Comparing Means of Two or More Groups with PROC GLM

by Valérie Boquia

So far, we have discussed how to know whether differences between two groups are large enough to be statistically significant (using proc TTEST). In this newsletter, we want to compare several groups. Suppose we ran an experiment to compare the effects of 3 treatments (control, Thyroxin and Thiouracil) on rat growth. This can be translated in testing the null hypothesis that the mean weight of rats is equal whatever the treatment they received (Thyroxin, Thiouracil or the control) against the alternative that at least two means are different concluding that the 3 treatments do not have the same effect on rat weight.

$H_0 : \mu_{CONTROL} = \mu_{THYROXIN} = \mu_{THIOURACIL}$

$H_a : \exists\ i, j , i \neq j , \mu_i \neq \mu_j$

Based on the experiment, we would like to know whether we could reject the null hypothesis or not. Indeed, small differences between sample means are usually present but are they significant? Notice that the general alternative doesn't specify which means are different from one other but only that some differences exist.

In a statistical test to compare several groups, the proceeding hypotheses are tested by partitioning the total variation in the data into variation due to differences between groups and variation due to error. The error variation doesn't refer to mistakes in the data but to the natural variation within a group and possibly variation due to other factors that weren't considered in the experiment. Because this test consists of analyzing the variation (or variance) in the data, it is referred to as an analysis of variance and is abbreviated as ANOVA. The specific case of considering only the variation due to one factor (in this case the treatment) is called one-way ANOVA. The idea behind this test is that the error variation represents the natural variation that would be expected by chance, so if the variation between groups is large relative to the error variation, the group means are likely to be different.

As for comparing two group means, there are 3 assumptions for an analysis of variance:

- Observations are independent
- Observations are sampled from a normal distribution
- Groups have equal variances

To perform an analysis of variance with SAS, you can use either proc ANOVA or proc GLM. The following code is using proc GLM on the data:

```
data ratgrwth;
   length trt $ 10;
   input trt$ weight @@;
datalines;
Control     153    Control    140 Thyroxin    144    Thyroxin   177
Control     164    Control    141 Thyroxin    191    Thyroxin   165
Control     185    Control    151 Thyroxin    172    Thyroxin   160
Control     169    Control    172 Thiouracil  119    Thiouracil 107
Control     177    Control    154 Thiouracil  133    Thiouracil 140
Thyroxin    171    Thyroxin   140 Thiouracil  108    Thiouracil 122
Thyroxin    138    Thyroxin   189 Thiouracil  122    Thiouracil 138
Thiouracil  129    Thiouracil 142
;
run;

proc glm data=ratgrwth ;
    class trt;
    model weight=trt;
    means trt /* / hovtest */;
    title 'Comparing the effect of 3 treatments on rat weight';
run;
```

The CLASS statement lists the variable that classifies the data into groups (in this case the 3 treatments control, Thyroxin and Thiouracil); it must precede the MODEL statement. The MODEL statement describes the relationship you want to investigate. In this case, you want to find out how much of the variation in the rat weights (`WEIGHT`) is due to differences among the treatments (`TRT`). The name of the continuous dependent variable (`WEIGHT`) is placed to the left of an equal sign. The name of the categorical independent variable (`TRT`) is placed to the right. The MEANS statement is optional but very useful. It instructs proc GLM to print the number of observations, the mean and the standard deviation of the dependent variable (`WEIGHT`) for each of the groups defined by the independent variable (`TRT`).

Interpreting the output may look complicated but let's start at the beginning. The first section of the output lists the name of the independent variable (`TRT`), the number of levels it has and the numeric or character code for each level and the number of observations in the data set under the heading '`Class Level Information`'. If either the dependent or the independent variable on an observation has a missing value, proc GLM will not include that observation in the analysis and will print a note that tells you how many observations were used in the analysis (the note does not appear in this case as there are no missing value).

The next section of the output begins with the name of the dependent variable (`WEIGHT`) just to the right of the label '`Dependent Variable:`'. Underneath this line is an ANOVA table that gives the sum of squares explained by the model and associated statistics in the row labeled '`Model`', the error sum of squares in the row labeled '`Error`', and the total sum of squares in the row labeled '`Corrected Total`'. To find out the results of the hypothesis test for comparing groups, find the row of output labeled '`Model`' and look at the column labeled '`F Value`' for the Fisher statistic and '`Pr>F`' for the associated p-value. Just behind, you will find some measures of fit of the linear model for the ANOVA to the data, these include '`R-Square`' statistic that represents the proportion of the total variation explained by the model.

For one-way ANOVA, the explained sum of squares is due to only one variable or factor. Thus, the information on the sources of variation in the model is identical to that in the '`Model`' row of the ANOVA table.

Let's come back to the results for this example, the P-value for the comparison of means is 0.0001 (in the ANOVA table, the column labeled '`Pr>F`' in the row labeled '`Model`'). When you compare it to the reference probability value 0.05, it is smaller so you reject the null hypothesis of equality of means and conclude that the mean rat weight is different among the different treatments. So at least one treatment (control, Thyroxin or Thiouracil) has a different effect on rat weights. A next newsletter will cover how to see, based on the fact we know there is at least one difference between groups, which groups are significantly different from which others.

Be careful that the conclusion of the ANOVA is only valid if the 3 assumptions made are acceptable namely that the observations are independent, observations for each group are a random sample from a population with a normal distribution (this should be verified by a test for normality), variances for the two independent groups are equal (this can be checked within the proc GLM using the option '`hovtest`' in comment in the MEANS statement).