Analysis of variance is a special case of regression (see Unit 2, Regression and Correlation). The distinction is that, in analysis of variance, all of the predictors are categorical and are called factors. Recall that in regression, generally, the predictors can be either continuous or discrete or a mix. The common feature of analysis of variance and normal theory linear regression is that the outcome variable is continuous and assumed distributed normal.

* 1 factor with 3 or more levels is a one way analysis of variance.
* 1 factor with just 2 levels is a two sample t-test.
* 2 factors (regardless of the number of levels) is a two way analysis of variance.
* 3 factors is a three-way analysis of variance. And so on.

In a factorial design, there are observations at every combination of levels of the factors. The analysis is used to explore interactions (effect modification) between factors. An interaction between factor I and factor II is said to exist when the response to factor II depends on the level of factor I and vice versa.

In a nested or hierarchical design, such as a two-level nested design, the analysis is of units (eg-patients) that are clustered by level of factor I (eg-hospital) which are in turn clustered by level of factor II (eg-city).

A special type of nested design is the longitudinal or repeated measurements design. Repeated measurements are clustered within subjects and the repeated measurements are made over a meaningful dimension such as time (eg-growth over time in children) or space. The analysis of repeated measurements data is discussed elsewhere.

Unit 7 is an introduction to the basics of analysis of variance.
# Table of Contents

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>4</td>
<td>10</td>
<td>14</td>
<td>20</td>
<td>20</td>
<td>24</td>
<td>26</td>
<td>26</td>
<td>27</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Learning Objectives

When you have finished this unit, you should be able to:

- Explain how analysis of variance is a special case of normal theory linear regression.
- Perform and interpret a one way analysis of variance.
- Explain what is meant by a multi-way analysis of variance.
- Explain what is meant by a factorial design analysis of variance.
- Explain the meaning of interaction of two factors.
- Explain what is meant by a nested design analysis of variance.
- Perform and interpret a two-way factorial analysis of variance, including an assessment of interaction.
1. The Logic of Analysis of Variance

- Analysis of variance is an analysis of the variability of means.
- Consider the following picture that represents two scenarios.
- In scenario 1 (left), the means are different.
- In scenario 2 (right), the means are the same.

<table>
<thead>
<tr>
<th>Scenario 1</th>
<th>Scenario 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>means are different</td>
<td>means are the same</td>
</tr>
<tr>
<td>μ₁ ≠ μ₂</td>
<td>μ₁ = μ₂</td>
</tr>
</tbody>
</table>

\[ S_1^2 \text{ and } S_2^2 \text{ each summarize “noise” controlling for location.} \]

The size of \(|\bar{X}_1 - \bar{X}_2|\) is larger than “noise”

\[ S^2 \text{ is larger than } S_1^2 \text{ and } S_2^2 \text{ because it is made larger by the extra variability among individuals due to change in location.} \]

\[ S_1^2 \text{ and } S_2^2 \text{ each summarize “noise” controlling for location.} \]

\[ |\bar{X}_1 - \bar{X}_2| \text{ is within the neighborhood of “noise”.} \]

\[ S^2 \text{ is similar in size to } S_1^2 \text{ and } S_2^2 \text{ because it does not have to accommodate an extra source of variability because of location differences between the two groups.} \]
A Special Case That is Convenient for Illustration is Balanced Data (Equal sample sizes in each group) -

When the sample sizes are the same in each group, it is easier to “see” how analysis of variance is an analysis of the variability of the means. Consider the following example (note – the algebra is not so tidy when the sample sizes are not equal)

**Example**

It has been hypothesized that the companionship of a pet provides psychological relief to its owners when they are experiencing stress. An experiment was conducted to address this question. Consenting participants were randomized to one of three conditions: 1- Pet Present, 2-Friend Present, or 3-Neither friend nor pet present. Each participant was then exposed to a stressor (it happened to be mental arithmetic). The outcome measured was heart rate.

Selected Summary Statistics, by Group:

<table>
<thead>
<tr>
<th></th>
<th>Group 1 Pet Present</th>
<th>Group 2 Friend Present</th>
<th>Group 3 Neither Pet nor Friend</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>n₁ = 15</td>
<td>n₂ = 15</td>
<td>n₃ = 15</td>
</tr>
<tr>
<td>(\bar{X})</td>
<td>(\bar{X}_1 = 73.48)</td>
<td>(\bar{X}_2 = 91.33)</td>
<td>(\bar{X}_3 = 82.52)</td>
</tr>
<tr>
<td>S</td>
<td>S₁ = 9.97</td>
<td>S₂ = 8.34</td>
<td>S₃ = 9.24</td>
</tr>
<tr>
<td>(S^2)</td>
<td>S₁² = 99.40</td>
<td>S₂² = 69.57</td>
<td>S₃² = 85.41</td>
</tr>
<tr>
<td>((n - 1)S^2)</td>
<td>((n₁ - 1)S₁² = )</td>
<td>((n₂ - 1)S₂² = )</td>
<td>((n₃ - 1)S₃² = )</td>
</tr>
<tr>
<td>(\sum(x - \bar{x})^2)</td>
<td>(\sum(X_{1j} - \bar{X}_1)^2 = )</td>
<td>(\sum(X_{2j} - \bar{X}_2)^2 = )</td>
<td>(\sum(X_{3j} - \bar{X}_3)^2 = )</td>
</tr>
<tr>
<td></td>
<td>= 1,391.57</td>
<td>= 974.05</td>
<td>= 1,195.70</td>
</tr>
</tbody>
</table>

- Do these data provided statistically significant evidence that the means are different, ie – that \(\mu_1\), \(\mu_2\), and \(\mu_3\) are not equal?
The Reasoning in an Analysis of Variance

**Proof by Contradiction**

**Signal-to-Noise**

We illustrate with a one way fixed effects analysis of variance.

1. *Begin with the null hypothesis assumption that the means are the same*
   - We’ll assume the null hypothesis is true, then apply this model to the observed data, and look to see if its application has led to an unlikely result, warranting rejection of the null hypothesis of equality of $\mu_1$, $\mu_2$, and $\mu_3$.

2. **State the assumptions necessary for computing probabilities.**
   - $X_{11}$ $\cdots$ $X_{1n_1}$ are distributed Normal ($\mu_1$, $\sigma^2$)
   - $X_{21}$ $\cdots$ $X_{2n_2}$ are distributed Normal ($\mu_2$, $\sigma^2$)
   - $X_{31}$ $\cdots$ $X_{3n_3}$ are distributed Normal ($\mu_3$, $\sigma^2$)
   - The variances are all equal to $\sigma^2$.
   - The observations are all independent.

3. **Specify $H_0$ and $H_A$.**
   
   $H_0$: $\mu_1 = \mu_2 = \mu_3$
   
   $H_A$: not

4. **“Reason” an appropriate test statistic (Signal-to-Noise).**

   **“Within Group Variability = Noise”** It is the variability of individuals within each group.

   - Noise: Obtain separate estimates of the common $\sigma^2$, one from each of the 3 samples: $S_1^2$, $S_2^2$, and $S_3^2$
   - Noise estimates the variability among individuals, controlling for location.

   - Obtain an estimate of the common variance, $\sigma^2$ by combining the separate estimates $S_1^2$, $S_2^2$, and $S_3^2$ into a weighted average. The weights are the degrees of freedom of each of $S_1^2$, $S_2^2$, and $S_3^2$

   $$\text{Estimate of } \sigma^2 = \hat{\sigma}^2_{\text{within}} = \frac{\sum_{i=1}^{3} (n_i - 1) S_i^2}{\sum_{i=1}^{3} (n_i - 1)}$$
“Between Group Variability = Signal”. It is the variability among means.

- Next, consider a special sample of size = 3. The 3 “data points” in this special sample are

\[
\begin{align*}
X_1^* &= \sqrt{n_1} \bar{X_1} = \sqrt{15} \times (73.4831) = 284.59882 \\
X_2^* &= \sqrt{n_2} \bar{X_2} = \sqrt{15} \times (91.3251) = 353.70059 \\
X_3^* &= \sqrt{n_3} \bar{X_3} = \sqrt{15} \times (82.5241) = 319.61446
\end{align*}
\]

The sample mean of these means is \( \bar{X}^* = 319.3046 \)

- We call the expected value of the sample variance of \( X_1^* \), \( X_2^* \), and \( X_3^* \) the \( \sigma_{\text{between}}^2 \).

- **When the null hypothesis \( H_0 \) is true, and only when \( H_0 \) is true**, the sample variance of \( X_1^* \), \( X_2^* \), and \( X_3^* \) is an estimate of \( \sigma^2 \).

\[
E \left[ \frac{\sum_{i=1}^{3} (X_i^* - \bar{X}^*)^2}{(3-1)} \right] = \sigma_{\text{between}}^2 = \sigma^2
\]

- **When the alternative hypothesis \( H_A \) is true**, the sample variance of \( X_1^* \), \( X_2^* \), and \( X_3^* \) is an estimate of a quantity \( (\sigma_{\text{between}}^2) \) that is larger than \( \sigma^2 \).

\[
E \left[ \frac{\sum_{i=1}^{3} (X_i^* - \bar{X}^*)^2}{(3-1)} \right] = \sigma_{\text{between}}^2 = \sigma^2 + \Delta \quad \text{where}
\]

\( \Delta = \text{function}(\mu_1, \mu_2, \mu_3) > 0 \)

- Thus, the “signal” is \( \Delta = \text{function}(\mu_1, \mu_2, \mu_3) > 0 \)
The “Signal-to-Noise” analysis compares the Between group means variability to the Within groups variability. In the analysis of variance application, the comparison that is made is actually

- \[
\frac{\text{Noise} + \text{Signal}}{\text{Noise}} = \frac{\text{Variability among a function of the group means}}{\text{Variability of individuals within groups}}
\]

- \[
\text{Var}(X_1^*, X_2^*, X_3^*) = \frac{\text{“df weighted sum of } S_1^2, S_2^2, \text{ and } S_3^2}{\hat{\sigma}_{\text{between}}^2}
\]

- \[
= \left[ \frac{\hat{\sigma}_{\text{between}}^2}{\hat{\sigma}_{\text{within}}^2} \right]
\]

- \[
= \left[ \frac{\sum (X_i^* - \overline{X}^*)^2/(3-1)}{\sum (n_i - 1)S_i^2/(\sum (n_i - 1))} \right]
\]

5. **Perform the calculations.**

Using the values in the table on page 5, we have

- \[
\hat{\sigma}_{\text{within}}^2 = \frac{\sum (n_i - 1)S_i^2}{\sum (n_i - 1)} = \frac{(1391.57 + 974.05 + 1195.70)}{(14 + 14 + 14)}
\]

- \[
= 84.79
\]

Using the values of the \(X_i^*\) on page 7, we also have

- \[
\hat{\sigma}_{\text{between}}^2 = \frac{\sum (X_i^* - \overline{X}^*)^2}{(3-1)} = 1,193.836
\]
We use the F distribution to compare the two variances.

- Luckily, the two variances are independent. When the null hypothesis $H_0$ is true:

$$F = \frac{\hat{\sigma}^2_{\text{between}} / \sigma^2_{\text{between}}}{\hat{\sigma}^2_{\text{within}} / \sigma^2_{\text{within}}}$$

Numerator degrees of freedom = $(k - 1) = (3-1) = 2$
Denominator degrees of freedom = $\sum (n_i - 1) = (3)(15-1)= 42$

<table>
<thead>
<tr>
<th>$H_0$ true</th>
<th>$\hat{\sigma}^2_{\text{within}}$</th>
<th>$\hat{\sigma}^2_{\text{between}}$</th>
<th>$\hat{\sigma}^2_{\text{between}} / \hat{\sigma}^2_{\text{within}}$</th>
<th>$F$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2$</td>
<td>$\sigma^2$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Large</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>$\sigma^2 + \Delta$</td>
<td>&gt; 1</td>
<td>&gt; 1</td>
<td></td>
<td>Small</td>
</tr>
</tbody>
</table>

For our data, $F = \frac{1193.836}{84.79} = 14.08$

The accompanying p-value is $\text{Prob} \left[ F_{df=2,42} \geq 14.08 \right] = .00002$.

6. “Evaluate” findings and report.

The assumption of the null hypothesis of equal means has led to an extremely unlikely result! The null hypothesis chances were approximately, 2 chances in 100,000 of obtaining 3 means of groups that are as different from each other as are 73.48, 91.33, and 82.52. The null hypothesis is rejected.

7. Interpret in the context of biological relevance.

This analysis provides statistically significant evidence of group differences in heart rate, depending on companionship by pet or by friend. But we do not know which, or if both, provides the benefit!!
2. Introduction to Analysis of Variance Modeling

**Preliminary note** - In the pages that follow, I am using the notation $X$ to refer to the *outcome variable* in analysis of variance.

Analysis of variance models, like regression models, have an identifiable basic structure.

<table>
<thead>
<tr>
<th>Structure of Analysis of Variance Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed = mean + random error</td>
</tr>
</tbody>
</table>

\[
X = \mu + \epsilon
\]

- $\mu$ - This is the expected value of $X$ which we write as $\mu = E[X] = \text{linear model}(\text{stuff})$
- $\epsilon$ - This is the idea of “random error”, “error in measurement”, “noise”
- **Subscripts** – Subscripts keep track of group membership and persons within groups. E.g. - $X_{ij} = \text{Observed value for j}^{\text{th}} \text{ person in the i}^{\text{th}} \text{ group}$.

A special feature of analysis of variance models are their use of subscripts.

**Example** – One way fixed effects analysis of variance.

<table>
<thead>
<tr>
<th>Structure of One Way Fixed Effects Analysis of Variance Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed = mean + random error</td>
</tr>
</tbody>
</table>

Subscripts:
- “$i$” keeps track of the group.
- “$j$” keeps track of the individual within the group.

\[
X_{ij} = \mu_i + \epsilon_{ij}
\]

- Observed data value for $j^{\text{th}}$ person in $i^{\text{th}}$ group
- Expected value is the mean of $i^{\text{th}}$ group
- Random error for $j^{\text{th}}$ observation of the $i^{\text{th}}$ group mean.
### Introduction to Defining an ANOVA Model

#### The One Way Fixed Effects ANOVA

- In a one way fixed effects analysis of variance (anova) model, the “model” permits the means to be different from one group to the next however they may differ. There is no straight line model nor any sort of curvy model assumption for that matter. Instead, the mean outcome for each group \( E(X_{ij}) = \mu_i \) is completely general. \( E(X_{ij}) = \mu_i \) is not related in any way to an assumed functional form (beyond a departure from an intercept).

- **Example – Heart rate and stress, continued** - In this analysis, the null hypothesis was that the means are all the same. The alternative hypothesis was the completely general hypothesis \( \mu_1 \neq \mu_2 \neq \mu_3 \).

- More generally, suppose the number of groups = \( K \), instead of 3 in the heart rate example.

- Define subscripting as follows. The first subscript will be “i” and will index the groups using \( i=1, \ldots, K \). The second subscript will be “j” and will index the jth individual in the ith group using \( j = 1, \ldots, n_i \).

- Keep track of the group specific sample sizes. Let \( n_i \) represent the sample size for the “ith” group.

- Let

\[
\mu_i = \text{mean for persons in the } i^{\text{th}} \text{ group} \\
\mu = \text{overall mean, without regard for group membership}
\]

- **Deviation from means model.** This is a nifty re-write that rewrites \( \mu_i \) as a new expression that is equal to itself. This is done by adding and subtracting \( \mu \) to \( \mu_i \).

\[
\mu_i = \mu + (\mu_i - \mu)
\]

\[
\uparrow \quad \uparrow \quad \uparrow
\]

mean for overall deviation from

\[
\text{group “i” mean specific to ith group}
\]

- Same nifty trick to obtain a rewrite of the observed \( X_{ij} \). Notice (below) that we are adding and subtracting two things this time:

\[
X_{ij} = \bar{X} + (\bar{X}_i - \bar{X}) + (X_{ij} - \bar{X}_i)
\]
• One more manipulation lets us express the variability of individuals about the overall mean as the sum of two contributions. This is useful for analysis purposes.

\[(X_{ij} - \bar{X}) = (\bar{X}_i - \bar{X}) + (X_{ij} - \bar{X}_i)\]

• Each source (between or within) contributes its own share to the total variability via the following (wonderful) result.

\[
\sum_{i=1}^{K} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2 = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\bar{X}_i - \bar{X})^2 + \sum_{i=1}^{K} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2
\]

We keep track of all this in an analysis of variance table.

<table>
<thead>
<tr>
<th>Source</th>
<th>df⁴</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>Variance Ratio = F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between groups</td>
<td>(K-1)</td>
<td>[\sum_{i=1}^{K} \sum_{j=1}^{n_i} (\bar{X}<em>i - \bar{X})^2 ][\sum</em>{i=1}^{K} \sum_{j=1}^{n_i} (\bar{X}_i - \bar{X})^2 ] (\frac{(K-1)}{(K-1)}]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[\sum_{i=1}^{K} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2]</td>
<td>[\sum_{i=1}^{K} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2]</td>
<td>(\frac{\hat{\sigma}<em>{\text{between}}^2}{\hat{\sigma}</em>{\text{within}}^2})</td>
</tr>
<tr>
<td>Within Groups</td>
<td></td>
<td>[\sum_{i=1}^{K} (n_i - 1)]</td>
<td>[\sum_{i=1}^{K} (n_i - 1)]</td>
<td>(\hat{\sigma}_{\text{within}}^2)</td>
</tr>
<tr>
<td>Total</td>
<td>N - 1</td>
<td>[\sum_{i=1}^{K} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2]</td>
<td>[\sum_{i=1}^{K} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2]</td>
<td>[\sum_{i=1}^{K} (n_i - 1)]</td>
</tr>
</tbody>
</table>

⁴ degrees of freedom
Example – continued from page 16: Stress and Heart Rate

Data

<table>
<thead>
<tr>
<th>Pet Present</th>
<th>Friend Present</th>
<th>Neither Pet, Nor Friend</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>69.17</td>
<td>99.69</td>
<td>84.74</td>
</tr>
<tr>
<td>68.86</td>
<td>91.35</td>
<td>87.23</td>
</tr>
<tr>
<td>70.17</td>
<td>83.40</td>
<td>84.88</td>
</tr>
<tr>
<td>64.17</td>
<td>100.88</td>
<td>80.37</td>
</tr>
<tr>
<td>58.69</td>
<td>102.15</td>
<td>91.75</td>
</tr>
<tr>
<td>79.66</td>
<td>89.82</td>
<td>87.45</td>
</tr>
<tr>
<td>69.23</td>
<td>80.28</td>
<td>87.78</td>
</tr>
<tr>
<td>75.98</td>
<td>98.20</td>
<td>73.28</td>
</tr>
<tr>
<td>86.45</td>
<td>101.06</td>
<td>84.52</td>
</tr>
<tr>
<td>97.54</td>
<td>76.91</td>
<td>77.80</td>
</tr>
<tr>
<td>85.00</td>
<td>97.05</td>
<td>70.88</td>
</tr>
<tr>
<td>69.54</td>
<td>88.02</td>
<td>90.02</td>
</tr>
<tr>
<td>70.08</td>
<td>81.60</td>
<td>99.05</td>
</tr>
<tr>
<td>72.26</td>
<td>86.98</td>
<td>75.48</td>
</tr>
<tr>
<td>65.45</td>
<td>92.49</td>
<td>62.65</td>
</tr>
</tbody>
</table>

Stata Illustration
.oneway stress treatment

<table>
<thead>
<tr>
<th>Source</th>
<th>df(^a)</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>Variance Ratio = F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Among groups</td>
<td>2</td>
<td>2387.69</td>
<td>(\hat{\sigma}^2_{between} = 1193.84)</td>
<td>(F = \frac{\hat{\sigma}^2_{between}}{\hat{\sigma}^2_{within}} = 14.08)</td>
</tr>
<tr>
<td>Within Groups</td>
<td>42</td>
<td>3561.30</td>
<td>(\hat{\sigma}^2_{within} = 84.79)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>44</td>
<td>5948.99</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) degrees of freedom

MATCH! – My hand calculation of the one way anova matches the Stata results
3. The One Way Fixed Effects Analysis of Variance

Fixed versus Random Effects are Introduced in Section 5b

The example used to introduce the logic of analysis of variance is a one way fixed effects analysis of variance. Thus, we have a feel for the one way analysis of variance already. A two sample t-test is a one way analysis of variance where the number of groups $K=2$.

<table>
<thead>
<tr>
<th>The PubHlth 540 View</th>
<th>The ANOVA View</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is the “signal” $(\bar{X}_1 - \bar{X}_2)$ large relative to “noise” where “noise” = $SE(\bar{X}_1 - \bar{X}_2)$?</td>
<td>Is the variability of $(\bar{X}_1, \bar{X}_2)$ large relative to “noise” where “noise” = weighted average of $S_1^2, S_2^2$</td>
</tr>
<tr>
<td>$t = \frac{(\bar{X}_1 - \bar{X}_2)}{SE(\bar{X}_1 - \bar{X}_2)}$</td>
<td>$F = \frac{function of variability of data \bar{X}_1, \bar{X}_2}{function of variability of &quot;noise&quot; S_1^2, S_2^2}$</td>
</tr>
<tr>
<td>= measure of distance of $(\bar{X}_1 - \bar{X}_2)$ from 0, expressed on SE scale.</td>
<td>= measure of variability among $\bar{X}_1, \bar{X}_2$ (“signal”) to variability of individuals within groups (“noise”)</td>
</tr>
</tbody>
</table>

Setting

1. **Normality.** The observed outcomes are distributed normal.

   Group 1: $X_{11}, \ldots, X_{1n_1}$ are a simple random sample from a Normal($\mu_1, \sigma^2$)
   Group 2: $X_{21}, \ldots, X_{2n_2}$ are a simple random sample from a Normal($\mu_2, \sigma^2$)
   Etc.
   Group $K$: $X_{K1}, \ldots, X_{Kn_K}$ are a simple random sample from a Normal($\mu_K, \sigma^2$)

2. **Constant variance.** The $K$ separate variance parameters are equal

3. **Independence** The observations are independent
The K separate population means are

\[ \mu_1, \mu_2, \ldots, \mu_K \]

\[ H_0: \mu_1 = \ldots = \mu_K \]

\[ H_A: \text{At least some are unequal} \]

### One Way Fixed Effects Analysis of Variance

**Model of** \( E[X_{ij}] = \mu_i \)

Recall -

“\( i \)” keeps track of the group.
“\( j \)” keeps track of the individual within the group.

Model for the mean in the \( i \)th group -

\[ (1) \quad E[X_{ij}] = \mu_i = [\mu + \tau_i] \quad \text{where} \]

\[ (2) \quad \sum_{i=1}^{K} \tau_i = 0 \]

**Key:** The “different-ness” of each mean is captured in the \( \tau_1, \ldots, \tau_K \). Notice the following:

- **Group 1:** \( \mu_1 = \mu + [\mu_1 - \mu] = \mu + \tau_1 \) says that \( \mu_1 - \mu = \tau_1 \)

- **...**

- **Group K:** \( \mu_K = \mu + [\mu_K - \mu] = \mu + \tau_K \) says that \( \mu_K - \mu = \tau_K \)

- By definition, \( \sum_{i=1}^{K} \tau_i = 0 \)

- If the means are NOT EQUAL, then at least one \( \tau_i = [\mu_i - \mu] \neq 0 \)
One Way Analysis of Variance
Fixed Effects Model

Setting:
- $K$ groups indexed $i = 1, 2, \ldots, K$
- Group specific sample sizes: $n_1, n_2, \ldots, n_K$
- $X_{ij} =$ Observation for the $j$th individual in the $i$th group

The one way analysis of variance fixed effects model of $X_{ij}$ is defined as follows:

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

where

$$\mu = \text{grand mean}$$
$$\tau_i = [\mu_i - \mu]$$
$$\sum_{i=1}^{K} \tau_i = 0$$

and

$$\varepsilon_{ij} \text{ is random error distributed Normal}(0, \sigma^2)$$

Estimation

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate using Sample Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>$\bar{X}$</td>
</tr>
<tr>
<td>$\tau_i$</td>
<td>$[\bar{X}_i - \bar{X}]$</td>
</tr>
<tr>
<td>$\mu + \tau_i$</td>
<td>$\bar{X}_i$</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>$\hat{\sigma}^2_{\text{within}} = \frac{\sum_{i=1}^{K} (n_i - 1)S_i^2}{\sum_{i=1}^{K} (n_i - 1)}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Expected Value of</th>
<th>$F$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ true</td>
<td>$\sigma^2$</td>
<td>$\sigma^2$</td>
</tr>
<tr>
<td>$H_A$ true</td>
<td>$\sigma^2$</td>
<td>$\sigma^2 + \Delta$</td>
</tr>
</tbody>
</table>


Example
Three groups of physical therapy patients were subjected to different treatment regimens. At the end of a specified period of time each was given a test to measure treatment effectiveness. The sample size was 4 in each group. The following scores were obtained.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>75</td>
<td>25</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>75</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>25</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>75</td>
<td>40</td>
</tr>
</tbody>
</table>

\[
\bar{X}_1 = 70 \quad \bar{X}_2 = 50 \quad \bar{X}_3 = 80
\]

\[
X_1^* = \sqrt{n} [\bar{X}_1] = (2) (70) = 140 \quad X_2^* = \sqrt{n} [\bar{X}_2] = (2) (50) = 100 \quad X_3^* = \sqrt{n} [\bar{X}_3] = (2) (80) = 160
\]

\[
\bar{X}^* = \frac{1}{3} \left( \sum_{i=1}^{3} X_i^* \right) = 133.3333
\]

\[H_0: \mu_1 = \mu_2 = \mu_3\]
\[H_A: \text{not} \]

**Step 1:** Test the assumption of equality of variances.

Tests of the assumption of equality of variances are discussed in Section 4a. These are of limited usefulness for two reasons:

1. Tests of equality of variance tend to be sensitive to the assumption of normality.
2. Analysis of variance methodology is pretty robust to violations of the assumption of a common variance.

**Step 2:** Estimate the within group variance (“noise”). This will be a weighted average of the k separate sample variances.

\[
\hat{\sigma}_{\text{within}}^2 = \frac{\sum_{i=1}^{K} (n_i - 1) S_i^2}{\sum_{i=1}^{3} (n_i - 1)} = \frac{5450.00}{9} = 605.56
\]
**Step 3:**  Estimate the between group variance ("noise + signal"). This will be a sample variance calculation for a special data set comprised of \( X_1^* = \sqrt{n \bar{X}_1} \), \( X_2^* = \sqrt{n \bar{X}_2} \), \( X_3^* = \sqrt{n \bar{X}_3} \).

\[
\hat{\sigma}_{\text{between}}^2 = \left[ \frac{\sum_{i=1}^{3} (X_i^* - \bar{X}^*)^2}{(3-1)} \right] = \frac{1866.67}{2} = 933.33
\]

**Step 4:** Summarize in an analysis of variance table. Perform F-test of null hypothesis of equal means.

<table>
<thead>
<tr>
<th>Source</th>
<th>df(^a)</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>Variance Ratio = F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between groups</td>
<td>(K-1)</td>
<td>( \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\bar{X}_i - \bar{X})^2 )</td>
<td>( \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\bar{X}_i - \bar{X})^2 / (K-1) )</td>
<td>( F = \frac{\hat{\sigma}<em>{\text{between}}^2}{\hat{\sigma}</em>{\text{within}}^2} = 1.54 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1866.67</td>
<td>933.33</td>
<td></td>
</tr>
<tr>
<td>Within Groups</td>
<td>( \sum_{i=1}^{K} (n_i - 1) )</td>
<td>( \sum_{i=1}^{K} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 )</td>
<td>( \sum_{i=1}^{K} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}<em>i)^2 / \sum</em>{i=1}^{K} (n_i - 1) )</td>
<td>( \hat{\sigma}_{\text{within}}^2 = 605.56 )</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>5450.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>N – 1</td>
<td>( \sum_{i=1}^{K} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2 )</td>
<td>( \sum_{i=1}^{K} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2 )</td>
<td>( \hat{\sigma}_{\text{within}}^2 = 605.56 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7316.67</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) degrees of freedom

p-value = \( \Pr [F_{DF=2,9} \geq 1.54] = .27 \)

**Conclusion.** The null hypothesis is not rejected. These data do not provide statistically significant evidence that effectiveness of treatment is different, depending on the type of treatment received (“1” versus “2” versus “3”).
**Step 5: Don’t forget to look at your data!**

![Graph illustrating the response to treatment among physical therapy patients, by treatment regime.](image)

**Response to Treatment Among Physical Therapy Patients, by Treatment Regime (N=12)**

<table>
<thead>
<tr>
<th>Treatment</th>
<th>n</th>
<th>Average</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>70</td>
<td>13.54</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>50</td>
<td>28.87</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>80</td>
<td>28.28</td>
</tr>
</tbody>
</table>

*a One way analysis of variance, p=.27; Bartlett test of equality of variance, p=.46*

**Stata command**

- `.dotplot score, over(txgroup) center nogroup nx(30)`
- `.graph box score, over(txgroup)`

**Notes**

- For small samples, try a side-by-side dot plot using the Stata command `dotplot`. For larger samples, do a side-by-side box and whisker plot using the command `graph box`.

- Notice that the scatter of the data seems to differ among the groups (in particular, the magnitude of the scatter is noticeably smaller among patients receiving treatment #1). The small sample sizes may explain its lack of statistical significance.

- Also because of small sample size (possibly), the discrepancy in mean responses (70 versus 50 versus 80) also fails to reach statistical significance.
4. Checking Assumptions of the One Way Analysis of Variance

a. Tests for Homogeneity of Variance

Violation of the assumption of homogeneity of variances is sometimes, *but not always*, a problem.
- Recall that the denominator of the overall F statistic in a one way analysis of variance is the “within group mean square.” It is a weighted average of the separate within group variance estimates $S^2$ and, as such, is an estimate of the assumed common variance.
  - When the within group variance parameters $\sigma_1^2, \sigma_2^2, ..., \sigma_K^2$ are at least reasonably similar, then the within group mean square is a good summary of the within group variability.
- The overall F test for equality of means in an analysis of variance is *reasonably robust* to moderate violations of the assumption of homogeneity of variance.
- However, pairwise t-tests and hypothesis tests of contrasts are not robust to violations of homogeneity of variance, as when the $\sigma_1^2, \sigma_2^2, ..., \sigma_K^2$ are very unequal.

Tests of homogeneity of variance are appropriate for the one way analysis of variance only. There are a variety of tests available.

- **F Test for Equality of Two Variances** – This was introduced in PubHlth540 Unit 7, Section 12 *Hypothesis Testing* (See p 50 under [http://www-unix.oit.umass.edu/~biep540w/pdf/testing.pdf](http://www-unix.oit.umass.edu/~biep540w/pdf/testing.pdf))
- **Bartlett’s test** - This test has high statistical power when the assumption of normality is met. However, it is very sensitive to the assumption of normality.
- **Levene’s test** – This test has the advantage of being much less sensitive to violations of normality. Its disadvantage is that it has less power than Bartlett’s test.
- **Brown-Forsythe test, also called Levene (med) test** – Similar to Levene’s Test.
**Bartlett’s Test**

- \( H_0: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_K^2 \)
- \( H_A: \) At least one \( \sigma_i^2 \) is unequal to the others

- Obtain the \( K \) separate sample variances \( S_1^2, \ldots, S_K^2 \)

- Obtain \( \hat{\sigma}_{within}^2 = \frac{\sum_{i=1}^{K} (n_i - 1)S_i^2}{\sum_{i=1}^{K} (n_i - 1)} \) the estimate of the (null hypothesis) common \( \sigma^2 \)

- Compute \( B = [\ln(\hat{\sigma}_{within}^2)] \left( \sum_{i=1}^{K} (n_i - 1) \right) - \sum_{i=1}^{K} (n_i - 1) \ln(S_i^2) \) \( \text{note – Some texts use this as the test.} \)

- Compute \( C = \left( \frac{1}{3(K-1)} \right) \left( \sum_{i=1}^{K} \frac{1}{(n_i - 1)} \right) - \frac{1}{\sum_{i=1}^{K} (n_i - 1)} \) \( \text{note – This is a correction factor} \)

- Compute \( \text{Bartlett Test Statistic} = \frac{B}{C} \) \( \text{note – the distribution of B/C is better approximated by chi square.} \)

- When the null hypothesis is true, Bartlett Test Statistic is distributed Chi square (df=K-1)
  - Reject null for large values of Bartlett Test

**Stata illustration**
Bartlett’s test of equal variance is provided with results of command `oneway`

```
.* oneway yvariable groupvariable
.oneway score txgroup
```
Levene’s Test (“Dispersion variable Analysis of Variance”)

The idea of Levene’s test and its modification is to create a new random variable, a *dispersion random variable* which we will represent as *d* and that is a measure of how the variances are different. Levene’s test (and its modifications) are a one way analysis of variance on a dispersion random variable *d*.

- **H₀**: \( \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 \)
- **H₁**: At least one \( \sigma_i^2 \) is unequal to the others

- Compute \( d_{ij} = |X_{ij} - \bar{X}_i| \)

- Perform a one way analysis of variance of the \( d_{ij} \)

- When the null hypothesis is true, the Levene Test One Way Analysis of Variance is distributed *F* (numerator df = \( K-1 \), denominator df = \( N-K \))
  - Reject null for large values of Levene Test One Way Anova *F*

**Stata illustration**

Levene test is `W0` in results of command `robvar`

```stata
. * robvar yvariable, by(groupvariable)
. robvar score, by(txgroup)
```

**Brown and Forsythe Modification of Levene’s Test (“Dispersion variable Analysis of Variance”)**

The Brown and Forsythe modification of Levene’s test utilizes as its dispersion random variable *d* the absolute deviation from the group median.

- **H₀**: \( \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 \)
- **H₁**: At least one \( \sigma_i^2 \) is unequal to the others

- Compute \( \tilde{d}_{ij} = |X_{ij} - \text{median}(X_{i1}, X_{i2}, \ldots, X_{in_i})| \)

- Perform a one way analysis of variance of the \( \tilde{d}_{ij} \)
• When the null hypothesis is true, the Brown and Forsythe One Way Analysis of Variance is distributed F (numerator df = K - 1, denominator df = N - K)
  ▪ Reject null for large values of Brown and Forsythe Test One Way Anova F

**Stata illustration**
Brown-Forsythe test is **w50** in results of command **robvar**

```
.* robvar yvariable, by groupvariable)
. robvar score, by(txgroup)
```

**Statistics in Practice: Guidelines for Assessing Homogeneity of Variance**

• **Look at the variances (or standard deviations) for each group first!**
  ▪ Compare the numeric values of the variances (or standard deviations)
    
    *If the ratio of the standard deviations is less than 3 (or so), it’s okay not to worry about homogeneity of variances*
  ▪ Construct a side-by-side box plot of the data and have a look at the sizes of the boxes.

• **Second, assess the reasonableness of the normality assumption.** This is important to the validity of the tests of homogeneity of variances.

• **Levene’s test** of equality of variances is least affected by non-normality; it’s a good choice.

• **Bartlett’s test** should be used with caution, given its sensitivity to violations of normality.
b. Graphical Assessments and Tests of Normality

Graphical assessments and tests of normality were introduced previously. See again PubHlth 640 Unit 2, Regression & Correlation, Section 5b, pp 75-77.

Analysis of Variance methods are reasonably robust to violation of the assumption of normality in analysis of variance.

- While, strictly speaking, the assumption of normality in a one way analysis of variance is that within each of the K groups the individual $X_{ij}$ for $j = 1, 2, \ldots, n_i$ are assumed to be a simple random sample from a $\text{Normal}(\mu_i, \sigma^2)$,

- The analysis of variance is actually relying on normality of the sampling distribution of the means $\bar{X}_i$ for $i = 1, 2, \ldots, K$. This is great because we can appeal to the central limit theorem.

- Thus, provided that the sample sizes in each group are reasonable (say 20-30 or more), and provided the underlying distributions are not too too different from normality, then the analysis of variance is reasonably robust to violation of the assumption of normality.

As with normal theory regression, assessments of normality are of two types in analysis of variance.

- Preliminary is to calculate the residuals:
  - For $X_{ij} =$ observation for “$j$”th person in group=$i$
  - Residual $r_{ij} = (X_{ij} - \bar{X}_i)$ difference between observed and mean for group

- 1. Graphical Assessments of the distribution of the residuals:
  - Dot plots with overlay normal
  - Quantile-Quantile plots using referent = normal

- 2. Numerical Assessments:
  - Calculation of skewness and kurtosis statistics
  - Shapiro Wilk test of normality
  - Kolmogorov Smirnov/Lillefors tests of normality
  - Anderson Darling/Cramer von Mises tests of normality
Statistics in Practice: Guidelines for the Handling of Violations of Normality

- **Small sample size setting (within group sample size n < 20, approximately):**
  - Replace the normal theory one way analysis of variance with a Kruskal Wallis nonparametric one way analysis of variance.  **This will be introduced in Topic 9, Nonparametrics**

- **Large sample size setting:**
  If you *really must*, consider a normalizing data transformation. Possible transformations include the following:

  1. **Logarithmic Transformation:** \[ X^* = \ln(X+1) \] helps positive skewness

  2. **Square Root Transformation:** \[ X^* = \sqrt{X + 0.5} \] helps heteroscedasticity

  3. **Arcsine Transformation:** \[ p^* = \arcsine \sqrt{p} \] for outcome 0 to 100 percentage

  4. If your data are actual proportions of the type \( X/n \) and you have \( X \) and \( n \) consider **Anscombe Arcsine Transformation**:

\[
p^* = \arcsine \sqrt{\frac{3}{8} \left( \frac{X}{n} + \frac{3}{4} \right)}
\]
5. Introduction to More Complicated Designs

So far we have considered just one analysis of variance design: the one way analysis of variance design.

- 2 or more groups were compared (e.g. – we compared 4 groups)
- However, the groups represented levels of just 1 factor (e.g. race/ethnicity)

Analysis of variance methods can become more complicated for a variety of reasons, including but not limited to the following-

(1) The model may have more terms, including interactions and/or adjustment for confounding – Fitting and interpretation become more challenging.

(2) One or more of the terms in the model might be measured with error instead of being fixed – Estimates of variance, their interpretation and confidence interval construction are more involved.

(3) The partitioning of total variability might not be as straightforward as what we have seen so far – Understanding and working with analysis of variance tables and, especially, knowing which F test to use, can be hard.

a. Balanced versus Unbalanced

The distinction pertains to the partitioning of the total variability and, specifically, the complexity involved in the variance components and their estimation.

BALANCED

The sample size in each cell is the same.

Equality of sample size makes the analysis easier.

Specifically, the partitioning of SSQ is straightforward

A 2 way balanced anova with n=1 is called the randomized block design
UNBALANCED

The sample sizes in the cells are different.

The partitioning of SSQ is no longer straightforward.

Here, a regression approach (reference cell coding) is sometimes easier to follow

b. Fixed versus Random Effects

“Fixed” versus “random” is more complicated than you might think.

- The distinction has to do with how the inferences will be used
- The formal analysis of variance is largely unchanged.

There exist a number of definitions of “fixed” versus “random” effects. Among them are the following.

- **Fixed effects** are levels of effects chosen by the investigator, whereas **Random effects** are selected at random from a larger population

- **Fixed effects** are either (1) levels chosen by the investigator or (2) all the levels possible
  - **Random effects** are a random sample from some universe of all possible levels.

- **Fixed effects** are effects that are interesting in themselves, whereas
  Effects are investigated as **random** if the underlying population is of interest.
Examples

- **Examples of Fixed Effects** – (1) Treatment (medicine v surgery) (2) Gender (all genders)

- **Examples of Random Effects** – (1) litter of animals (this is an example of a random block) (2) interviewer in a data collection setting where there might be multiple interviewers or raters.

Illustration -

In “fixed” versus “random”, the ways we think about the null hypothesis are slightly different. Consider a one way anova analysis which explores variations in SAT scores, depending on the university affiliation of the students.

- Outcome is $X_{ij} = \text{SAT score for “}j\text{”th individual at University “}i\text{”}$

- Factor is University with

  $i = 1$ if University is Massachusetts  
  $2$ if University is Wisconsin  
  $3$ if University is Alaska

- Subscript “$j$” indexes student within the University

  - **FIXED effects perspective**

    Interest pertains only to the 3 Universities (MA, WI, and AK)

    $H_0: \mu_{\text{Massachusetts}} = \mu_{\text{Wisconsin}} = \mu_{\text{Alaska}}$

  - **RANDOM effects perspective**

    Massachusetts, Wisconsin, and Alaska are a random sample from the population of universities in the US.

    $H_0$: Mean SAT scores are equal at all American Universities
c. Factorial versus Nested

The distinction “factorial” versus “nested” is an important distinction pertaining to the discovery of interaction or effect modification (factorial) versus control for confounding (nested).

Consider the context of a two way analysis of variance that explores Factor A at “a” levels and Factor B at “b” levels.

**FACTORIAL**

All combinations of factor A and factor B are investigated.

Factorial design permits investigation of A x B interaction

Thus, good for exploration of effect modification, synergism, etc.

Frequently used in public health, observational epidemiology

**Example**
Factor A = Plant at a=3 levels and Factor B = CO2 at b=2 levels

Note: All (3)(2) = 6 combinations are included

<table>
<thead>
<tr>
<th>Pea plant</th>
<th>Ambient CO₂</th>
<th>Double CO₂</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>![Tree Image]</td>
<td>![Tree Image]</td>
</tr>
<tr>
<td>Bean plant</td>
<td>![Tree Image]</td>
<td>![Tree Image]</td>
</tr>
<tr>
<td>Corn plant</td>
<td>![Tree Image]</td>
<td>![Tree Image]</td>
</tr>
</tbody>
</table>

Nature | Population/ Sample | Observation/ Data | Relationships/ Modeling | Analysis/ Synthesis
NESTED

The levels of the second factor are nested in the first.

*Confounding (by Factor A) of the Factor B- Outcome relationship* is controlled through use of stratification on Factor A.

Familiar examples are *hierarchical, split plot, repeated measures, mixed models*.

Nested designs are frequently used in biology, psychology, and complex survey methodologies.

Factor A, the stratifying variable, is sometimes called the “primary sampling unit.”

**Example**
Factor A = Trees at 3 levels and Factor B = Leaf at 5 levels, nested

<table>
<thead>
<tr>
<th>Trees</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leaves</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>Leaves</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
</tbody>
</table>
6. Some Other Analysis of Variance Designs

What design should you use?
In brief, the answer depends on (1) the research question (2) knowledge of underlying biology and, specifically, knowledge of external influences that might be effect modifying or confounding or both and (3) availability of sample size.

Briefly, three other analysis of variance designs are introduced here

a. The Randomized Complete Block Design  
b. Two Way Fixed Effects Analysis of Variance – Equal cell numbers  
c. The Two Way Hierarchical or Nested Design

a. The Randomized Complete Block Design

Randomized Complete Block Design Model

Setting:
I groups or treatments indexed i = 1, 2, …, I  
J blocks indexed j = 1, 2, …., J  
Sample size is 1 in each treatment x block combination  

\[ X_{ij} = \text{Observation for the one individual in the } j\text{th block of the } i\text{th group/treatment} \]

The randomized complete block design model of \( X_{ij} \) is defined as follows:

\[ X_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \]

where

\[ \mu = \text{grand mean} \]

\[ \alpha_i = [\mu_i - \mu] \text{ and } \sum_{i=1}^{I} \alpha_i = 0 \]

\[ \beta_j = [\mu_j - \mu] \text{ and } \sum_{j=1}^{J} \beta_j = 0 \]

and \( \epsilon_{ij} \) is random error distributed Normal(0,\( \sigma^2 \))
Example -

An investigator wishes to compare 3 treatments for HIV disease. However, it is suspected that response to treatment might be confounded by baseline cd4 count. The investigator seeks to control (better yet, eliminate) confounding. To accomplish this, consenting subjects are grouped into 8 “homogeneous” blocks according to cd4 count. Within each block, baseline cd4 counts are assumed to be similar.

Within each block, there are 3 subjects, one per treatment. Assignment of subject to treatment within each block is randomized.

### Data Layout –

<table>
<thead>
<tr>
<th>Patient</th>
<th>Block is stratum of cd4 count</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1st</td>
<td>Drug 3</td>
</tr>
<tr>
<td>2nd</td>
<td>Drug 1</td>
</tr>
<tr>
<td>3rd</td>
<td>Drug 2</td>
</tr>
</tbody>
</table>

While there are two factors ..

- Only one factor is of interest – Treatment (drug) at 3 levels.
- The other factor is called a blocking factor; its influence on outcome is not of interest. But we do want to control for its possible confounding effect.

A characteristic of a randomized complete block design is that the sample size is ONE in each Treatment x Block combination.
Randomized Complete Block Design Analysis of Variance

“i” indexes treatment, i=1 …. I
“j” indexes block, j = 1 …. J

\[ \mu_{ij} = \text{Mean [ Outcome ] for drug “i” in block “j”} \]

\[
i = 1 \ldots I \quad \text{In this example, I = 3}
\]
\[
j = 1, 2, \ldots, J \quad \text{In this example, J = 8}
\]

\[
E[X_{ij}] = \mu_{ij} = \mu_i + \alpha_i + \beta_j \quad \text{where}
\]

\[
X_{ij} = \mu_i + \alpha_i + \beta_j + \varepsilon_{ij}
\]

\[
\begin{align*}
(1) \quad \alpha_i &= \mu_i - \mu \\ &\quad \text{and} \quad \sum_{i=1}^{I} \alpha_i = 0
\end{align*}
\]

\[
(2) \quad \beta_j &= \mu_j - \mu \\ &\quad \text{and} \quad \sum_{j=1}^{J} \beta_j = 0
\]

\[
X_{ij} = \bar{X}_i + [\bar{X}_i - \bar{X}_.] + [\bar{X}_j - \bar{X}_.] + [X_{ij} - \bar{X}_i - \bar{X}_j + \bar{X}_.]
\]

**algebraic identity**

- \( \varepsilon_{ij} \) is assumed distributed Normal(0, \( \sigma^2 \))

- Because \( n=1 \) in each cell, block x treatment interactions, if they exist, cannot be estimated.
Total SSQ and its Partitioning

\[ X_{ij} = \bar{X}_{..} + [\bar{X}_i - \bar{X}_{..}] + [\bar{X}_j - \bar{X}_{..}] + [X_{ij} - \bar{X}_i - \bar{X}_j + \bar{X}_{..}] \rightarrow \]

\[ [X_{ij} - \bar{X}_{..}] = [\bar{X}_i - \bar{X}_{..}] + [\bar{X}_j - \bar{X}_{..}] + [X_{ij} - \bar{X}_i - \bar{X}_j + \bar{X}_{..}] . \]

Squaring both sides and summing over all observations yields (because the cross product terms sum to zero!)

\[ \sum_{i=1}^{I} \sum_{j=1}^{J} [X_{ij} - \bar{X}_{..}]^2 \]

\[ = \sum_{i=1}^{I} \sum_{j=1}^{J} [\bar{X}_i - \bar{X}_{..}]^2 + \sum_{i=1}^{I} \sum_{j=1}^{J} [\bar{X}_j - \bar{X}_{..}]^2 + \sum_{i=1}^{I} \sum_{j=1}^{J} [X_{ij} - \bar{X}_i - \bar{X}_j + \bar{X}_{..}]^2 \]

\[ = I \sum_{i=1}^{I} [\bar{X}_i - \bar{X}_{..}]^2 + J \sum_{j=1}^{J} [\bar{X}_j - \bar{X}_{..}]^2 + I \sum_{j=1}^{J} [\bar{X}_i - \bar{X}_j + \bar{X}_{..}]^2 \]

**Analysis of Variance Table**

<table>
<thead>
<tr>
<th>Source</th>
<th>df(^a)</th>
<th>Sum of Squares</th>
<th>E (Mean Square)</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Due <strong>treatment</strong></td>
<td>(I-1)</td>
<td>J ( \sum_{i=1}^{I} (\bar{X}<em>i - \bar{X}</em>{..})^2 )</td>
<td>( \sigma^2 + J \left[ \sum_{i=1}^{I} \alpha_i^2 \right] ) ( (I-1) )</td>
<td>F = MSQ(<em>{treatment} ) / MSQ(</em>{residual} )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>df = (I-1), (I-1)(J-1)</td>
</tr>
<tr>
<td>Due <strong>block</strong></td>
<td>(J-1)</td>
<td>J ( \sum_{j=1}^{J} (\bar{X}<em>j - \bar{X}</em>{..})^2 )</td>
<td>( \sigma^2 + I \left[ \sum_{j=1}^{J} \beta_j^2 \right] ) ( (J-1) )</td>
<td>F = MSQ(<em>{block} ) / MSQ(</em>{residual} )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>df = (J-1), (I-1)(J-1)</td>
</tr>
<tr>
<td>Residual</td>
<td>(I-1)(J-1)</td>
<td>( \sum_{i=1}^{I} \sum_{j=1}^{J} (X_{ij} - \bar{X}_i - \bar{X}<em>j + \bar{X}</em>{..})^2 )</td>
<td>( \sigma^2 )</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>IJ - 1</td>
<td>( \sum_{i=1}^{I} \sum_{j=1}^{J} (X_{ij} - \bar{X}_{..})^2 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) degrees of freedom
b. The Two Way Fixed Effects Analysis of Variance

---

Two Way Analysis of Variance

Fixed Effects Model

Setting:
- I levels of Factor #1 and are indexed i = 1, 2, ..., I
- J levels of Factor #2 and are indexed j = 1, 2, ..., J
The sample size for the group defined by Factor 1 at level “i” and Factor 2 at level “j” is n_{ij}
“k” indexes the kth observation in the “ij”th group
X_{ijk} = Observation for the kth individual in the jth block of the ith group/treatment

The two way analysis of variance fixed effects model of X_{ijk} is defined as follows:

\[ X_{ijk} = \mu + \alpha_i + \beta_j + (\alpha \beta)_{ij} + \varepsilon_{ijk} \]

where
- \( \mu \) = grand mean
- \( \alpha_i = [\mu_i - \mu]\) and \( \sum_{i=1}^{I} \alpha_i = 0 \)
- \( \beta_j = [\mu_j - \mu]\) and \( \sum_{j=1}^{J} \beta_j = 0 \)
- \( (\alpha \beta)_{ij} = \mu + [\mu_i - \mu] + [\mu_j - \mu] + [\mu_{ij} - (\mu_i - \mu) - (\mu_j - \mu) + \mu] \)
- \( \sum_{i=1}^{I} (\alpha \beta)_{ij} = 0 \) and \( \sum_{j=1}^{J} (\alpha \beta)_{ij} = 0 \)

and \( \varepsilon_{ij} \) is random error distributed Normal(0,\( \sigma^2 \))
Example -
An investigator wishes compare the fish growth means among 6 groups in a 2 way factorial design factor #1 being light at 2 levels and factor #2 being water temperature at 3 levels.

Two Way Fixed Effects Analysis of Variance

Let \( i \) index light, \( i = 1, 2 \).
\( j \) index water temperature, \( j = 1, 2, 3 \).
\( k \) index individual in the \((i, j)^{th}\) group, \( k = 1, \ldots, n_{ij} \).

\[ \mu_{ij} = \text{Expected mean growth at 6 weeks for fish raised under conditions of light = “i” and water temperature = “j”}. \]

\[ \mu_{ij} = \mu + [\mu_i - \mu] + [\mu_j - \mu] + [\mu_{ij} - (\mu_i - \mu) - (\mu_j - \mu) + \mu] \]

\[ \mu = \text{Overall population mean} \]

\( \beta_i = [\mu_i - \mu] \) is the light effect. It is estimated by

\[ \beta_i = \frac{\bar{X}_{.i} - \bar{X}_{..}}{n} \]

\( \tau_j = [\mu_j - \mu] \) is the water temperature effect. It is estimated by

\[ \tau_j = \frac{\bar{X}_{.j} - \bar{X}_{..}}{n} \]

\( (\beta \tau)_{ij} = [\mu_{ij} - (\mu_i - \mu) - (\mu_j - \mu) + \mu] \) is the extra, joint, effect of the \( i^{th} \) light level and \( j^{th} \) water temperature. It is estimated by

\[ (\beta \tau)_{ij} = \frac{\bar{X}_{ij} - (\bar{X}_{.i} - \bar{X}_{..}) - (\bar{X}_{.j} - \bar{X}_{..}) + \bar{X}_{..}}{n} \]

Thus, an individual response \( X_{ij} \) is modeled

\[ X_{ijk} = \mu_{ij} + \epsilon_{ijk} \]

\[ = \mu + [\mu_i - \mu] + [\mu_j - \mu] + [\mu_{ij} - (\mu_i - \mu) - (\mu_j - \mu) + \mu] + \epsilon_{ijk} \]

\[ = \mu + \alpha_i + \beta_j + (\alpha \beta)_{ij} + \epsilon_{ijk} \]
Assumptions

- $\sum_{i=0}^{1} \alpha_i = 0$
- $\sum_{j=1}^{J} \beta_j = 0$
- $\sum_{i=1}^{I} (\alpha \beta)_{ij} = 0$
- $\sum_{j=1}^{J} (\alpha \beta)_{ij} = 0$

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>df(^a)</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Due light (I-1)</td>
<td>(I-1)</td>
<td>$\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n} (\bar{X}<em>{ij} - \bar{X}</em>{..})^2$</td>
<td>$\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n} (\bar{X}<em>{ij} - \bar{X}</em>{..})^2/(I-1)$</td>
<td>$F = \frac{\hat{\sigma}_{\text{light}}^2}{\hat{\sigma}^2}$</td>
</tr>
<tr>
<td>Due temperature (J-1)</td>
<td>(J-1)</td>
<td>$\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n} (\bar{X}<em>{..} - \bar{X}</em>{ij})^2$</td>
<td>$\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n} (\bar{X}<em>{..} - \bar{X}</em>{ij})^2/(J-1)$</td>
<td>$F = \frac{\hat{\sigma}_{\text{temp}}^2}{\hat{\sigma}^2}$</td>
</tr>
<tr>
<td>Due interaction (I-1)(J-1)</td>
<td>(I-1)(J-1))</td>
<td>$\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n} (\bar{X}<em>{ij} - \bar{X}</em>{..} - \bar{X}_{ij} + \bar{X})^2$</td>
<td>$\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n} (\bar{X}<em>{ij} - \bar{X}</em>{..} - \bar{X}_{ij} + \bar{X})^2/(I-1)(J-1)$</td>
<td>$F = \frac{\hat{\sigma}_{\text{light*temp}}^2}{\hat{\sigma}^2}$</td>
</tr>
<tr>
<td>Within Groups</td>
<td></td>
<td>$\sum_{i=1}^{I} \sum_{j=1}^{J} (n_{ij} - 1)$</td>
<td>$\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n} (X_{ij} - \bar{X}_{ij})^2$</td>
<td>$\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n} (X_{ij} - \bar{X}_{ij})^2$</td>
</tr>
<tr>
<td>Total</td>
<td>(N-1)</td>
<td>$\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n} (X_{ijk} - \bar{X}_{..})^2$</td>
<td>$\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n} (X_{ijk} - \bar{X}_{..})^2$</td>
<td>$\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n} (X_{ijk} - \bar{X}_{..})^2$</td>
</tr>
</tbody>
</table>

\(^a\) degrees of freedom
Example, continued –
Recall that it is of interest to study the effect of water temperature and light on fish growth.

- Outcome $X =$ fish growth at six weeks
- Factor I is Light at 2 levels (low and high)
- Factor II is Water Temperature at 3 levels (cold, lukewarm, warm)

Following are the data.

<table>
<thead>
<tr>
<th>Light</th>
<th>Water Temp</th>
<th>$X =$ Fish Growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>1=low</td>
<td>1=cold</td>
<td>4.55</td>
</tr>
<tr>
<td>1=low</td>
<td>1=cold</td>
<td>4.24</td>
</tr>
<tr>
<td>1=low</td>
<td>2=lukewarm</td>
<td>4.89</td>
</tr>
<tr>
<td>1=low</td>
<td>2=lukewarm</td>
<td>4.88</td>
</tr>
<tr>
<td>1=low</td>
<td>3=warm</td>
<td>5.01</td>
</tr>
<tr>
<td>1=low</td>
<td>3=warm</td>
<td>5.11</td>
</tr>
<tr>
<td>2=high</td>
<td>1=cold</td>
<td>5.55</td>
</tr>
<tr>
<td>2=high</td>
<td>1=cold</td>
<td>4.08</td>
</tr>
<tr>
<td>2=high</td>
<td>2=lukewarm</td>
<td>6.09</td>
</tr>
<tr>
<td>2=high</td>
<td>2=lukewarm</td>
<td>5.01</td>
</tr>
<tr>
<td>2=high</td>
<td>3=warm</td>
<td>7.01</td>
</tr>
<tr>
<td>2=high</td>
<td>3=warm</td>
<td>6.92</td>
</tr>
</tbody>
</table>

There are 3 analysis questions and, thus, three null hypotheses of interest:

(1) $H_o$: No effect due to light,
e.g. mean length is the same over the two levels of light

(2) $H_o$: No effect due to temperature
e.g. mean length is the same over the three levels of temperature

(3) $H_o$: Not a differential effect
of one treatment over the levels of the other (e.g. no interaction)
A meaningful analysis might proceed in this order.

**Step 1. Test for no interaction**

- If there is interaction, this means that the effect of light on growth depends on the water temperature and vice versa.

- Accordingly, the meaning that can be given to an analysis of effects of light (Factor I) or an analysis of the effects of water temperature (Factor II) depend on an understanding of interaction.

- The correct F statistic is “due interaction” 
  \[ F = \frac{\hat{\sigma}^2_{\text{light*temperature}}}{\hat{\sigma}^2} \]

- Numerator df = (I-1)(J-1).

- Denominator df = \( \sum_{i=1}^{1} \sum_{j=1}^{1} (n_{ij}-1) \)

**If interaction is NOT significant**

**Step 2. Test for main effect of Factor I**

Use \[ F = \frac{\hat{\sigma}^2_{\text{factor I}}}{\hat{\sigma}^2} \]

Numerator df = (I-1).

Denominator df = \( \sum_{i=1}^{1} \sum_{j=1}^{1} (n_{ij}-1) \)

**Repeat for Factor II.**

**If interaction is SIGNIFICANT**

Analyst must decide whether there is meaning to test for main effects, as the response to one level of a factor (e.g., Factor II = water temperature) depends on the level of another factor (e.g., Factor I = level of light).

If the analyst decides to collapse the data to a one way analysis of variance, then the meaning of the analysis of a main effect is that it yields an estimate of the average main effect, taken over the levels of the other factor. This may or may not be of interest.
### Example – continued

<table>
<thead>
<tr>
<th>Source</th>
<th>df(^a)</th>
<th>Sum Squares</th>
<th>Mean Square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Due light</td>
<td>(2-1) = 1</td>
<td>2.98</td>
<td>2.98</td>
<td>(F = \frac{\hat{\sigma}<em>\text{light}^2}{\hat{\sigma}</em>\text{error}^2} = 10.39)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(P = .018)</td>
</tr>
<tr>
<td>Due water temp</td>
<td>(3-1) = 2</td>
<td>3.984</td>
<td>1.992</td>
<td>(F = \frac{\hat{\sigma}<em>\text{temp}^2}{\hat{\sigma}</em>\text{error}^2} = 6.95)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(p = .027)</td>
</tr>
<tr>
<td>Due interaction</td>
<td>(I-1)(J-1) = 2</td>
<td>1.268</td>
<td>0.634</td>
<td>(F = \frac{\hat{\sigma}<em>\text{light*temp}^2}{\hat{\sigma}</em>\text{error}^2} = 2.21)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(p = .191)</td>
</tr>
<tr>
<td>Error</td>
<td>(\sum_{i=1}^{I} \sum_{j=1}^{J} (n_{ij} - 1))</td>
<td>1.721</td>
<td>0.2868</td>
<td>(\hat{\sigma}_\text{error}^2)</td>
</tr>
<tr>
<td>Total</td>
<td>(N - 1 = 11)</td>
<td><strong>9.953</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) degrees of freedom
The analysis of variance table tells us the following.
1. Fail to reject the hypothesis of no interaction. Thus we can test for main effects using the MSE.
2. Reject the null hypothesis of no temperature effect.
3. Reject the null hypothesis of no light effect

Don’t forget to look at your data!

Stata Illustration:
```
. use "http://people.umass.edu/biep640w/datasets/unit7_page48.dta"
. label define tempf 1 "1=cold" 2 "2=lukewarm" 3 "3=warm"
. label values Temp tempf
. label define lightf 1 "1=low" 2 "2=high"
. label values Light lightf
. label var growth "Growth at 6 Weeks"
. sort temp
. graph box growth, over(temp)
. sort light
. graph box growth, over(light)
```
c. The Two Way Hierarchical (Nested) Design

The Two Way Hierarchical (Nested) Design Model

Setting:
- \( I \) groups or treatments indexed \( i = 1, 2, ..., I \)
- \( J \) primary sampling units nested within each treatment and indexed \( j=1, 2, \ldots, J \)
- Sample size is \( n \) in each treatment x block combination; these are secondary sampling units
- \( k \) indexes the secondary sampling units and are indexed \( k = 1, 2, \ldots, n \)
- \( X_{ijk} \) = Observation for the \( k^{th} \) secondary sampling unit of the \( j^{th} \) primary sampling unit in the \( i^{th} \) group/treatment

The two way hierarchical (nested) design model of \( X_{ijk} \) is defined as follows:

\[
X_{ijk} = \mu + \alpha_i + b_{(i)j} + \epsilon_{(ij)k}
\]

where
- \( \mu \) = grand mean
- \( \alpha_i = [\mu_i - \mu] \) and \( \sum_{i=1}^{I} \alpha_i = 0 \)

and
- \( b_{(ij)} \) is random and distributed Normal\((0, \sigma_b^2)\)
- \( \epsilon_{ij} \) is random error distributed Normal\((0, \sigma^2)\)
- \( b_{(ij)} \) and \( \epsilon_{(ij)k} \) are mutually independent

Example -
An investigator wishes to compare 3 sprays applied to leaves on trees. Each treatment is applied to 6 leaves of 4 trees. Thus, the total number of observations is \((3 \text{ treatments}) \times (4 \text{ trees}) \times (6 \text{ leaves}) = 72\)

In an example such as this, sampling is done in multiple stages. Here – (1) In stage 1, a random sample of trees is selected (thus “tree” is the primary sampling unit) and (2) in stage 2, within each tree, a random sample of leaves is selected for measurement.
In a two way hierarchical (nested) design,

\[ I = \# \text{ treatments.} \quad \text{In this example, } I=3 \]
\[ J = \# \text{ primary sampling units} \quad \text{In this example, } J=4 \]
\[ n = \# \text{ secondary sampling units, nested} \quad \text{In this example, } n=6 \]

- Only one factor is of interest – Treatment (spray) at 3 levels.
- The effect of tree is random and is \textit{not} of interest.
- Similarly, the effect of leaf is random and is \textit{not} of interest.

\textbf{Two Way Hierarchical (Nested) Analysis of Variance}

The nested model, for the reason of having \textit{random effects}, looks a little different from a fixed effects model.

\[ \mu_i = \text{Mean } \{ \text{Outcome} \} \text{ for spray } \text{“}i\text{”} \]
\[ = \mu + \alpha_i \]

\[ X_{ijk} = \mu + \alpha_i + b_{(ij)} + \varepsilon_{(ijk)} \quad \text{where} \]

- The parenthesis notation “(i)j” tells us that tree “j” is nested in spray “i”
- The parenthesis notation “(ij)k” tells us that leaf “k” is nested in the “jth” tree receiving spray “i”

\[ \alpha_i = [ \mu_i - \mu ] \quad \text{and} \quad \sum_{i=1}^{I} \alpha_i = 0 \]

\[ X_{ijk} = \bar{X}_{..} + [ \bar{X}_{i..} - \bar{X}_{..} ] + [ \bar{X}_{ij.} - \bar{X}_{i..} ] + [ X_{ijk} - \bar{X}_{ij.} ] \quad \text{algebraic identity} \]

\textbf{Assumptions}

The \( b_{(ij)} \) are independent and distributed Normal(0, \( \sigma_b^2 \))

The \( \varepsilon_{(ijk)} \) are independent and distributed Normal(0, \( \sigma_c^2 \))

The \( b_{(ij)} \) and \( \varepsilon_{(ijk)} \) are mutually independent
Total SSQ and its Partitioning

\[ X_{ijk} = \bar{X}_{..} + [\bar{X}_{i..} - \bar{X}_{..}] + [\bar{X}_{ij.} - \bar{X}_{i..}] + [X_{ijk} - \bar{X}_{ij.}] \]

\[ [X_{ijk} - \bar{X}_{..}] = [\bar{X}_{i..} - \bar{X}_{..}] + [\bar{X}_{ij.} - \bar{X}_{i..}] + [X_{ijk} - \bar{X}_{ij.}] \]

Squaring both sides and summing over all observations yields (because the cross product terms sum to zero!)

\[ \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} [X_{ijk} - \bar{X}_{..}]^2 \]

\[ = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} [\bar{X}_{i..} - \bar{X}_{..}]^2 + \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} [\bar{X}_{ij.} - \bar{X}_{i..}]^2 + \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} [X_{ijk} - \bar{X}_{ij.}]^2 \]

\[ = Jn \sum_{i=1}^{I} [\bar{X}_{i..} - \bar{X}_{..}]^2 + n \sum_{i=1}^{I} \sum_{j=1}^{J} [\bar{X}_{ij.} - \bar{X}_{i..}]^2 + \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} [X_{ijk} - \bar{X}_{ij.}]^2 \]
### Analysis of Variance Table

<table>
<thead>
<tr>
<th>Source</th>
<th>df[a]</th>
<th>Sum of Squares</th>
<th>E (Mean Square)</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Due treatment</td>
<td>(I-1)</td>
<td>$\frac{1}{Jn} \sum_{i=1}^{I} (\bar{X}<em>{i..} - \bar{X}</em>{..})^2$</td>
<td>$\sigma_c^2 + \frac{n\sigma_b^2}{I}$</td>
<td>$F = \frac{\text{MSQ}<em>{\text{treatment}}}{\text{MSQ}</em>{\text{within treatment}}}$</td>
</tr>
<tr>
<td>Within treatment Among samples</td>
<td>I (J-1)</td>
<td>$\frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{J} (\bar{X}<em>{ij.} - \bar{X}</em>{..})^2$</td>
<td>$\sigma_c^2 + \frac{n\sigma_b^2}{J}$</td>
<td>$F = \frac{\text{MSQ}<em>{\text{within treatment}}}{\text{MSQ}</em>{\text{residual}}}$</td>
</tr>
<tr>
<td>Residual</td>
<td>IJ(n-1)</td>
<td>$\frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n} (X_{ijk} - \bar{X}_{ij.})^2$</td>
<td>$\sigma_c^2$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>IJn – 1</td>
<td>$\frac{1}{IJn} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n} (X_{ijk} - \bar{X}_{..})^2$</td>
<td>$\sigma_c^2$</td>
<td></td>
</tr>
</tbody>
</table>

[a] degrees of freedom

**Note:** The correct F test for treatment has in the denominator the mean square for within treatment among samples. This can be appreciated as the correct definition by looking at the expected mean squares.
7. Introduction to Variance Components and Expected Mean Squares

For the advanced student – PubHlh 640 students can skip this section

Why? An understanding of variance components and expected mean squares guides us in understanding hypothesis testing and, in particular, F-test definition.

- When the design contains only fixed effects + measurement error, then all F-tests are correctly defined as

  \[
  F = \frac{\text{Mean square (effect of interest)}}{\text{Mean square(error)}}
  \]

- However, if the design contains any random effects, then the denominator of an F-test is not necessarily mean square (error).

- The question is then: What is the correct definition of the F-test?

Beyond the scope of this course is the algebra involved in the solution for the correct definition of an F test in a model containing one or more random effects. However, the idea is the following.

- Using a deviation from means thinking, re-express the observed outcome \(X_{ij}\) (or \(X_{ijk}\) or \(X_{(ij)}\) or whatever, depending on the design) using an algebraic identity that corresponds to the analysis of variance model and that reveals the “effects” in the model. For example –

  - **One Way Analysis of Variance**
    \[
    X_{ij} = \bar{X}_. + [\bar{X}_i - \bar{X}_.] + [\bar{X}_j - \bar{X}_.] \Rightarrow
    [X_{ij} - \bar{X}_.] = [\bar{X}_i - \bar{X}_.] + [\bar{X}_j - \bar{X}_.]
    \]

  - **Two Way Analysis of Variance**
    \[
    X_{ijk} = \bar{X}_. + [\bar{X}_i - \bar{X}_.] + [\bar{X}_j - \bar{X}_.] + [\bar{X}_k - \bar{X}_.] + [\bar{X}_{ijk} - \bar{X}_{ij}] \Rightarrow
    [X_{ijk} - \bar{X}_.] = [\bar{X}_i - \bar{X}_.] + [\bar{X}_j - \bar{X}_.] + [\bar{X}_k - \bar{X}_.] + [\bar{X}_{ijk} - \bar{X}_{ij}] + [\bar{X}_{ijk} - \bar{X}_{ij}]
    \]

  - **Randomized Complete Block Design**
    \[
    X_{ij} = \bar{X}_. + [\bar{X}_i - \bar{X}_.] + [\bar{X}_j - \bar{X}_.] + [\bar{X}_k - \bar{X}_.] + [\bar{X}_{ijk} - \bar{X}_{ij}] \Rightarrow
    [X_{ij} - \bar{X}_.] = [\bar{X}_i - \bar{X}_.] + [\bar{X}_j - \bar{X}_.] + [\bar{X}_k - \bar{X}_.] + [\bar{X}_{ijk} - \bar{X}_{ij}] + [\bar{X}_{ijk} - \bar{X}_{ij}]
    \]
• **Hierarchical or Nested Design**

\[
X_{ijk} = \overline{X}_{..} + [\overline{X}_{i..} - \overline{X}_{..}] + [\overline{X}_{.i.} - \overline{X}_{i..}] + [X_{ijk} - \overline{X}_{ij.}] 
\]

\[
[ X_{ijk} - \overline{X}_{..} ] = [ \overline{X}_{i..} - \overline{X}_{..} ] + [ \overline{X}_{.i.} - \overline{X}_{i..} ] + [ X_{ijk} - \overline{X}_{ij.} ]
\]

• Square both sides and solve for expected value – a fair amount of tedious algebra.

**How to Obtain Expected Mean Squares Without Having to do the Whole Algebraic Derivation**

**Example – Two Way Factorial Design (Balanced)** - Consider the two way factorial analysis of variance design with Factor A at “a” levels and Factor B at “b” levels and an equal number of replicates = n at each combination of Factor A x Factor B.

**Step 1 – Pretend that all the factors are random.**

Thus, associated with each random effect is a variance component.

In this scenario -

\[
X_{ijk} - \mu = a_i + b_j + (ab)_{ij} + \varepsilon_{ijk}\text{ where}
\]

The \( a_i \) are independent and distributed Normal(0, \( \sigma_a^2 \))

The \( b_j \) are independent and distributed Normal(0, \( \sigma_b^2 \))

The \( (ab)_{ij} \) are independent and distributed Normal(0, \( \sigma_{ab}^2 \))

The \( \varepsilon_{ijk} \) are independent and distributed Normal(0, \( \sigma_e^2 \))

The \( a_i, b_j, (ab)_{ij} \) and \( \varepsilon_{ijk} \) are mutually independent
Step 2 – Partition degrees of freedom, using the following rules.

- df, total “corrected” = Total sample size – 1
- df for each main effect of factor = # levels – 1
- df for each two way interaction = product of main effect df
- df for error (replication) = sum of the (within group sample size – 1)

Example, continued -
A has “a” levels, B has “b” levels, all combinations are represented, yielding ab cells, each with a “within group” sample size of n. Thus,

<table>
<thead>
<tr>
<th>Effect</th>
<th>Df</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>(a-1)</td>
</tr>
<tr>
<td>B</td>
<td>(b-1)</td>
</tr>
<tr>
<td>AB</td>
<td>(a-1)(b-1)</td>
</tr>
<tr>
<td>Error (Replication)</td>
<td>ab(n-1)</td>
</tr>
<tr>
<td>Total, “corrected”</td>
<td>(N-1) = abn-1</td>
</tr>
</tbody>
</table>

Step 3 – Construct a table shell with variance components across the top and sources of variance as rows along the side. Omit row for total “corrected”

<table>
<thead>
<tr>
<th>Source \ Variance</th>
<th>$\sigma^2_c$</th>
<th>$\sigma^2_{AB}$</th>
<th>$\sigma^2_A$</th>
<th>$\sigma^2_B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error (Replication)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Step 4 – Correct the column headings by writing in the correct coefficients of the variance components across the top using the following rules.

- $\sigma^2_c$ always has coefficient = 1
- Otherwise, the coefficients are the letters that are not in the subscripts

Example -

<table>
<thead>
<tr>
<th>Source \ Variance</th>
<th>$1 \sigma^2_c$</th>
<th>$n \sigma^2_{AB}$</th>
<th>$nb \sigma^2_A$</th>
<th>$na \sigma^2_B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error (Replication)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Step 5 – Each expected mean square will include $\sigma_e^2$

<table>
<thead>
<tr>
<th>Source \ Variance</th>
<th>1 $\sigma_e^2$</th>
<th>n $\sigma_{AB}^2$</th>
<th>nb $\sigma_A^2$</th>
<th>na $\sigma_B^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$\sigma_e^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>$\sigma_e^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AB</td>
<td>$\sigma_e^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error (Replication)</td>
<td>$\sigma_e^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Step 6 – For each “source of variation” row of your table, include in the expected mean square those column headings for which the subscript letterings include the letters of the “source”. Work from the “bottom” up.

Example -
For source “AB”, the lettering is “ab”. Include the one column heading w “ab” in the subscript

<table>
<thead>
<tr>
<th>Source \ Variance</th>
<th>1 $\sigma_e^2$</th>
<th>n $\sigma_{AB}^2$</th>
<th>nb $\sigma_A^2$</th>
<th>na $\sigma_B^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$\sigma_e^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>$\sigma_e^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AB</td>
<td>$\sigma_e^2 + n \sigma_{AB}^2$</td>
<td>$\sigma_A^2$</td>
<td>na $\sigma_B^2$</td>
<td></td>
</tr>
<tr>
<td>Error (Replication)</td>
<td>$\sigma_e^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For source “B”, the lettering is “b”. Include the two column headings w “b” in the subscript

<table>
<thead>
<tr>
<th>Source \ Variance</th>
<th>1 $\sigma_e^2$</th>
<th>n $\sigma_{AB}^2$</th>
<th>nb $\sigma_A^2$</th>
<th>na $\sigma_B^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$\sigma_e^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>$\sigma_e^2 + n \sigma_{AB}^2 + na \sigma_B^2$</td>
<td>$\sigma_A^2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AB</td>
<td>$\sigma_e^2 + n \sigma_{AB}^2$</td>
<td>$\sigma_A^2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error (Replication)</td>
<td>$\sigma_e^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For source “A”, the lettering is “a”. Include the two column headings w “a” in the subscript

<table>
<thead>
<tr>
<th>Source \ Variance</th>
<th>1 $\sigma_e^2$</th>
<th>n $\sigma_{AB}^2$</th>
<th>nb $\sigma_A^2$</th>
<th>na $\sigma_B^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$\sigma_e^2 + n \sigma_{AB}^2 + nb \sigma_A^2$</td>
<td>$\sigma_A^2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>$\sigma_e^2 + n \sigma_{AB}^2 + na \sigma_B^2$</td>
<td>$\sigma_A^2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AB</td>
<td>$\sigma_e^2 + n \sigma_{AB}^2$</td>
<td>$\sigma_A^2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error (Replication)</td>
<td>$\sigma_e^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Step 7 – For the fixed effects, correct your expected mean squares solution that you developed under the assumption of random effects using the following approach.

The coefficients remain the same

Keep \( \sigma^2_c \) in the solution. Keep the last term in the solution. Drop from the solution all the terms in between.

Replace the \( \sigma^2 \) with \( \sum \text{something} \)

The numerator “something” is the sum of the fixed effects squared (note squaring)

The denominator “something” is the degrees of freedom

Example, continued – Suppose that, in actuality, \( A \) and \( B \) are both fixed

In step 1, we pretended that \( A \) and \( B \) were both random -

\[
X_{ijk} - \mu = a_i + b_j + (ab)_{ij} + \varepsilon_{ijk}
\]

where

The \( a_i \) are independent and distributed Normal(0, \( \sigma^2_a \))

The \( b_j \) are independent and distributed Normal(0, \( \sigma^2_b \))

The \( (ab)_{ij} \) are independent and distributed Normal(0, \( \sigma^2_{ab} \))

The \( \varepsilon_{ijk} \) are independent and distributed Normal(0, \( \sigma^2_c \))

The \( a_i, b_j, (ab)_{ij} \) and \( \varepsilon_{ijk} \) are mutually independent

Now, in step 7, write down the correct fixed effects model -

\[
X_{ijk} - \mu = \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}
\]

with

\[
\sum_{i=1}^{a} \alpha_i = 0 \quad \text{and} \quad \sum_{j=1}^{b} \beta_j = 0 \quad \text{and} \quad \sum_{i=1}^{a} (\alpha\beta)_{ij} = 0 \quad \text{and} \quad \sum_{j=1}^{b} (\alpha\beta)_{ij} = 0
\]
Expected Mean Squares for Two Way Factorial Design with equal sample size = n/cell

<table>
<thead>
<tr>
<th>Source \ Variance</th>
<th>All random Effects</th>
<th>All Fixed Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$\sigma_c^2 + n\sigma_{AB}^2 + nb\sigma_A^2 \to$</td>
<td>$\sigma_c^2 + nb\sum\sigma_{i}^2$</td>
</tr>
<tr>
<td></td>
<td>keep drop keep</td>
<td>(a-1)</td>
</tr>
<tr>
<td>B</td>
<td>$\sigma_c^2 + na\sigma_B^2 \to$</td>
<td>$\sigma_c^2 + na\sum\beta_j^2$</td>
</tr>
<tr>
<td></td>
<td>keep drop keep</td>
<td>(b-1)</td>
</tr>
<tr>
<td>AB</td>
<td>$\sigma_c^2 + n\sigma_{AB}^2 \to$</td>
<td>$\sigma_c^2 + n\sum\sum(\alpha\beta)_{ij}^2$</td>
</tr>
<tr>
<td></td>
<td>keep keep</td>
<td>(a-1)(b-1)</td>
</tr>
<tr>
<td>Error (Replication)</td>
<td>$\sigma_c^2$</td>
<td>$\sigma_c^2$</td>
</tr>
</tbody>
</table>
8. Introduction to Variance Components

How to Construct F Tests

For the advanced student – PubHlth 640 students can skip this section

The expected mean squares are useful in solving for the correct F-test in analysis of variance hypothesis testing. This section is a little “how to”.

Example –

Consider again the two way factorial analysis of variance design with Factor A at “a” levels and Factor B at “b” levels and an equal number of replicates = n at each combination of Factor A x Factor B. **Factor A is random. Factor B is fixed**

\[ X_{ijk} - \mu = a_i + \beta_j + (a\beta)_{ij} + \varepsilon_{ijk} \]

Define the F test for the null hypothesis

\[ H_0: \beta_1 = \beta_2 = \ldots = \beta_b = 0 \] (no main effect of Factor B).

Step 1 – Write down the expected mean squares.

*Note – An interaction [ random effect] x [ fixed effect] is a random effect*

<table>
<thead>
<tr>
<th>Source \ Variance</th>
<th>E(MSQ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>( \sigma_e^2 + n \sigma_{AB}^2 + nb \sigma_A^2 )</td>
</tr>
<tr>
<td>B</td>
<td>( \sigma_e^2 + n \sigma_{AB}^2 + na \sum \beta_j^2 ) ( b-1 )</td>
</tr>
<tr>
<td>AB</td>
<td>( \sigma_e^2 + n \sigma_{AB}^2 )</td>
</tr>
<tr>
<td>Error (Replication)</td>
<td>( \sigma_e^2 )</td>
</tr>
</tbody>
</table>
Step 2 – Locate in the table of expected mean squares the source for which
the expected mean square that contains the terms you want to test.

<table>
<thead>
<tr>
<th>Source \ Variance</th>
<th>E(MSQ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$\sigma_c^2 + n\sigma_{AB}^2 + nb\sigma_A^2$</td>
</tr>
<tr>
<td>B</td>
<td>$\sigma_c^2 + n\sigma_{AB}^2 + na\sum\beta_j^2/(b-1)$</td>
</tr>
<tr>
<td>AB</td>
<td>$\sigma_c^2 + n\sigma_{AB}^2$</td>
</tr>
<tr>
<td>Error (Replication)</td>
<td>$\sigma_e^2$</td>
</tr>
</tbody>
</table>

Step 3 – Consider the assumption that the null hypothesis is true. What happens
to this particular expected mean square when the null is true?.

<table>
<thead>
<tr>
<th>Source \ Variance</th>
<th>E(MSQ)</th>
<th>E(MSQ) when Null True</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$\sigma_c^2 + n\sigma_{AB}^2 + nb\sigma_A^2$</td>
<td>$\sigma_c^2 + n\sigma_{AB}^2 + nb\sigma_A^2$</td>
</tr>
<tr>
<td>B</td>
<td>$\sigma_c^2 + n\sigma_{AB}^2 + na\sum\beta_j^2/(b-1)$</td>
<td>$\sigma_c^2 + n\sigma_{AB}^2 + 0$</td>
</tr>
<tr>
<td>AB</td>
<td>$\sigma_c^2 + n\sigma_{AB}^2$</td>
<td>$\sigma_c^2 + n\sigma_{AB}^2$</td>
</tr>
<tr>
<td>Error (Replication)</td>
<td>$\sigma_e^2$</td>
<td>$\sigma_e^2$</td>
</tr>
</tbody>
</table>

Step 4 – The denominator mean square will be the mean square for the other source
that has the same expected mean square.

<table>
<thead>
<tr>
<th>Source \ Variance</th>
<th>E(MSQ) when Null True</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$\sigma_c^2 + n\sigma_{AB}^2 + nb\sigma_A^2$</td>
</tr>
<tr>
<td>B</td>
<td>$\sigma_c^2 + n\sigma_{AB}^2$</td>
</tr>
<tr>
<td>AB</td>
<td>$\sigma_c^2 + n\sigma_{AB}^2$</td>
</tr>
<tr>
<td>Error (Replication)</td>
<td>$\sigma_e^2$</td>
</tr>
</tbody>
</table>

Thus,

$$F_{(b-1);(a-1)(b-1)} = \frac{MSQ (B)}{MSQ (AB)}$$

with degrees of freedom = (b-1), (a-1)(b-1)
Appendix 1
Review: Statistical Comparison of Two Groups

Recall some tools that are related to the Normal distribution.

1) The Student’s t-Distribution
2) The Chi Square Distribution
3) The F Distribution
4) The Distribution of the Sum of Two Independent Normals

Given: We have one sample of $n_x$ independent observations from a distribution that is Normal($\mu_x, \sigma_x^2$): $X_1 \cdots X_{nx}$ and

We have a second sample of $n_y$ independent observations from a distribution that is Normal($\mu_y, \sigma_y^2$): $Y_1 \cdots Y_{ny}$

From these samples, we calculate the following:

$$\bar{X}_{nx} = \frac{1}{n_x} \sum_{i=1}^{n_x} X_i$$
$$\bar{Y}_{ny} = \frac{1}{n_y} \sum_{j=1}^{n_y} Y_j$$
$$S_x^2 = \frac{n_x}{(n_x - 1)} \sum_{i=1}^{n_x} (X_i - \bar{X})^2$$
$$S_y^2 = \frac{n_y}{(n_y - 1)} \sum_{j=1}^{n_y} (Y_j - \bar{Y})^2$$

1) The Student’s t-Distribution

- One definition of the Student’s t-distribution is in the context of “standardization.” The idea is very similar to the idea of standardization to a z-score. Here, however, the standardization formula uses $S_x^2$ instead of $\sigma_x^2$. The result looks very much like a z-score. However, it is called a t-score.

$$t_{df = n_x - 1} = \left[ \frac{(\bar{X} - \mu_x)}{S_x / \sqrt{n_x}} \right]$$ is distributed Student’s t with degrees of freedom = ($n_x - 1$)

- Similarly, we can “standardize” $\bar{Y}$ using $\mu_y$ and $S_y^2$ to obtain t-score random variables that are distributed Student’s t with degrees of freedom = ($n_y - 1$).
2) The Chi Square Distribution

- This distribution is useful in analyses of variability. For the variability in the X’s:

\[
\frac{(n_x - 1)S_x^2}{\sigma_x^2}
\]

is distributed Chi square with degrees of freedom = \((n_x - 1)\).

- Similarly, we can investigate the variability in Y.

3) The F Distribution

- The F distribution derives in several ways. We’ll consider just one.

\[
F = \frac{S_x^2 / \sigma_x^2}{S_y^2 / \sigma_y^2}
\]

is distributed F with two degree of freedom specifications.

Numerator degrees of freedom = \((n_x - 1)\)
Denominator degrees of freedom = \((n_y - 1)\)

4) The Distribution of the Sum of Two Independent Normals

- If X is distributed Normal\((\mu_x, \sigma_x^2)\) and Y is distributed Normal\((\mu_y, \sigma_y^2)\) then, under independence of X and Y

\[
(X + Y)
\]

is also distributed Normal\([ \mu_x + \mu_y, \sigma_x^2 + \sigma_y^2 ]\)

- The application of this result that we actually use here is the following.

If \(\bar{X}\) has distribution Normal\((\mu_x, \sigma_x^2 / n_x)\) and \(\bar{Y}\) has distribution Normal\((\mu_y, \sigma_y^2 / n_y)\) then, under independence

\[
(\bar{X} + \bar{Y})
\]

is also distributed Normal\([ \mu_x + \mu_y, \sigma_x^2 / n_x + \sigma_y^2 / n_y ]\)
Test of Equality of Two Independent Means, $\mu_1$ and $\mu_2$

Equal Variances ($\sigma_1^2 = \sigma_2^2 = \sigma_{\text{COMMON}}^2$)

**Example:** *(Note: These data are hypothetical.)*

Functional status scores among patients receiving zidovudine for the treatment of AIDS were compared with those of patients not receiving zidovudine. Here, we assume that the scores are normally distributed with distributions that have the same variance $\sigma^2$. However, $\sigma^2$ is unknown. Data are:

<table>
<thead>
<tr>
<th>Zidovudine</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_1 = 15$</td>
<td>$n_2 = 22$</td>
</tr>
<tr>
<td>$\bar{X}_1 = 120$</td>
<td>$\bar{X}_2 = 96$</td>
</tr>
<tr>
<td>$S_1 = 40$</td>
<td>$S_2 = 35$</td>
</tr>
</tbody>
</table>

1. **Research Question.**
   
   Do patients receiving zidovudine have higher functional status scores?

2. **Assumptions.**
   
   $\bar{X}_1$ is distributed Normal ($\mu_1$, $\sigma^2/15$) and $\bar{X}_2$ is distributed Normal ($\mu_2$, $\sigma^2/22$)

3. **$H_0$ and $H_A$.**
   
   $H_0 : \mu_1 = \mu_2$
   
   $H_A : \mu_1 > \mu_2$

4. **Test statistic.**
   
   $t_{\text{score}} = \left[ \frac{(\bar{X}_1 - \bar{X}_2) - E[(\bar{X}_1 - \bar{X}_2) | H_{o\text{true}}]}{SE[(\bar{X}_1 - \bar{X}_2) | H_{o\text{true}}]} \right]$
When $\sigma^2$ is unknown, but assumed to be the same in both groups, the standard error of $\left(\bar{X}_1 - \bar{X}_2\right)$ is estimated using the following:

$$S\hat{E}(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{S^2_{pool}}{n_1} + \frac{S^2_{pool}}{n_2}}$$

For the interested reader, here is why.

$$S\hat{E}(\bar{X}_1 - \bar{X}_2) = \sqrt{\text{var}(\bar{X}_1 - \bar{X}_2)} = \sqrt{\frac{\hat{\sigma}^2}{n_1} + \frac{\hat{\sigma}^2}{n_2}} = \sqrt{\frac{S^2_{pool}}{n_1} + \frac{S^2_{pool}}{n_2}}$$

where $S^2_{pool} = \hat{\sigma}^2$.

$S^2_{pool}$ is our best guess of the unknown $\sigma^2$. It is a weighted average of the separate $S^1_2$ and $S^2_2$:

$$S^2_{pool} = \frac{(n_1 - 1)S^2_1 + (n_2 - 1)S^2_2}{(n_1 - 1) + (n_2 - 1)}$$

The degrees of freedom (df) associated with this guess is the sum of those associated with $S^1_2$ and $S^2_2$. Thus, $df = (n_1 - 1) + (n_2 - 1)$.

For these data:

$$\hat{\sigma}^2 = S^2_{pool} = \frac{(n_1 - 1)S^2_1 + (n_2 - 1)S^2_2}{(n_1 - 1) + (n_2 - 1)} = \frac{(15 - 1)40^2 + (22 - 1)35^2}{(15 - 1) + (22 - 1)} = 1375$$

$$S\hat{E}(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{S^2_{pool}}{n_1} + \frac{S^2_{pool}}{n_2}} = \sqrt{\frac{1375}{15} + \frac{1375}{22}} = 12.42$$

Degrees of freedom = $(n_1 - 1) + (n_2 - 1) = (15 - 1) + (22 - 1) = 35$. 

The likelihood of these findings or ones more extreme under the assumption that $H_0$ is true is $p$-value $= \Pr\left[ \left( \bar{X}_1 - \bar{X}_2 \right) \geq (120 - 96) | H_{o\text{true}} \right]$.

5. Calculations.

\[
p\text{-value} = \Pr\left[ \left( \bar{X}_1 - \bar{X}_2 \right) \geq (120 - 96) \right]
= \Pr\left[ \left( \frac{\bar{X}_1 - \bar{X}_2 - E[(\bar{X}_1 - \bar{X}_2)|H_0]}{SE(\bar{X}_1 - \bar{X}_2)} \right) \geq \frac{(120 - 96) - (0)}{12.42} \right]
= \Pr\left[ t_{\text{score}} \geq 1.93 \right] \quad \text{where degrees of freedom} = 35
= .03
\]

Note: $t_{\text{score}}=1.93$ says “the observed difference in average functional status scores equal to (120-96) = 24 is 1.93 standard error units greater than the expected difference of 0 when the null hypothesis is true.”

6. “Evaluate”.

The assumption of the null hypothesis, examined in light of the observed data has led to an “unlikely” finding (a 3% chance). Specifically, the null hypothesis chances that the 15 patients in the zidovudine treated group would have a mean score that is (120-96)=24 points higher than the average of the 22 scores among the control group is 3 in 100. This is a small likelihood, small enough to prompt rejection of $H_0$.

7. Interpret.

The investigator infers a benefit of zidovudine on functional status.
Test of Equality of Two Independent Means, $\mu_1$ and $\mu_2$

Unequal Variances ($\sigma_1^2 \neq \sigma_2^2$)

The analysis is slightly different when the variances are unequal.

- The estimated SE should reflect the dissimilarity of the variances.
- With a larger # of unknowns, because we have to estimate more things, our remaining degrees of freedom will be smaller.

Example: (same example.)

Data are:

<table>
<thead>
<tr>
<th>Zidovudine</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_1 = 15$</td>
<td>$n_2 = 22$</td>
</tr>
<tr>
<td>$\bar{X}_1 = 120$</td>
<td>$\bar{X}_2 = 96$</td>
</tr>
<tr>
<td>$S_1 = 40$</td>
<td>$S_2 = 35$</td>
</tr>
</tbody>
</table>

Our test statistic is still a t-score and has the following form:

$$t_{score} = \left[ \frac{\bar{X}_1 - \bar{X}_2 - E[(\bar{X}_1 - \bar{X}_2) \mid H_{true}]}{\hat{SE}[(\bar{X}_1 - \bar{X}_2) \mid H_{true}]} \right]$$

Here, because the variances are not assumed to be equal, the standard error of $(\bar{X}_1 - \bar{X}_2)$ is estimated using the following:

$$\hat{SE}(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}.$$  For these data,

$$\hat{SE}(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{40^2}{15} + \frac{35^2}{22}} = 12.74$$
Here is why:

\[
SE(\bar{X}_1 - \bar{X}_2) = \sqrt{\text{var}(\bar{X}_1 - \bar{X}_2)} = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}.
\]

We have to modify our calculation of the degrees of freedom, however.

Degrees of freedom = \[\frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left(\frac{S_1^2}{n_1}\right)^2 + \left(\frac{S_2^2}{n_2}\right)^2}\].

In this example we get

\[
\frac{\left(\frac{40^2}{15} + \frac{35^2}{22}\right)^2}{\left(\frac{40^2}{15}\right)^2 + \left(\frac{35^2}{22}\right)^2} = 27.44 \approx 27 \text{ note – round down.}
\]

Thus,

\[
p\text{-value} = \Pr\left[(\bar{X}_1 - \bar{X}_2) \geq (120 - 96)\right]
\]

\[
= \Pr\left[\frac{\bar{X}_1 - \bar{X}_2 - E[(\bar{X}_1 - \bar{X}_2) | H_0]}{SE(\bar{X}_1 - \bar{X}_2)} \geq \frac{(120 - 96) - (0)}{12.74}\right]
\]

\[
= \Pr\left[t_{\text{score}; DF=27} \geq 1.88\right]
\]

\[= .035\]
Test of Equality of Two Proportions

A test of the equality of two independent proportions can be obtained using either a z-score or a chi square test. The z-score method is reviewed here.

- Recall - this test derives from the Binomial distribution.
- It is an application of the central limit theorem.

Example -

Consider the Needle Exchange Trial and its interim analysis. Among the preliminary aims was an analysis to identify variables that are associated with both randomization assignment and outcome. Such variables are potential confounders of response to intervention.

The literature suggests that women might respond differently to intervention than men. Therefore, an interim analysis sought to determine if there are gender differences in randomization assignment.

Among n=101 eligible and followed as of the interim analysis:

<table>
<thead>
<tr>
<th>Pharmacy Sales</th>
<th>Pharmacy Sales + Needle Exchange</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n_1 = 53 )</td>
<td>( n_2 = 48 )</td>
</tr>
<tr>
<td># women = 9 = X_1</td>
<td># women = 13 = X_2</td>
</tr>
<tr>
<td>% women = 17.0 = ( \bar{X}_1 )</td>
<td>% women = 27.1 = ( \bar{X}_2 )</td>
</tr>
</tbody>
</table>

1. Research Question.

Is the proportion of women in the pharmacy sales + needle exchange condition (27.1%) significantly greater than the proportion of women in the pharmacy sales condition (17.0), considering the enrollment sizes of 53 and 48, respectively?
2. Assumptions.

\(X_1\) is distributed Binomial \((n_1=53, \pi_1)\) and \(X_2\) is distributed Binomial \((n_2=53, \pi_2)\) where

- \(\pi_1 = \text{Proportion women in Pharmacy Sales}\)
- \(\pi_2 = \text{Proportion women in Pharmacy Sales + Needle Exchange}\)

3. \(H_0\) and \(H_A\).

\(H_0: \pi_1 = \pi_2\)
\(H_A: \pi_1 \neq \pi_2\)  (Note that this is a 2 sided alternative)

4. Test statistic.

- Instead of using \((X_1 - X_2)\), the difference in counts, the test statistic is based on the difference in the observed proportions. This is \((\bar{X}_1 - \bar{X}_2)\).

- The test statistic compares the observed value of the difference \((\bar{X}_1 - \bar{X}_2)\) to its null hypothesis expected value, and expresses this difference in SE units.

\[
z_{\text{score}} = \left[ \frac{(\bar{X}_1 - \bar{X}_2) - E[(\bar{X}_1 - \bar{X}_2) | H_{o,\text{true}}]}{SE[(\bar{X}_1 - \bar{X}_2) | H_{o,\text{true}}]} \right]
\]
Solution for null hypothesis value, \( E\left[ (\bar{X}_1 - \bar{X}_2) \mid H_0 \text{true} \right] \)

When \( H_0 \) is true, \( \pi_1 = \pi_2 = \pi \). Thus, \( (\pi_1 - \pi_2) = 0 \). Thus,

\[
E\left[ (\bar{X}_1 - \bar{X}_2) \mid H_0 \text{true} \right] = (\pi_1 - \pi_2) = 0.
\]

Solution for null hypothesis value, \( S\hat{E}\left[ (\bar{X}_1 - \bar{X}_2) \mid H_0 \text{true} \right] \)

\[
S\hat{E}\left[ (\bar{X}_1 - \bar{X}_2) \mid H_0 \right] = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n_1} + \frac{\hat{\pi}(1-\hat{\pi})}{n_2}}
\]

where \( \hat{\pi} \) is our best guess of the common \( \pi \) and is defined \( \hat{\pi} = \frac{X_1 + X_2}{n_1 + n_2} \)

Here is why,

\[
S\hat{E}(\bar{X}_1 - \bar{X}_2) = \sqrt{\text{var}(\bar{X}_1 - \bar{X}_2)} = \sqrt{\text{var}(\bar{X}_1) + \text{var}(\bar{X}_2)} = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n_1} + \frac{\hat{\pi}(1-\hat{\pi})}{n_2}}
\]

when \( H_0 \) is true.

\( \hat{\pi} \) is our best guess of the unknown \( \pi \). It is the overall proportion, regardless of group membership.

\[
\hat{\pi} = \frac{X_1 + X_2}{n_1 + n_2}
\]

For these data:

\[
\hat{\pi} = \left[ \frac{X_1 + X_2}{n_1 + n_2} \right] = \left[ \frac{9 + 13}{53 + 48} \right] = .218
\]

\[
S\hat{E}(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n_1} + \frac{\hat{\pi}(1-\hat{\pi})}{n_2}} = \sqrt{\frac{.218(1-.218)}{53} + \frac{.218(1-.218)}{48}} = .0823
\]

In the Needle Exchange Trial data, interest is in the likelihood of obtaining a magnitude of difference as great or greater than $|0.271 - 0.170| = 0.1010$ assuming the null hypothesis to be true.

The required p-value calculation is thus

\[
p-value = 2 \Pr[|\bar{X}_2 - \bar{X}_1| \geq |0.271 - 0.170|].
\]

6. Calculations.

\[
p-value = 2 \Pr[(\bar{X}_2 - \bar{X}_1) \geq (0.271 - 0.170)]
\]

\[
= 2 \Pr \left[ \frac{(\bar{X}_2 - \bar{X}_1) - E(\bar{X}_2 - \bar{X}_1)}{SE(\bar{X}_2 - \bar{X}_1)} \geq \frac{(0.271 - 0.170) - (0)}{0.0823} \right]
\]

\[
= 2 \Pr[z-score \geq 1.23] = 2[.10935]
\]

\[
= 0.22
\]

Here: \(z_{score}=1.23\) says “the observed difference in % women in the two randomization groups equal to (.271-.170) = .1010 is 1.23 standard error units greater than the expected difference of 0 when the null hypothesis is true.”

7. “Evaluate”.

Here, the assumption of the null hypothesis, examined in light of the observed data has led to a reasonably likely finding. Specifically, under the null hypothesis the chances were 22% chance of obtaining a discrepancy in the % women in the two groups equal to 10 percentage points or more.

8. Interpret.

The null hypothesis is not rejected. We conclude that there are not significant differences in the proportion of women in the two study conditions among the 101 available for interim analysis.
Appendix 2
“Reference Cell Coding” versus ”Deviation from Mean”

“Reference cell coding” and “deviation from mean” coding are two ways of expressing the same model. In analysis of variance, we use “deviation from mean”. In regression, often, we used “reference cell”. Each parameterizes group differences in means without the use of an explicit functional form (such as a linear model); that is, the change in mean response with change in group is “completely general” or “model free”.

<table>
<thead>
<tr>
<th>Deviation from mean.</th>
<th>Reference cell coding.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good when cell sample sizes are equal</td>
<td>Especially good when cell sample sizes are NOT equal</td>
</tr>
<tr>
<td>Permits straightforward partitioning of sums of squares</td>
<td>Permits straightforward interpretation of estimated betas</td>
</tr>
<tr>
<td>Good for understanding analysis of variance</td>
<td></td>
</tr>
</tbody>
</table>

Illustration for the ONE WAY Analysis of Variance

Deviation from means
\[ \mu_i = \mu_i + (\mu - \mu) \]
\[ = \mu + (\mu_i - \mu) \]
\[ = \mu + \tau_i \text{ for } i = 1 \text{ to } K \]

Reference cell coding.
\[ \mu_i = \beta_0 + \sum_{i=2}^{K} \beta_i \text{(Indicator for Group "i")} \]

Nature  Population/  Observation/  Relationships/  Analysis/  
Sample  Data  Modeling  Synthesis
Appendix 3
Multiple Comparisons Adjustment Procedures

Opinions vary on the appropriateness of *multiple comparisons adjustments*.

- It is quite correct to say that "if many tests are performed, then the chances of drawing an incorrect inference increases".

- However, it is also true that over the course of a day, or a project, or a year, or a lifetime, analysts perform many tests. *Under what circumstances of multiple hypothesis tests is it necessary to acknowledge and adjust for the performance of many statistical tests?*

- An important consideration in the performance of multiple hypothesis tests is to understand the correct probability model framework. For example - we sometimes pursue a statistical hypothesis test comparison *after* having seen the data (hence the term “data driven”) and *after* having “noticed” some group differences. This is a *post-hoc* analysis situation. The probability models which were appropriate for the overall F test may no longer be appropriate. This confuses the theory of multiple comparisons adjustments. *(For the interested reader – A more appropriate probability model framework in this situation utilizes the theory of order statistics)*

**Analysis of variance** is a setting for which the issue of multiple comparisons is transparent.

- Typically the analyst begins with a global assessment to learn if there is any discrepancy among the group means and then proceeds to identify where these discrepancies are.

- In the analysis of variance setting, it is sometimes suggested that, with respect to the issue of multiple comparisons, one should proceed as follows …

  IF $\mu_1 \approx \mu_2 \approx \ldots \approx \mu_K$, at least reasonably so, based on the overall F test,
  THEN there’s no need to compare individual means.

  **BUT … IF** at least some $\mu_i \neq \mu_i$, based on the overall F test
  **THEN** , sure, go ahead and compare means
  **AND WHEN YOU DO**, take care to adjust for multiple comparisons.

- Others have different views. This section is an overview of the commonly encountered multiple adjustments procedures.
Recall that the meaning of a type I or II error is based on what is actually true in nature:

<table>
<thead>
<tr>
<th>Type I Error (α)</th>
<th>Type II Error (β)</th>
</tr>
</thead>
<tbody>
<tr>
<td>H₀ is true We incorrectly infer Hₐ</td>
<td>Hₐ is true We incorrectly infer H₀</td>
</tr>
</tbody>
</table>

The chances of making a mistake increases as the number of tests performed increases.

The likelihood of an incorrect inference increases w # tests when null is true

- Several tests are performed, #=k
- The tests are independent
- The null hypothesis is true in every test
- Suppose each test utilizes a type I error=.05 rule. Then
  Overall Type I error
  \[ = 1 - \text{Probability} \text{ [ all accept the null ]} \]
  \[ = 1 - \Pr[\text{test #1 accepts}] \times \Pr[\text{test #2 accepts}] \times \cdots \times \Pr[\text{test #k accepts}] \]
  \[ = 1 - (.95) \times (.95) \times \cdots \times (.95) \]
  \[ = 1 - (.95)^k \]

<table>
<thead>
<tr>
<th># Independent Tests = k</th>
<th>Overall Type I Error = 1 – [.95]^k</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.05</td>
</tr>
<tr>
<td>2</td>
<td>.0975</td>
</tr>
<tr>
<td>10</td>
<td>.4013</td>
</tr>
<tr>
<td>30</td>
<td>.7854</td>
</tr>
</tbody>
</table>

And the likelihood of an incorrect inference increases w # tests when the alternative is true

IF

- Several tests are performed, #=k
- The tests are mutually independent
- The alternative hypothesis is true in every test
- Suppose that the power to detect the alternative is = .90 in every case.
- This means that each test has type II error=.10.

Then

Overall Type II error
\[ = \text{Probability} \text{ [ at least one incorrect acceptance of the null ]} \]
\[ = 1 - \text{Probability} \text{ [EVERY test correctly rejects the null]} \]
\[ = 1 - \{ \text{prob[test 1 rejects]} \times \text{prob[test 2 rejects]} \times \cdots \times \text{prob[test k rejects]} \} \]
\[ = 1 - (.90) \times (.90) \times \cdots \times (.90) \]
\[ = 1 - [.90]^k \]
### Currently available multiple comparisons procedures protect against type I error only. They afford no protection against type II error.

- **Scenario #2 receives little attention.** One reason for this is that the alternative hypothesis represents multiple possibilities, thus rendering estimation of type II error difficult. Another possible reason is the opinion that it is worse to make a type I error (rejecting the null hypothesis by mistake) than to make a type II error (retaining the null hypothesis by mistake).

- **Scenarios #1 and #2 represent extremes that are unlikely to occur in reality.** It is more often the case that the tests performed are NOT independent and that only some of the $H_A$ are true while the remainder are not.

<table>
<thead>
<tr>
<th># Independent Tests = k</th>
<th>Overall Type II Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.10</td>
</tr>
<tr>
<td>2</td>
<td>.19</td>
</tr>
<tr>
<td>10</td>
<td>.6513</td>
</tr>
<tr>
<td>30</td>
<td>.9576</td>
</tr>
</tbody>
</table>
a. LSD Procedure: A t-Test for the Comparison of Two Groups
The following is NOT a multiple comparisons adjustment procedure

“LSD” stands for “least significant difference”.

Setting
- The one way analysis of variance
  Group 1: \( X_{11}, \ldots, X_{1n_1} \) are Normal\((\mu_1, \sigma^2)\)
  Group 2: \( X_{21}, \ldots, X_{2n_2} \) are Normal\((\mu_2, \sigma^2)\)
  
  
  Group K: \( X_{11}, \ldots, X_{Kn_k} \) are Normal\((\mu_K, \sigma^2)\)
- Constant variance \( \sigma^2 \)
- The overall F-test is statistically significant
- Question - Which group means differ from the others?

The null and alternative hypotheses compare mean “i” versus mean “j”

- \( H_0: \mu_i = \mu_j \) versus \( H_A: \mu_i \neq \mu_j \)

The LSD test statistic is a t-score that utilizes a standard error that derives from the one way analysis of variance

\[
t = \frac{\left[ \bar{X}_i - \bar{X}_j \right] - [0]}{SE\left[ \bar{X}_i - \bar{X}_j \right]} \quad \text{where} \quad SE\left[ \bar{X}_i - \bar{X}_j \right] = \sqrt{\left[ \frac{\hat{\sigma}^2_{within}}{n_i} + \frac{\hat{\sigma}^2_{within}}{n_j} \right]} \quad \text{and}
\]

\[
\hat{\sigma}^2_{within} = \frac{\sum_{i=1}^{K} (n_i - 1)S_i^2}{\sum_{i=1}^{K} (n_i - 1)} \quad \text{is the estimate of the assumed common variance}
\]

\[
df = \sum_{i=1}^{K} (n_i - 1)
\]
**Assessment of statistical significance**

- Under the assumption that the null hypothesis is true, and under the assumptions of the one way analysis of variance, this t-score statistic is distributed Student’s t.
  - The degrees of freedom is \( df = \sum_{i=1}^{K} (n_i - 1) \) because we have “k” guesses of the assumed common variance, based on degrees of freedom equal to \( (n_1 - 1), (n_2 - 1), \ldots, (n_K - 1) \), respectively, and we have pooled the “k” guesses into one “weighted” estimate (with weights given by \( df \)).

**Confidence interval estimate of \( [\mu_i - \mu_j] \)**

- Point Estimate: \( \bar{X}_i - \bar{X}_j \)
- SE [ Point Estimate ]: \( SE[\bar{X}_i - \bar{X}_j] = \sqrt{\frac{\hat{\sigma}^2_{within}}{n_j} + \frac{\hat{\sigma}^2_{within}}{n_j}} \)
- Confidence. Coefficient: Percentile from Student’s t Distribution with \( df = \sum_{i=1}^{K} (n_i - 1) = N-K \)
- 95% CI is thus \( \bar{X}_i - \bar{X}_j \pm (t_{N-k; .975}) \cdot SE[\bar{X}_i - \bar{X}_j] \)
b. t-Test for a Linear Contrast
The following is also NOT a multiple comparisons adjustment procedure

The tool of linear contrast is a useful mechanism for making selected comparisons of groups.

- The investigator may not want to look at groups two at a time.
- It may instead be of interest to compare collections of groups.
- If the comparison can be defined in terms of a linear contrast, then the construction of a hypothesis test is straightforward.
  - This is because the expected value of the linear contrast under the null hypothesis is zero, and because the Student's t-distribution can be used.

The setting is the ONE WAY FIXED EFFECTS ANALYSIS of Variance.

$Y_{11}, ..., Y_{1,n_1}$ is a random sample from $\text{Normal}(\mu_1, \sigma^2)$

... $Y_{K1}, ..., Y_{K,n_K}$ is a random sample from $\text{Normal}(\mu_K, \sigma^2)$

**Definition Linear Contrast, L**

- A linear contrast is a linear combination of the $K$ sample means in which the multipliers add up to exactly zero.

\[
L = [a_1]Y_1 + [a_2]Y_2 + ... + [a_i]Y_i + ... + [a_j]Y_j + ... + [a_k]Y_K \quad \text{with}
\]

\[
[a_1] + [a_2] + ... + [a_i] + ... + [a_j] + ... + [a_k] = 0
\]

- The notation for the estimate of the linear contrast is $L$. 

### Example 1 -

- We wish to compare the "i"th and "j"th group means only.
- The required linear contrast estimator is defined:
  \[ L = [0]Y_i + [0]Y_2 + ... + [+1]Y_i + ... + [-1]Y_j + ... + [0]Y_K \]
- The sum of the multipliers add up to zero, as required:
  \[ \text{Sum of multipliers} = [0] + [0] + ... + [+1] + ... + [-1] + ... + [0] = 0 \]
- Getting rid of all the “zeroes” and simplifying reveals
  \[ L = Y_i - Y_j, \] which is the comparison we wanted in the first place.

### Example 2 -

- We wish to compare group 1 with groups 2, 3.
- E.g – Patients in group #1 received a sugar pill for their headache (control), patients in group #2 received advil (active) and patients in group #3 received aspirin (active). The investigator wants to know if placebo is significantly different in its effects than any active medication.
- An appropriate linear contrast estimator here is:
  \[ L = [+1]Y_1 + \left[ \frac{-1}{2} \right]Y_2 + \left[ \frac{-1}{2} \right]Y_3 \]
- Checking … the sum of the multipliers add up to zero:
  \[ \text{Sum of multipliers} = [+1] + \left[ \frac{-1}{2} \right] + \left[ \frac{-1}{2} \right] = 0 \]
**t-Test for Linear Contrast**

Null and Alternative Hypotheses

\[ H_0: \ E[L] = [a_1] \mu_1 + [a_2] \mu_2 + \ldots + [a_i] \mu_i + \ldots + [a_j] \mu_j + \ldots + [a_k] \mu_K = 0 \]

\[ H_A: \text{ not (two sided).} \]

The test statistic/pivotal quantity is a t-score.

\[ t_{df=(N-K)} = \left[ \frac{L - E(L \mid H_{O \text{true}})}{S\hat{E}(L \mid H_{O \text{true}})} \right] = \left[ \frac{L - 0}{S\hat{E}(L \mid H_{O \text{true}})} \right] \]

where

1. \[ L = [a_1] \bar{Y}_1 + [a_2] \bar{Y}_2 + \ldots + [a_i] \bar{Y}_i + \ldots + [a_j] \bar{Y}_j + \ldots + [a_k] \bar{Y}_K \] and

2. \[ S\hat{E}[L \mid H_{O \text{true}}] = \hat{\sigma}_{\text{within}} \sqrt{\frac{a_1^2}{n_1} + \frac{a_2^2}{n_2} + \ldots + \frac{a_K^2}{n_K}} \]

3. **Degrees of freedom** = \( df = \sum_{i=1}^{K} (n_i-1) = N-K \)
Derivation of Solution for SE:

\[ \text{S\hat{E}}[L] = \text{SE}(\ [a_1]\bar{Y}_1 + [a_2]\bar{Y}_2 + ... + [a_j]\bar{Y}_j + ... + [a_k]\bar{Y}_k) \]

\[ = \sqrt{\text{VAR}(\ [a_1]\bar{Y}_1 + [a_2]\bar{Y}_2 + ... + [a_j]\bar{Y}_j + ... + [a_k]\bar{Y}_k)} \]

\[ = \sqrt{\text{var}(\ [a_1]\bar{Y}_1) + \text{var}(\ [a_2]\bar{Y}_2) + ... + \text{var}(\ [a_k]\bar{Y}_k)} \]

\[ = \sqrt{[a_1^2]\text{var}(\bar{Y}_1) + [a_2^2]\text{var}(\bar{Y}_2) + ... + [a_k^2]\text{var}(\bar{Y}_k)} \]

\[ = \sqrt{\left[ a_1^2 \left( \frac{\hat{\sigma}_\text{WITHIN}^2}{n_1} \right) + a_2^2 \left( \frac{\hat{\sigma}_\text{WITHIN}^2}{n_2} \right) + \ldots + a_k^2 \left( \frac{\hat{\sigma}_\text{WITHIN}^2}{n_K} \right) \right]} \]

\[ = \hat{\sigma}_\text{WITHIN} \sqrt{\left( \frac{a_1^2}{n_1} \right) + \left( \frac{a_2^2}{n_2} \right) + \ldots + \left( \frac{a_k^2}{n_K} \right)} \]

This is also written as follows:

\[ = \sqrt{\text{MSE} \left[ a_1^2 \left( \frac{\sigma^2}{n_1} \right) + a_2^2 \left( \frac{\sigma^2}{n_2} \right) + \ldots + a_k^2 \left( \frac{\sigma^2}{n_K} \right) \right]} \]
**Confidence Interval for Linear Contrast**

Suppose the parameter of interest is \( E[L] = [a_1] \mu_1 + [a_2] \mu_2 + ... + [a_i] \mu_i + ... + [a_j] \mu_j + ... + [a_k] \mu_k \)

**Solution for a \((1-\alpha/2)100\%\) Confidence Interval of \( E[L] \):**

1. **Point estimate:**
   \[
   L = [a_1] \bar{Y}_1 + [a_2] \bar{Y}_2 + ... + [a_i] \bar{Y}_i + ... + [a_j] \bar{Y}_j + ... + [a_k] \bar{Y}_K
   \]

2. **SE[Point estimate]:**
   \[
   \hat{\sigma}_{\text{within}} \sqrt{\frac{a_1^2}{n_1} + \frac{a_2^2}{n_2} + ... + \frac{a_k^2}{n_K}} = \sqrt{\text{mse} \left( \frac{a_1^2}{n_1} + \frac{a_2^2}{n_2} + ... + \frac{a_k^2}{n_K} \right)}
   \]

3. **Confidence coefficient:**
   \[
   (1-\alpha/2)100\% \text{ percentile of Student's } t
   \]
   with degrees of freedom \( \sum_{i=1}^{K} (n_i-1) = N-K \)

Putting these together yields

\[
L \pm t_{DF=N-K; (1-\alpha/2)} \sqrt{\text{mse} \left( \frac{a_1^2}{n_1} + \frac{a_2^2}{n_2} + ... + \frac{a_k^2}{n_K} \right)}
\]
c. The Bonferroni Approach
This is a multiple comparisons adjustment procedure!

The “Bonferroni” method is the most conservative.

- “Conservative” - It is difficult to infer the alternative because it is difficult to reject the null hypothesis.

- The Bonferroni approach says: Use as a cut-off alpha level for declaring significance the nominal level alpha divided by the total number of tests performed.

- Number of ways to choose two groups for comparison from a collection of K:

\[
\binom{K}{2} = \frac{K(K-1)}{2}
\]

- The multiple comparisons adjustment alpha level for each test is therefore, according to this approach:

\[
\alpha_{\text{Bonferroni Adjusted}} = \frac{\alpha_{\text{DESIR ED}}}{K(K-1)/2}
\]

Remarks

- A potential problem with the Bonferroni approach is that, even when the number of groups is moderate, it is extremely difficult to correctly reject \( H_0 \).

- Thus, the Bonferroni approach has low power.
d. The Tukey Procedure
Another multiple comparisons adjustment procedure

The Tukey Method is Appropriate in a Selected Setting ONLY

- The sample sizes \( n_i \) in each group \( i=1, \ldots, K \) are all equal; AND
- It is of interest to compare groups two at a time.

Introduction to the Studentized Range Distribution

- This distribution is used to assess significance when using the Tukey method
- It is related to the Student’s t-distribution.
- It’s the appropriate reference distribution when implementing the Tukey procedure because it derives from the correct probability framework, that of order statistics.

”Look at the maximum discrepancy” and then “work your way in” until the discrepancies are no longer significant

- **Order** the sample means from largest to smallest: \( \bar{Y}_{\text{MAX}} \geq \ldots \geq \bar{Y}_{\text{MIN}} \)

- **Consider largest mean:**
  - Construct CI to compare largest mean to smallest mean. If significant, then
  - Construct CI to compare largest mean to 2nd smallest mean. If significant, then
  - Construct CI to compare largest mean to 3rd smallest mean, etc
  - End with largest mean when first NON significant result is obtained.
  - Draw an overbar to connect the means that do not differ significantly

- **Now start over but with consideration of the 2nd largest mean:**
  - Construct CI to compare 2nd largest mean to smallest mean. If significant, then
  - Construct CI to compare 2nd largest mean to 2nd smallest mean. If significant, then
  - Construct CI to compare 2nd largest mean to 3rd smallest mean, etc
  - End with 2nd largest mean when first NON significant result is obtained.
  - Draw another overbar to connect the means that do not differ significantly
• Continue cycle of starting with consideration of the 3rd, 4th, etc largest mean etc.
  - Eventually, you’ll end up with a series of overbars that you’ve drawn.
  - These can be put together in a schematic that summarizes the comparisons

How to Construct Confidence Intervals Using the Studentized Range Distribution

• Preliminary: A Clarification of Notation

<table>
<thead>
<tr>
<th></th>
<th>Notation in Lecture</th>
<th>Notation in Kleinbaum et al</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common sample size in group</td>
<td>n</td>
<td>n*</td>
</tr>
<tr>
<td>Number of groups</td>
<td>K</td>
<td>k</td>
</tr>
<tr>
<td>Total sample size (summed)</td>
<td>N=nK</td>
<td>n=n*k</td>
</tr>
</tbody>
</table>

(1) Point estimate: \( \bar{Y}_i - \bar{Y}_j \)

(2) SE[Point estimate]: \( \sqrt{mse} \)

(3) Confidence coefficient: \( T = \left( \frac{1}{\sqrt{n}} \right) q_{k;N-K,1-\alpha} \)

Putting these together yields

\( \bar{Y}_i - \bar{Y}_j \pm [ T ]\sqrt{mse} \)
e. Scheffe Method

Another multiple comparisons adjustment procedure

Scheffe’s method is based on a result (not shown here or in your text) which tells you the maximum width of a confidence interval for a linear contrast.

Thus, Scheffe’s method permits “fishing expeditions”. However, the cost for this privilege is the accompanying large width of associated confidence intervals.

Setting

- The sample sizes $n_i$ in each group $i=1, ..., K$ are NOT equal; AND
- It is of interest to perform comparisons defined as linear contrasts.
- The one way analysis of variance
  - Group 1: $X_{11},...,X_{n_1}$ are Normal($\mu_1, \sigma^2$)
  - Group 2: $X_{21},...,X_{n_2}$ are Normal($\mu_2, \sigma^2$)
  - ... 
  - Group K: $X_{1k},...,X_{nk}$ are Normal($\mu_k, \sigma^2$)
- Constant variance $\sigma^2$
- The overall F-test is statistically significant
- Question – Does a selected linear contrast have expected value zero?

$$ E[L] = a_1\mu_1 + a_2\mu_2 + ... + a_K\mu_K = 0 $$

where

$$ a_1 + a_2 + ... + a_K = 0 $$

Hypotheses

$H_0$: $E[L] = 0$

$H_A$: $E[L] = 0$
The solution for a Scheffe multiple comparison follows an approach that is already familiar

- Estimate $L$ using the sample means
  $$\hat{L} = \tilde{a}_1 \bar{Y}_1 + \tilde{a}_2 \bar{Y}_2 + \ldots + \tilde{a}_i \bar{Y}_i + \ldots + \tilde{a}_j \bar{Y}_j + \ldots + \tilde{a}_K \bar{Y}_K$$

- Estimate the appropriate standard error
  $$\hat{SE}[\hat{L}] = \hat{\sigma}_{\text{within}} \sqrt{\frac{a_1^2}{n_1} + \frac{a_2^2}{n_2} + \ldots + \frac{a_K^2}{n_K}} = \sqrt{\text{mse}} \sqrt{\frac{a_1^2}{n_1} + \frac{a_2^2}{n_2} + \ldots + \frac{a_K^2}{n_K}}$$

- The required confidence coefficient is a Scheffe multiplier $S$
  $$S = \sqrt{(K-1) F_{K-1, N-K; (1-\alpha)}}$$

- Obtain confidence interval as estimate $\pm$ (confidence coefficient) (SE) machinery
  $$95\% \text{ CI for } L = \hat{L} \pm (S) \hat{SE}[\hat{L}]$$
Which Multiple Comparison Procedure Should Be Used?

(1) If only pairwise comparisons are of interest and all K sample sizes are equal, then

- Use Tukey procedure
- Even though Scheffe could be used, Tukey confidence intervals are narrower

(2) If the sample sizes are unequal and/or if contrasts are of interest,

- Use Scheffe's method

(3) The Bonferroni approach is the most conservative

- Note
  If you obtain statistical significance here ..

Then you would have obtained statistical significance by any other method.