Nested Factors

Design of Experiments - Montgomery
Chapter 13

Definitions

- Factors A and B are considered crossed if

  Every level of B occurs with every level of A

  A factorial model involves crossed factors

	Factor A			
Factor B	1	2	3	4
1	xx	xx	xx	xx
2	xx	xx	xx	xx
3	xx	xx	xx	xx

A	1			2			3			4		
B	1	2	3	1	2	3	1	2	3	1	2	3
	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	x	x	x	x	x	x	x	x

- Factors A and B considered nested if

  Levels of B occur with only one level of A

  Recall replicated Latin square designs

  One can arbitrarily number levels of B

A	1			2			3			4		
B	1	2	3	4	5	6	7	8	9	10	11	12
	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	x	x	x	x	x	x	x	x

Replication as a Nested Factor

- Consider CRD $y_{ij} = \mu + \tau_i + \epsilon_{ij}$
- Can write design where $a = 3$ and $n = 4$ as

Trt	1				2				3			
Rep	1	2	3	4	5	6	7	8	9	10	11	12
	x	x	x	x	x	x	x	x	x	x	x	x

- Thus could build replicate into model as factor
- Order of replicates unimportant $\rightarrow$ nested
- Brackets denote which factor its nested within

$$y_{ij} = \mu + \tau_i + r_{j(i)}$$

- Replication variability is used as error, $e_{ij} = r_{j(i)}$
- In SAS, omit lowest level term from model statement. Otherwise, all tests must be done using test option or statement (i.e., 0 df error).

Subsampling

- In many problems, difficult to measure EU response
- **Subsampling** - sampling EU numerous times
- Done to get more accurate measure of EU response
- Often use average of subsamples for analysis
- What if we include subsamples in analysis?

EU	1				2				3			
Sub	1	2	3	4	5	6	7	8	9	10	11	12
	x	x	x	x	x	x	x	x	x	x	x	x

- No association between subsamples across EUs
- Numbering of subsamples arbitrary
- Subsamples always a nested factor

Analysis with Subsamples

- If subsample added to model, results comparable to using the average of the subsamples
- Could also look at variance or median as summary

- Helps with design of future experiments
- Can check for consistency of measurements
- Protect against missing values and contamination
- Computational benefit if $\sigma^2_{\text{Sub}} > \sigma^2$
- Examples
  - Soil Samples within plot (e.g. moisture content, acidity)
  - Biochem analysis of animal tissue
  - Multiple plates of single agar batch

Simple Random Effects Model

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \quad \begin{cases} i = 1, 2 \\ j = 1, 2, \ldots n \end{cases}$$

- In some situations, can consider this as subsampling
- Primarily interested in μ or σ^2_τ
- Two stages of sampling
  - Randomly choose units of interest (σ^2_τ)
  - Obtain measurements on that unit of interest (σ^2)
- Use subsampling variability in test $H_0 : \sigma^2_\tau = 0$

- Var($\bar{\mu}$)=MS$_{Trt}/nk$
- Same variance based on averages of primary units

CRD with Subsampling

- Interested in effect of four methods of spreading mulch on soil moisture content. Have field of 16 plots - 4 for each spreading method.
  - Cannot measure moisture content directly
  - Choose 2 sites within plot to measure moisture
  - Samples averaged to obtain moisture content
  - Can view subsamples **nested** within plot
  - Introduces new source of variability (σ^2_δ)

$$y_{ijl} = \mu + \tau_i + \epsilon_{j(i)} + \delta_{k(ij)} \quad \begin{cases} i = 1, 2 \ldots a \\ j = 1, 2, \ldots n \\ k = 1, 2, \ldots s \end{cases}$$

$$\epsilon_{j(i)} \sim \text{N}(0, \sigma^2)$$
$$\delta_{k(ij)} \sim \text{N}(0, \sigma^2_\delta)$$

Source	DF	EMS
Trt	$a-1$	$s\sigma^2 + \sigma^2_\delta + ns\sum \tau_i^2/(a-1)$
Plot	$a(n-1)$	$s\sigma^2 + \sigma^2_\delta$
Subsampling	$an(s-1)$	σ^2_δ
Total	$ans-1$	

SAS Program

```
options nocenter ls=75;
\* Example soil.sas *\
data new;
input trt plot sub resp;  *** Before we ignored sub
cards;                    *** Plot is our replicate (nested within trt)
1 1 1 2                   *** Sub is nested within plot
1 1 2 3


;

proc sort;
 by trt plot;

proc means noprint;
 var resp;
 by trt plot;
 output out=new1 mean=respmn;

proc glm;
 class trt;
 model respmn=trt;

proc glm data=new;
 class trt plot sub;
 model resp=trt plot(trt);
 test h=trt e=plot(trt);
```

```
Dependent Variable: respmn
                                 Sum of
Source                     DF       Squares   Mean Square  F Value  Pr > F
Model                       3   26.04296875    8.68098958    40.41  <.0001
Error                      12    2.57812500    0.21484375
Corrected Total            15   28.62109375

R-Square     Coeff Var      Root MSE    respmn Mean
0.909922      12.73167      0.463512       3.640625

Source                     DF   Type III SS   Mean Square  F Value  Pr > F
trt                         3   26.04296875    8.68098958    40.41  <.0001


Dependent Variable: resp
                                 Sum of
Source                     DF       Squares   Mean Square  F Value  Pr > F
Model                      15   57.24218750    3.81614583    10.39  <.0001
Error                      16    5.87500000    0.36718750
Corrected Total            31   63.11718750

R-Square     Coeff Var      Root MSE     resp Mean
0.906919      16.64439      0.605960      3.640625

Source                     DF   Type III SS   Mean Square  F Value  Pr > F
trt                         3   52.08593750   17.36197917    47.28  <.0001
plot(trt)                  12    5.15625000    0.42968750     1.17  0.3772

Tests of Hypotheses Using the Type III MS for plot(trt) as an Error Term

Source                     DF   Type III SS   Mean Square  F Value  Pr > F
trt                         3   52.08593750   17.36197917    40.41  <.0001
```


RCBD with Subsampling

- Interested in studying the tenderizing methods of steak. Three animals are chosen and the 4 methods of treatment are applied to like portions of each animal. These portions are then divided up into five smaller portions and the tenderness is evaluated. Since method of treatment is applied to larger portion, the EU for tenderness are the larger portions. The individual evaluations relative to the method are subsamples.

Source	DF	EMS
Animal	$b-1$	
Trt	$a-1$	$s\sigma^2+\sigma^2_\delta+bs\sum\tau_i^2/(a-1)$
Animal*Trt	$(b-1)(a-1)$	$s\sigma^2+\sigma^2_\delta$
Subsampling	$ab(s-1)$	σ^2_δ
Total	$abs-1$	

Reasoning for Nested Factors

Consider the following two examples

1 Drug company interested in stability of product
   - Two manufacturing sites
   - Three batches from each site
   - Ten tablets from each batch

2 Stratified random sampling procedure
   - Randomly sample five states
   - Randomly select three counties
   - Randomly select two towns
   - Randomly select five households

More manageable experiment than factorial, CRD
   - Drug - Batches as an non-nested factor?
   - Sampling - more concentrated than CRD

Statistical Model

- Consider a two factor problem

$$y_{ijk}=\mu+\tau_i+\beta_{j(i)}+\epsilon_{k(ij)} \quad \begin{cases} i=1,2,\ldots,a \\ j=1,2,\ldots,b \\ k=1,2,\ldots,n \end{cases}$$

- Bracket notation represents nested factor
- Cannot include interaction
   - Not all levels of B appear with all levels of A
   - Cannot separate main effect of B and interaction AB
- Factors may be random or fixed
- Can use EMS algorithm to describe tests

Partitioning the Sum of Squares

- Rewrite observation as:

$$y_{ijk} = \overline{y}_{...} + (\overline{y}_{i..} - \overline{y}_{...}) + (\overline{y}_{ij.} - \overline{y}_{i..}) + (y_{ijk} - \overline{y}_{ij.})$$

- Can look at $\sum\sum\sum (y_{ijk} - \overline{y}_{...})^2$
- Right hand side simplifies to

$$bm\sum_i (\overline{y}_{i..} - \overline{y}_{...})^2 + n\sum_i\sum_j (\overline{y}_{ij.} - \overline{y}_{i..})^2 + \sum\sum\sum (\overline{y}_{ijk} - \overline{y}_{ij.})^2$$

- SS_A + $SS_{B(A)}$ + SS_E
- Under normality, all SS/σ^2 independent

Analysis of Variance Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
A	SS_A	$a-1$	MS_A	
B(A)	$SS_{B(A)}$	$a(b-1)$	$MS_{B(A)}$	
Error	SS_E	$ab(n-1)$	MS_E	
Total	SS_T	$abn-1$		

$SS_T = \sum\sum\sum y_{ijk}^2 - y_{...}^2/abn$

$SS_A = \frac{1}{bn}\sum y_{i..}^2 - y_{...}^2/abn$

$SS_{B(A)} = \frac{1}{n}\sum\sum y_{ij.}^2 - \frac{1}{bn}\sum y_{i..}^2$

$SS_E = \sum\sum\sum y_{ijk}^2 - \frac{1}{n}\sum\sum y_{ij.}^2$

- Use EMS to define proper tests

Example - 2-Factor Nested Model (Fixed)

τ_i

$\beta_{j(i)}$

$\epsilon_{k(ij)}$

Example - 2-Factor Nested Model (Random)

τ_i

$\beta_{j(i)}$

$\epsilon_{k(ij)}$

Example - 2-Factor Nested Model (Mixed)

τ_i

$\beta_{j(i)}$

$\epsilon_{k(ij)}$

Example

A company is interested in testing the uniformity of their film-coated pain tablets. A random sample of three batches were collected from each of their two blending sites. Five tablets were assayed from each batch.

Site	1			2		
Batch	1	2	3	4	5	6
	5.03	4.64	5.10	5.05	5.46	4.90
	5.10	4.73	5.15	4.96	5.15	4.95
	5.25	4.82	5.20	5.12	5.18	4.86
	4.98	4.95	5.08	5.12	5.18	4.86
	5.05	5.06	5.14	5.05	5.11	5.07

- What are the factors?
- Are any nested?
- Which are random and which are fixed?

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Site	.01825	1	.01825	
Batch(Site)	.45401	4	.11350	
Error	.29020	24	.01209	
Total	.76246	29		

$y_{11.} = 25.41$ $y_{12.} = 24.20$ $y_{13.} = 25.67$
$y_{21.} = 25.30$ $y_{22.} = 26.08$ $y_{23.} = 24.64$
$\sum\sum\sum y_{ijk}^2 = 763.8188$

$SS_T = 763.8188 - 151.3^2/30 = .76247$

$SS_A = (75.28^2 + 76.02^2)/15 - 151.3^2/30 = .01825$

$SS_{B(A)} = (25.41^2 + 24.20^2 + \ldots + 24.64^2)/5 - 763.07459 = .45401$

$SS_E = 763.8188 - 763.5286 = .2902$

Results

- Site: $F = .01825/.1135 = .1608$. There is not enough evidence to suggest that the two coating sites are different.
- Batch: $F = .1135/.0121 = 9.39$. Compare to $F_{4,24}$. There is significant batch-to-batch variability.

$$\hat{\sigma}^2 = .0121 \qquad \hat{\sigma}_\beta^2 = \frac{.1135 - .0121}{5} = .0203$$

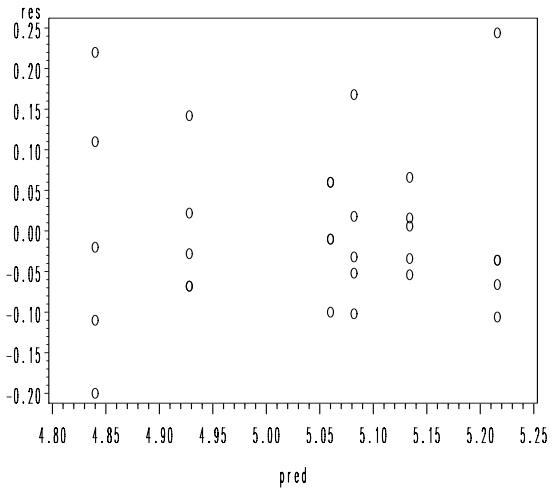
- Batch variability is $.0203/(.0203 + .0121) = 62.7\%$ of the total variability. It appears that efforts should be made to eliminate the batch-to-batch variability. Investigate what goes into coating a batch and see where the variability could be.

SAS Program

```
options nocenter ls=75;

data new;
infile "coating.dat";
input site batch tablet resp;

proc glm;
class site batch;
model resp=site batch(site);
random batch;
test h=site e=batch(site);
output out=new1 p=pred r=res;

symbol1 v=circle;
proc gplot;
plot res*pred;

proc glm data=new;
class site batch;
model resp=site batch site*batch;
random batch site*batch;
test h=site e=batch*site;
run;
```



```
Dependent Variable: RESP
                                  Sum of          Mean
Source                    DF       Squares        Square  F Value    Pr > F
Model                      5     0.4722667     0.0944533     7.81    0.0002
Error                     24     0.2902000     0.0120917
Corrected Total           29     0.7624667

Source                    DF     Type I SS  Mean Square  F Value    Pr > F
SITE                       1     0.0182533    0.0182533     1.51    0.2311
BATCH(SITE)                4     0.4540133    0.1135033     9.39    0.0001


Tests of Hypotheses using the Type III MS for
BATCH(SITE) as an error term

Source                    DF   Type III SS  Mean Square  F Value    Pr > F
SITE                       1     0.0182533    0.0182533     0.16    0.7089

---------------------------------------------------------------------------

Dependent Variable: RESP
                                  Sum of          Mean
Source                    DF       Squares        Square  F Value    Pr > F
Model                      5     0.4722667     0.0944533     7.81    0.0002
Error                     24     0.2902000     0.0120917
Corrected Total           29     0.7624667

Source                    DF     Type I SS  Mean Square  F Value    Pr > F
SITE                       1     0.0182533    0.0182533     1.51    0.2311
BATCH                      2     0.0115267    0.0057633     0.48    0.6266
SITE*BATCH                 2     0.4424867    0.2212433    18.30    0.0001
```

Nested Model as Factorial

- Suppose we treat design as two factor factorial
- Naively interpret SAS results
  - Significant batch*site variability
  - No longer significant batch-to-batch variability
- What does interaction mean?
- We're assuming batch 1 effect similar across sites
- Can't separate interaction from main effect
- Notice $SS_{AB} + SS_B = SS_{B(A)}$

$$df_{AB} + df_B = df_{B(A)}$$

- Could use factorial results and properly analyze model

General m-Stage Nested Design

- Consider 3 factor nested design

$$y_{ijkl} = \mu + \tau_i + \beta_{j(i)} + \gamma_{k(ij)} + \epsilon_{l(ijk)}$$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
A	SS_A	$a-1$	MS_A	
B(A)	$SS_{B(A)}$	$a(b-1)$	$MS_{B(A)}$	
C(B)	$SS_{C(B)}$	$ab(c-1)$	$MS_{C(B)}$	
Error	SS_E	$abc(n-1)$	MS_E	
Total	SS_T	$abcn-1$		

- Problem with nested designs
  - Few df for non-nested factor
  - In mixed/random situation, less power for non-nested factor

Staggered Nested Design

- Improve sampling efficiency with unbalanced design
- Consider A fixed, B and C are random
- Staggered Nested Design
  - a samples of non-nested factor (a levels of A)
  - 2 samples of first nested factor
  - 1 sample of second nested factor (except two from one)
  - Continue
- Results in $a-1$ and a degrees of freedom

Crossed and Nested Factors Model

- Can have design with crossed and nested factors
- These factors can be fixed or random
  - Nested-factorial designs
  - Repeated measure designs
- Example: Investigator interested in improving the number of rounds per minute fired from a Navy gun. Believes a new method of loading the gun will increase the number of rounds fired. Needs a team of people to use this gun. Divided teams into groups based on physique (slight, average, and heavy). Selected three teams from each of these groupings for the experiment. Each team was presented with both methods of loading and used each method twice in a random order.

  Method (L_i) - Fixed

  Group (G_j) - Fixed

  Team within Group ($T_{k(j)}$) - Random

$$y_{ijkl} = \mu + L_i + G_j + LG_{ij} + T_{k(j)} + LT_{ik(j)} + \epsilon_{l(ijk)}$$

Example - EMS

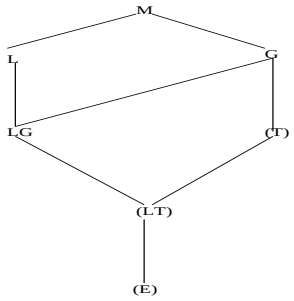
L_i
G_j
LG_{ij}
$T_{k(j)}$
$LT_{ik(j)}$
$\epsilon_{l(ijk)}$

Groups	Slight			Normal			Heavy		
Teams	1	2	3	4	5	6	7	8	9
Method 1	x	x	x	x	x	x	x	x	x
	x	x	x	x	x	x	x	x	x
Method 2	x	x	x	x	x	x	x	x	x
	x	x	x	x	x	x	x	x	x

- What are degrees of freedom?

Source of Variation	Degrees of Freedom
L	$(a-1)=1$
G	$(b-1)=2$
LG	$(a-1)(b-1)=2$
T(G)	$b(c-1)=6$
LT(G)	$(a-1)b(c-1)=6$
Error	$abc(n-1)=18$
Total	35

- To compute df for nested effects
  - Treat as factorial and pool df
  - $T_{k(j)} = T_k + GT_{kj}$
  - $LT_{ik(j)} = LT_{ik} + LGT_{ijk}$
  - Could pool SS in similar manner



Restricted Model:
L: Leading random term is LT
G: Leading random term is T
LG: Leading term is LT
T: Leading term is E because LT has fixed L
LT: Leading term is E

Unrestricted Model:
L: Leading random term is LT
G: Leading random term is T
LG: Leading term is LT
T: Leading term is LT
LT: Leading term is E

SAS Program

```
options nocenter ls=75;
data new;
infile 'guns.dat';
input method group team resp;

proc glm;
class group method team;
model resp = group|method team(group) method*team(group);
random team(group) method*team(group);
test h=group e=team(group);
test h=method e=method*team(group);
test h=group*method e=method*team(group);
means group / duncan lines e=team(group);
means method / duncan lines e=method*team(group);
means group*method;

proc mixed;
class group method team;
model resp=group|method;
random team(group) method*team(group);
lsmeans method group / adjust=duncan tdiff;
```



```
Dependent Variable: RESP

Source                    DF    Sum of Squares     F Value      Pr > F
Model                     17      229.25000000       11.29      0.0001
Error                     18       21.50000000
Corrected Total           35      250.75000000

Source                    DF       Type III SS     F Value      Pr > F
GROUP                      2      146.16666667       61.19      0.0001
METHOD                     1       26.69444444       22.35      0.0002
GROUP*METHOD               2       12.05555556        5.05      0.0182
TEAM(GROUP)                6       11.83333333        1.65      0.1906
METHOD*TEAM(GROUP)         6       32.50000000        4.53      0.0057

Source                 Type III Expected Mean Square
GROUP                  Var(Error) + 2 Var(METHOD*TEAM(GROUP))
                       + 4 Var(TEAM(GROUP)) + Q(GROUP,GROUP*METHOD)
METHOD                 Var(Error) + 2 Var(METHOD*TEAM(GROUP))
                       + Q(METHOD,GROUP*METHOD)
GROUP*METHOD           Var(Error) + 2 Var(METHOD*TEAM(GROUP))
                       + Q(GROUP*METHOD)
TEAM(GROUP)            Var(Error) + 2 Var(METHOD*TEAM(GROUP))
                       + 4 Var(TEAM(GROUP))
METHOD*TEAM(GROUP)     Var(Error) + 2 Var(METHOD*TEAM(GROUP))
```

$$\hat{\sigma}_T^2 = \frac{11.833/6 - 21.5/18}{4} = 0.194 \qquad \hat{\sigma}_{LT}^2 = \frac{32.5/6 - 21.5/18}{2} = 1.365$$

$$\hat{\sigma}^2 = \frac{21.5}{18} = 1.194$$

Tests of Hypotheses using the Type III MS for
TEAM(GROUP) as an error term

Source	DF	Type III SS	F Value	Pr > F
GROUP	2	146.16666667	37.06	0.0004

Tests of Hypotheses using the Type III MS for
METHOD*TEAM(GROUP) as an error term

Source	DF	Type III SS	F Value	Pr > F
METHOD	1	26.69444444	4.93	0.0682

Source	DF	Type III SS	F Value	Pr > F
GROUP*METHOD	2	12.05555556	1.11	0.3881

Duncan's Multiple Range Test for variable: RESP

NOTE: This test controls the type I comparisonwise error rate, not
the experimentwise error rate
Alpha= 0.05 df= 6 MSE= 1.972222

Number of Means	2	3
Critical Range	1.403	1.454

Means with the same letter are not significantly different.

Duncan Grouping	Mean	N	GROUP
A	19.0000	12	3
B	16.1667	12	2
C	14.0833	12	1

REML Estimation Iteration History

Iteration	Evaluations	Objective	Criterion
0	1	64.32842424	
1	3	59.67562606	0.00134581
2	1	59.63177037	0.00003263
3	1	59.63078094	0.00000002
4	1	59.63078028	0.00000000

Convergence criteria met.

Covariance Parameter Estimates (REML)

Cov Parm	Estimate
TEAM(GROUP)	0.00000000
METHOD*TEAM(GROUP)	1.24999961
Residual	1.19444458

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
GROUP	2	6	19.78	0.0023
METHOD	1	6	7.23	0.0361
GROUP*METHOD	2	6	1.63	0.2718

Least Squares Means

Effect	GROUP	METHOD	LSMEAN	Std Error	DF	t
METHOD		1	15.55555556	0.45304181	6	34.34
METHOD		2	17.27777778	0.45304181	6	38.14
GROUP	1		14.08333333	0.55486063	6	25.38
GROUP	2		16.16666667	0.55486063	6	29.14
GROUP	3		19.00000000	0.55486063	6	34.24

Least Squares Means

Pr > |t|	Alpha	Lower	Upper
0.0001	.	.	.
0.0001	.	.	.
0.0001	0.017	12.2680	15.8986
0.0001	0.017	14.3514	17.9820
0.0001	0.017	17.1847	20.8153

Differences of Least Squares Means

Effect	GROUP	METHOD	_GROUP	_METHOD	Difference	Std Error
METHOD		1		2	-1.72222222	0.64069787
GROUP	1		2		-2.08333333	0.78469143
GROUP	1		3		-4.91666667	0.78469143
GROUP	2		3		-2.83333333	0.78469143

Differences of Least Squares Means

DF	t	Pr > |t|	Alpha	Lower	Upper
6	-2.69	0.0361	0.050	.	.
6	-2.65	0.0378	0.017	-4.6505	0.4839
6	-6.27	0.0008	0.017	-7.4839	-2.3495
6	-3.61	0.0112	0.017	-5.4005	-0.2661

Improving Power

- What if we're interested in G
  - Since tested over $T_{k(j)}$, increase number of teams
- What if interested in L
  - Since tested over $LT_{ik(j)}$, increase number of teams
- What if interested in LT, T
  - Since tested over error, increase number of replications