SAS 7. Analysis of variance

When you have more than two groups, a t-test (or the nonparametric equivalent) is no longer applicable. Instead, we use a technique called analysis of variance. This chapter covers analysis of variance designs with one or more independent variables, as well as more advanced topics such as interpreting significant interactions, and unbalanced designs.

1. One-Way Analysis of Variance

The method used today for comparisons of three or more groups is called analysis of variance (ANOVA). This method has the advantage of testing whether there are any differences between the groups with a single probability associated with the test. The hypothesis tested is that all groups have the same mean. Before we present an example, notice that there are several assumptions that should be met before an analysis of variance is used.

Essentially, we must have independence between groups (unless a repeated measures design is used); the sampling distributions of sample means must be normally distributed; and the groups should come from populations with equal variances (called homogeneity of variance).

***Example**:* 15 Subjects in three treatment groups X,Y and Z.

X	Y	Z
700	480	500
850	460	550
820	500	480
640	570	600
920	580	610

The null hypothesis is that the mean(X)=mean(Y)=mean(Z). The alternative hypothesis is that the means are not all equal. How do we know if the means obtained are different because of difference in the reading programs(X,Y,Z) or because of random sampling error? By chance, the five subjects we choose for group X might be faster readers than those chosen for groups Y and Z.

We might now ask the question, "What causes scores to vary from the grand mean?" In this example, there are two possible sources of variation, the first source is the training method (X,Y or Z). The second source of variation is due to the fact that individuals are different.
SUM OF SQUARES total;
SUM OF SQUARES between groups;
SUM OF SQUARES error （within groups）;

F ratio = MEAN SQUARE between groups/MEAN SQUARE error
= (SS between groups/(k -1)) / (SS error/(N-k))

SAS codes:

```
DATA READING;
      INPUT GROUP $ WORDS @@;
DATALINES;
X 700 X 850 X 820 X 640 X 920
Y 480 Y 460 Y 500 Y 570 Y 580
Z 500 Z 550 Z 480 Z 600 Z 610
;
PROC ANOVA DATA=READING;
      TITLE 'ANALYSIS OF READING DATA';
      CLASS GROUP;
      MODEL WORDS=GROUP;
      MEANS GROUP;
RUN;
```

The ANOVA Procedure
Dependent Variable: words

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	215613.3333	107806.6667	16.78	0.0003
Error	12	77080.0000	6423.3333		
Corrected Total	14	292693.3333			

* Following the word MODEL is our dependent variable or a list of dependent variable. MEANS GROUP will give us the mean value of the dependent variable(words) for each level of group

Now that we know the reading methods are different, we want to know what the differences are. Is X better than Y or Z? Are the means of groups Y and Z so close that we cannot consider them different? In general , methods used to find group differences after the null hypothesis has been rejected are called **post hoc**, or **multiple comparison test**. These include Duncan's multiple-range test, the Student-Newman-Keuls' multiple-range test, least significant-difference test, Tukey's studentized range test, Scheffe's multiple-comparison procedure, and others.

To request a post hoc test, place the SAS option name for the test you want, following a slash (/) on the MEANS statement. The SAS names for the post hoc tests previously listed are DUNCAN, SNK, LSD, TUKEY, AND SCHEFFE, respectively.

For our example we have:

```
      MEANS GROUP / DUNCAN;
```

Or

```
      MEANS GROUP / SCHEFFE ALPHA=.1
```

At the far left is a column labeled "Duncan Grouping." Any groups that are not significantly different from one another will have the same letter in the Grouping column.

The ANOVA Procedure
Duncan's Multiple Range Test for words

NOTE: This test controls the Type I comparison wise error rate, not the experiment wise error rate.

Alpha	0.05
Error Degrees of Freedom	12
Error Mean Square	6423.333

Number of Means	2	3
Critical Range	110.4	115.6

```
Means with the same letter are not significantly different.

Duncan Grouping          Mean      N    group

               A        786.00      5    x

               B        548.00      5    z
               B
               B        518.00      5    y
```

On the right are the group identification, The order is determined by the group means from highest to lowest. At the far left is a column labeled ”Duncan Grouping”. Any groups that are not significantly different from one another will have the same letter in the grouping column. In our example, the Y and Z groups both have the letter “B” in the groping column and are therefore not significantly different. The letter “B” between Group Z and group Y is there for visual effect. It helps us realize that groups Y and Z are not significantly different (at 0.5 level). Group X has an A in the grouping column and is therefore significantly different from Y and Z groups.
From Duncan multiple-range test, WE conclude that

1. method X is superior to both methods Y and Z
2. method Y and Z are not significantly different

2. Computing Contrasts

Suppose you want to make some specific comparisons. For example, if method X is a new method and methods Y and Z are more traditional methods, you may decide to compare method X to the mean of method Y and method Z to see if there is a difference between the new and traditional methods. You may also want to compare method Y to method Z to see if there is a difference. These comparisons are called **contrasts**, **planned comparisons**, or a **priori comparisons**.

To specify comparisons using SAS software, you need to use PROC GLM (General Linear Model) instead of PROC ANOVA. PROC GLM is similar to PROC ANOVA and uses many of the same options and statements. However, PROC GLM is a more generalized program and can be used to compute contrasts or to analyze unbalanced designs.

```
PROC GLM DATA=READING;
      TITLE 'ANALYSIS OF READING DATA -- PLANNED COMPARIONS';
      CLASS GROUP;
      MODEL WORDS = GROUP;
      CONTRAST 'X VS. Y AND Z' GROUP -2 1  1;
      CONTRAST 'Method Y VS. Z' GROUP 0 1 -1;
RUN;
```

* the rules are simple: (1) the sum of the coefficients must add to zero (2) the order of the coefficients must match the alphanumeric order of the levels of the Class variable if it is not formatted (3) A zero coefficient means that you do not want to include the corresponding level in the comparison (4) levels with negative coefficient are compared to levels with positive coefficients
The first contrast statement in the previous program gives you a comparison of method X against the mean of methods Y and Z. the second contrast statement will only perform a comparison between methods Y and Z.

The GLM Procedure

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
X VS. Y AND Z	1	213363.3333	213363.3333	33.22	<.0001
METHOD Y VS Z	1	2250.0000	2250.0000	0.35	0.5649

*Notice that method X is shown to be significantly different from methods Y and Z combined, and there is no difference between methods Y and Z at the 0.05 level