Topic 11. Unbalanced Designs [ST&D section 9.6, page 219; chapter 18]

11.1 Definition of missing data

Accidents often result in loss of data. Crops are destroyed in some plots, plants and animals die, volunteers quit a study before it is finished, etc. Life happens. Please note that if crops are destroyed on some plots as a result of the treatments or if animals die as a result of the treatment, these are not cases of missing data. The appropriate measurement for these experimental units should be made and included with the rest of the dataset.

Indeed, in the standard methods for handling missing data, it is assumed that *missing data are due to mistakes* and not to either a failure of a treatment or an effect of a treatment. To put it another way, any missing observation is assumed, if it had been made, to abide by the same mathematical model as the observations that are present.

11.1.1 Missing data in single-factor designs

In a one-way design, the imbalance resulting from a missing data causes no serious problems. The only effect is a reduction of in r, the sample size(s) of the affected class(es). This reduction in r will affect tests for means separation because, as you saw before, the minimum significant differences used by those methods depend on r.

Recall the expression for the minimum significant difference (w) used in the Tukey fixed-range method for means separation:

$$w = q_{\alpha,p,df_{MSE}} \sqrt{\frac{MSE}{2}\left(\frac{1}{r_1}+\frac{1}{r_2}\right)} \quad \text{for unequal r}$$

The implication of this expression is that each and every pairwise comparison require a different value of w, depending on the number of experimental units within the two treatment levels under consideration. When a dataset is unbalanced, R will no longer generate a nice mean separation table like before, in which a single value for w is used for all comparisons:

```
Tukey's Studentized Range (HSD) Test for Nlevel

Minimum Significant Difference        5.0499

Tukey Grouping           Mean       N    Culture

            A          28.800       5    3DOk1
      B     A          23.940       5    3DOk5
      B     C          19.880       5    3DOk7
      D     C          18.700       5    Comp
      D     E          14.600       5    3DOk4
            E          13.260       5    3DOk13
```

Instead, the output will look like this:

Culture Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
3DOk1 - 3DOk5	4.450	-2.052	10.952	
3DOk1 - Comp	9.325	3.305	15.345	***
3DOk1 - 3DOk7	10.450	3.077	17.823	***
3DOk1 - 3DOk4	13.250	7.539	18.961	***
3DOk1 - 3DOk13	14.850	8.830	20.870	***
3DOk5 - 3DOk1	-4.450	-10.952	2.052	
3DOk5 - Comp	4.875	-1.627	11.377	
3DOk5 - 3DOk7	6.000	-1.772	13.772	
3DOk5 - 3DOk4	8.800	2.583	15.017	***
3DOk5 - 3DOk13	10.400	3.898	16.902	***
Comp - 3DOk1	-9.325	-15.345	-3.305	***
Comp - 3DOk5	-4.875	-11.377	1.627	
Comp - 3DOk7	1.125	-6.248	8.498	
Comp - 3DOk4	3.925	-1.786	9.636	
Comp - 3DOk13	5.525	-0.495	11.545	
3DOk7 - 3DOk1	-10.450	-17.823	-3.077	***
3DOk7 - 3DOk5	-6.000	-13.772	1.772	
3DOk7 - Comp	-1.125	-8.498	6.248	
3DOk7 - 3DOk4	2.800	-4.323	9.923	
3DOk7 - 3DOk13	4.400	-2.973	11.773	
3DOk4 - 3DOk1	-13.250	-18.961	-7.539	***
3DOk4 - 3DOk5	-8.800	-15.017	-2.583	***
3DOk4 - Comp	-3.925	-9.636	1.786	
3DOk4 - 3DOk7	-2.800	-9.923	4.323	
3DOk4 - 3DOk13	1.600	-4.111	7.311	
3DOk13 - 3DOk1	-14.850	-20.870	-8.830	***
3DOk13 - 3DOk5	-10.400	-16.902	-3.898	***
3DOk13 - Comp	-5.525	-11.545	0.495	
3DOk13 - 3DOk7	-4.400	-11.773	2.973	
3DOk13 - 3DOk4	-1.600	-7.311	4.111	

Variable-sized confidence intervals are used throughout the analysis; so, in the spirit of full disclosure, the program is presenting the complete results of that analysis. You are then at liberty to organize these results into a single table, if you wish, declaring significance groupings where appropriate.

11.2 Missing data in two-factor designs

Missing values begin to cause more serious problems once you have crossed classifications. The simplest example of this is found in two-way classifications (e.g. two-factor experiments or RCBD's). In such cases, the missing values destroy the symmetry (i.e. the balance) of the design. With this loss of symmetry goes the simplicity of the analysis as well. And as more and more values are missing, the analysis becomes more and more complex.

Example: An RCBD taken from Snedecor & Cochram (1980), page 275. The data table below shows the yields of four breeding lines of wheat. An accident with the thresher during harvest led to loss of yield data for one of the plots, as indicated by the missing value for Y_{41} (line 4 in block 1).

Line	Block 1	2	3	4	5	Means	Totals
A	32.3	34.0	34.3	35.0	36.5	34.42	172.1
B	33.3	33.0	36.3	36.8	34.5	34.78	173.9
C	30.8	34.3	35.3	32.3	35.8	33.70	168.5
D		26.0	29.8	28.0	28.8	28.15	112.6
Means	32.13	31.83	33.93	33.03	33.90	32.76	
Totals	96.4	127.3	135.7	132.1	135.6		627.1

The strategy for dealing with such an imbalance is to replace the missing datapoint with its best estimate and proceed with the analysis. So what is the best estimate of this missing value?

Contrary to what one might think at first, the "best estimate" is not simply the predicted value of the cell, based on the effect of Line D and the effect of Block 1. The reason for this is that the values of each of these effects ($\tau_D = 28.15 - 32.76$; $\beta_1 = 32.13 - 32.76$) are themselves already affected by the loss of the datapoint.

The better approach is to assign a value to the cell that will minimize the error sum of squares. This is what the predicted value *would have* accomplished, if we had unbiased estimates of the effects of Line D and Block 1. Since we have no such unbiased estimates, this value is found using a least-squares approach.

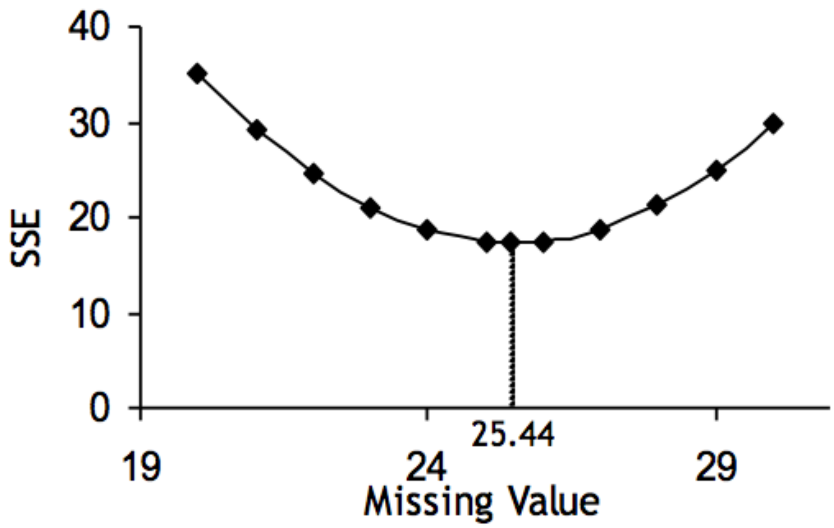
If the missing value is in row i and column j of this two-way classification, and "t" is the number of treatments and "b" is the number of blocks, the least-squares estimate to be inserted is given by the following formula:

$$\boxed{\text{Estimated } Y_{ij} = (tY_{i.} + bY_{.j} - Y_{..}) / [(t-1)(b-1)]}$$

For this particular dataset, the value to be inserted is:

$$\text{Estimated } Y_{41} = [4 * 112.6 + 5 * 96.4 - 627.1] / (3 * 4) = \mathbf{25.44}$$

This is called the "least-squares" estimate of the missing value because it minimizes the error sum of squares. That is, if different ANOVAs are performed on this dataset, using different values to replace the missing datapoint, and you plot the SSE for each of these analyses as a function of these values, a minimum is found at **25.44**. See plot on next page.



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

Once this value is determined, it is entered in the table as the missing plot and the ANOVA is computed as usual:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	206.1710600	29.4530086	20.39	<.0001
Error	12	17.3299600	1.4441633		
Corrected Total	19	223.5010200			
Block	4	35.2113200	8.8028300	6.10	0.0065
Treatment	3	170.9597400	56.9865800	39.46	<.0001

At this point, two additional corrections are required:

1. **The degrees of freedom in the total and error sums of squares must be adjusted.** In fact, we do not have 20 independent measurements in this dataset, we have only 19. So we really only have 18 df for the total and 11 df for the error sums of squares (i.e. the df for the total and for the error must each be reduced by 1).

2. **The sums of squares for both treatment and block must also be adjusted by a correction factor before their mean squares are computed.**

The corrections to be subtracted from each of these sums of squares:

$$\text{Correction for SST} = [Y_{.j} - (t-1)^*\text{estimated } Y_{ij}]^2 / t^*(t-1)$$
$$\text{Correction for SSB} = [Y_{i.} - (b-1)^*\text{estimated } Y_{ij}]^2 / b^*(b-1)$$

In this particular example:

Correction for SST = $[96.4 - 3*25.44]^2 / 4*3 = 33.601$

So, Corrected SST = 170.95974 – 33.601 = 137.36

Correction for SSB = $[112.6 - 4*25.44]^2 / 5*4 = 5.875$

So, Corrected SSB = 35.21132 – 5.875 = 29.34

The correct ANOVA is therefore:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	206.1710600	29.4530086	20.39	<.0001
Error	**11**	17.3299600	**1.5754509**		
Corrected Total	**18**	223.5010200			
Block	4	**29.34**	**7.335**	**4.66**	**0.0191**
Treatment	3	**137.36**	**45.79**	**29.06**	**<.0001**

11.2.1 Same RCBD Example using R

Missing data are indicated in R by a "NA". The dataset for importing into R would thus be:

trtmt	block	yield
1	1	32.3
1	2	34.0
1	3	34.3
1	4	35.0
1	5	36.5
2	1	33.3
2	2	33.0
2	3	36.3
2	4	36.8
2	5	34.5
3	1	30.8
3	2	34.3
3	3	35.3
3	4	32.3
3	5	35.8
4	1	NA
4	2	26.0
4	3	29.8
4	4	28.0
4	5	28.8

And the R code for the previous example could be:

```
miss1_mod<-lm(yield ~ trtmt + block, miss_dat)
anova(miss1_mod)
              Df  Sum Sq Mean Sq F value    Pr(>F)
trtmt          3  122.46   40.82  25.911 2.75e-05 ***
block          4   29.34    7.33   4.655   0.0192 *
Residuals     11   17.33    1.58
```

```
miss2_mod<-lm(yield ~ block + trtmt, miss_dat)
anova(miss2_mod)
              Df  Sum Sq Mean Sq F value    Pr(>F)
block          4   14.44    3.61   2.291    0.125
trtmt          3  137.36   45.79  29.062 1.59e-05 ***
Residuals     11   17.33    1.58
```

Recall, from the historical, "least squares" approach:

Block	4	29.34	7.335	4.66	0.0191
Treatment	3	137.36	45.79	29.06	<.0001

For the last factor in each model, we obtain *exactly the same result* as when we replaced the missing value with its least-squares estimate!

What is going on here? The results should not depend on the order of the factors in our model! It's time to talk about sums of squares.

By default, R produces a certain type of sums of squares called *sequential* or *incremental* SS (or Type I SS), in which variation is assigned to each variable in the model *sequentially*, in the order they are specified in the model statement. Depending on the situation, this strategy may or may not be desirable.

There is another type of SS, however, called partial SS (or Type II SS). In this approach, variation is assigned to each variable in the model *as though* it were entered last in the model. In this way, each variable accounts *only* for that variation that is independent of the other variables in the model. That is, the effect of each variable is evaluated after all other factors have been accounted for; and the only variation assigned to a variable is that variation which we can be certain is due to that variable alone.

In R, partial (or Type II) SS can be obtained with the Anova() function in the "car" package:

```
#library(car)
miss1_mod<-lm(yield ~ trtmt + block, miss_dat)
Anova(miss1_mod, type=2)
```

	Sum Sq	Df	F value	Pr(>F)	
trtmt	137.4	3	29.0624	1.586e-05	***
block	29.3	4	4.6552	0.01919	*
Residuals	17.3	11			

```
miss2_mod<-lm(yield ~ block + trtmt, miss_dat)
Anova(miss2_mod, type=2)
```

	Sum Sq	Df	F value	Pr(>F)	
block	29.3	4	4.6552	0.01919	*
trtmt	137.4	3	29.0624	1.586e-05	***
Residuals	17.3	11			

Note that the Type II SS are insensitive to the order of terms in the model. Moreover, the Type II SS exactly match our result using a least squares approach.

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

The computational formulas within lm() that make use of treatment means provide correct statistics for *balanced* or *orthogonal* data (i.e. data with an equal number of observations: $r_{ij} = r$ for all i and j). When data are not balanced, the sums of squares computed from these means can contain functions of (i.e. become contaminated by) other parameters in the model.

To illustrate the effects of unbalanced data on the estimation of differences between means and computation of sums of squares, consider the data in these two-way tables (the first table features the original data, the second table features the means of the original data):

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

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

Consider the table on the right: Within level 1 of factor 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. Thus we may conclude that there is no evidence in the table of a difference between the levels of A.

Now consider the table on the left: The marginal means for A are 7 and 6. The difference between these marginal means (7 – 6 = 1) may be interpreted as measuring an overall effect of the factor A. This conclusion would be incorrect. The problem is that, because the design is unbalanced, the effect of factor B influences the calculation of the effect of factor A. Orthogonality has been broken.

The observed difference between the marginal means for the two levels of A is a measure of the effect of factor B in addition to the effect of factor A.

This statement can be illustrated in the following way. Let's express the observations in the left-hand table in terms of the linear model:

$$y_{ij} = \mu + \alpha_i + \beta_j$$

For simplicity, the interaction and error terms have been left out of the model. You can think of this as an RCBD with one rep per cell.

Data	B	
	1	2
A 1	$7 = \mu + \alpha_1 + \beta_1$ $9 = \mu + \alpha_1 + \beta_1$	$5 = \mu + \alpha_1 + \beta_2$
A 2	$8 = \mu + \alpha_2 + \beta_1$	$4 = \mu + \alpha_2 + \beta_2$ $6 = \mu + \alpha_2 + \beta_2$

A little algebra shows the difference between marginal means for A_1 and A_2 to be:

Mean A_1 – Mean A_2 = 1/3 (7 + 9 + 5) – 1/3 (8 + 4 + 6)

$= 1/3\,[(\alpha_1 + \beta_1) + (\alpha_1 + \beta_1) + (\alpha_1 + \beta_2)] - 1/3[(\alpha_2 + \beta_1) + (\alpha_2 + \beta_2) + (\alpha_2 + \beta_2)]$

$= (\alpha_1 - \alpha_2) + \mathbf{1/3\,(\beta_1 - \beta_2)}$

So, instead of estimating the difference between the effects of A_1 and A_2 (what we would expect the difference in marginal means to represent), the difference between the marginal means of A estimates $(\alpha_1 - \alpha_2)$ **PLUS** a function of the factor B parameters: $1/3\,(\beta_1 - \beta_2)$.

In other words:

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

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

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

However, the sum of squares for A computed by Type I SS actually tests the hypothesis:

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

This null hypothesis involves the factor B difference in addition to the factor A difference. In summary, the problem with unbalanced designs in multifactor analyses is that the **factors get mixed up with each other in the calculations**.

11.3.1 Effects of unbalanced data on the estimation of the marginal means

Let's continue with the simple dataset discussed in the previous section. In terms of the model $y_{ij} = \mu_{ij} + \varepsilon_{ijk}$, we usually want to estimate the marginal means of A:

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

In this particular dataset, however, since it is unbalanced, the A marginal means actually estimate:

$$(2\mu_{11} + \mu_{22}) / 3 \qquad \text{and} \qquad (\mu_{21} + 2\mu_{22}) / 3$$

These estimates are functions of the usually irrelevant cell frequencies and may, for that reason, be useless.

For example, the expected marginal mean for A_1 is:

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

which can be simplified:

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

In R, we use the lsmeans() function (library "lsmeans") to produce the least-squares estimates of class variable means. Because these means adjust for the contamination effects of other factors in the model, these means are sometimes referred to as *adjusted means*. Least-squares, or adjusted, means should not, in general, be confused with ordinary means, which are available through the methods we've discussed to this point in the class. Those previous methods (e.g. HSD.test(), etc.) produce simple, unadjusted means for all observations in each class or treatment. Except for one-way designs and some nested and balanced factorial structures, these unadjusted means are generally not equal to the least-squares means when data is missing.

For the RCBD example we were using at the start of this chapter, the least-squares means can be obtained in R with the following code:

```
#library(lsmeans)
lsmeans(miss1_mod, "trtmt")
lsmeans(miss1_mod, "block")
```

Then, for all subsequent lsmeans comparison, you must first assign the computed lsmeans to an object. Here, I am assigning it to an object I've called "miss_lsm":

```
miss_lsm <- lsmeans(miss1_mod, "trtmt")
```

This object can then be acted upon by the contrast() function within the "lsmeans" package.

Examples:

To perform Tukey (HSD) pairwise comparisons among the adjusted means:

```
contrast(miss_lsm, method = "pairwise", adjust = "tukey")
```

To perform a Dunnett test (comparisons to a control) among the adjusted means:

```
contrast(miss_lsm, method = "trt.vs.ctrl")
```

To compare the adjusted means using orthogonal contrasts (group comparisons):

```
contrast(miss_lsm, list("A vs. B" = c(1,-1,0,0), "AB vs. CD" = c(1,1,-1,-
      1), "C vs. D" = c(0,0,1,-1)))
```

To conduct a trend analysis among the levels of the factor of interest:

```
contrast(miss_lsm, method = "poly")
```

In all these cases, notice that the estimates are the differences in the LSMeans.

The following table presents a comparison of ordinary means and least-squares (i.e. adjusted) means, using data from the previous RCBD example. The right pair of columns corresponds to the case where the missing data has been replaced by its least squares estimate (25.44); the left pair of columns corresponds to the case where there is a missing data point.

	Missing value as "25.44166"		**Missing value as "NA"**	
	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
Treatment D	27.6083	27.6083	28.1500	27.6083
Block 1	30.4604	30.4604	32.1333	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

You will notice that the only means that are affected by the missing data point are those for Treatment D and Block 1, the cell of the missing data.

When the missing value is replaced by its least-squares estimate (25.44166), the balance of the design is "restored" and the means and LS means are identical. When the missing value is not replaced (i.e. when "NA" is used), the unadjusted means are not equal to the least-squares means for Treatment D and Block 1, where the missing data is located. The means of unbalanced data are a function of sample sizes (i.e. cell frequencies); the LS means are not. Said another way, the lsmeans() function produces values that are identical to those obtained by replacing the missing data by its least-squares estimate.

In summary, a major problem in the analysis of unbalanced data is the contamination of means, and thus the differences among means, by effects of other factors. The solution to these problems is to replace missing data by their least squares estimates and to remove the contaminating effects of other factors through a proper adjustment of means. With R, all this means is that you should use the Type II SS (**A**nova()) and the lsmeans() function.

11.4 More Sums of Squares

If there is a Type I SS and a Type II SS, it begs the question: Might there be more? You're in luck! In fact, there are four types of sums of squares, each with their associated statistics. These four types, of course, are called Types I, II, III, and IV (Goodnight 1978). Though we are going to use only Type I and Type II SS during this course, here's a brief description of all four.

11.4.1 Type I (sequential or incremental SS)

Type I sums of squares are determined by considering each source (factor) sequentially, in the order they are listed in the model. The Type I SS may not be particularly useful for analyses of unbalanced, multi-way structures but may be useful for balanced data and nested models. Type I SS are also useful for parsimonious polynomial models (i.e. regressions), allowing the simpler components (e.g. linear) to explain as much variation as possible before resorting to models of higher complexity (e.g. quadratic, cubic, etc.). Also, comparing Type I and other types of sums of squares provides some information regarding the magnitude of imbalance in the data.

Types II and III SS are also know as *partial* sums of squares,
in which each effect is adjusted for other effects.

11.4.3 Type III

Type III is also a partial SS approach, but it's a little easier to explain than Type II; so we'll start here. In this model, every effect is adjusted for *all other effects*. The Type III SS will produce the same SS as a Type I SS for a data set in which the missing data are replaced by the least-squares estimates of the values. The Type III SS correspond to Yates' weighted squares of means analysis. One use of this SS is in situations that require a comparison of main effects even in the presence of interactions (something the Type II SS does not do and something, incidentally, that many statisticians say should not be done anyway!).

In particular, the main effects A and B are adjusted for the interaction A*B, as long as all these terms are in the model. If the model contains only main effects, then you will find that the Type II and Type III analyses are the same.

11.4.2. Type II

Type II partial SS are a little more difficult to understand. Generally, the Type II SS for an effect U, which may be a main effect or interaction, is adjusted for an effect V *if and only if* V does not contain U. Specifically, for a two-factor structure with interaction, the main effects A and B are *not* adjusted for the A*B interaction because the interaction contains both A and B. Factor A is adjusted for B because the symbol B does not contain A. Similarly, B is adjusted for A. Finally, the A*B interaction is adjusted for each of the two main effects because neither main effect contains both A and B. Put another way, the Type II SS are adjusted for all factors that do not contain the **complete** set of letters in the effect. In some ways, you could think of it as a sequential, partial SS; in that it allows lower-order terms explain as much variation as possible, adjusting for one another, before letting higher-order terms take a crack at it.

11.4.4 Type IV

The Type IV functions were designed primarily for situations where there are empty cells, also known as "radical" data loss. The principles underlying the Type IV sums of squares are quite involved and can be discussed only in a framework using the general construction of estimable functions. It should be noted that the Type IV functions are not necessarily unique when there are empty cells but are identical to those provided by Type III when there are no empty cells.