

## Topic 11 Unbalanced designs (ST&D Section 9.6, p.219 & Chapter 18)

### 11. 1. The problem of missing data

- Accidents often result in **loss of data**. It is assumed that missing items are due to mistakes and not to a failure of a treatment: any missing observation  $Y_{ij}$  is assumed to follow the same mathematical model as the observations that are present.
- In a **one-way design**, the imbalance resulting from a missing data is not a problem.
- Missing values pose a **problem for two-way classifications**. Missing items destroy the symmetry and simplicity of the analysis, which becomes more complex if several  $Y_{ij}$  are missing.

### 11. 2. RCBD Example with one missing data

Yield of four breeding lines of wheat in which we have supposed that the yield  $Y_{41}$  for **line D in block 1 is missing**.

Line	Block					Total
	1	2	3	4	5	
A	32.3	34.0	34.3	35.0	36.5	172.1
B	33.3	33.0	36.3	36.8	34.5	173.9
C	30.8	34.3	35.3	32.3	35.8	168.5
D		26.0	29.8	28.0	28.8	112.6
Total	96.4	127.3	135.7	132.1	135.6	627.1

- **Insert the least-squares estimate of the missing value** in the vacant cell and analyze the complete data. This method gives least-squares estimates of every treatment mean and correct residual sum of squares.

If the missing value is in row  $i$  column  $j$ , and “ $I$ ” is the number of treatments and “ $J$ ” the number of blocks, the value to be inserted is calculated by the following formula:

$$\text{Estimated } Y_{ij} = (IY_{i.} + JY_{.j} - Y_{..}) / [(I-1)(J-1)]$$

From the totals in table 1, the value to be inserted is:

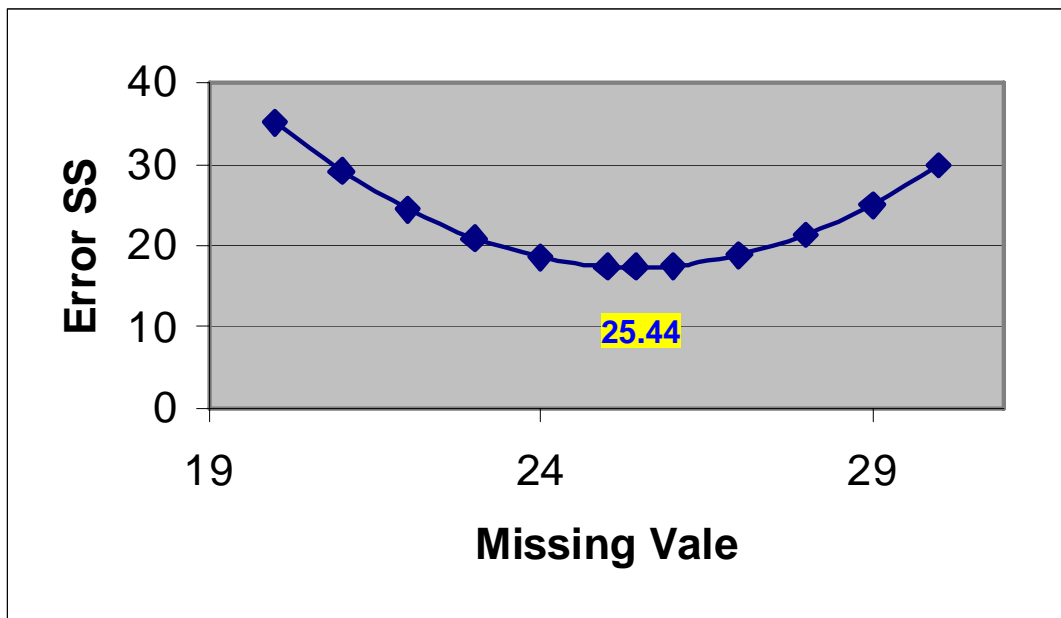
$$[4 * 112.6 + 5 * 96.4 - 627.1] / 3 * 4 = \mathbf{25.441666}$$

This value is entered in the table as the missing plot. ANOVA is computed as usual.

## Effect of replacing different missing values on the **Error Sum of Squares**

Missing Value	Error SS
20	35.097
21	29.167
22	24.437
23	20.907
24	18.577
25	17.447
<b>25.44</b>	17.331
26	17.517
27	18.787
28	21.257
29	24.927
30	29.797

25.44 is the least-squares estimate of the missing value.



The error SS has a minimum at 25.44

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	206.74650	29.53521	20.45	0.0001
Error	12	17.33100	1.44425		
C. Total	<b>19</b>	224.07750			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
TRTMNT	3	171.36150	57.12050	39.55	0.0001
BLOCK	4	35.38500	8.84625	6.13	0.0063

**Two additional corrections are required:**

- **The d.f. in the total and error sums of squares are both reduced by 1:** 18 d.f. for the total and 11 d.f. for the error sums of squares,
- **Row SS and Column SS are both adjusted by a special correction before their MS is computed.**

Correction to be subtracted from the Treatment SS:

$$\text{Correction SS treatment} = [Y_{.j} - (I-1) \cdot \text{estimated } Y_{ij}]^2 / I \cdot (I-1)$$

$$\text{Correction SS blocks} = [Y_{i.} - (J-1) \cdot \text{estimated } Y_{ij}]^2 / J \cdot (J-1)$$

In the example from Table 1:

$$SS_{\text{trtmnt}} = [96.4 - 3 \cdot 25.4]^2 / 4 \cdot 3 = 34.0033 \Rightarrow SS_{\text{trtmnt}} = 171.361 - 34.003 = \mathbf{137.36}$$

$$SS_{\text{block}} = [112.6 - 4 \cdot 25.4]^2 / 5 \cdot 4 = 6.05 \Rightarrow SS_{\text{block}} = 35.38 - 6.05 = \mathbf{29.33}$$

The corrected ANOVA is:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	206.74650	29.53521	20.45	0.0001
Error	<b>11</b>	17.33100	<b>1.57554</b>		
Corr. Total	<b>18</b>	224.07750			

Source	DF	Type I SS	MS	F Value	Pr > F
TRTMNT	3	<b>137.36</b>	45.78	29.06	0.0001
BLOCK	4	<b>29.33</b>	7.33	4.66	0.0192

### 11. 2. 1 Same RCBD Example using SAS

Missing data are indicated in SAS by a “.”

The SAS program for the previous example is:

```

Data lec_11SC;
  do trtmnt= 1 to 4;
    do block= 1 to 5;
      input yield @@;
      output;
    end;
  end;
cards;
32.3      34.0      34.3      35.0      36.5
33.3      33.0      36.3      36.8      34.5
30.8      34.3      35.3      32.3      35.8
.          26.0      29.8      28.0      28.8
;
proc glm;
class trtmnt block;
model yield=trtmnt block;
run; quit;

```

The output for this program is:

```

Class      Levels      Values
TRTMNT      4      1 2 3 4
BLOCK      5      1 2 3 4 5
Number of observations in data set = 20

```

**NOTE: Due to missing values, only 19 observations can be used in this analysis.**

Dependent Variable: YIELD

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	151.79952	21.68565	13.76	0.0001
Error	11	17.32996	1.57545		
Corrected Total	18	169.12947			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
TRTMNT	3	122.46347	40.82116	25.91	0.0001
BLOCK	4	29.33604	7.33401	4.66	0.0192

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TRTMNT	3	137.35921	45.78640	29.06	0.0001
BLOCK	4	29.33604	7.33401	4.66	0.0192

Note that the **Type III SS** produce exactly the **same result** as the one we obtained by replacing the missing value with its **least-squares estimate** based on block and column totals.

## 11. 2. 2 Effect of the order of the factors in the model statement: differences between Type I and Type III sum of squares.

In the previous SAS program if we replace

model yield=**trtmnt block**;

Source	DF	Type I SS	Mean Square	F Value	Pr > F
<b>TRTMNT</b>	3	122.46347	40.82116	25.91	0.0001
<b>BLOCK</b>	4	29.33604	7.33401	4.66	0.0192

**By**

model yield= **block trtmnt**;

Source	DF	Type I SS	Mean Square	F Value	Pr > F
<b>BLOCK</b>	4	14.44031	3.61008	2.29	0.1248
<b>TRTMNT</b>	3	137.35921	45.78640	29.06	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
<b>BLOCK</b>	4	29.33604	7.33401	4.66	0.0192
<b>TRTMNT</b>	3	137.35921	45.78640	29.06	0.0001

- Type III SS produce exactly the same result as before, but the TYPE I SS is different.
- When the block effect is the **last factor in the model**, the TYPE I SS<sub>blocks</sub> is equal to the TYPE III SS<sub>blocks</sub>.
- When the treatment effect is the last factor in the model, the TYPE I SS<sub>trtmnt</sub> is equal to the TYPE III SS<sub>trtmnt</sub>.

### TYPE I sum of squares:

- Lists the SS for each variable as if it were entered one at a time into the model, in the order they are specified in the model statement.
- **Is an incremental SS.**
- If there is any **variance that is common to two or more variables**, the variance will be attributed to one variable.

### TYPE III sum of squares:

- Gives the sum of squares that would be obtained for each variable **as if it were entered last into the model**.
- The effect of each variable is evaluated after all other factors have been accounted for.

**Effects of unbalanced data on the estimation of differences between means**

- The computational formulas for PROC GLM provide correct statistics for *balanced* or *orthogonal* data.
- When data are **unbalanced**, sums of squares computed from these means can contain functions of the other parameters of the model.

**Example** of the effects of unbalanced data on the estimation of differences between means and computation of sums of squares:

Data		B		
		1	2	Mean
A	1	7, 9	5	7
	2	8	4, 6	6
		8	5	

Means		B		
		1	2	Mean
A	1	8	5	6.5
	2	8	5	6.5
		8	5	

- Within level 1 of B, the cell mean for each level of A is 8, hence there is no evidence of a difference between the levels of A within level 1 of B.
- Likewise, there is no evidence of a difference between levels of A within level 2 of B because both means are 5.
- However, the difference between marginal means for A is  $7 - 6 = 1$

**Unbalance:** the B effect gets mixed up in the calculation of the A effect

This can be verified by expressing the observations in terms of the ANOVA model  $y_{ij} = \mu + \alpha_i + \beta_j$ .

For simplicity, interaction and error terms have been left out of the model

		B		The difference between marginal means for $A_1$ and $A_2$ is  $\frac{1}{3} [(\alpha_1 + \beta_1) + (\alpha_1 + \beta_1) + (\alpha_1 + \beta_2)] - \frac{1}{3} [(\alpha_2 + \beta_1) + (\alpha_2 + \beta_2) + (\alpha_2 + \beta_2)]$ $= (\alpha_1 - \alpha_2) + \frac{1}{3} (\beta_1 - \beta_2)$
		1	2	
A	1	$7 = \mu + \alpha_1 + \beta_1$ $9 = \mu + \alpha_1 + \beta_1$	$5 = \mu + \alpha_1 + \beta_2$	
	2	$6 = \mu + \alpha_2 + \beta_1$	$4 = \mu + \alpha_2 + \beta_2$ $6 = \mu + \alpha_2 + \beta_2$	

The observed difference between the marginal means for the two levels of A measures **the effect of factor B** in addition to the effect of factor A.

The null hypothesis about A we would normally wish to test is:

$$H_0: \alpha_1 - \alpha_2 = 0$$

However, SS for A computed by Type I SS in PROC GLM actually tests:

$$H_0: \alpha_1 - \alpha_2 + 1/3 (\beta_1 - \beta_2) = 0$$

The difference between the marginal means of A estimates ( $\alpha_1 - \alpha_2$ ) plus a function of the factor B parameters:  $1/3 (\beta_1 - \beta_2)$ .

**The difference between the A marginal means is biased by factor B effects.**

### 11. 3. 1. Effects of unbalanced data on the estimation of the marginal means

In terms of the  $\mu$  model  $y_{ij} = \mu_{ij} + \varepsilon_{ijk}$ , we usually want to estimate

$$(\mu_{11} + \mu_{12})/2 \quad \text{and} \quad (\mu_{21} + \mu_{22})/2.$$

However, the A marginal means estimate  $(2\mu_{11} + \mu_{22})/3$  and  $(\mu_{21} + 2\mu_{22})/3$

For example the expected marginal mean for A1 is:

$$[(\mu + \alpha_1 + \beta_1) + (\mu + \alpha_1 + \beta_1) + (\mu + \alpha_1 + \beta_2)]/3 = \mu + \alpha_1 + 2/3\beta_1 + 1/3\beta_2$$

The means of factor A are contaminated by effects of other factors

#### The **LSMEANS** statement:

**LSMEANS** produces the least-squares estimates of class variable means

**MEANS** produces, unadjusted means for all observations in each class.

Except for one-way designs, and some nested and balanced factorial structures, these unadjusted means are generally not equal to the least-squares means.

MEANS and LSMEANS from Table 1 can be obtained in SAS by:

```
model yield=trtmnt block;  
means trtmnt block;  
lsmeans trtmnt block / pdiff;
```

- The PDIFF option after the slash prints all possible probability values for the hypothesis  $H_0: LSM_i = LSM_j$ .
- These tests are analogous to the LSD in the balanced case.
- To compare **LSMEANS** using other multiple comparison techniques use

```
lsmeans trtmnt block / pdiff adjust= tukey;  
or = Dunnett or = Scheffe
```

**Table 2.** Comparison of **means** and **LS means** using data from Table 1 with the missing data and with the missing data replaced by its mean squares estimate.

	Missing value as 25.44166		Missing value as “.”	
	Means	LS Means	Means	LS Means
Treatment A	34.4200	34.4200	34.4200	34.4200
Treatment B	34.7800	34.7800	34.7800	34.7800
Treatment C	33.7000	33.7000	33.7000	33.7000
<b>Treatment D</b>	27.6083	27.6083	<b>28.1500</b>	27.6083
<b>Block 1</b>	30.4604	30.4604	<b>32.1333</b>	30.4604
Block 2	31.8250	31.8250	31.8250	31.8250
Block 3	33.9250	33.9250	33.9250	33.9250
Block 4	33.0250	33.0250	33.0250	33.0250
Block 5	33.9000	33.9000	33.9000	33.9000

- Left columns: the design is “**balanced**”. **Means** = **LS means**.
- Right columns unadjusted means are not equal to the least-squares means for **treatment D and block 1**, where the **missing data** is located.
- Means of unbalanced data are a function of sample sizes; LS means are not.
- The **LS means** produce values that are identical to those obtained by replacing the missing data by its least-squares estimate.

A major problem in the analysis of unbalanced data is the **contamination of means and differences between means by effects of other factors**.

The **solution** to this problem is to adjust the means to remove the contaminating effects using **LSMEANS** and the use of **Type III SS**.

#### 11. 4. Sums of Squares Computed by PROC GLM

PROC GLM recognizes different theoretical approaches to the analysis of variance by providing **four types of sums of squares**.

Type I SS, Type II SS, Type III SS, and Type IV SS.

Though we are going to use only **Type I** and **Type III** SS during this course a description of all four types is included.

#### 4. 1. Type I

- Type I SS correspond to adding each source (factor) **sequentially** to the model in the order listed.
- The Type I SS may **not** be particularly **useful for analysis of unbalanced multi way structures** but may be useful for nested models, polynomial models, and certain tests involving the homogeneity of regression coefficients.
- Comparing Type I and other types of sums of squares provides some information on the effect of the lack of balance.

#### 11. 4. 2. Type II

- Type II SS is adjusted for all factors that do not contain the **complete** set of letters in the effect
- Type II SS for an effect U, is adjusted for an effect V if and only if V does not contain U.
- For a two-factor structure with interaction, the main effect **A is adjusted by B but not for the A\*B interactions**. A\*B is adjusted for A & B.

#### 11. 4. 3. Type III

- In this model **every effect is adjusted for all other effects**.
- This is the closest thing to a "**standard**" for ANOVA.
- Type III sums of squares will produce the same SS as a Type I SS for a data set in which the missing data are replaced by least-squares estimates.

**Type III** sums of squares are **partial sums of squares**: each effect is adjusted for all other effects.

Balanced data  $\Rightarrow$  no difference between partial (I) or sequential (III) SS

#### 11. 4. 4. Type IV

- The Type IV functions are useful when there are empty cells.
- Type IV functions are not necessarily unique when there are empty cells
- They are = to those provided by Type III when there are no empty cells.

**PROC GLM** produces **Type I** and **Type III** SS as default. The 4 SS can be requested in PROC GLM as options in the MODEL statement.

The following SAS statement specifies the printing of all 4 sums of squares.

```
model . . . / ss1 ss2 ss3 ss4;
```

## 11. 5. Unbalanced nested designs

The **unbalance in subsample number** in a nested design generates additional **problems** with the **Expected Mean Squares** (ST&D page 168).

**Example:** Specific gravity of boards from several trees in three locations

Location	Location 1						Location 3						Location 4			
Tree	1023		1096		1153		3008		3015		3020		4053		4067	
100xSG	55	53	50	51	54	58	45	48	52	48	52	62	59	55	60	

- Trees and locations are RANDOM
- Trees are nested in location.
- **RANDOM statement, by default produces Type III EMS model**
- Type I method is valid only in pure nested designs and is used here because it is easier to understand.
- If there are crossed factors used the default **MIVQUE(0)** Estimates,

```

data STD170;
input location tree data @@;
cards;
1 1023 55 1 1096 53 1 1096 50 1 1096 51
1 1153 54 1 1153 58 3 3008 45 3 3008 48
3 3015 52 3 3015 48 3 3020 52 4 4053 62
4 4067 59 4 4067 55 4 4067 60
;
proc glm;
class location tree;
model data = location tree (location);
random tree(location) location /test;
proc varcomp method=Type1;
class location tree;
model data = location tree (location);
run; quit;

```

## Output

Dependent Variable: data

Source	DF	SS	MS	F Value	Pr > F
Model	7	286.57	40.94	7.32	0.0088
Error	7	39.17	<b>5.60</b>		
Corrected Total	14	325.73			

Source	DF	Type III SS	MS	F Value	Pr > F
location	2	198.95	99.48	17.78	0.0018
tree(location)	5	64.33	12.87	2.30	0.1539

Source	Type III	Expected Mean Square
location	Var(Error) + <b>1.54</b>	Var(tree(location)) + <b>4.05</b>
tree(location)	Var(Error) + <b>1.67</b>	Var(tree(location))

## Tests of Hypotheses for Random Model ANOVA

Source	DF	Type III SS	MS	F Value	Pr > F
location	2	198.95	99.48	8.09	0.0239
Error	5.3744	66.07	<b>12.29</b>		

Error: **0.921**\*MS(tree(location)) + **0.079**\*MS(Error)

Note that MSE **12.29** = **5.6** + **1.54**\***4.35**

Source	DF	Type III SS	MS	F Value	Pr > F
tree(loc)	5	64.33	12.87	2.30	0.1539
Error: MS(Error)	7	39.17	<b>5.60</b>		

## Variance component estimation using Method=Type 1

Source	DF	SS	MS	Expected Mean Square
loc	2	222.2	111.1	Var(Error)+2.22Var(tree(loc))+ <b>4.9</b>
tree(loc)	5	64.3	<b>12.87</b>	Var(Error)+ <b>1.67</b> Var(tree(loc))
Error	7	39.2	<b>5.6</b>	Var(Error)
Corr. Total	14	325.7		

$$(\mathbf{12.87} - \mathbf{5.60}) / \mathbf{1.67} = \mathbf{4.35}$$

Variance Component	Estimate	MIVQUE(0)
Var(location)	<b>19.44</b>	<b>16.72</b>
Var(tree(location))	<b>4.35</b>	<b>3.92</b>
Var(Error)	<b>5.60</b>	<b>7.89</b>