Analysis of Variance (ANOVA)

Purpose:

To look for a statistically significant relationship between two variables where the independent variable is nominal or categorical and the dependent variable is typically interval/ratio

What do we mean by "statistically" significant?)

Example of a One Way ANOVA: Educational Attainment (Measured in Years) for Four Demographic Groups (GSS data)

White Males	Black Males	White Females	Black Females
16	16	16	14
18	12	12	10
14	11	14	12
14	14	14	13
16		11	11
16		11	

What might be our hypothesis? Our null hypothesis?

Could we use t-tests to examine this? Explain.

Why would ANOVA be better?

Educational Attainment (Measured in Years) for Four Demographic Groups (GSS data)

White Males	Black Males	White Females	Black Females
16	16	16	14
18	12	12	10
14	11	14	12
14	14	14	13
16		11	11
16		11	

Which group appears to have most education? Least?

Educational Attainment (Measured in Years) for Four Groups (GSS data)

White Males	Black Males	White Females	Black Females
16	16	16	14
18	12	12	10
14	11	14	12
14	14	14	13
16		11	11
16		11	
$\bar{Y}$ = 15.67	13.25	13.00	12.00
S = 1.51	2.22	2.00	1.58
S^2 = 2.27	4.92	4.00	2.50

After examining the Mean, SD, and Variance is it easier to find differences?

Are the differences we see true differences in the population or only due to sampling error?

Affect of Demographic Group on Educational Attainment (Measured in Years)

White Males	Black Males	White Females	Black Females
16	16	16	14
18	12	12	10
14	11	14	12
14	14	14	13
16		11	11
16		11	

Why is this considered a "one-way" ANOVA?

There is one dependent (education) and one independent (group membership) variable.

Assumptions of ANOVA:

1. Independent random samples (selection of one sample has no effect on another)
2. Dependent variable is interval-ratio (ordinal is sometimes used)
3. Population is normally distributed
4. Population variances are equal (in our example variances are close enough)

Statement of Null Hypothesis:

There is no difference in education between demographic groups.

$u_1 = u_2 = u_3 = u_4$

Statement of Hypothesis:

Not all demographic groups have equal education.

How does ANOVA work?

ANOVA examines the difference between the samples (or groups) as well as the difference within a single sample (or group).

These differences can also be referred to as the variance or variation.

When considering our example, is there variation among the scores within one of the samples (or groups)?

That is, are all the scores alike (no variation) or is there a broad variation in scores?

How does ANOVA work?

ANOVA allows us to determine whether the variance between samples (or groups) is larger than the variance within the samples (or groups).

Why is this information valuable?

Calculating an ANOVA means that we want to calculate the *F* statistic so we can determine the likelihood of obtaining the scores (data) by chance.

There are six steps to calculating the *F* statistic:

(Calculating ANOVA)

(1) "sum of squares" between the groups (i.e., sum of squared deviations),

(2) "sum of squares" within the groups,

(3) degrees of freedom for each.

(Calculating ANOVA)

(4) using the information collected in steps 1 – 3 to calculate the "mean square between" and "mean square within"

(5) using this info. to create the *F* ratio (or *F* statistic)

(6) Making a decision

(Step 1) Between-group sum of squares (SSB) measures the (squared) difference in average years of education between our four groups.

Calculating SSB is done by (a – e):

(a) determining the mean education for the four groups, that is, the overall mean

(b) determining the difference between each group mean and the overall mean,

(c) squaring these differences between each sample mean and the overall mean,

(d) multiplying each squared difference by the number of scores or cases in its respective group, and

(e) summing the squared differences.

This tells us the sum of squared deviations between the sample means and the overall mean score.

SSB measures the difference in average years of education between the groups.

Educational Attainment (Measured in Years) for Four Groups (GSS data)

	White Males	Black Males	White Females	Black Females
$\bar{Y}$ =	15.67	13.25	13.0	12.0
S =	1.51	2.22	2.00	1.58
S^2 =	2.27	4.92	4.00	2.50

(15.67 + 13.25 +13.0 +12.0)/4 = 13.57 = overall mean

$6(15.67 - 13.57)^2 = 26.46$

$4(13.26 - 13.57)^2 = .38$

$6(13.0 - 13.57)^2 = 1.95$

$5(12.0 - 13.57)^2 = 12.32$

41.11 (SSB or the sum of squared deviations between each group mean and the overall mean)

(2) Within-group sum of squares (SSW) measures the variation of scores within each single sample or group (i.e., the variation between each individual score and its group mean).

It is also thought of as the amount of unexplained variation after considering the variation found between each sample (or group) and the dependent variable.

Educational Attainment (Measured in Years) for Four Groups (GSS data)

White Males	Black Males	White Females	Black Females
$(16-15.67)^2$	16 (same)	16 (same)	14 (same)
$(18-15.67)^2$	12	12	10
$(14-15.67)^2$	11	14	12
14 .	14	14	13
16 .		11	11
16 .	sum $(y - \bar{y})^2$	sum $(y - \bar{y})^2$	11
sum $(y - \bar{y})^2$			sum $(y - \bar{y})^2$
$\bar{Y}$ = 15.67	13.25	13.0	12.0
S = 1.51	2.22	2.00	1.58
S^2 = 2.27	4.92	4.00	2.50

Calculating the SSW = sum of the four group's sum of squared deviations or the unexplained variation after considering SSB

The total sum of squares (SST) refers to the variation found within each group plus the variation found between the groups.

SST = SSB + SSW

Decomposing the SST for a single case helps to see that SSW is the variation left after considering SSB. That is, it can help us see that: SSW + SSB = SST.

Let's take the fifth white male with 16 yrs of education. The difference between his score and the overall mean (13.57) yrs is 2.43 (16 – 13.57). This is equal to the SSB + SSW. More specifically:

The difference between his group mean (15.67) and the overall mean is 2.10 (SSB = 15.67 – 13.57)

Looking within his group, the difference between his score and his group's mean is .33 years (SSW = 16-15.67)

When adding his SSB (2.10) and SST (.33), we get the total sum of squares (SST) for this case (2.43).

(4) In order to use this information to calculate the *F* statistic we must determine the mean square between and mean square within.

$$F \text{ statistic} = \frac{\text{mean square between}}{\text{mean square within}} = \frac{MSB}{MSW}$$

$$MSB = \frac{\text{Sum of Squares Between}}{\text{Degrees of Freedom Between}}$$

$$MSW = \frac{\text{Sum of Squares Within}}{\text{Degrees of Freedom Within}}$$

To calculate the degrees of freedom between and the degrees of freedom within:

DFW = (# of cases) – (# of groups)

DFB = (# of groups) – 1

For our example when calculating the mean square between = SSB/DFB:

$$\frac{SSB = 41.11}{DFB = 4 - 1 = 3} = 13.70$$

Mean square within = SSW/DFW:

$$\frac{SSW = 56.08}{DFW = 21 - 4 = 17} = 3.30$$

(5) Create the *F* statistic

$$F = \frac{\text{Mean Sq Between}}{\text{Mean Sq Within}} = \frac{SSB/DFB}{SSW/DFW}$$

(for our example: $F = 13.70/3.30 = 4.15$)

A larger F statistic means that there is more variation between groups than within groups.

A larger F statistic supports the hypothesis that the groups are affecting the dependent variable.

Summary:

$$F = \frac{SSB/DFB}{SSW/DFW} = \frac{\text{Mean Sq Between}}{\text{Mean Sq Within}}$$

$$F = \frac{41.11/3}{56.08/17} = \frac{13.70}{3.30} = 4.15$$

(6) Making a decision

1. Determine a specific alpha (such as .05) and refer to the appropriate alpha *f* distribution table (such as the .05 *f* table or the .01 *f* table) to determine the probability of obtaining a particular *F* statistic.
2. To read the table you must select the two degrees of freedom used to calculate the *f* statistic (in our example these were 3 & 17).

(6) Making a decision

3. Determine the "critical E" found in the table for the corresponding degrees of freedom and alpha (eg., 3.20 is found for 3 and 17 DF)
4. If the calculated F is larger than the corresponding F, then the hypothesis is supported.

(6) Making a decision

Is our hypothesis supported at the .05 level?

Our F statistic would need to be larger than 3.20.
Since it is, the answer is Yes.

(6) Making a decision

Is our hypothesis supported at the .01 level?

Our F statistic would need to be larger than 5.18. Since it is not, the answer is NO.

The F statistic doesn't advise us about which groups are different, only that educational attainment does differ significantly by demographic group members.

That is, we reject the null hypothesis and conclude that the years of education do vary by group membership.

While we know that "years of education" does vary by demographic group, the F test doesn't give us a measure of association.

How do we determine the strength of the relationship between education and demographic group?

Eta^2

$$Eta^2 = \frac{SSB}{SST}$$

or

$$\frac{41.11}{97.19} = .42$$

Thus, 42% of the variation in educational attainment can be attributed to demographic group membership.

Or, 42% of the variation in the dependent variable (educational attainment) can be explained by the independent variable (group membership)