

## **Analysis of Variance (ANOVA)**

### **Required Statements:**

**CLASS or CLASSES:** The CLASS statement is used to define variables which represent groupings or classifications of the data. Examples would be treatment and replication ID numbers or letters. Since the values of these variables represent the levels of an effect, SAS will accept either numeric or alphabetic data in the CLASS statement variables. Note: Data do not have to be sorted according to the CLASS variables.

**MODEL:** The model statement specifies the response and independent effects to be used. The general form is `MODEL Dependent var. = Independent var.`, where the *Independent var.* is a list of all or some of the variable names given in the CLASS statement. One good feature of GLM is the ability to easily give interaction and nested terms in the MODEL statement. Interaction effects can be specified by the \* symbol between two or more terms (e.g. A\*B). This notation can become cumbersome if many interactions are present so a shorthand version exists. The vertical bar between effects signifies all possible interactions. An example would be: A|B|C = A, B, A\*B, C, A\*C, B\*C, and A\*B\*C in that order. Nesting occurs when one effect exists completely within another effect either in time or space. These effects are denoted with parentheses such as REP(A) which is read as 'REP within A'. The \*, |, and () can be used alone or in combination to give the desired results.

## **Additional Statements and Options:**

**MEANS and LSMEANS:** The MEANS statement allows the user to obtain the means for effects or interactions along with their std errors. Popular uses of ANOVA includes mean separation techniques (multiple comparison procedures) such as LSD, TUKEY, DUNCAN, etc. SAS provides a large selection of these methods. They are specified as options to the MEANS statement. For example to get the mean, std error, and Fisher's LSD for two effects, the following could be used: MEANS A B / LSD LINES. Note that the options are given after a slash, /, (this is standard for all SAS statements). The LINES option forces SAS to group means based on the LSD using the letters A, B, C, etc. Without this option the LSD option may use asterisk notation to signify differences between all possible pairs of means or it will use the LINE notation. SAS chooses between these based on whether the data is balanced or unbalanced.

Which comparison procedure to use depends on many criteria such as: 1) the type of error to be controlled (experiment-wise or comparison-wise), 2) the objectives of the research, 3) the publication outlet if any, 4) personal preference.

The Least Squares Mean (LSMEANS) statement is used when there are missing values or covariates within the data. A short explanation of LSMEANS is given in handout # 2.1. SAS provides for comparison of LSMEANS by the PDIFF option which gives a table of p-values for all possible pairwise comparisons. It is important to remember that the probabilities associated with PDIFF are applicable to a limited number of preplanned comparisons only. Std errors for LSMEANS can be requested with the STDERR option.

Another use of LSMEANS is with interactions. SAS will calculate means for interactions, but will not compute any multiple comparison statistics. In this case LSMEANS with the PDIFF option can provide means comparisons.

**CONTRAST:** An alternative to multiple comparison procedures is the use of a limited number of single degree of freedom contrasts to test hypotheses of interest. SAS provides for these with the CONTRAST statement. A single df contrast has the general form of:

CONTRAST '*any label*' *Factor name* {*coefficients*}. The label portion allows you to give a name to the hypothesis being tested so it can be identified on the printout. The factor name identifies what the contrast is working on and the coefficients specify the contrast itself. An example might be CONTRAST 'Trts vs Ctrl' A 2 -1 -1. Here the average of the last 2 levels of A are contrasted to the first level. Some contrasts test more than one hypothesis at a time (multiple df contrasts) and these are separated by a comma in the CONTRAST statement; EX: CONTRAST 'Testit' A 2 -1 -1 , A 0 1 -1. This would be a 2 degree of freedom contrast.

## Example 1 - Factorial CRD:

```
PROC GLM;  
  CLASS VAR FERT;  
  MODEL YIELD = VAR FERT VAR*FERT;  
  
  MEANS VAR FERT VAR*FERT / LSD LINES;  
  LSMEANS VAR FERT VAR*FERT / PDIFF STDERR;
```

NOTES: In the balanced case (no missing values) both the MEANS and LSMEANS statements produce the same results. In that situation only one statement type would be required (probably MEANS). When missing values do occur the LSMEANS statement should be used. Also it should be noted that GLM does not produce LSD statistics for the interaction VAR\*FERT. To compare these means use the LSMEANS statement with the PDIFF and STDERR options.

## **Example 2 - Factorial RCB:**

```
PROC GLM;  
  CLASS VAR FERT BLOCK;  
  MODEL YIELD = BLOCK VAR FERT VAR*FERT;  
  MEANS VAR FERT VAR*FERT / LSD LINES;
```

NOTES: The RCB design is specified by the addition of a term for blocks. The MEANS statement works as before.

## **Example 3 - Split Plot:**

```
PROC GLM;  
  CLASS VAR FERT BLOCK;  
  MODEL YIELD = BLOCK VAR VAR*BLOCK FERT VAR*FERT;  
  
  TEST H=VAR E=VAR*BLOCK;  
  
  CONTRAST 'VAR1 vs VAR2' VAR 1 -1 / E=VAR*BLOCK;  
  
  MEANS VAR / DUNCANS E=VAR*BLOCK;  
  MEANS FERT VAR*FERT / DUNCANS;
```

NOTES: An error term for main plots has been added to the model (VAR\*BLOCK). A TEST statement is now required so SAS will test the main plot effect VAR with the new error term. Also the MEANS statement must now be done in two pieces because of the more complex error structure. This would also be true for LSMEANS. In this case the DUNCANS separation procedure has been used. A CONTRAST statement has also been added to demonstrate a single df contrast between the two varieties. Here again the correct error term must be specified.

## **Example 4 - Split Block:**

```
PROC GLM;  
  CLASS VAR FERT BLOCK;  
  MODEL YIELD = BLOCK VAR VAR*BLOCK FERT FERT*BLOCK  
           VAR*FERT;  
  
  TEST H=VAR E=VAR*BLOCK;  
  TEST H=FERT E=FERT*BLOCK;  
  
  MEANS VAR / LSD LINES E=VAR*BLOCK;  
  MEANS FERT / LSD LINES E=FERT*BLOCK;  
  MEANS VAR*FERT;
```

NOTES: In this last example the number of error terms has been increased one more level due to an additional restriction on the randomization and both the TEST and MEANS statements reflect this.