Lecture 7: General Linear Model and the Analysis of Variance

• 7. 1 Introduction to ANOVA

- Technique used to test differences between sample means

- Can be used to test whether any number of means differ

- Can be used to look at the interacting effects of two or more variables

- Compares *variability* within and between experimental groups to test differences between means

- Why not multiple t tests?

- Increase in the probability of a type-I error

- ANOVA yields an accurate and known Type-I error probability

- The *t*-tests are not independent

- A multiple *t*-test approach is not powerful: if H_0 is false, it is less likely to be rejected

- A multiple *t*-test cannot assess the effects of two or more independent variables simultaneously

- Basic Idea of ANOVA

- If all scores in different groups were simply randomly selected from a ***single*** population of scores, the group means would likely differ due to sampling variability

  - How much they would be expected to differ would depend on the ***variability of the population***

  - Is the variability ***between*** groups greater than that expected on the basis of chance?

  - Is the variability ***between*** groups greater than that expected on the basis of the ***within***-group variability?

- ANOVA Nomenclature

	$Group_1$	$Group_2$	$Group_3$	
	X_{11}	X_{12}	X_{13}	
	X_{21}	X_{22}	X_{23}	
	X_{31}	X_{32}	X_{33}	
	X_{41}	X_{42}	X_{43}	
	X_{51}	X_{52}	X_{53}	
	$\vdots$	$\vdots$	$\vdots$	
Mean	$\overline{X}_1$	$\overline{X}_2$	$\overline{X}_3$	$\overline{X}.$
SD	SD_1	SD_2	SD_3	$SD.$

• 7. 2 ANOVA Computation

- We need a way of comparing the variability of sample means and variability within samples

- We also need a way of deciding whether the variation among the sample means is large relative to the variation within the samples

ANOVA = $\dfrac{Between - Group\ Variability}{Within - Group\ Variability}$

- Sum of Squares Between SS_B

$\alpha_j = \overline{X}_j - \overline{X}.$ effect of treatment

$SS_B = \Sigma_j\, n_i\, \alpha^2{}_j = \Sigma_j\, n_i\, (\overline{X}_j - \overline{X}.)^2$, recall that $s^2 = \dfrac{(X_i - \overline{X})^2}{n-1}$

If the ANOVA design is balanced (n's are equal across groups), then

$SS_B = n\ \Sigma_j\, \alpha^2{}_j = n\ \Sigma_j\ (\overline{X}_j - \overline{X}.)^2$

- Sum of Squares Within SS_W

$SS_W = \Sigma_j\, \Sigma_i\, (X_{ij} - \overline{X_j})^2 = SS_{W1} + SS_{W2} + \ldots\ SS_{WJ}$

- Total Sum of Squares SS_{TOTAL}

$SS_{TOTAL} = \Sigma_j\Sigma_i\, (X_{ij} - \overline{X}.)^2$

$SS_{TOTAL} = SS_B + SS_W$ (In one-factor ANOVA)

It reflects all sources of variation

$\Sigma_j\Sigma_i\, (X_{ij} - \overline{X}.)^2 = \Sigma_j\, n_i\, (\overline{X}_j - \overline{X}.)^2 + \Sigma_j\, \Sigma_i\, (X_{ij} - \overline{X_j})^2$

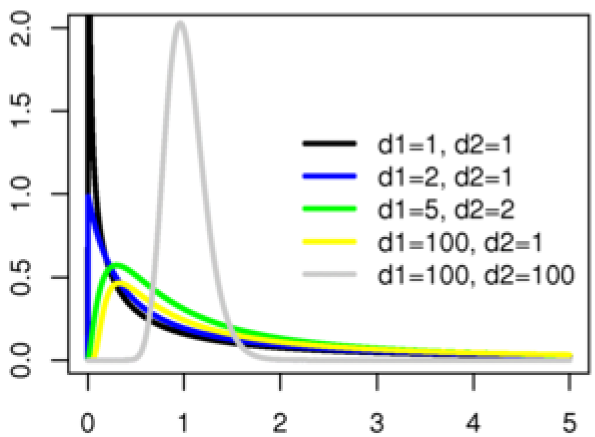
- The F-test

$$F = \frac{SS_B / J-1}{SS_W / N-J} = \frac{MS_B}{MS_W}$$ (this is the ratio of two independence variance estimates)

$$df = \frac{J-1}{N-J}$$

- If H_0 is true, both variance estimates are estimating the same parameter σ^2, F = 1 $\rightarrow$ No treatment effects (sample means are drawn from same population)

- If H_0 is false, F > 1 $\rightarrow$ Means are different (sample means are from different populations)



- Summary of Logic

- Calculate two estimates of the population variance, MS_B (based on variability *between* groups, dependent on H_0), and MS_W (based on variability *within* groups, independent of H_0)

$$F = \frac{Between - Group\ Variability}{Within - Group\ Variability} = \frac{MS_B}{MS_W}$$

If they agree, no reason to reject H_0

If $MS_B > MS_W$, then difference between group means must have contributed to MS_B and we should reject H_0

- Two separate estimates of population variance

- MS_W is an unbiased estimate *regardless of the presence of treatment effects*

- MS_B is an unbiased estimate of σ^2 *only if there are no treatment effects* (H_0 is true)

- When *systematic differences between groups* exist along with the random variability among individuals, MS_B tends to be larger than σ^2 and hence larger than MS_W

- When the hypothesis that all the treatment effects are zero is exactly true, the numerator of the F estimates only the population error variance

- Otherwise, the numerator is estimating some larger value, with the particular value depending on just how large the treatment effects are

• 7. 3 A Statistical Model

$X_{ij} = \mu + \alpha_j + \varepsilon_{ij}$

X_{ij} = score for person i in group j

μ = population mean

α_j = effect of treatment j ($\alpha_j = \mu_j - \mu$)

ε_{ij} = error for score X_{ij}, or residual of the score X_{ij} when predicted from μ and α_j

($\varepsilon_{ij} = X_{ij} - \mu - \alpha_j$)

$\varepsilon_{ij} \sim \text{NID}(0, \sigma^2)$

- *Assumptions of the Model*

- *Normality*: Assume that scores in each group are normally distributed

- *Homogeneity of Variance*: The scores in each group have the same variance

- *Independence of Observations*: Knowing one score in an experimental group tells us nothing about the other scores

- However, ANOVA is robust with respect to mild violations of normality and homogeneity of variance except with small and/or unequal sample sizes

- ## 7. 4 ANOVA Example

- Experiment to examine the effect of different drugs on anxiety

Drug1	Drug2	Drug3	
40	34	12	
30	75	02	
11	40	32	
22	51	05	
55	72	14	
$\overline{X}_1 = 31.60$	$\overline{X}_2 = 54.40$	$\overline{X}_3 = 13.00$	$\overline{X}. = 33.00$
$SD_1 = 16.86$	$SD_2 = 18.50$	$SD_3 = 11.70$	$SD. = 22.92$

- Calculations

- In order to calculate MS_B and MS_W we need to calculate the appropriate sums of squares

- SS_B: Represents Sum of squared deviations of group means from the grand mean. In effect, a measure of differences between groups

$SS_B = n \Sigma_j (\overline{X}_j - \overline{X}.)^2$, where n = sample size

$$= 5[(31.60 - 33.00)^2 + (54.40 - 33.00)^2 + (13.00 - 33.00)^2] = 4299.60$$

- SS_W: Sum of squared deviations within each group (it can be obtained by subtraction)

$$SS_W = \Sigma_j \Sigma_i (X_{ij} - \overline{X}_j)^2 = (40 - 31.60)^2 + (30 - 31.60)^2 + \ldots + (14 - 13)^2 = 3053.4$$

- SS_T: Represents sum of squared deviations of all observations from the grand mean

$$SS_T = SS_B + SS_W$$

$$SS_T = 4299.6 + 3054.4 = 7354$$

Alternatively, $SS_T = \Sigma_j\Sigma_i (X_{ij} - \overline{X}.)^2 = \Sigma_j\Sigma_i X^2 - \frac{(\Sigma X)^2}{N}$, where N = number of observations

$$= (40^2 + 30^2 + \ldots + 14^2 - \frac{(495)^2}{15} = 7354$$

- *Degrees of Freedom*

$df_T = N - 1$ (where N = number of observations)

$= 15 - 1 = 14$

$df_B = J - 1$ (where J is number of groups)

$= 3 - 1 = 2$

$df_W = df_T - df_B$

$= 14 - 2 = 12$

- *Mean Squares and F-value*

$$MS_B = \frac{SS_B}{df_B} = \frac{4299.60}{2} = 2149.8$$

$$MS_W = \frac{SS_W}{df_W} = \frac{3054.40}{12} = 254.43$$

$$F = \frac{MS_B}{MS_W} = \frac{2149.80}{254.43} = 8.45$$

- *ANOVA Summary Table*

Source	*SS*	*df*	*MS*	*F*	*Sig.*
Between	4299.60	2	2149.80	8.45	<.01
Within	3054.40	12	254.53		
Total	7354.00	14			

- *Conclusions*

- Between groups estimate of the population variance is much larger than the within groups estimate → F value > 1

- Critical F-values corresponding to the df of the two mean squares (df_B and df_W)

- From tables: ($F_{.05} = 3.89$ and $F_{.01} = 6.93$); Because $F_{obt} > F_{crit}$ we can reject H_0 and conclude that the groups were sampled from populations with different means

- **7. 5 Estimating Model Parameters**

$X_{ij} = \mu + \alpha_j + \varepsilon_{ij}$

X_{ij} = score for person i in group j

μ = population mean

α_j = effect of treatment j ($\alpha_j = \mu_j - \mu$)

ε_{ij} = error for score X_{ij}, or residual of the score X_{ij} when predicted from μ and α_j

($\varepsilon_{ij} = X_{ij} - \mu - \alpha_j$)

$\varepsilon_{ij} \sim NID(0, \sigma^2)$

- Estimation of the terms of the model – Example

X_1	X_2	X_3	
3	4	5	
4	5	6	
5	6	7	
$\overline{X}_1 = 4$	$\overline{X}_2 = 5$	$\overline{X}_3 = 6$	$\overline{X}. = 5$

$\hat{\mu} = \overline{X}. = 5$

$\hat{\alpha}_j = \hat{\mu}_j - \hat{\mu} = \overline{X}_j - \overline{X}.$

$\hat{\alpha}_1 = 4 - 5 = -1$

$\hat{\alpha}_2 = 5 - 5 = 0$

$\hat{\alpha}_3 = 6 - 5 = 1$

- *Residuals of the model* (also called noise or leftover, after fitting the model). These are very important for analyses of models – goodness of fit

$\hat{\varepsilon}_{ij} = x_{ij} - \hat{\mu} - \hat{\alpha}_j$

$\hat{\varepsilon}_{11} = x_{ij} - 5 - (-1) = 3 - 5 + 1 = -1$ (residual of particular case)

$\hat{\varepsilon}_{21} = 4 - 5 - (-1) = 0$; $\hat{\varepsilon}_{31} = 5 - 5 - (-1) = 1$

$\sum_i \sum_j \hat{\varepsilon}_{ij} = 0$

• 7. 6 Partitioning the Variability

- We want to ask if the estimates (estimators of σ^2) are independent from each other

$$\underbrace{X_{ij}}_{a} - \underbrace{\overline{X}_{\cdot}}_{b} = \underbrace{(\overline{X}_j - \overline{X}_{\cdot})}_{c} + \underbrace{(X_{ij} - \overline{X}_j)}_{d}$$

a = individual score

b = grand mean

c = distance from group mean to grand mean

d = distance from raw score to group mean

$$\Sigma_j\Sigma_i\,(X_{ij} - \overline{X}_{\cdot})^2 = SS_T$$

$$= \Sigma_j\Sigma_i\,[(\overline{X}_j - \overline{X}_{\cdot}) + (X_{ij} - \overline{X}_j)]^2$$

$$= \Sigma_j\Sigma_i\,(\overline{X}_j - \overline{X}_{\cdot})^2 + \Sigma_j\,\Sigma_i\,(X_{ij} - \overline{X}_j)^2 + 2\Sigma_j\,(\overline{X}_j - \overline{X}_{\cdot}) \cdot \Sigma_i\,(X_{ij} - \overline{X}_j)$$

$\Sigma_i\,(X_{ij} - \overline{X}_j) = 0$ (deviations from the mean), so we obtain

$$\Sigma_j\Sigma_i\,(X_{ij} - \overline{X}_{\cdot})^2 = \Sigma_j\, n\,(\overline{X}_j - \overline{X}_{\cdot})^2 + \Sigma_j\,\Sigma_i\,(X_{ij} - \overline{X}_j)^2$$

$SS_T = SS_B + SS_w$

$\therefore$ SS_B and SS_w are independent

$df_T = df_B + df_W$

$N - 1 = (N - J) + (J - 1)$

$\therefore$ df_B and df_W are independent

$\therefore$ the variance estimates are independent

• 7. 7 Magnitude of Effect

- Eta-squared $\eta^2 = \dfrac{SS_B}{SS_T}$

    - It is the proportion of the total variability of the data that is accounted for by the treatment effect (also called R^2)

    - It varies from 0 (no effect) to 1 (no error)

        in the example of drugs and anxiety $\eta^2 = \dfrac{4299.6}{7354} = .58$

    η^2 is positively biased (overestimates the true effect), with larger bias for a larger number of groups and smaller sample sizes

- Omega-squared $\omega^2 = \dfrac{SS_B - (J-1)MS_W}{SS_T + MS_W}$

    - It is the proportion of variance accounted for – with a correction factor

        in the example of $\eta^2 = \dfrac{4299.6 - (3-1)254.43}{7354 + 254.43} = .50$

- But …next day

- But…what to do after a large F-value in ANOVA?
- F-test is a non-directional omnibus test
- We need more focused comparisons
- Planned orthogonal contrasts
- Post-Hoc tests
- Effect size