other variables you aren’t measuring, haven’t controlled for or are unaware of.
  - Error like A will have different effects on scores but this happens independently of A.
  - If error gets too large it will mask the effects of A and make it impossible to analyze the effects of A
  - Most of the effort in research designs is done to try and minimize error to make sure the effect of A is not “buried” in the noise.
  - The error term is important because it gives us a “yard stick” with which to measure the variability cause by the A effect. We want to make sure that the variability attributable to A is greater than the naturally occurring variability (error)

GLM

- Example of GLM – ANOVA backwards
  - We can generate a data set using the GLM formula
  - We start off with every subject at the GM (e.g. μ=5)

a_1		a_2	
Case	Score	Case	Score
s_1	5	s_6	5
s_2	5	s_7	5
s_3	5	s_8	5
s_4	5	s_9	5
s_5	5	s_{10}	5

GLM

- Then we add in the effect of A (a1 adds 2 points and a2 subtracts 2 points)

a_1		a_2	
Case	Score	Case	Score
s_1	$5 + 2 = 7$	s_6	$5 - 2 = 3$
s_2	$5 + 2 = 7$	s_7	$5 - 2 = 3$
s_3	$5 + 2 = 7$	s_8	$5 - 2 = 3$
s_4	$5 + 2 = 7$	s_9	$5 - 2 = 3$
s_5	$5 + 2 = 7$	s_{10}	$5 - 2 = 3$
$\sum Y_{a_1} = 35$		$\sum Y_{a_2} = 15$	
$\sum Y_{a_1}^2 = 245$		$\sum Y_{a_2}^2 = 45$	
$\overline{Y}_{a_1} = 7$		$\overline{Y}_{a_3} = 3$	

GLM

- Changes produced by the treatment represent deviations around the GM

$$n\sum(\bar{Y}_j - GM)^2 = n[(7-5)^2 + (3-5)^2] =$$

$$5(2)^2 + 5(-2)^2 \, or \; 5[(2)^2 + (-2)^2] = 40$$

GLM

- Now if we add in some random variation (error)

a1		a2		SUM
Case	Score	Case	Score	
s1	5 + 2 + 2 = 9	s6	5 – 2 + 0 = 3	
s2	5 + 2 + 0 = 7	s7	5 – 2 – 2 = 1	
s3	5 + 2 – 1 = 6	s8	5 – 2 + 0 = 3	
s4	5 + 2 + 0 = 7	s9	5 – 2 + 1 = 4	
s5	5 + 2 – 1 = 6	s10	5 – 2 + 1 = 4	
$\sum Y_{a_1} = 35$		$\sum Y_{a_2} = 15$		$\sum Y = 50$
$\sum Y_{a_1}^2 = 251$		$\sum Y_{a_2}^2 = 51$		$\sum Y^2 = 302$
$\bar{Y}_{a_1} = 7$		$\bar{Y}_{a_3} = 3$		$\bar{Y} = 5$

GLM

- Now if we calculate the variance for each group:

$$s_{N-1}^{2}=\frac{\sum Y_{a_1}^{2}-\frac{(\sum Y)^{2}}{N}}{N-1}=\frac{251-\frac{35^{2}}{5}}{4}=1.5$$

$$s_{N-1}^{2}=\frac{\sum Y_{a_2}^{2}-\frac{(\sum Y)^{2}}{N}}{N-1}=\frac{51-\frac{15^{2}}{5}}{4}=1.5$$

- The average variance in this case is also going to be 1.5 (1.5 + 1.5 / 2)

GLM

- We can also calculate the total variability in the data regardless of treatment group

$$s_{N-1}^2 = \frac{\sum Y^2 - \frac{(\sum Y)^2}{N}}{N-1} = \frac{302 - \frac{50^2}{10}}{9} = 5.78$$

  - The average variability of the two groups is smaller than the total variability.

Analysis – deviation approach

- The total variability can be partitioned into between group variability and error.

$$\left(Y_{ij} - GM\right) = \left(Y_{ij} - \overline{Y}_j\right) + \left(\overline{Y}_j - GM\right)$$

Analysis – deviation approach

- If you ignore group membership and calculate the mean of all subjects this is the grand mean and total variability is the deviation of all subjects around this grand mean
- Remember that if you just looked at deviations it would most likely sum to zero so…

Analysis – deviation approach

$$\sum_i \sum_j \left(Y_{ij} - GM\right)^2 = n\sum_j \left(\bar{Y}_j - GM\right)^2 + \sum_i \sum_j \left(Y_{ij} - \bar{Y}_j\right)^2$$

$$SS_{total} = SS_{bg} + SS_{wg}$$

$$SS_{total} = SS_A + SS_{S/A}$$

Analysis – deviation approach

A	Score	$\left(Y_{ij}-GM\right)^2$	$\left(\bar{Y}_j-GM\right)^2$	$\left(Y_{ij}-\bar{Y}_j\right)^2$
	9	16		4
	7	4		0
a_1	6	1	$(7 – 5)^2 = 4$	1
	7	4		0
	6	1		1
	3	4		0
	1	16		4
a_2	3	4	$(3 – 5)^2 = 4$	0
	4	1		1
	4	1		1
	$\sum Y = 50$			
	$\sum Y^2 = 302$	$\sum=52$	$\sum = 8$	$\sum = 12$
	$\bar{Y} = 5$		$n\sum = 5(8) = 40$	

52 = 40 + 12

Analysis – deviation approach

- degrees of freedom
  - $DF_{total} = N – 1 = 10 -1 = 9$
  - $DF_A = a – 1 = 2 – 1 = 1$
  - $DF_{S/A} = a(S – 1) = a(n – 1) = an – a = N – a = 2(5) – 2 = 8$

Analysis – deviation approach

- Variance or Mean square
  - $MS_{total} = 52/9 = 5.78$
  - $MS_A = 40/1 = 40$
  - $MS_{S/A} = 12/8 = 1.5$
- Test statistic
  - $F = MS_A/MS_{S/A} = 40/1.5 = 26.67$
  - Critical value is looked up with df_A, $df_{S/A}$ and alpha. The test is always non-directional.

Analysis – deviation approach

- ANOVA summary table

Source	SS	df	MS	F
A	40	1	40	26.67
S/A	12	8	1.5	
Total	52	9		

Analysis – computational approach

- Equations

$$SS_Y = SS_T = \sum Y^2 - \frac{\left(\sum Y\right)^2}{N} = \sum Y^2 - \frac{T^2}{an}$$

$$SS_A = \frac{\sum\left(\sum a_j\right)^2}{n} - \frac{T^2}{an}$$

$$SS_{S/A} = \sum Y^2 - \frac{\sum\left(\sum a_j\right)^2}{n}$$

  - Under each part of the equations, you divide by the number of scores it took to get the number in the numerator

Analysis – computational approach

- Analysis of sample problem

$$SS_T = 302 - \frac{50^2}{10} = 52$$

$$SS_A = \frac{35^2 + 15^2}{5} - \frac{50^2}{10} = 40$$

$$SS_{S/A} = 302 - \frac{35^2 + 15^2}{5} = 12$$

Analysis – regression approach

Levels of A	Cases	Y	X	YX
a_1	S_1	9	1	9
	S_2	7	1	7
	S_3	6	1	6
	S_4	7	1	7
	S_5	6	1	6
a_2	S_6	3	-1	-3
	S_7	1	-1	-1
	S_8	3	-1	-3
	S_9	4	-1	-4
	S_{10}	4	-1	-4
Sum		50	0	20
Squares Summed		302	10	
N		10		
Mean		5		

Analysis – regression approach

- $Y = a + bX + e$
- $e = Y – Y'$

Analysis – regression approach

- Sums of squares

$$SS(Y) = \sum Y^2 - \frac{\left(\sum Y\right)^2}{N} = 302 - \frac{50^2}{10} = 52$$

$$SS(X) = \sum X^2 - \frac{\left(\sum X\right)^2}{N} = 10 - \frac{0^2}{10} = 10$$

$$SP(YX) = \sum YX - \frac{(\sum Y)(\sum X)}{N} = 20 - \frac{(50)(0)}{10} = 20$$

Analysis – regression approach

$$SS_{(Total)} = SS(Y) = 52$$

$$SS_{(regression)} = \frac{[SP(YX)]^2}{SS(X)} = \frac{20^2}{10} = 40$$

$$SS_{(residual)} = SS_{(total)} - SS_{(regression)} = 52 - 40 = 12$$

- Slope

$$b=\frac{\sum YX-\frac{\left[(\sum Y)(\sum X)\right]}{N}}{\sum X^{2}-\frac{\left(\sum X\right)^{2}}{N}}=\frac{SP(YX)}{SS(X)}=\frac{20}{10}=2$$

- Intercept

$$a=\bar{Y}-b\bar{X}=5-2(0)=5$$

Analysis – regression approach

$$Y' = a + bX$$

For a_1 :

$$Y' = 5 + 2(1) = 7$$

For a_2 :

$$Y' = 5 + 2(-1) = 3$$

Analysis – regression approach

- Degrees of freedom
  - $df_{(reg.)}$ = # of predictors
  - $df_{(total)}$ = number of cases – 1
  - $df_{(resid.)}$ = df(total) – df(reg.) = 9 – 1 = 8

Statistical Inference and the F-test

- Any type of measurement will include a certain amount of random variability.
- In the F-test this random variability is seen in two places, random variation of each person around their group mean and each group mean around the grand mean.
- The effect of the IV is seen as adding further variation of the group means around their grand mean so that the F-test is really:

Statistical Inference and the F-test

$$F = \frac{effect + error_{BG}}{error_{WG}}$$

- If there is no effect of the IV than the equation breaks down to just:

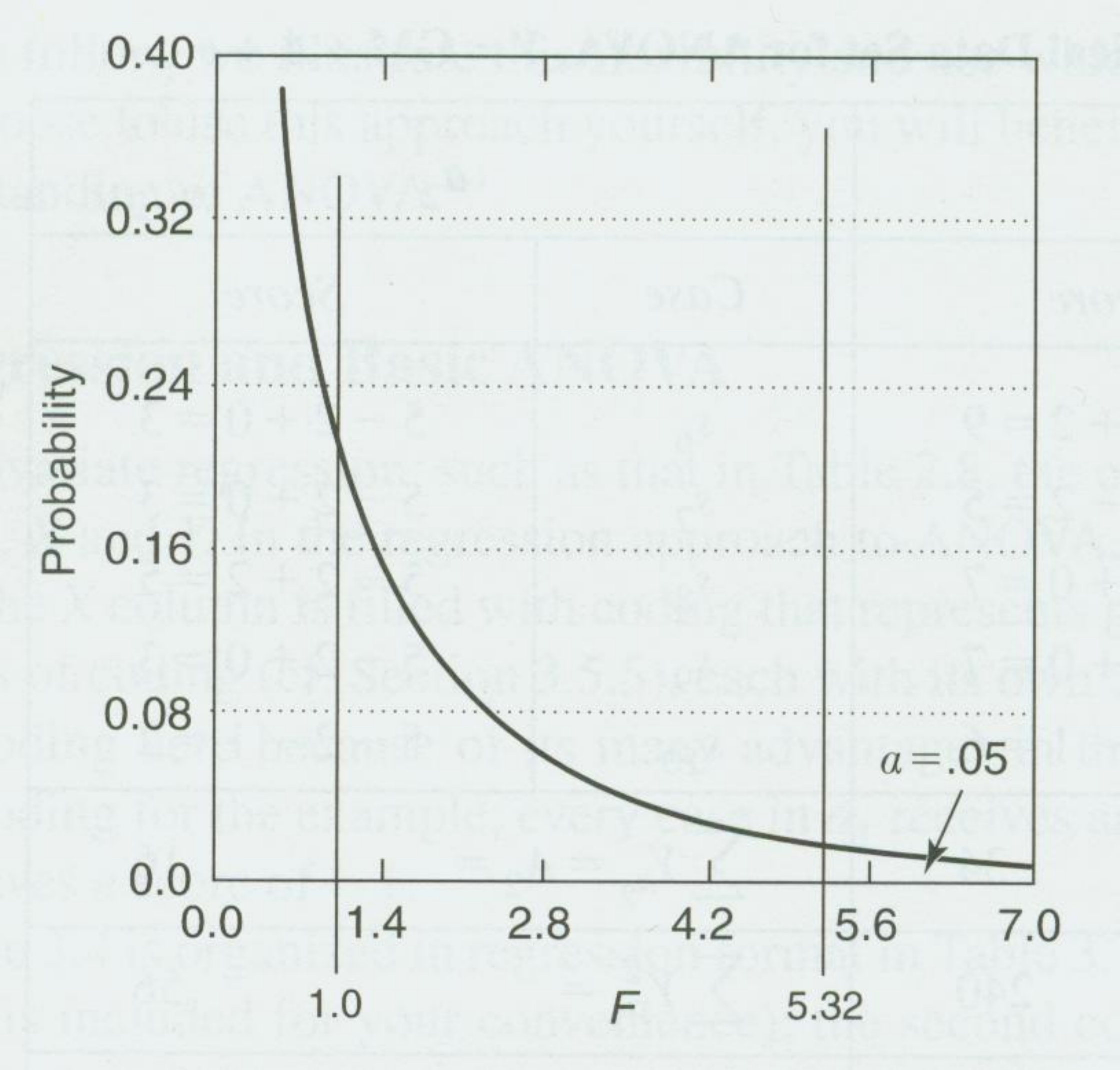
$$F = \frac{error_{BG}}{error_{WG}} \approx 1$$

  - which means that any differences between the groups is due to chance alone.

Statistical Inference and the F-test

- The F-distribution is based on having a between groups variation due to the effect that causes the F-ratio to be larger than 1.
- Like the t-distribution, there is not a single F-distribution, but a family of distributions. The F distribution is determined by both the degrees of freedom due to the effect and the degrees of freedom due to the error.

Statistical Inference and the F-test



Assumptions of the analysis

- Robust – a robust test is one that is said to be fairly accurate even if the assumptions of the analysis are not met. ANOVA is said to be a fairly robust analysis. With that said…

Assumptions of the analysis

- Normality of the sampling distribution of means
    - This assumes that the sampling distribution of each level of the IV is relatively normal.
    - The assumption is of the sampling distribution not the scores themselves
    - This assumption is said to be met when there is relatively equal samples in each cell and the degrees of freedom for error is 20 or more.

Assumptions of the analysis

- Normality of the sampling distribution of means
  - If the degrees of freedom for error are small than:
    - The individual distributions should be checked for skewness and kurtosis (see chapter 2) and the presence of outliers.
    - If the data does not meet the distributional assumption than transformations will need to be done.

Assumptions of the analysis

- Independence of errors – the size of the error for one case is not related to the size of the error in another case.
    - This is violated if a subject is used more than once (repeated measures case) and is still analyzed with between subjects ANOVA
    - This is also violated if subjects are ran in groups. This is especially the case if the groups are pre-existing
    - This can also be the case if similar people exist within a randomized experiment (e.g. age groups) and can be controlled by using this variable as a blocking variable.

Assumptions of the analysis

- Homogeneity of Variance – since we are assuming that each sample comes from the same population and is only affected (or not) by the IV, we assume that each groups has roughly the same variance
  - Each sample variance should reflect the population variance, they should be equal to each other
  - Since we use each sample variance to estimate an average within cell variance, they need to be roughly equal

Assumptions of the analysis

- Homogeneity of Variance
  - An easy test to assess this assumption is:

$$F_{\max} = \frac{S^2_{\text{largest}}}{S^2_{\text{smallest}}}$$

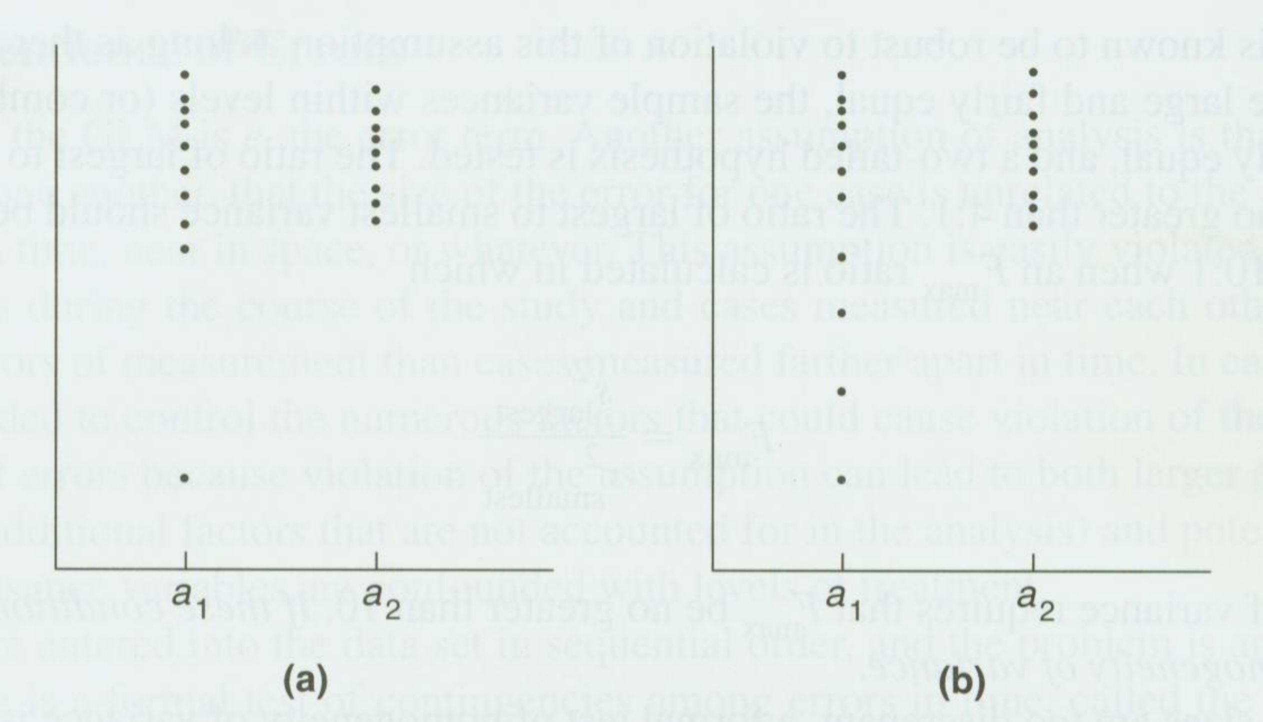
$$F_{\max} \le 10,\ \text{than the variances are roughly homogenous}$$

Assumptions of the analysis

- Absence of outliers
  - Outliers – a data point that doesn’t really belong with the others
    - Either conceptually, you wanted to study only women and you have data from a man
    - Or statistically, a data point does not cluster with other data points and has undue influence on the distribution
    - This relates back to normality

Assumptions of the analysis

- Absence of outliers



(a) (b)

FIGURE 3.3 Assessing Outliers through Disconnectedness Where the Low Score at a_1 in (a) Is More Likely to Be a True Outlier than the Same Low Score in (b)