Analysis of Variance (ANOVA)

Much of statistical inference centers around the ability to distinguish between two or more groups in terms of some underlying response variable \( y \). Sometimes, there are but two groups (males vs. females or timing at PFI vs. 24 hours before PFI), sometimes more groups, and sometimes multiple sets of groups from more than one factor (categorical explanatory variable). These different situations and their corresponding analyses are given below.

1. **2 groups**: Use a (Welch’s) t-test or Wilcoxon rank sum test if normality assumptions are not met.

2. **3 or more groups, 1 factor**: Use a one-way ANOVA or the Kruskal-Wallis test if normality assumptions are not met.

3. **2 or more factors**: Use an n-way ANOVA to address the effects of the factors and their interactions on the response \((y)\).

The type of ANOVA performed depends on how the data being analyzed were collected. In general, there are many types of designs used in practice, all of which can be analyzed by some type of ANOVA. Some of these are:

1. **Observational studies**: weakest inference, data collected without formal randomization

2. **Completely randomized designs (CRD)**: treatments (groups) randomly assigned to units

3. **Randomized block designs (RBD)**: CRD’s performed within heterogeneous blocks of similar units, can remove variation in the treatment by blocking

4. **Factorial (crossed) designs**: CRD’s or RBD’s on two or more factors

5. **Nested designs**: Designs where levels of one factor are nested within (i.e.: ”only have meaning within” the levels of another factor

6. **Repeated measures designs**: Repeated measures on the same units over time

7. **Split plot designs, latin squares**, etc.

Regardless of what type of ANOVA is needed for a given problem, the basic idea of ANOVA is the same in all cases. Ultimately, all ANOVA does is partition the total variation in the response variable \( y \) (called SST) into those parts explained by the various factors under study. In other words, it quantifies how much of the variability in \( y \) is due to the factors of interest.

**Conceptual ANOVA**: Why is it called “analysis of variance?”
Example: Insects such as fruit flies rest, but is this rest sleep? One experiment to investigate this question gave caffeine to fruit flies to see if it affected their rest. We know that caffeine reduces sleep in mammals, so if it reduces rest in fruit flies, this might be a hint that the rest is really sleep. The amounts of rest (in minutes) for the four caffeine treatment groups appear below:

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>1 mg/ml</th>
<th>2.5 mg/ml</th>
<th>5 mg/ml</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>542</td>
<td>482</td>
<td>500</td>
<td>250</td>
</tr>
<tr>
<td></td>
<td>627</td>
<td>479</td>
<td>450</td>
<td>236</td>
</tr>
<tr>
<td></td>
<td>520</td>
<td>486</td>
<td>466</td>
<td>315</td>
</tr>
<tr>
<td></td>
<td>529</td>
<td>575</td>
<td>377</td>
<td>299</td>
</tr>
<tr>
<td></td>
<td>556</td>
<td>464</td>
<td>491</td>
<td>323</td>
</tr>
</tbody>
</table>

|                |          |          |          |
| Sample Mean    | $\bar{y}_1 = 554.8$ | $\bar{y}_2 = 497.2$ | $\bar{y}_3 = 456.8$ | $\bar{y}_4 = 284.6$ |
| Sample SD      | $s_1 = 42.588$ | $s_2 = 44.280$ | $s_3 = 48.823$ | $s_4 = 39.259$ |
| Sample Size    | $n_1 = 5$ | $n_2 = 5$ | $n_3 = 5$ | $n_4 = 5$ |

- We want to test for any differences among the population rest times for the 4 groups, i.e.: we want to test:

Two Types of Variation Present

1. **Within-sample Variation** - variation within a sample from one group. How would you measure this? Letting $n$ be the total sample size and $t$ be the number of treatments (groups):

   $$\text{MSW} = \frac{\sum(n_i - 1)s_i^2}{n - t} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2 + (n_4 - 1)s_4^2}{n - t}$$

   $$= \frac{4(42.588)^2 + \cdots + 4(39.259)^2}{20 - 4} = \frac{30797.6}{16} = 1924.85 \text{ with } 16 \text{ d.f.}$$

2. **Between-sample Variation** - variation between the samples. How would you measure this?

   $$\text{MSB} = \frac{1}{t - 1} \sum_{i=1}^{t} n_i(\bar{y}_i - \bar{y}.)^2 = \frac{1}{3} \left[5(554.8 - 448.35)^2 + \cdots + 5(284.6 - 448.35)^2\right]$$

   $$= \frac{1}{3} \left[203016.95\right] = 67672.32 \text{ with } 3 \text{ d.f.}$$

- Note that, as expected, the between group mean sum of squares is considerably larger than the within group mean SS. This indicates emphatically that there are significant differences in mean rest time between the groups relative to the variation in rest time within a group. How do we formally test for such a difference?

- We wish to test:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad \Leftrightarrow \quad H_0 : \begin{cases} \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ \mu_1 = \mu_3 = \mu_2 = \mu_4 \end{cases}$$

   vs. $H_a : \text{At least one of these is not true.}$
• Suppose $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$ (homogeneity of variance assumption).

Aside: We could test each of the 6 cases ($H_0 : \mu_i = \mu_j$ vs. $H_a : \mu_i \neq \mu_j, i \neq j$) with a t-test, at $\alpha = 0.05$. For example, we could test $H_0 : \mu_1 = \mu_2$ vs. $H_a : \mu_1 \neq \mu_2$ with the t-test statistic:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)}} = \frac{554.8 - 497.2}{\sqrt{\left(\frac{42.588^2}{5}\right) + \left(\frac{44.280^2}{5}\right)}} = \frac{57.6}{27.475} = 2.096.$$  

The p-value for this test is: $P(|T_8| > 2.096) = 0.069$ indicating suggestive evidence of a significant difference in the mean rest time between the control group and and the 1 mg/ml group.

If we were to fail to reject $H_0$ for each case (at $\alpha = 0.05$), can we conclude that $\mu_1 = \mu_2 = \mu_3 = \mu_4$?

As the number of test increases, the probability of falsely rejecting $H_0$ in at least one of those test increases.

$\Rightarrow$ Combined Type I error $\alpha$ for 6 t-tests is larger than the $\alpha$-level for a single test.

• This test hinges on comparing the within-sample variance ($s_{W}^2$) to the between-sample variance ($s_{B}^2$).

• Fact: Under $H_0$, both $s_{B}^2$ & $s_{W}^2$ are independent unbiased estimates of the population variance $\sigma^2$ (since then, all 16 observations are from the same population). Hence, we expect $s_{B}^2 / s_{W}^2 = 1$ under $H_0$.

• If $s_{B}^2 >> s_{W}^2$ (as in the fly example), we have evidence against $H_0$. How do we quantify the degree of evidence against $H_0$?

Result: Under $H_0$, the test statistic $F = \frac{s_{B}^2}{s_{W}^2} \sim F \left(t - 1, \sum_{i=1}^{t} n_i - t\right)$ distribution.

• So, large values of $F$ $\Rightarrow$

• To test $H_0$: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ vs. $H_a$: At least one is different, the test statistic is:

$$F = \frac{s_{B}^2}{s_{W}^2} = \frac{67672.32}{1924.85} = 35.16 >> 9.01 = F_{.001}(3, 16),$$

so the p-value < .001, and we reject $H_0$ concluding there are highly significant differences in the mean rest times. This is called the main effects F-test.
Example 2: An observational study was performed in a large statistics class where the hours of sleep the previous night ($y$) and the seating position (front, middle, back) in the lecture hall ($x_1$) were recorded. The teacher wanted to see if there was any relationship between these two variables. For this example, I want to demonstrate how to use MatLab to run an ANOVA model.

An ANOVA model was run using the `anovan` function below and produced the following ANOVA table:

```matlab
[p1,tab1,stats1,terms1] = anovan(sleep.hours,{sleep.seat},1);
```

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum Sq.</th>
<th>d.f.</th>
<th>Mean Sq.</th>
<th>F</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>9.44</td>
<td>2</td>
<td>4.71993</td>
<td>3.21</td>
<td>0.0429</td>
</tr>
<tr>
<td>Error</td>
<td>250.135</td>
<td>170</td>
<td>1.47138</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>259.575</td>
<td>172</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Constrained (Type III) sums of squares.

What does this indicate about any differences in mean sleeping time according to seating position?

Since there appears to be an overall effect of seating position on mean sleeping time, we can also look at the magnitude and significance of all pairwise differences between the three means using the `multcompare` function to perform multiple comparisons of the factor level means as follows:

```matlab
[ci,means] = multcompare(stats1,'ctype','hsd');
```

- The "[ci means]" piece returns the group means and 95% (default) confidence intervals for the group means. The 'ctype','hsd' piece on the RHS tells MatLab to use the Tukey Honest Significance Difference (HSD) 95% joint confidence intervals for the three group means.

- There are other methods of performing multiple comparisons among the group means. Some of the more popular choices are:
  1. Fisher’s least significant difference (LSD).
  2. Bonferroni method
  3. Scheffe’s method

The population marginal means of groups X1=Back and X1=Front are significantly different.
• The Tukey HSD method tends not to be too conservative so is generally the preferred choice of comparison among any two groups when multiple comparisons are to be made.

• What assumptions are being made in running this ANOVA model?
  
  1. 
  2. 
  3. 

• A residual plot and normal quantile plot are shown below with the MatLab code following. What do these tell you?

```matlab
figure(1) % 1st Figure
yhat1 = sleep.hours-stats1.resid;
plot(yhat1,stats1.resid,'ko'); % Plots residuals vs. predicted y-values
xlabel('Predicted Values','fontsize',14,'fontweight','b');
ylabel('Residuals','fontsize',14,'fontweight','b');
title('Residual Plot','fontsize',14,'fontweight','b');
figure(2); % 2nd Figure
qqplot(stats1.resid); % Normal quantile plot of residuals
xlabel('Standard Normal Quantiles','fontsize',14,'fontweight','b');
ylabel('Residuals','fontsize',14,'fontweight','b');
title('Normal Quantile Plot','fontsize',14,'fontweight','b');
```
Back to the Sleep Example: Suppose in addition to the mean hours of sleep and the seating position, the gender of the student was also recorded. Here then, we have two factors (seat, gender) and want to examine their effect and possible interactive effect on mean hours of sleep. In this case, we can use a 2-way ANOVA. What does such a model really look like?

Let:

\[ y_i = \text{the mean hours of sleep for the } i^{th} \text{ student}, \ i = 1, 2, \ldots, 173, \]

\[ x_{1i} = \begin{cases} 1 & \text{if the } i^{th} \text{ seat is in the Front} \\ 0 & \text{otherwise} \end{cases}, \]

\[ x_{2i} = \begin{cases} 1 & \text{if the } i^{th} \text{ seat is in the Middle} \\ 0 & \text{otherwise} \end{cases}, \]

\[ x_{3i} = \begin{cases} 1 & \text{if the } i^{th} \text{ student is Female} \\ 0 & \text{otherwise} \end{cases}, \]

\[ x_{4i} = x_{1i} \times x_{3i}, \]

\[ x_{5i} = x_{2i} \times x_{3i}. \]

Then, this 2-way ANOVA model can be written as:

\[ y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \epsilon_i. \]

Although there were 173 subjects in this sleep study, suppose for a moment that there were but 12 subjects with 2 at each of the seat/gender combinations. In matrix form, the model would appear as follows:

\[
\begin{bmatrix}
  y_1 \\
  y_2 \\
  y_3 \\
  y_4 \\
  y_5 \\
  y_6 \\
  y_7 \\
  y_8 \\
  y_9 \\
  y_{10} \\
  y_{11} \\
  y_{12}
\end{bmatrix} =
\begin{bmatrix}
  1 & 1 & 0 & 1 & 1 & 0 \\
  1 & 1 & 0 & 1 & 1 & 0 \\
  1 & 1 & 0 & 0 & 0 & 0 \\
  1 & 1 & 0 & 0 & 0 & 0 \\
  1 & 0 & 1 & 1 & 0 & 1 \\
  1 & 0 & 1 & 1 & 0 & 1 \\
  1 & 0 & 1 & 0 & 0 & 0 \\
  1 & 0 & 1 & 0 & 0 & 0 \\
  1 & 0 & 0 & 1 & 0 & 0 \\
  1 & 0 & 0 & 1 & 0 & 0 \\
  1 & 0 & 0 & 0 & 0 & 0 \\
  1 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
  \beta_0 \\
  \beta_1 \\
  \beta_2 \\
  \beta_3 \\
  \beta_4 \\
  \beta_5 \\
  \epsilon_1 \\
  \epsilon_2 \\
  \epsilon_3 \\
  \epsilon_4 \\
  \epsilon_5 \\
  \epsilon_6 \\
  \epsilon_7 \\
  \epsilon_8 \\
  \epsilon_9 \\
  \epsilon_{10} \\
  \epsilon_{11} \\
  \epsilon_{12}
\end{bmatrix}.
\]

- Hence, any "ANOVA model" is nothing more than a linear model with the design matrix defined in the "right way". Specifically, continuous variables (such as mean hours of sleep) contain the values themselves, whereas categorical variables (such as seat & gender) can be expressed as indicator variables.

- If a categorical variable has t levels, then we need t − 1 indicator variables to describe the possible outcomes for that variable.
Note that in the sleep example, a first-order model with an interaction requires 6 parameters to be estimated. In general, for 2 factors at \(a\) levels and \(b\) levels respectively, the model contains \(ab\) parameters, and hence there are \(ab\) degrees of freedom in the explanatory part of the model.

Although we could run any ANOVA model as a general linear model, software is written to construct these indicator variables behind the scenes and simply report the ANOVA table with the relevant components to allow the user to examine the effects of the factors and their interactions on the response variable. Yeah!!!

Before running the 2-way ANOVA model, we first compute some basic group statistics and look at some plots of the main effects as well as an interaction plot expressing the relationship between seat and gender on mean hours of sleep.

% Computes group statistics for both seat location and gender
[order numel means se] = grpstats(sleep.hours,{sleep.seat sleep.gender}, ... 
    {'gname','numel','mean','sem'},.05);

order = numel = means = se =

'Front' 'Female' 24 7.1250 0.2762
'Front' 'Male' 20 6.0000 0.2206
'Middle' 'Female' 37 7.0541 0.1547
'Middle' 'Male' 31 6.8710 0.2086
'Back' 'Female' 29 7.0690 0.1836
'Back' 'Male' 32 7.3594 0.2733

% Constructs plots of the group means for both groups separately
maineffectsplot(sleep.hours,{sleep.seat,sleep.gender}, ... 
    'varnames',{'Seat Location','Gender'});

% Constructs an interaction plot of seat location and gender on sleep hours
interactionplot(sleep.hours,{sleep.seat sleep.gender}, ... 
    'varnames',{'Seating Location','Gender'});
Finally, the 2-way ANOVA is run with the following MatLab code and produces the ANOVA table which follows. Are the results surprising given the plots created?

% Runs the 2-way ANOVA with an interaction (2)
[p1,tab1,stats1,terms1] = anovan(sleep.hours,{sleep.seat,sleep.gender}, ...'
'varnames',{'Seat','Gender'}, ...'
'model','interaction');

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum Sq.</th>
<th>d.f.</th>
<th>Mean Sq.</th>
<th>F</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seat</td>
<td>10.809</td>
<td>2</td>
<td>5.40474</td>
<td>3.85</td>
<td>0.023208</td>
</tr>
<tr>
<td>Gender</td>
<td>4.78</td>
<td>1</td>
<td>4.7796</td>
<td>3.4</td>
<td>0.066806</td>
</tr>
<tr>
<td>Seat*Gender</td>
<td>12.87</td>
<td>2</td>
<td>6.43489</td>
<td>4.58</td>
<td>0.011543</td>
</tr>
<tr>
<td>Error</td>
<td>234.48</td>
<td>167</td>
<td>1.40407</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>259.575</td>
<td>172</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Constrained (Type III) sums of squares.

What can we say about a gender effect?

Do the various sums of squares sum to the total sum of squares?

- Multiple comparisons among the different levels of individual groups and between levels of combinations of groups can be made to compare group means.

- If some of the explanatory variables are continuous, we can include these as additional variables in the model as well. Formally, if an experiment is performed where units are randomly assigned to combinations of the (categorical) model factors and then measurements are made on the response variable $y$ and other continuous covariates, the resulting analysis is known as an **analysis of covariance** (ANCOVA).