Random Factor in a Single Factor Experiment

Design of Experiments - Montgomery
Section 3-9
Random Factor vs Fixed Factor

• Fixed factor: Want to draw inference on the $a$ level effects used in the experiment

• Random factor: Want to draw inference on the population of factor effects $\rightarrow$ population mean & variance

• Examples of each in a CRD (1=fixed, 2=random)
  1. Want to compare the proboscis length of 10 honey bee colonies in Utah
     - Randomly choose $n$ bees from each of the $a = 10$ colonies of interest.
     - Want to compare the $\tau_i$’s of the $a$ colonies
  2. Want to study the proboscis length among all honey bee colonies in Utah
     - Randomly choose $a = 10$ colonies from the population of honey bee colonies in Utah. Then randomly choose $n$ bees from each of the selected colonies.
     - Want to assess $\sigma^2_\tau$. 
Random Factor vs Fixed Factor

- Inference broader in random effects case because it focuses on the population of effects, not just those effects in the experiment.

- The random selection of factor effects is done to allow valid conclusions about the entire population.

- However, levels are often not randomly chosen but commonly considered as if they were (selection bias?)
Random Effects Model (CRD)

• Same model as the fixed case with additional assumptions

\[ y_{ij} = \mu + \tau_i + \epsilon_{ij} \]

\[ \begin{align*}
\mu & \text{ - grand mean} \\
\tau_i & \text{ - } i\text{th treatment effect} \\
\epsilon_{ij} & \sim \text{N}(0, \sigma^2), \text{ independent}
\end{align*} \]

• Additionally assume

\[ \tau_i \sim \text{N}(0, \sigma^2_{\tau}) \]

\[ \{\tau_i\} \text{ and } \{\epsilon_{ij}\} \text{ independent} \]

• The results in \( V(y_{ij}) = \sigma^2_{\tau} + \sigma^2 \), \( \text{Cov}(y_{ij}, y_{ik}) = \sigma^2_{\tau} \)
Inference

• The primary hypotheses are:

\[ H_0 : \sigma^2_\tau = 0 \]
\[ H_1 : \sigma^2_\tau > 0 \]

• Partitioning of Total Sum of Squares identical but expected mean squares change

\[ E(MS_E) = \sigma^2 \]
\[ E(MS_{\text{Treatment}}) = \sigma^2 + n\sigma^2_\tau \]

• However, under \( H_0 \), the statistic \( F_0 \sim F_{\alpha,a-1,N-a} \)

• Thus, same test as before but different scope of inference

• Conclusions directed at the population of levels
Model Estimates (ANOVA method)

- Usually interested in estimating variances
- Using mean squares from ANOVA table
  \[ \hat{\sigma}^2 = \text{MS}_E \]
  \[ \hat{\sigma}_\tau^2 = (\text{MS}_{\text{Treatment}} - \text{MS}_E)/n \]
  If unbalanced, replace \( n \) with
  \[ n_0 = ((\sum n_i)^2 - \sum n_i^2)/((a - 1) \sum n_i) \]
- Estimate of \( \sigma_\tau^2 \) can be negative using this approach
  - Supports \( H_0 \)? Use zero as estimate? Be careful!!
  - Validity of model? Normal distribution assumption?
  - Consider a Bayesian approach (nonnegative prior)?
Confidence intervals

• $\sigma^2$: Same as the fixed case

\[
\frac{(N - a)\text{MS}_E}{\sigma^2} \sim \chi^2_{N-a} \leq \sigma^2 \leq \frac{(N - a)\text{MS}_E}{\chi^2_{\alpha/2,N-a}}
\]
Confidence intervals

• $\sigma^2_T$: Linear combination of $\chi^2$

$$\frac{(a - 1)MS_{Trt}}{\sigma^2 + n\sigma^2_T} \sim \chi^2_{a-1}$$

SO

$$f(\sigma^2_T) = \frac{\sigma^2 + n\sigma^2_T}{n(a - 1)} \chi^2_{a-1} - \frac{\sigma^2}{n(N - a)} \chi^2_{N-a}$$

No closed form expression for this distribution

Approximations available (Section 13-6)
Confidence intervals

- Ratio of $\sigma^2_\tau$ to $V(y_{ij})$

  Common estimate if goal is to reduce variance

  Uses ratio of two $\chi^2$ distributions (i.e., $F$ dist)

  \[
  \frac{L}{L+1} \leq \frac{\sigma^2_\tau}{\sigma^2 + \sigma^2_\tau} \leq \frac{U}{U+1}
  \]

  where

  \[
  L = \frac{1}{n} \left( \frac{\text{MST}_{\text{Trt}}}{\text{MSE} \, F_{\alpha/2, a-1, N-a}} - 1 \right)
  \]

  \[
  U = \frac{1}{n} \left( \frac{\text{MST}_{\text{Trt}}}{\text{MSE} \, F_{1-\alpha/2, a-1, N-a}} - 1 \right)
  \]

  or

  \[
  \frac{F_0 - F_{\alpha/2, a-1, N-a}}{F_0 + (n-1)F_{\alpha/2, a-1, N-a}} \leq \frac{\sigma^2_\tau}{\sigma^2 + \sigma^2_\tau} \leq \frac{F_0 - F_{1-\alpha/2, a-1, N-a}}{F_0 + (n-1)F_{1-\alpha/2, a-1, N-a}}
  \]
Confidence intervals

- Grand mean \( \mu \)

**Example:** Average proboscis length of all bees in Utah. Experimental unit is the colony. Selection of bees is subsampling. The bees within a colony are correlated but bees across colonies are not.

\[
\bar{y}_{..} = \frac{1}{a}(\bar{y}_{1.} + \bar{y}_{2.} + ... + \bar{y}_{a.})
\]

\( \bar{y}_{i.} \) iid Normal

Variance can be expressed in terms of \( MS_{Trt} \) and \( MSE \)

CI for \( \mu \) : 

\[
\bar{y}_{..} \pm t \sqrt{\text{Var}(\bar{y}_{..})}
\]
Likelihood Estimation Approach

- Linear mixed model expressed

\[ Y = X\beta + Z\delta + \varepsilon \]

- \( \beta \) is a vector of fixed-effect parameters
- \( \delta \) is a vector of random-effect parameters
- \( \varepsilon \) is the error vector

- \( \delta \) and \( \varepsilon \) assumed uncorrelated
  - means 0
  - covariance matrices \( G \) and \( R \)
Likelihood Estimation Approach

- $\text{Cov}(Y) = \Sigma = ZGZ' + R$
- If $R = \sigma^2 I$ and $Z = 0$, back to fixed effects model
- SAS Proc Mixed allows one to specify $G$ and $R$
- $G$ through RANDOM, $R$ through REPEATED
- When there are random effects, better to use likelihood estimation rather than least squares estimation.
- However, will often get same estimates
Likelihood Estimation Approach

• For known $G$ and $R$, use generalized least-squares

$$\hat{\beta} = (X'\Sigma^{-1}X)^{-1} X'\Sigma^{-1}Y$$
$$\hat{\delta} = GZ'\Sigma^{-1}(Y - X\hat{\beta})$$

• For unknown $G$ and $R$, their REML estimates can be substituted into these expressions

• REML maximizes likelihood that takes into account loss of degrees of freedom

$$-2 \log L = (n - p) \log(2\pi) + \log(|\Sigma|) + r'\Sigma^{-1}r + \log(|X'\Sigma^{-1}X|)$$
where \( r = Y - X\hat{\beta} \)
Example

A supplier delivers several hundred batches of raw material to a company each year. The company is interested in a high yield from each batch of raw material (percent usable). Therefore, to investigate the consistency of this supplier, an experiment is done where five batches were selected at random and three yield determinations were made on each batch.

<table>
<thead>
<tr>
<th>Batch</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>74</td>
<td>68</td>
<td>75</td>
<td>72</td>
<td>79</td>
</tr>
<tr>
<td></td>
<td>76</td>
<td>71</td>
<td>77</td>
<td>74</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>72</td>
<td>77</td>
<td>73</td>
<td>79</td>
</tr>
</tbody>
</table>
## Example Analysis

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sum of Squares</th>
<th>Degrees of Freedom</th>
<th>Mean Square</th>
<th>$F_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>147.73</td>
<td>4</td>
<td>36.93</td>
<td>20.5</td>
</tr>
<tr>
<td>Within</td>
<td>18.00</td>
<td>10</td>
<td>1.80</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>165.73</td>
<td>14</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Highly significant result ($F_{0.05,4,10} = 3.48$)
- $\hat{\sigma}^2_\tau = (36.93 - 1.80)/3 = 11.71$
- 86.7% ($11.71/(11.71+1.80)$) of total variance attributable to batch differences
- Conclusion: Time to improve consistency of the batches
Example: CI’s

- 95% CI for $\sigma^2$

$$\frac{SS_E}{\chi^2_{.025,10}} \leq \sigma^2 \leq \frac{SS_E}{\chi^2_{.975,10}} = (18.00/20.48, 18.00/3.25)$$

$$= (0.879, 5.538)$$

- 95% CI for Intraclass Correlation

$$\left( \frac{20.52-4.47}{20.52+(3-1)4.47}, \frac{20.52-(1/8.84)}{20.52+(3-1)(1/8.84)} \right)$$

$$(0.545, 0.984)$$

using property that $F_{1-\alpha/2,v_1,v_2} = 1/F_{\alpha/2,v_2,v_1}$
Using SAS

data example;
   input batch percent;
cards;
   1 74  1 76  1 75
   2 68 ...
;

proc glm;
   class batch;
   model percent=batch;
   random batch;
   output out=diag r=res p=pred;

proc mixed cl;
   class batch;
   model percent = ;
   random intercept / subject=batch vcorr=1;
run;
GLM Output

Dependent Variable: PERCENT

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>147.73333</td>
<td>36.93333</td>
<td>20.52</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Error</td>
<td>10</td>
<td>18.00000</td>
<td>1.80000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>14</td>
<td>165.73333</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>BATCH</td>
<td>4</td>
<td>147.73333</td>
<td>36.93333</td>
<td>20.52</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

Source Type III Expected Mean Square

BATCH Var(Error) + 3 Var(BATCH)

Using ANOVA method

<table>
<thead>
<tr>
<th>Variance Component</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Var(BATCH)</td>
<td>11.7111111</td>
</tr>
<tr>
<td>Var(Error)</td>
<td>1.80000000</td>
</tr>
</tbody>
</table>
MIXED Output

The MIXED Procedure

Iteration History

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Evaluations</th>
<th>-2 Res Log Like</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>77.03284751</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>62.75265372</td>
<td>0.00000000</td>
</tr>
</tbody>
</table>

Convergence criteria met.

Estimated V Correlation
Matrix for batch 1

<table>
<thead>
<tr>
<th>Row</th>
<th>Col1</th>
<th>Col2</th>
<th>Col3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0000</td>
<td>0.8668</td>
<td>0.8668</td>
</tr>
<tr>
<td>2</td>
<td>0.8668</td>
<td>1.0000</td>
<td>0.8668</td>
</tr>
<tr>
<td>3</td>
<td>0.8668</td>
<td>0.8668</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Covariance Parameter Estimates (REML)

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Estimate</th>
<th>Alpha</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>BATCH</td>
<td>11.71111111</td>
<td>0.05</td>
<td>4.0450</td>
<td>114.2090</td>
</tr>
<tr>
<td>Residual</td>
<td>1.80000000</td>
<td>0.05</td>
<td>0.8788</td>
<td>5.5436</td>
</tr>
</tbody>
</table>
Example: Negative $\sigma^2_\tau$ Estimate

data new;
input course subj score @@;
cards;
1 1  74.62 1 2  73.90 1 3  72.27 1 4  71.60 1 5  73.80
1 6  77.42 1 7  72.16 1 8  76.69 1 9  75.84 1 10 70.35
2 1  72.55 2 2  71.44 2 3  72.67 2 4  72.59 2 5  71.25
2 6  68.99 2 7  69.61 2 8  77.44 2 9  73.99 2 10 73.90
3 1  76.66 3 2  74.76 3 3  70.47 3 4  75.38 3 5  68.32
3 6  76.69 3 7  73.34 3 8  68.24 3 9  69.33 3 10 78.22

proc glm;
  class course; model score = course;
  random course / test;

proc mixed cl;
  class course; model score = ;
  random course;
run;
General Linear Models Procedure

Dependent Variable: SCORE

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>10.11154667</td>
<td>5.0557733</td>
<td>0.60</td>
<td>0.5557</td>
</tr>
<tr>
<td>Error</td>
<td>27</td>
<td>227.34895000</td>
<td>8.4203315</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>29</td>
<td>237.46049667</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLASS</td>
<td>2</td>
<td>10.11154667</td>
<td>5.0557733</td>
<td>0.60</td>
<td>0.5557</td>
</tr>
</tbody>
</table>

Source          | Type III Expected Mean Square
CLASS           | Var(Error) + 10 Var(CLASS)

Tests of Hypotheses for Random Model Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>course</td>
<td>2</td>
<td>10.111547</td>
<td>5.055773</td>
<td>0.60</td>
<td>0.5557</td>
</tr>
<tr>
<td>Error: MS(Error)</td>
<td>27</td>
<td>227.348950</td>
<td>8.420331</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

****Estimate of Var(Class) = (5.056-8.420)/10 = -0.336****
The MIXED Procedure

REML Estimation Iteration History

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Evaluations</th>
<th>Objective</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>93.37965543</td>
<td>0.00000000</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>93.37965543</td>
<td>0.00000000</td>
</tr>
</tbody>
</table>

Convergence criteria met.

Covariance Parameter Estimates (REML)

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Estimate</th>
<th>Alpha</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLASS</td>
<td>0.00000000</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>Residual</td>
<td>8.18829299</td>
<td>0.05</td>
<td>5.1935</td>
<td>14.7977</td>
</tr>
</tbody>
</table>

**** Above is result without NOBOUND option ****

Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Estimate</th>
<th>Alpha</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>course</td>
<td>-0.3365</td>
<td>0.05</td>
<td>-1.4244</td>
<td>0.7515</td>
</tr>
<tr>
<td>Residual</td>
<td>8.4203</td>
<td>0.05</td>
<td>5.2634</td>
<td>15.6003</td>
</tr>
</tbody>
</table>

**** This is result with the NOBOUND option ****
Use of the NOBOUND option

- When variances are assumed to be nonnegative, MIXED may set some variance estimates equal to zero
- This will affect other variances estimates
- Will also affect $F$ tests $\rightarrow$ inflate Type I error
- Be wary of this!!

- Use of the NOBOUND option reduces the impact on $F$ tests and other variance estimates
- For CI, uses standard Wald large sample $Z$ CI