

1

An introduction to analysis of variance

1.4 The fertiliser dataset: an example of one-way ANOVA

In this first example, we wish to compare the efficacy of three fertilisers. A field experiment has produced the yield of a crop from 30 plots, each fertiliser having been applied to 10 of these plots. Which fertiliser produces the greatest yields?

Step 1: The data

SAS will require the data in columns as illustrated in Table 1.2 in the main text; that is, with levels of fertiliser represented as subscripts one to three in one column called `FERTIL`, and the yields from each plot in a second column called `YIELD`. We described in detail, in the introduction, how to import the *fertiliser* dataset. Alternatively, you can enter the data manually. To do this, you must be in Analyst mode (by choosing Solution > Analysis > Analyst from the menus). An empty datasheet then appears in a window. Data may be entered by simply typing in the cells and pressing return. Once the worksheet contains data, you need to name the columns. First select the column by clicking on the top row, where it has a default name of A or similar. Then type in the column name. Once you have finished, don't forget to save the file (by choosing File > Save from the menus).

Step 2: The question

The relevant procedure is `PROC GLM`, followed by the model formula (see SAS COMMANDS Box 1.1). This may either be entered into the Editor window, or via menus. In the Editor window, notice that each line must end in a semi-colon, with the semi-colon after 'run' being crucial. The first line specifies the command name, and the dataset. Because we have placed our *fertiliser* dataset in the work library, we only have to specify its name. Later, we will discuss how to place all datasets in a separate library, so that they do not need to be imported every time. The second line tells SAS that `FERTIL` is a class variable, that is, the values of the variable divide the dataset into classes. This is SAS's name for what in the main text is called a categorical variable. If you do not specify, SAS assumes that a variable is continuous (for which the preferred SAS term is 'Quantitative'). The third line specifies the model formula. The final line of a set of commands that are to be executed must be 'run;', otherwise they won't happen. (Don't forget to select Run>Submit from the menu bar or click on the running figure).

8 An introduction to analysis of variance

When using the menu route, to transfer variable names to panes, you first need to highlight the variable name, then click on the title box of the destination pane. So to transfer YIELD to the dependent variable pane, highlight YIELD then click on Dependent

Finally, don't forget that you need to be in the 'Analyst' environment to use menus, and the initial environment to type your commands in the editor window.

| SAS COMMANDS 1.1 Analysis of variance | |
|--|---|
| Commands | <pre>proc glm data=fertiliser; class FERTIL; model YIELD=FERTIL; run;</pre> |
| Menu route | Statistics > Anova > Factorial Anova... YIELD → Dependent FERTIL → Independent |

SAS produces its output in a large and expansive format, which derives from the old-fashioned lineprinter. It takes practice to find all the information on a small window on screen.

If you compare the output with Box 1.1 in the main text, you can see that the essence of the output, the ANOVA table, is the same. The heading Pr > F is the SAS version of the p-value, indicating explicitly that this is the probability of the F ratio being this large or greater (under the null hypothesis). It also provides some extra pieces of information, such as the grand mean (YIELD Mean 4.643667). Root MSE is the square root of the error mean square, used to construct the confidence intervals as described in the main text. The other statistics (R-Square and the meaning of Type III SS for example) will be explained more fully in later chapters.

Basic descriptive statistics may also be obtained, which are useful for presenting the results and having a preliminary inspection of the within-group variances. The menu route involves moving response and explanatory variables into appropriate boxes. You select the variable from the list in the left hand pane, and click on the button above the pane to which you want to move it.

| SAS OUTPUT BOX 1.1 Analysis of the <i>fertiliser</i> dataset: one-way ANOVA | | | | | |
|---|----------|----------------|-------------|------------|--------|
| The GLM Procedure | | | | | |
| Class Level Information | | | | | |
| Class | Levels | Values | | | |
| FERTIL | 3 | 1 2 3 | | | |
| The GLM Procedure | | | | | |
| Dependent Variable: YIELD | | | | | |
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 2 | 10.82274667 | 5.41137333 | 5.70 | 0.0086 |
| Error | 27 | 25.62215000 | 0.94896852 | | |
| Corrected Total | 29 | 36.44489667 | | | |
| | R-Square | Coeff Var | Root MSE | YIELD Mean | |
| | 0.296962 | 20.97804 | 0.974150 | 4.643667 | |
| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
| FERTIL | 2 | 10.82274667 | 5.41137333 | 5.70 | 0.0086 |

| SAS COMMANDS 1.2 Descriptive statistics | |
|---|--|
| Commands | <pre>proc means data=fertiliser; by FERTIL; var YIELD; run;</pre> |
| Menu route | Statistics > Descriptive > Summary Statistics... YIELD → Analysis FERTIL → Class |

SAS BOX 1.2 (PROVIDES THE INFORMATION FOR TABLE 1.5)

Descriptive statistics for the *fertiliser* data set

The MEANS Procedure

Analysis Variable : YIELD

| FERTIL | N | | Mean | Std Dev | Minimum | Maximum |
|--------|-----|----|-----------|-----------|-----------|-----------|
| | Obs | N | | | | |
| 1 | 10 | 10 | 5.4450000 | 0.9759810 | 4.0700000 | 7.1400000 |
| 2 | 10 | 10 | 3.9990000 | 0.9717504 | 3.0700000 | 6.2800000 |
| 3 | 10 | 10 | 4.4870000 | 0.9747142 | 3.5300000 | 7.0000000 |

These two output boxes provide you with all the information you require to construct confidence intervals for each group mean (namely s , n and \bar{y}).

1.7 Exercises

Melons

This experiment compares the yield from four melon varieties. It was designed so that each variety was grown in six plots—but two plots growing variety three were accidentally destroyed. There is just one categorical explanatory variable. After the dataset *melons* has been read into the work library it would be analysed as follows:

```
Commands  proc glm data=melons;
           class VARIETY;
           model YIELDM = VARIETY;
           lsmeans VARIETY;c
           run;
```

Menu route Statistics > Anova > Factorial Anova...

YIELDM → Dependent

VARIETY → Independent

Means.. with LS Means tab

VARIETY → LS Mean

Note the alternative way of producing the group means as part of the analysis, rather than obtaining the descriptive statistics separately. This would produce the following output.

| SAS OUTPUT FOR TABLE 1.7 Analysis of variance for melons | | | | | | |
|--|----------|----------------|-------------|-------------|--------|--|
| The GLM Procedure | | | | | | |
| Class Level Information | | | | | | |
| Class | Levels | Values | | | | |
| VARIETY | 4 | 1 | 2 | 3 | 4 | |
| Number of observations 22 | | | | | | |
| The GLM Procedure | | | | | | |
| Dependent Variable: YIELDM | | | | | | |
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F | |
| Model | 3 | 1115.280536 | 371.760179 | 23.80 | <.0001 | |
| Error | 18 | 281.189142 | 15.621619 | | | |
| Corrected Total | 21 | 1396.469677 | | | | |
| | R-Square | Coeff Var | Root MSE | YIELDM Mean | | |
| | 0.798643 | 14.28765 | 3.952419 | 27.66318 | | |
| Source | DF | Type III SS | Mean Square | F Value | Pr > F | |
| VARIETY | 3 | 1115.280536 | 371.760179 | 23.80 | <.0001 | |
| Least Squares Means | | | | | | |
| | VARIETY | YIELDM LSMEAN | | | | |
| | 1 | 20.4900000 | | | | |
| | 2 | 37.4033333 | | | | |
| | 3 | 20.4625000 | | | | |
| | 4 | 29.8966666 | | | | |

Note that SAS states the levels of any categorical variable, before producing the ANOVA table.

Dioecious trees

See SAS output for this exercise in the answers for exercises.