Print

# Summary: Lesson 3: Analysis of Variance

This summary contains topic summaries, syntax, and sample programs.

## Topic Summaries

*To go to the movie where you learned a task or concept, select a link.*

### ANOVA Review

Analysis of variance, or ANOVA, is a statistical technique that is used to compare the means of two or more groups of observations or treatments. You should have a continuous dependent variable and one or more discrete independent variables (also called predictor or explanatory variables). They separate your data into several groups.

There are three assumptions for ANOVA that you need to verify before you perform the hypothesis test. Remember that the assumption of independent observations means that there is no correlation between any one observation in the data set and another. The assumption that the data for each group is approximately normal can be verified by examining plots of the data. The assumption of constant variances can be checked by looking at descriptive statistics and plots of the data and by conducting a test for constant variances. If these assumptions are not valid, the probability of drawing incorrect conclusions from the analysis might be increased.

When an ANOVA interaction is not statistically significant you can analyze the main effects with the model in its current form. This is generally the method that you use when you analyze designed experiments. However, even when you analyze designed experiments, some statisticians might suggest that if the interaction is not significant, you can delete the interaction effect from your model, rerun the model, and then analyze the main effects only. This increases the power of the main effects test. The approach that you choose might depend on your subject-matter knowledge of the data and whether you think you should include the nonsignificant interaction term. If the interaction term is significant, it is good practice to keep the main effect terms that make up the interaction in the model, whether they are significant or not. This preserves model hierarchy.

When you fit a two-way ANOVA model, you examine the effects of two predictor variables simultaneously. At the same time, you might also be interested in determining whether the two predictor variables interact with respect to their effects on the response variable. Interaction is said to occur when the effect of one variable on the response variable depends on the levels of the other variable.

PROC GLM uses the method of least squares to fit a general linear model of which ANOVA is a special case. The basic syntax of PROC GLM for fitting a two-way ANOVA model includes the following options. The PLOTS= option controls the plots produced through ODS Graphics. The CLASS statement enables you to specify the classification variables for the analysis and the MODEL statement specifies dependent and independent variables.

The MEANS statement computes the arithmetic means and standard deviations of all continuous variables in the model (both dependent and independent) for each effect listed in the MEANS statement. The LSMEANS statement computes least squares means for each effect listed in the LSMEANS statement. The STORE statement enables you to store the results of your analysis for additional processing by PROC PLM.

```
PROC GLM <options> PLOTS(global-plot-options)=
                    (plot-request (specific-plot-options));
    CLASS variables;
    MODELdependents=independents </ options>;
    MEANS effects </ options>;
    LSMEANS effects </ options>;
    STORE OUT= item-store-name </LABEL='label'>;
RUN;
```

## Postfitting Analyses

The STORE statement requests that the procedure save the context and results of the statistical analysis into an item store. An item store is a binary file format that cannot be modified by the user. The contents of the item store can be processed with the PLM procedure. The syntax of the STORE statement requires you to mention the name of the item store, which is a one-or two-level SAS name, and similar to a SAS data set name. A two-level SAS name, such as libname.membername, is preferred so that the item store can be referred to from the respective library at a later stage.

The PROC PLM statement invokes the PLM procedure. The PLM procedure, unlike most SAS/STAT procedures, does not operate primarily on an input data set. Instead, the procedure requires you to specify an item store with the RESTORE= option in the PROC PLM statement.

The LSMEANS statement computes and compares the least squares means of fixed effects. The LSMESTIMATE statement provides a mechanism for obtaining custom hypothesis tests among least squares means. The SHOW statement uses the Output Delivery System to display contents of the item store. Use the WHERE statement in the PLM procedure when the item store contains BY-variable information and you want to apply the PROC PLM statements to only a subset of the BY groups.

```
PROC PLM RESTORE=item-store-specification <options>;
    LSMEANS <model-effects> </ options>;
    LSMESTIMATE model-effect <'label'> values
        < divisor=n><,...<'label'> values
        < divisor=n> ;
    SHOW options;
    WHERE expression;
RUN;
```

## Evaluations of Model Assumptions and Remedial Measures

Broadly speaking, the goal of testing statistical hypotheses is to determine whether a claim or conjecture about some feature of the population, a parameter, is strongly supported by the information obtained from the sample data. In the language of statistics, the claim or the research hypothesis that we want to establish is called the alternative hypothesis $H_1$. The opposite statement, one that nullifies the research hypothesis, is called the null hypothesis, $H_0$. The word "null" in this context means that the assertion that we are seeking to establish is actually void.

In hypothesis testing, we can identify two types of potential errors, labeled Type I error and Type II error. The Type I error rate, often denoted as $\alpha$, is the probability of wrongly rejecting the null hypothesis when $H_0$ is true. The Type I error rate is also called the significance level of a test. A customary level for alpha for a hypothesis test is 0.05.

The Type II error rate, often denoted as $\beta$, is the probability of failing to reject the null hypothesis when $H_0$ is false. The power of a statistical test is equal to $1-\beta$. This is the probability that you correctly reject the null hypothesis, or the probability of detecting a true effect. You prefer that your tests have a low Type I error rate and a high power. However, $\alpha$ and $\beta$ cannot be determined independently of each other. They also depend on the sample size and the standard errors of the test of interest.

Violation of the independence assumption affects the accuracy of the standard errors and thus the results of significance tests. If the observations are positively correlated then standard errors might be underestimated which can lead to test statistics that are too large. Significance might be detected when it is not truly there which results in an increase in Type I error rate. If the observations are negatively correlated, then the opposite situation arises. That is, standard errors are overestimated. Test statistics might be too small and true significance is not detected. Thus, the result for negatively correlated data is that power might suffer and the ability of the statistical test to detect a true difference is reduced.

ANOVA is robust against departures from normality especially with a large enough sample size. This means that the probability of incorrectly rejecting the null hypothesis is not appreciably increased over the set alpha value. But power might suffer when the normality assumption is violated. This means that the probability of rejecting the null hypothesis, when it is false, decreases, which is nothing but the power of the test. Hence, the ability of the test to detect a true difference is reduced.

When the sample sizes are equal ANOVA is robust against non-constant variances. However, when sample sizes of the groups are unequal, the effect of non-constant variances is more pronounced. If the variances of the groups with a larger sample size are larger, the ANOVA loses power. (That is, the ability of the test to detect a true difference is reduced.) On the other hand, if the variances of the groups with the smaller sample sizes are larger then the Type I error rate might increase. (That is, the probability of incorrectly rejecting the null hypothesis is increased.)

Correlated observations can arise in data from a complex survey design, any type of clustered data, repeated measures on a given subject, or data gathered over time. Data can be correlated even when measurements are not taken on the same subjects.

There are several ways of checking for the normality assumption. Normal probability plots of the residuals and histograms of the residuals are graphical methods to evaluate the normality of the data and are available as part of the ODS Graphics output for PROC GLM. Normal probability plots graph the distribution of the residuals against how the residuals would be distributed if they were normally distributed.

PROC UNIVARIATE can be used to determine whether your data is normal. This procedure generates a variety of summary statistics, such as the mean and median, as well as numerical representations of properties such as skewness and kurtosis. The NORMAL option in PROC UNIVARIATE produces a table with tests for normality. In general, if the *p*-values are less than 0.05, then the data should be considered non-normally distributed.

```
PROC UNIVARIATE <options>;
```

For a one-way analysis of variance, several formal statistical tests were developed to evaluate the homogeneity of the variances. Among the tests that are available in PROC GLM are tests developed by Bartlett, Brown, and Forsythe, Levene, and O'Brien. It should be noted that Bartlett's test should be used only if the data is normally distributed, and the default Levene’s test is considered to be the standard test.

For data gathered from a complex survey design, you can use PROC SURVEYREG with a CLASS statement to perform the ANOVA. If the assumption of normality is violated, transformation of the response variable often normalizes the data and an ANOVA can then be conducted on the transformed data. Transformations might also correct unequal variance problems. When appropriate, you can use PROC GENMOD or PROC GLIMMIX with the appropriate distribution and link function to fit a generalized linear model to nonnormal data.

```
PROC SURVEYREG <options>;
    CLASS variables;
```

```
PROC GENMOD <options>;
```

```
PROC GLIMMIX <options>;
```

When data is extremely skewed or there are extreme outliers, transformation of the data might not correct the problem. In this case, nonparametric ANOVA might be appropriate. The NPAR1WAY procedure can be used to fit a nonparametric one-way ANOVA.

```
PROC NPAR1WAY <options>;
```

For two-way or higher-ordered nonparametric ANOVA, you might consider ranking your dependent variable and use PROC GLM to perform ANOVA on ranks (Iman 1988 and Iman 1982). When variances are unequal, you can use PROC GENMOD, PROC MIXED, or PROC GLIMMIX to model the nonconstant variances. If the variances are unequal and it is a one-way ANOVA, then Welch's variance-weighted ANOVA can be used. You might also use nonparametric ANOVA for this situation. When variances are unequal, another approach is to transform the response variable to stabilize the variances.

PROC GLM performs statistical analyses for general linear models and assumes independence, normality, and constant variance. However, this procedure has only one tool to account for nonconstant variance: Welch's ANOVA for one-way analysis of variance. On the other hand, PROC GLIMMIX and PROC GENMOD provide a more general approach and do not require the assumptions of normality, independence, or constant variance.

PROC GLIMMIX fits statistical models to data with correlations or nonconstant variability and where the response is not necessarily normally distributed. These models are known as generalized linear models.

The PROC GLIMMIX statement invokes the GLIMMIX procedure and enables you to specify a data set and various options. The CLASS statement enables you to specify the classification variables to be used in the analysis. It must appear before the MODEL statement and can appear only once. The MODEL statement is required and names the dependent variable and the fixed effects. This statement can appear only once. The RANDOM statement with the _RESIDUAL_ keyword defines R, the residual variance-covariance matrix. You can specify multiple RANDOM statements. The GROUP= option in the RANDOM statement estimates the covariance parameters by groups. The COVTEST statement provides a mechanism to obtain statistical inferences for the covariance parameters. You can specify multiple COVTEST statements.

**PROC GLIMMIX** *<options>***;**
  **CLASS** *variables***;**
  **MODEL** *response<(response-options)>=<fixed-effects>*
    *</ options>***;**
  **RANDOM_RESIDUAL_** *</ options>***;**
  **COVTEST** *<'label'> <test-specification> </ options>***;**
**RUN;**

**RANDOM** *effect* **/ GROUP=** *variables***;**

## Sample Programs

### Performing a Two-Way Analysis of Variance

```
title "MYDATA.school DATA SET";
proc glm data=mydata.school plots(unpack)=all;
   class school gender;
   model reading3=school|gender;
run;
quit;
title;

proc glm data=mydata.school;
   class school gender;
   model reading3=school|gender;
   lsmeans school*gender / pdiff adjust=tukey cl;
run;
quit;

proc glm data=mydata.school;
   class school gender;
   model reading3=school|gender;
   lsmeans school*gender / slice=gender slice=school;
run;
quit;
```

### Using the LSMESTIMATE Statement to Estimate Relationships of Interest

```
ods select none;
proc glm data=mydata.school;
```

```
   class school gender;
   model reading3 = school|gender;
   store out=mydata.schoolstore;
run;
quit;
ods select all;
proc plm restore=mydata.schoolstore;
   lsmestimate school*gender 'Female Cottonwood&Dogwood vs. Female
                      Maple&Pine'
                      .5 0 .5 0 -.5 0 -.5 0 / elsm;
   lsmestimate school 'Cottonwood vs. Dogwood, Maple and Pine'
                       1 -0.333333 -0.333333 -0.333333;
   lsmestimate school 'Cottonwood vs. Dogwood, Maple and Pine'
                       3 -1 -1 -1 / divisor=3;
run;
```

### Identifying Violations of ANOVA Assumptions

```
proc glm data=mydata.school plots(unpack)=diagnostics;
   class gender semesters school;
   model reading3=gender semesters school gender*school;
   output out=check r=residuals p=predicted;
run;
quit;

goptions reset=all;
ods select moments BasicMeasures GoodnessOfFit;
proc univariate data=check;
   var residuals;
   histogram / normal;
run;

data school;
   set mydata.school;
   group=compress(gender||school||semesters);
run;

ods select classlevels hovftest means;
proc glm data=school;
   class group;
   model reading3=group;
   means group / hovtest;
run;
quit;
```

### Looking for Unequal Variances

```
ods graphics / imagemap=on;
ods select modelanova overallanova QQPlot ResidualHistogram boxplot
   HOVFTest;
proc glm data=mydata.pressure2 plots (unpack)=all;
   class drug;
   model bpchange=drug;
   means drug / hovtest;
   id drug;
run;
quit;
```

Close