ANALYSIS OF VARIANCE

A. Let us take a moment to get some perspective on where we have been and where we are going. When you select a statistical procedure, your choice begins by determining whether each of your variables' "level of measurement" is nominal, ordinal, or continuous (interval or ratio).

1. If your variables are ordinal-level (and if you do not wish to treat them as continuous), you must use nonparametric statistics (statistics that are based on the relative rankings of observations). Although we have mentioned a couple of nonparametric statistics (i.e., Fisher's Exact Test and the Wilcoxon test), only parametric statistics (statistics that are point estimates of population parameters) are covered in this course.

2. The following table lists the appropriate parametric procedures for testing whether or not an independent variable (or set of independent variables) and a dependent variable are statistically independent:

<table>
<thead>
<tr>
<th>Level of measurement of the Independent variable(s)</th>
<th>Nominal</th>
<th>Both</th>
<th>Continuous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level of measurement of the dependent variable</td>
<td>t-test ($\hat{x}_1 - \hat{x}_2$) or Log-linear analysis</td>
<td>Logistic Regression</td>
<td>Logistic Regression</td>
</tr>
<tr>
<td>Nominal</td>
<td>t-test ($\bar{x}_1 - \bar{x}_2$) or ANOVA</td>
<td>ANACOVA</td>
<td>Linear Regression</td>
</tr>
</tbody>
</table>

3. It should be clear from the table that this course has emphasized tests of independence in analyses in which the dependent variable is...
continuous. There are two exceptions to this:

a. chi-square test for contingency tables

The chi-square from a contingency table can be partitioned in much the same way as can $SS_{\text{TOTAL}}$ in a regression analysis. In the former case, specific cells' contributions to chi-square are associated with log-linear parameter estimates; in the latter case, proportions of the explained variance (contributions to $R$-square, if you will) are associated with linear regression parameter estimates. Thus, log-linear analysis is to the chi-square test what regression analysis is to the $F$-test: Each supplements a global significance test with parameter estimates that isolate the sources of statistical dependencies.

b. t-test for a difference in proportions

Like log-linear analysis, the t-test for $H_0: \pi_1 = \pi_2$ (in which the true difference in population proportions is estimated) is a parametric test between independent and dependent variables that are nominal-level. Unlike the t-test, log-linear analysis can be performed with independent variables having more than two 2-levels.

4. It should also be pointed out that although ANACOVA (analysis of covariance) has not been taught "in name," it has been taught in fact. In the previous two sections of the Lecture Notes, illustrations of multiple regressions were given in which income (a continuous variable) was dependent, and in which gender (a nominal-level variable) and either education or prestige (both continuous) were independent. Gender was used instead of a second continuous variable to simplify depictions of
data in two-dimensional plots. What distinguishes an ANACOVA from multiple regression is that the nominal-level variables in an ANACOVA are converted to dummy variables prior to entering them into a regression equation. Also (as already introduced in conjunction with Lab 7), the slopes associated with these dummy variables are interpreted differently than slopes associated with continuous variables in a regression equation.

B. This course now concludes with a brief introduction to analysis of variance (ANOVA). Analysis of variance is the statistical procedure most commonly used in controlled EXPERIMENTS (i.e., in research designs in which the researcher has control over who does versus does not have a particular "treatment").

C. ANOVAs are used to test the null hypothesis that the means of k groups are equivalent. That is, they are used to test

\[ H_0: \mu_1 = \mu_2 = \ldots = \mu_k \]

\[ H_A: \mu_i \neq \mu_j, \text{ for some } i \neq j. \]

Note this null hypothesis's similarity to that evaluated in a t-test (viz., \( H_0: \mu_1 = \mu_2 \)). In fact, the ANOVA is a direct extension of the t-test.

D. An ILLUSTRATION

1. Let's consider an example in which there are two treatment groups and a control group. Five people are randomly assigned to each of the three groups. Thus, \( n_1 = n_2 = n_3 = 5 \) and \( n = \sum_{i=1}^{3} n_i = 15 \).

NOTE: Whenever \( n_1 = n_2 = \ldots = n_m \), one's research design is said to be a BALANCED DESIGN.
2. You begin with a list of 15 movies. Before asking subjects which movies they would attend, subjects from different treatment groups are (for each of the 15 movies) presented with one of three hypothetical situations. Subjects in the first treatment group \((T_1)\) were told that they had lost tickets; subjects in the second treatment group \((T_2)\) were told they had lost money; subjects in the third treatment group \((T_3)\) were told they were given tickets. In this experiment, the third treatment group serves as a control (i.e., a point of comparison) for the other two groups. The exact wording of the hypothetical situations is as follows (repeated for each of 15 movie titles):

\(T_1:\) "On your way to the theater to see \((\text{name of movie})\), you discover that you have lost the $10 tickets."

\(T_2:\) "On your way to the theater to see \((\text{name of movie})\), you discover that you have lost $10."

\(T_3:\) "Someone offers you a free ticket to see \((\text{name of movie})\).

Repeated for all groups: "Would you attend the movie?"

3. Let \(Y_{ij}\) be the number of movies that the \(j^{th}\) person in the \(i^{th}\) group says he/she would attend. Thus, \(i\) takes the values 1,2,3 (\(m = 3\) is the number of groups) and \(j\) takes the values 1,2,3,4,5 (since \(n_i = 5\) for \(i = 1,2,3\)).

NOTE how a nominal-level independent variable (viz., group) is involved in the analysis only by the addition of a subscript to the dependent variable. By not introducing a distinct independent variable with numbers arbitrarily assigned to the discrete groups, this notation helps
keep the statistician from mistaking the data's level of measurement. That is, the notation keeps one from mistaking the nominal-level group variable for an ordinal-, interval-, or ratio-level variable.

4. The movie data are listed in the left of each pair of columns below:

<table>
<thead>
<tr>
<th>T₁ (lost tickets)</th>
<th>T₂ (lost money)</th>
<th>T₃ (control group)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y₁₁ = 5</td>
<td>-1</td>
<td>Y₂₁ = 8</td>
</tr>
<tr>
<td>Y₁₂ = 6</td>
<td>0</td>
<td>Y₂₂ = 11</td>
</tr>
<tr>
<td>Y₁₃ = 8</td>
<td>2</td>
<td>Y₂₃ = 13</td>
</tr>
<tr>
<td>Y₁₄ = 7</td>
<td>1</td>
<td>Y₂₄ = 8</td>
</tr>
<tr>
<td>Y₁₅ = 4</td>
<td>-2</td>
<td>Y₂₅ = 10</td>
</tr>
</tbody>
</table>

5. Note that subscripts change to dots (•) when data are averaged over a subscript. For example, Y₁. is the notation used to represent the mean number of movies mentioned by the five subjects in the first treatment group. The three group means and the overall mean are computed as follows:

\[
Y_{1.} = \frac{\sum_{j=1}^{n₁} Y₁j}{n₁} = \frac{5}{5} = \frac{5 + 6 + 8 + 7 + 4}{5} = \frac{30}{5} = 6
\]

\[
Y_{2.} = \frac{\sum_{j=1}^{n₂} Y₂j}{n₂} = \frac{50}{5} = 10
\]

\[
Y_{3.} = \frac{\sum_{j=1}^{n₃} Y₃j}{n₃} = \frac{55}{5} = 11
\]
\[
Y_{..} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{ni} Y_{ij}}{\sum_{i=1}^{m} ni} = \frac{30 + 50 + 55}{5 + 5 + 5} = \frac{135}{15} = 9
\]

6. The data and means can be graphed as follows:

7. RECALL that the objective here is to test \( H_0: \mu_1 = \mu_2 = \mu_3 \). The strategy used in making this test is as follows: If the variance between the group means is large in comparison to the variance within the groups, it is NOT likely that the group averages differ by chance.

a. This is the motivation for partitioning the variance in the dependent variable (here, number of movies) into the variance between the groups (\( SS_{\text{BETWEEN}} \)) and the variance within the groups (\( SS_{\text{WITHIN}} \)). Consider the following observation, \( Y_{34} = 15 \). As with each of the other 14 observations, this one can be divided into three parts:
15 = 9 + (11 - 9) + (15 - 11)

the overall mean the difference between the group mean and the overall mean the difference between the observation and the group mean

b. AN ASIDE: A clear parallel with bivariate regression should be becoming clear at this point. The group mean is the best point estimate of the value on the dependent variable for all subjects in the group. That is, \( \hat{Y}_i \) in an analysis of variance is the same as \( \hat{Y} \) within a regression analysis. Therefore, the general form of the above equivalence is written in regression notation as follows:

\[
Y_i = \bar{Y} + (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)
\]

More detail will be presented shortly on how \( Y_i \) can be computed using regression analysis.

c. The general form of the equivalence is written in analysis of variance notation as follows:

\[
Y_{ij} = Y_{..} + (Y_i - Y_{..}) + (Y_{ij} - Y_i)
\]

This formula can be rewritten as

\[
(Y_{ij} - Y_{..}) = (Y_i - Y_{..}) + (Y_{ij} - Y_i)
\]

Squaring both sides,

\[
(Y_{ij} - Y_{..})^2 = [(Y_i - Y_{..}) (Y_{ij} - Y_i)]^2
\]

\[
= (Y_i - Y_{..})^2 + (Y_{ij} - Y_i)^2 + 2(Y_i - Y_{..})(Y_{ij} - Y_i)
\]

Then by summing across all "i" and "j" to get SS\_TOTAL,

\[
\sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - Y_{..})^2 = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_i - Y_{..})^2 + \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - Y_i)^2
\]

229
\[ + 2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - Y_{..})(Y_{ij} - Y_{i..}) \]

Now considering just the last term in the equation,

\[ 2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - Y_{..})(Y_{ij} - Y_{i..}) = 2 \sum_{i=1}^{m} (Y_{i..} - Y_{..}) \sum_{j=1}^{n_i} (Y_{ij} - Y_{i..}) \]

(Note here that when there is no j-subscript in an expression, the expression is constant with respect to the j-summation.)

Note that for each group (i.e., for each i) the last part of this expression equals zero, because it is the sum of the deviations from the group mean. That is,

\[ \sum_{j=1}^{n_i} (Y_{ij} - Y_{i..}) = 0 \]

Thus the cross-product term equals zero, and the final formula for \( SS_{\text{TOTAL}} \) is as follows:

\[ \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - Y_{..})^2 = \sum_{i=1}^{m} n_i (Y_{i..} - Y_{..})^2 + \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - Y_{i..})^2 \]

**TOTAL SUM OF SQUARES** = **BETWEEN GROUPS** + **WITHIN GROUPS**

**SS_{TOTAL}** = **SS_{BETWEEN}** + **SS_{WITHIN}**

**SS_{TOTAL}** = **SS_{TREATMENT}** + **SS_{ERROR}**

E. Thus, here we find another parallel between ANOVA and regression analysis.

In particular, whereas regression models partition \( SS_{\text{TOTAL}} \) into \( SS_{\text{REGRESSION}} \) and \( SS_{\text{ERROR}} \), an ANOVA partitions \( SS_{\text{TOTAL}} \) into \( SS_{\text{BETWEEN}} \) and \( SS_{\text{WITHIN}} \).
a. Returning to the table of data on movies, the within-group sum of squares is computed by squaring and summing the deviations from group means. Accordingly, one finds that $\Sigma(Y_{1j} - Y_{1.})^2 = 10$, $\Sigma(Y_{2j} - Y_{2.})^2 = 18$, $\Sigma(Y_{3j} - Y_{3.})^2 = 30$, and thus

$$SS_{\text{WITHIN}} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - Y_{i.})^2 = 10 + 18 + 30 = 58.$$ 

Also,

$$SS_{\text{BETWEEN}} = \sum_{i=1}^{m} n_i (Y_{i.} - Y_{..})^2$$

$$= 5*(6 - 9)^2 + 5*(10 - 9)^2 + 5*(11 - 9)^2$$

$$= 45 + 5 + 20 = 70.$$ 

Thus, $SS_{\text{TOTAL}} = 128 = 70 + 58 = SS_{\text{TREATMENT}} + SS_{\text{ERROR}}$.

1. With this information an ANOVA TABLE can now be constructed.

a. The ANOVA table for the movie data is as follows:

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>2</td>
<td>70</td>
<td>35.00</td>
<td>7.24</td>
</tr>
<tr>
<td>Error</td>
<td>12</td>
<td>58</td>
<td>4.83</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>128</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

b. More generally, the generic form of an ANOVA table is as follows:


2. As in regression analysis, the F-test can now be used to test whether a significant amount of variance has been explained.

a. Notice the logic here: At some point, the differences between the group means (as measured by $\sum n_i \left(Y_{i.} - Y_{..}\right)^2$) is going to be large enough in comparison to the within-group variance so that you will no longer believe that the differences are due to chance. In the movie illustration,

$$F_{12}^2 = \frac{SS_{BETWEEN}}{\text{# groups} - 1} \frac{70}{2} = \frac{SS_{WITHIN}}{n - \text{# groups}} = \frac{58}{12} = 7.24 .$$

b. From the F-table (Table D for $\alpha = .05$), $F_{12,.05}^2 = 3.89 .

c. Because the calculated F of 7.24 falls in the rejection region (i.e., it is larger than 3.89, which is the critical value from the F-table), you can conclude that there is significantly more variance between groups than within groups on the number of movies that
subjects say that they would watch. If you take a moment to think about what this means, you will notice that little variance within groups implies that subjects within each group have values close to their group mean. Likewise, large between-group variance implies that group means are distant from each other. Following this reasoning, it should be clear that the F-test performed in the ANOVA is a test of equivalence among group means. In other words, the finding that F falls into the rejection region is evidence that you should reject \( H_0 \) from the following hypotheses:

\[
\begin{align*}
H_0: \mu_1 &= \mu_2 = \mu_3 \\
H_A: \mu_i \neq \mu_j \text{ for at least one pair of } i \neq j \quad (i=1,2,3, \quad j=1,2,3)
\end{align*}
\]

d. Thus the rejection rule is, "Reject \( H_0 \) if \( F > 3.89 \)." Given that \( F = 7.24 \), we reject \( H_0 \) and conclude that at least two means are significantly different from each other.

"You say that one of your eyes is different from the other? Which one?"
e. **Multiple comparison tests** comprise a class of statistical tests for
deciding (at a given \( \alpha \)-level) whether specific pairs or combinations
of means are significantly different from each other. Although these
tests are commonly covered in courses on analysis of variance, they
are beyond the scope of STAT 401. Instead, we shall link ANOVA back
to regression analysis—a method that we know a lot more about at
this point.

F. In fact, ANOVA can be thought of as a type of regression analysis. (In
contrast, there are many researchers who argue that regression is a special
case of ANOVA—namely one in which only linear polynomial relations among
means are considered.) The simplest way to perform an ANOVA using
regression analysis is to do so using dummy variables. Consider the
following data matrix:

\[
\begin{pmatrix}
Y & D_1 & D_2 \\
5 & 1 & 0 \\
6 & 1 & 0 \\
8 & 1 & 0 \\
7 & 1 & 0 \\
4 & 1 & 0 \\
8 & 0 & 1 \\
11 & 0 & 1 \\
13 & 0 & 1 \\
8 & 0 & 1 \\
10 & 0 & 1 \\
9 & 0 & 0 \\
12 & 0 & 0 \\
11 & 0 & 0 \\
15 & 0 & 0 \\
8 & 0 & 0
\end{pmatrix}
\]

1. Note that in this data matrix the dual-subscript notation is done away
with and the following two dummy variables are used:

\[
D_1 = \begin{cases} 
1 & \text{if had Treatment #1} \\
0 & \text{if did NOT have Treatment #1}
\end{cases}
\]

and
\[ D_2 = \begin{cases} 1 & \text{if had Treatment #2} \\ 0 & \text{if did NOT have Treatment #2} \end{cases} \]

2. The next step is to estimate group differences in projected movie attendance by estimating, \( \hat{Y} = \hat{a} + \hat{b}_1D_1 + \hat{b}_2D_2 \).

3. Probably the easiest way to calculate the regression coefficients in a dummy variable analysis (like this one) is as follows:

   a. First, recall that \( \hat{Y}_{ij} = Y_i \) (i.e., the best estimate for each group member is the mean of the member's group). Accordingly, for each group set \( \hat{Y}_i = Y_i \) for each subject within the \( i \)th group.

   b. When \( i = 1 \), \( D_1 = 1 \) and \( D_2 = 0 \), thus \( \hat{Y}_1 = \hat{a} + \hat{b}_1 = Y_{1.} = 6 \).

   c. When \( i = 2 \), \( D_1 = 0 \) and \( D_2 = 1 \), thus \( \hat{Y}_2 = \hat{a} + \hat{b}_2 = Y_{2.} = 10 \).

   d. When \( i = 3 \), \( D_1 = D_2 = 0 \) and thus \( \hat{Y}_3 = \hat{a} = Y_{3.} = 11 \).

   e. Thus \( \hat{a} = 11 \), \( \hat{b}_1 = 6 - 11 = -5 \), and \( \hat{b}_2 = 10 - 11 = -1 \). And the resulting regression equation is \( \hat{Y} = 11 - 5D_1 - D_2 \).

4. In words, "Subjects who were (hypothetically) given tickets said, on average, that they would attend 11 movies. The average number of movies that subjects who had (hypothetically) lost tickets said that they would attend was 5 movies less than this. On average, subjects who had (hypothetically) lost money said they would attend one movie less than the average number said by subjects who were (hypothetically) given tickets."

5. WARNING: Do not attempt to interpret slopes associated with dummy variables as you would slopes from other types of multiple regression.
equations. That is, do not speak of the effects of one dummy variable adjusted for (controlling for, holding constant, net of, etc.) the effects of another dummy variable. Dummy variables must be taken as a group that in combination comprise a single nominal-level variable. Coefficients associated with dummy variables should be interpreted as deviations from the constant in the equation, where the constant equals the mean value of the dependent variable for the group for which all dummy variables equal zero.

6. Finally, it should be clear at this point that analysis of variance differs from regression only in the way that estimated values on the dependent variable are determined.

a. In an analysis of variance, the mean for a subject’s group is the best estimate of the subject’s value on the dependent variable.

b. In a regression analysis, this best estimate is calculated using a formula for the linear association between the dependent variable and a continuous independent variable.

G. Assumptions in regression and ANOVA

1. The assumptions of linear regression are as follows:

a. Homoscedasticity—the conditional variance of Y (estimated by $\hat{\sigma} = \text{MSE}$) is the same for all values of X.

b. Normality—the conditional distribution of Y for each fixed value of X is normal.

c. Randomness—the Ys are statistically independent (i.e., randomly sampled).
d. Linearity—X and Y are linearly related.

2. The assumptions of analysis of variance are the same EXCEPT in that

a. the Ys within each treatment group must be statistically independent (i.e., there are as many random samples as treatments.) and

b. X and Y need not be linearly related.

1) Insofar as the groups specified in an ANOVA are merely different (i.e., not "more" or "less" than each other in some way), it is meaningless to speak of the dependent variable as varying linearly from one group to the next.

2) In an ANOVA the variance explained by a nominal-level (or grouping) variable is calculated without first having assigned values to each level of (i.e., without operationalizing) the nominal variable.