Ch.7 ANOVA

Outline

1. One-Way Analysis of Variance
   (a) **Using** PROC GLM and PROC ANOVA
   (b) **Using** PROC NPAR1WAY
   (c) Post-Hoc Comparisons for One-Way ANOVA
   (d) Computing Contrasts
2. Two-Way Analysis of Variance
3. Interpreting Significant Interactions
4. N-way Factorial Designs
5. Analysis of Covariance
   This material covers sections 7BCDEFGH and 6D.
PROC GLM

- **PROC GLM (General Linear Models)** is the general regression and ANOVA procedure in SAS. It is appropriate for any univariate analysis, and there is a facility for MANOVA (multivariate analysis of variance).
- **Although PROC GLM** can be used for multiple linear regression, **PROC REG** is more efficient for this purpose.
- **PROC ANOVA** is more computationally efficient for analysis of variance problems involving balanced designs (equal sample sizes for each treatment group).
- **For general ANOVA and ANCOVA** (analysis of covariance) problems, **PROC GLM** must be used.
One-Way Analysis of Variance
Unbalanced Case – PROC GLM

- To compare the means for more than two independent, normally distributed samples of unequal size, PROC GLM should be used.

- Example: The distance (in m) required to stop a car going 50 km/hr on wet pavement was measured several times for each of three brands of tires to compare the traction of each brand. The same vehicle was used for each measurement. The resulting distances were recorded in a file called tract.txt.
<table>
<thead>
<tr>
<th>BRAND</th>
<th>DISTANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>41</td>
</tr>
<tr>
<td>M</td>
<td>43</td>
</tr>
<tr>
<td>M</td>
<td>44</td>
</tr>
<tr>
<td>M</td>
<td>44</td>
</tr>
<tr>
<td>B</td>
<td>44</td>
</tr>
<tr>
<td>B</td>
<td>43</td>
</tr>
<tr>
<td>G</td>
<td>39</td>
</tr>
<tr>
<td>G</td>
<td>41</td>
</tr>
<tr>
<td>G</td>
<td>40</td>
</tr>
<tr>
<td>G</td>
<td>41</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BRAND</th>
<th>DISTANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>46</td>
</tr>
<tr>
<td>M</td>
<td>40</td>
</tr>
<tr>
<td>M</td>
<td>42</td>
</tr>
<tr>
<td>M</td>
<td>49</td>
</tr>
<tr>
<td>B</td>
<td>46</td>
</tr>
<tr>
<td>B</td>
<td>44</td>
</tr>
<tr>
<td>G</td>
<td>42</td>
</tr>
<tr>
<td>G</td>
<td>44</td>
</tr>
<tr>
<td>G</td>
<td>43</td>
</tr>
</tbody>
</table>
Read data

* To read this data into a data set called TRACT, use

FILENAME TRACTION 'TRACT.TXT';
    DATA TRACT;
    INFILE TRACTION;
    INPUT BRAND $ DISTANCE;
    RUN;

* We wish to compare the stopping distance of the three brands. In other words, we wish to know whether mean DISTANCE depends on BRAND.
* Plot the data first. The plot may indicate that DISTANCE does not depend upon BRAND. It may also indicate departures from the model assumptions – look for outliers and indications of nonconstant variance.

PROC PLOT;
   PLOT DISTANCE*BRAND;
RUN;

* The plot gives some indication that the stopping distance distributions are not all the same.
Calculating means

* The means for each brand will be different. We can calculate them, noting that **BRAND** is a CLASSification variable (or factor) and **DISTANCE** is a response (or dependent) variable.

PROC MEANS MEAN;
   VAR DISTANCE;
   CLASS BRAND;
RUN;
The analysis of variance will help us to decide whether the observed differences among the three brands are significant. We must use PROC GLM, because the sample sizes for the different brands are unequal.

PROC GLM DATA=TRACT;
    CLASS BRAND;
    MODEL DISTANCE=BRAND;
RUN; QUIT;
Output

* The first page of output is a summary of the levels of the classification variable and the total number of experimental units in the study.
* The second page of output contains the analysis of variance table.
* The numerator and denominator degrees of freedom for the F-ratio are given in the $DF$ column:
  ** No. of treatment groups $= 3$, so numerator $DF = 3-1=2$.
  ** Total No. of observations $= 19$, so denominator $DF = 19-3= 16$. 
The sums of squares and mean squares:

* The sums of squares:

  ** SSW = SSE = Sum of Squares Error = 65.4 When divided by its degrees of freedom, this summarizes the variability observed within each treatment group.
  ** SSB = SSModel = Sum of Squares Model = 41.6 When divided by its degrees of freedom, this summarizes the variability observed between each treatment mean.

* The mean squares:

  ** MSE = SSE/DF = 4.09
  ** MSModel = SSModel/DF = 20.8
Discussion and Conclusion

* The $F$ ratio is the statistic used for testing the hypothesis that the mean DISTANCE does not differ for the different brands. $F = \frac{MS_{Model}}{MSE} = 5.09$.

* The p-value $= P(F > F)$, if the true population means are actually equal. A small p-value implies strong evidence against this hypothesis.

* The p-value for our test is .0195 so we reject the null hypothesis at the 5% level; we have strong evidence that the mean distance depends on brand.
One-Way Analysis of Variance
Balanced Case – PROC ANOVA

- To compare the means of more than two independent, normally distributed samples of equal size, PROC ANOVA should be used.

- Example: The file thiamin.txt contains measurements of thiamin content for 6 samples of 4 different cereal grains.
* To save space, more than one observation has been stored in each record. To read in such data, use the @@ symbol at the end of the INPUT statement.

WHEAT 5.2 WHEAT 4.5 WHEAT 6.0 WHEAT 6.1 WHEAT 6.7
WHEAT 5.8 BARLEY 6.5 BARLEY 8.0 BARLEY 6.1
BARLEY 7.5 BARLEY 5.9 BARLEY 5.6 MAIZE 5.8 MAIZE 4.7
MAIZE 6.4 MAIZE 4.9 MAIZE 6.0 MAIZE 5.2 OATS 8.3
OATS 6.1 OATS 7.8 OATS 7.0 OATS 5.5 OATS 7.2

* The data set THIAMIN will contain 24 observations on 4 variables which is the same as if the data had been entered into the file in the standard case-by-variable format.
PROC ANOVA

OPTIONS PAGESIZE=40;
/* THIAMIN.SAS */
/* thiamin.txt contains measurements of thiamin content for 6 samples of 4 different cereal grains */
DATA THIAMIN;
   INFILE 'thiamin.txt';
   INPUT GRAIN $ CONTENT @@;
PROC PLOT;
   PLOT CONTENT*GRAIN;
PROC ANOVA; /* can be used since the sample sizes are equal (6) */
   CLASS GRAIN;
   MODEL CONTENT=GRAIN;
RUN;
Analysis

* The numerator and denominator degrees of freedom for the F-ratio are given in the DF column:
  ** No. of treatment groups = 4, so numerator DF = 4-1=3.
  ** Total No. of observations = 24, so denominator DF = 24-4= 20.
* The sums of squares:
  ** SSW = SSE = Sum of Squares Error = 15.1
  ** SSB = SSModel = Sum of Squares Model = 8.98
* The mean squares:
  ** MSE = SSE/DF = .76
  ** MSModel = SSModel/DF = 2.99
Discussion and Conclusion

* The F ratio is the statistic used for testing the hypothesis that the mean CONTENT does not differ for the different brands. \( F = \frac{MS_{\text{Model}}}{MSE} = 3.96. \)

* The p-value \( = P(F > F) \), if the true population means are actually equal. A small p-value implies strong evidence against this hypothesis.

* The p-value for our test is .023 so we reject the null hypothesis at the 5% level; we have strong evidence that the mean distance depends on GRAIN.
One-Way Analysis of Variance
Non-Normal Case – PROC NPARIWAY

- To compare the distributions of two or more independent samples of unequal size, PROC NPARIWAY should be used, if the data is clearly not normally distributed for each level of the classification variable.

- In such cases, we can use a nonparametric method to determine whether the distribution of the response variable depends on the classification variable.
An experiment was conducted to compare four varieties of sweet potatoes. The four varieties were randomly assigned to 28 fields. The average yield per acre for each field was recorded in POTATO.TXT.

```
8.3 9.1 10.1 7.8
9.4 9.0 10.0 8.2
9.1 8.1 9.6 8.1
9.1 8.2 9.3 7.9
9.0 8.8 9.8 7.7
8.9 8.4 9.5 8.0
8.9 8.3 9.4 8.1
```
Analysis

* Each column contains the yields for one of the four varieties.
* The classification variable is VARIETY which has 4 levels (each corresponding to a potato variety). YIELD is the response variable.
* This data set is not organized in the usual case-by-variable format, but we can use a DO loop to make the conversion:
DATA POTATO

/*EXAMPLE FROM P. 581, FREUND AND WILSON, STATISTICAL METHODS*/
/* POTATO.SAS */
DATA POTATO;
INFILE 'POTATO.TXT';
INPUT VAR1-VAR4;
DO VARIETY=1 TO 4;
  IF VARIETY=1 THEN YIELD=VAR1;
  ELSE IF VARIETY=2 THEN YIELD=VAR2;
  ELSE IF VARIETY=3 THEN YIELD=VAR3;
  ELSE YIELD=VAR4;
OUTPUT;
/* Note that this assigns the yield for each of VAR1-VAR4 to the YIELD variable, and outputs a single VARIETY-YIELD observation to the SAS data set POTATO at each step of the DO loop */ END;
* The Kruskal-Wallis test can be used to determine whether the distribution of YIELD depends on VARIETY.
* PROC NPAR1WAY WILCOXON performs this test.

PROC PLOT;
   PLOT YIELD*VARIETY;
PROC NPAR1WAY WILCOXON;
   CLASS VARIETY;
   VAR YIELD;
RUN; QUIT;
Discussion and Conclusion

* Note the outlier in the first variety. This is an indication of non-normality, so the use of the nonparametric test is recommended here.
* The null hypothesis is that the yield distribution is the same for each variety. The test uses an approximate $\chi^2$ statistic. The p-value is computed by PROC NPAR1WAY.
* The output gives the p-value as .0001 which is very strong evidence against the null hypothesis. Therefore, we conclude that YIELD depends on VARIETY. In particular, note that the fourth variety has a lower mean score than the other varieties, indicating that it will usually yield less than the other varieties.
Simulation Experiments

• The ANOVA assumptions boil down to the following model:

\[ Y_{ij} = \mu_i + \varepsilon_{ij} \]

where \( \mu_i \) is the mean for the \( i \)th treatment group (\( i \)th level of the factor), and \( \varepsilon_{ij} \) is the within treatment random error which is normally distributed with mean 0 and variance \( \sigma^2 \).

\( Y_{ij} \) is the \( j \)th response in the \( i \)th treatment group. If the design is balanced, then \( j = 1, 2, \ldots, n \), for each \( i = 1, 2, \ldots, k \).
Let’s simulate $k = 3$ samples of size 5, using $\sigma^2 = 1$, and the means are all 0. We should expect to accept the null hypothesis with high probability.

DATA ANOVASIM;
 /* ANOVASIM.sas */
    MU1=0;
    MU2=0;
    MU3=0;
    SIGMA1=1;
    SIGMA2=1;
    SIGMA3=1;
    DO I = 1 TO 5;
    F_LEVEL = 1;
    Y=MU1 + SIGMA1*RANNOR(0);
    OUTPUT;
Simulation Cont’d

F_LEVEL = 2;
Y=MU2 + SIGMA2*RANNOR(0);
OUTPUT;
F_LEVEL=3;
Y=MU3 + SIGMA3*RANNOR(0);
OUTPUT;
END;
PROC ANOVA DATA=ANOVASIM;
CLASS F_LEVEL;
MODEL Y = F_LEVEL;
RUN; QUIT;
Questions

- What is the effect of one of the variances being much larger than the others? Say, \( \text{SIGMA1} = 10.0 \).
- What if \( \text{MU1} = 1 \)? What if \( \text{MU1} = 2 \)?
- What if \( \text{MU1} = 2, \text{and } \text{SIGMA1} = 5 \)?
- What if the one of the error distributions is a centered exponential?
- Now, try \text{PROC NPAR1WAY} on such data.
Post-Hoc Comparisons for One-Way ANOVA

- The one-way or one factor ANOVA is used to test for differences among several population means.
- Suppose measurements on some variable $V$ are taken from $K$ normal populations which may or may not have different means.
- The null hypothesis is $\mu_1 = \mu_2 = \cdots = \mu_K$.
- The alternative hypothesis is that there is a difference among the means.
- PROC GLM (or ANOVA) is used to conduct the test.
- Outcomes:
  * No difference (p-value is large)
  * Difference (p-value is small)
- If no difference is found, stop.
Post-hoc Tests

- If there is a difference, we can determine which means are different, using a *post-hoc* or multiple comparison test.
- There are $K(K-1)/2$ possible pairwise mean comparisons that can be conducted. It is desirable to control the over-all Type I error rate so that the probability of rejecting the null hypothesis for at least one of these tests is around $\alpha$. A number of techniques have been devised to accomplish this: e.g. Student-Newman-Keuls, Scheffe, etc.
Example: Student-Newman-Keuls Comparisons

Data are on total iron content in 4 types of iron formation: carbonate (1), silicate (2), magnetite (3) and hematite (4)). Is there a difference in the mean amount of iron by type? The number of measurements per type is equal. We can use PROC ANOVA.

DATA IRON;
INFILE 'iron.txt' FIRSTOBS=2;
INPUT CONTENT TYPE;
PROC ANOVA;
   CLASS TYPE;
   MODEL CONTENT = TYPE;
   MEANS TYPE / SNK;
* SNK gives us the Student-Newman-Keuls comparisons;
RUN; QUIT;
Example: Student-Newman-Keuls Comparisons

Output: ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>509.122000</td>
<td>169.707333</td>
<td>10.85</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>36</td>
<td>563.134000</td>
<td>15.642611</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>39</td>
<td>1072.256000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Example: Student-Newman-Keuls Comparisons**

Means with the same letter are not significantly different.

<table>
<thead>
<tr>
<th>SNK Grouping</th>
<th>Mean</th>
<th>N</th>
<th>TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>33.840</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>B</td>
<td>29.950</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>26.080</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>24.690</td>
<td>10</td>
<td>2</td>
</tr>
</tbody>
</table>

**Interpretation:** Types 3 and 4 are significantly different from each other and from types 1 and 2.
Example: Scheffe Comparisons

PROC ANOVA;
  CLASS TYPE;
  MODEL CONTENT = TYPE;
  MEANS TYPE / SCHEFFE ALPHA = .05; * type I error rate;
RUN; QUIT;
Output:

Means with the same letter are not significantly different.

<table>
<thead>
<tr>
<th>Scheffe Grouping</th>
<th>Mean</th>
<th>N</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>33.840</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B A</td>
<td>29.950</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B C</td>
<td>26.080</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>24.690</td>
<td>10</td>
<td>2</td>
</tr>
</tbody>
</table>

Interpretation: type 3 and 4 are not discernible; type 3 and 1 are not discernible and type 1 and 2 are not discernible.
Other types of comparisons that can be made:
* Tukey
* Bonferroni
* Dunnett
* Duncan
Earlier, we found that mean stopping distance depended on tire brand. We can conduct pairwise tests to see which brands differ from each other. For example, the Tukey method is used here.

```sql
PROC GLM DATA=TRACT;
CLASS BRAND;
MODEL DISTANCE=BRAND;
LSMEANS BRAND/ PDIFF adjust = TUKEY;
```
Different methods

- Tukey’s method is recommended since it controls the Type I error rate, and simulation studies have shown it to be more sensitive than the other tests.
- Duncan’s multiple range test can be used with **PROC ANOVA**. It cannot be used with **PROC GLM**.
- If there had been no pairwise differences among the means, there is a 5% probability that Tukey’s method would have led to a rejection of the null hypothesis in one of these tests (at the 5% level.)
Contrasts

• Sometimes, we don’t simply want to compare population means, but instead, we might want to compare specific linear combinations of the means, called contrasts.
• A contrast is a linear combination of the means whose coefficients add to 0.
• E.g. $\mu_1 + \mu_2 - 2\mu_3$.
• E.g. $\mu_1 + 2\mu_2 - 3\mu_3$.
• Exercise: Which of the following are contrasts?
  1. $\mu_1 - \mu_2$.
  2. $\mu_1 + \mu_2 + \mu_3$
  3. $3\mu_1 - 2\mu_2 - \mu_3$
Example: Contrasts

PROC GLM DATA = IRON;
    CLASS TYPE;
    MODEL CONTENT = TYPE;
    CONTRAST '1st versus Others' TYPE -3 1 1 1;
    CONTRAST '3rd versus 2nd/4th' TYPE 0 1 -2 1;
RUN; QUIT;
## Contrast Output

<table>
<thead>
<tr>
<th>Contrast</th>
<th>DF</th>
<th>Contrast SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st versus Others</td>
<td>1</td>
<td>87.38133333</td>
<td>87.38133333</td>
<td>5.59</td>
<td>0.0236</td>
</tr>
<tr>
<td>3rd versus 2nd/4th</td>
<td>1</td>
<td>3.12816667</td>
<td>3.12816667</td>
<td>0.20</td>
<td>0.6574</td>
</tr>
</tbody>
</table>
Two-Way Analysis of Variance

- **PROC GLM** and **PROC ANOVA** can be used for analysis of variance involving more than one factor.

- In the case of two or more factors, one must check for interaction effects among the factors.
Example

The calorie content (ENERGY) of six different brands of orange juice were determined by three different machines. The numbers below (from OJ.TXT) are the determination in calories per 6 fluid ounces. We are interested in knowing whether the caloric content differs for the different brands, but we also would like to take into account differences in the machines’ ability to measure caloric content.

```
M1  89  97  92  105  100  91
M1  94  96  94  101  103  92
M2  92  101  94  110  100  95
M2  90  100  98  106  104  99
M3  90  98  94  109  99  94
M3  94  92  96  107  97  98
```
Orange Juice Data

Now, there are two classification variables: MACHINE and BRAND. The response variable is the ENERGY measurement. The data are read into the data set OJ from the file OJ.TXT as follows.

/* ORANGE JUICE DATA - P. 415 FREUND AND SIMON */
/* OJ.SAS */
DATA OJ;
  INFILE 'OJ.TXT';
  INPUT MACHINE $ A B C D E F;
  /* convert to standard case-by-variable format: */
  DO BRAND=1 TO 6;
    IF BRAND=1 THEN ENERGY=A;
Sources of variation

ELSE IF BRAND=2 THEN ENERGY=B;
ELSE IF BRAND=3 THEN ENERGY=C;
ELSE IF BRAND=4 THEN ENERGY=D;
ELSE IF BRAND=5 THEN ENERGY=E;
ELSE ENERGY=F;
OUTPUT;
END;

Now, there are 4 possible sources of variation in ENERGY: BRAND, MACHINE, MACHINE*BRAND interactions and within treatment group variation.
The model involves two classification variables and an interaction term. Thus, to perform the analysis of variance, we use the following. Note that the sample sizes for each treatment combination are 2. If any were different from 2, then we would use PROC GLM.

PROC ANOVA;
   CLASS MACHINE BRAND;
   MODEL ENERGY=MACHINE BRAND MACHINE*BRAND;
RUN; QUIT;
Output

* ANOVA table: DF, SS, MS, F, and p-value for overall model

* Detailed ANOVA table: DF for each factor, and interaction SS and MS for each factor and interaction F ratio for each factor and interaction – this is used to test the three null hypotheses that mean ENERGY does not depend on BRAND, MACHINE, or an interaction between the two factors.

The corresponding p-values for each test, given that the other factors are in the model already.
**Conclusion**

* **MACHINE** gives a p-value of .0253 indicating strong evidence that **MACHINE** has an effect on the **ENERGY** measurements.

* **BRAND** gives a p-value of .0001 indicating strong evidence that even with the effect of **MACHINE** accounted for, the different brands have different mean energy content.

* **BRAND*MACHINE** gives a p-value of .2724 indicating no evidence that there is an interaction effect on energy content.
Summary

- The analysis of variance is used to test for differences in the means among 3 or more populations.
- **PROC GLM** must be used if the samples coming from the different populations or treatment groups are of different sizes. Otherwise, **PROC ANOVA** can be used.
- If the data are clearly not normally distributed, and the treatment groups are defined by the different levels of a single factor (i.e. a 1-way layout), then **PROC NPAR1WAY** is to be used.
- In any case, the factor is identified to SAS using the **CLASS** statement.
- In a two way layout, there are 2 classification variables, and a possible interaction between the 2 factors must be tested before testing for main effects.
Interpreting Significant Interactions

- When there are two or more factors, it is possible for the effects to be due to either or both of the factors (main effects) or to an interaction between the factors.

- When only the main effects are significant, interpretation is easy.
Example (fixed effects)

In the orange juice example above (assuming fixed effects), we saw that the effect on ENERGY due to interaction between MACHINE and BRAND was not significant, while the main effects were found to be significant. Interpretation:

* The means of the energy measurements differ for the different machines.
* The means of the energy measurements differ among the different orange juice brands.
Interaction between factors

- When there are effects due to an interaction between the factors, the analysis is a bit more complicated.

- Example. In exercise 7.3 (on page 232) of the textbook, tennis balls are tested in a machine to see how many bounces they can withstand before they fail to bound 30% of their dropping height. Two brands are tested, and age is taken into account.
Interaction: Tennis example

DATA TENNIS; /* tennis.sas */
   INFILE 'tennis.txt';
   INPUT AGE $ BRAND $ NBOUNCE @@;
PROC ANOVA;
   CLASS AGE BRAND;
   MODEL NBOUNCE = BRAND|AGE;
   /* AGE and BRAND are FIXED effects */
RUN; QUIT;
## Interaction: Tennis ANOVA Output

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Anova SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRAND</td>
<td>1</td>
<td>352.800000</td>
<td>352.800000</td>
<td>22.08</td>
<td>0.0002</td>
</tr>
<tr>
<td>AGE</td>
<td>1</td>
<td>2205.000000</td>
<td>2205.000000</td>
<td>138.03</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>AGE*BRAND</td>
<td>1</td>
<td>352.800000</td>
<td>352.800000</td>
<td>22.08</td>
<td>0.0002</td>
</tr>
</tbody>
</table>
Interaction: Tennis example

* On the basis of the output from the above program, we see that the interaction between \textit{AGE} and \textit{BRAND} is highly significant.

* It is important that this interaction effect be understood clearly \textit{before} interpreting the main effects (\textit{AGE} and \textit{BRAND} alone). An \textit{interaction} plot is useful for this purpose.
/* CREATE AN INTERACTION PLOT: */
PROC MEANS NWAY NOPRINT DATA=TENNIS;
  CLASS AGE BRAND;
  VAR NBOUNCE;
  OUTPUT OUT = MEANS MEAN=;  
PROC PLOT DATA=MEANS;
  PLOT NBOUNCE*BRAND=AGE;
RUN; QUIT;

* This is a plot of the mean of the response variable NBOUNCE against one of the two factors BRAND, for each level of AGE.
Interpretation of Interaction Plot

* It indicates that new brand W tennis balls bounce better than new brand P tennis balls, but old brand P tennis balls bounce much better than old brand W tennis balls.
* It indicates that new brand W tennis balls bounce better than new brand P tennis balls, but old brand P tennis balls bounce much better than old brand P tennis balls.
* Thus, in order to really know which brand of ball bounces better, we must know the age of the ball as well. We need to study old and new balls separately. Two t-tests could be used to make these comparisons.
**Analyzing Two Factors with One Way ANOVA**

* A better approach is to convert the two-way analysis of variance problem into a one-way analysis of variance in which each factor-level combination is treated as a level of a new single factor.
* We will create a new classification variable called **AGEBRAND** which will contain the values **NEW-P, NEW-W, OLD-P, and OLD-W**.
* **The concatenation operator** `||` can be used to do this *in the data step*: 
/* tennis1.sas */
DATA TENNIS1;
   INFILE 'tennis.txt';
   INPUT AGE $ BRAND $ NBOUNCE @@;
   AGEBRAND = AGE || BRAND;
PROC ANOVA;
   CLASS AGEBRAND;
   MODEL NBOUNCE = AGEBRAND;
   MEANS AGEBRAND / DUNCAN;
   /* DO NOT USE THE MEANS STATEMENT WITH
   PROC GLM */
RUN; QUIT;
One-Way Output

Means with the same letter are not significantly different.

<table>
<thead>
<tr>
<th>Duncan Grouping</th>
<th>Mean</th>
<th>N</th>
<th>AGEBRAND</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>75.200</td>
<td>5</td>
<td>New P</td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>62.600</td>
<td>5</td>
<td>Old P</td>
</tr>
<tr>
<td>C</td>
<td>45.800</td>
<td>5</td>
<td>Old W</td>
</tr>
</tbody>
</table>
Interpretation of One-Way Output

- This performs the one-way analysis of variance and tests for differences among all pairs of means. From the output, we see that there is not a significant difference between the means for each brand in case the balls are new, but there is a significant difference between the brands, if the balls are old.
N-Way Factorial Designs

- The analysis of variance for more than two factors proceeds in a similar manner to that of one and two factors. PROC GLM and PROC ANOVA may be used, as appropriate.

- Example. An experiment was performed to investigate the surface finish of a metal part. The experiment was a $2^3$ factorial design in the factors feed rate (A), depth of cut (B), and tool angle (C). There were 2 replicates for each factor-level combination. The data are in finish.txt.
## Table of Surface Finish Data

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Surface Finish Rep. 1</th>
<th>Surface Finish Rep. 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>12</td>
<td>15</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>16</td>
<td>14</td>
</tr>
</tbody>
</table>
FINISH.DAT

This data is read into the SAS data set SURFACE as follows.

/* EXAMPLE - P. 504, MONTGOMERY, STAT. QUALITY CONTROL */
FILENAME FINISH 'finish.txt';
  DATA SURFACE;
  INFILE FINISH;
  INPUT A B C SF1 SF2;
  DO REP=1 TO 2;
    IF REP=1 THEN SF=SF1;
    ELSE SF=SF2;
  OUTPUT;
  END;
* The following lines are required to perform the analysis of variance. Note that there are first order and second order interaction terms in the MODEL statement.

PROC GLM;
   CLASS A B C;
   MODEL SF=A B C A*B A*C B*C A*B*C;
RUN; QUIT;
More than two factors: Output

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>45.562500000</td>
<td>45.56250000</td>
<td>18.69</td>
<td>0.0025</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>10.562500000</td>
<td>10.56250000</td>
<td>4.33</td>
<td>0.0709</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>3.062500000</td>
<td>3.06250000</td>
<td>1.26</td>
<td>0.2948</td>
</tr>
<tr>
<td>A*B</td>
<td>1</td>
<td>7.562500000</td>
<td>7.56250000</td>
<td>3.10</td>
<td>0.1162</td>
</tr>
<tr>
<td>A*C</td>
<td>1</td>
<td>0.062500000</td>
<td>0.06250000</td>
<td>0.03</td>
<td>0.8767</td>
</tr>
<tr>
<td>B*C</td>
<td>1</td>
<td>1.562500000</td>
<td>1.56250000</td>
<td>0.64</td>
<td>0.4465</td>
</tr>
<tr>
<td>A<em>B</em>C</td>
<td>1</td>
<td>5.062500000</td>
<td>5.06250000</td>
<td>2.08</td>
<td>0.1875</td>
</tr>
</tbody>
</table>
More than two factors: Output

- We test the three-factor interaction first. The p-value for the test is .1875 which means that we have very weak evidence against the null hypothesis. That is, the three-factor interaction effect is not significant at the 5% significance level. The two-factor interaction effects can then be tested. Since these are not significant, then the main effects can be tested.
More than two factors: Output Cont’d

• There is strong evidence that $A$ factor (feed rate) is related to surface finish, moderate evidence that $B$ is related and no evidence that $C$ is related.
• We can test pairwise differences of means for all 4 $A$-$B$ factor level combinations using
  
  \[
  \text{LSMEANS } A \mid B /\text{PDIFF ADJUST=TUKEY;}
  \]
  
  * We see that the 1 1 combination is significantly different from all other combinations.
## Output from LSMEANS

Adjustment for Multiple Comparisons: Tukey

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>SF</th>
<th>LSMEAN</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-1</td>
<td>9.25</td>
<td>9.2500000</td>
<td>1</td>
</tr>
<tr>
<td>-1</td>
<td>1</td>
<td>9.5</td>
<td>9.5000000</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>11.25</td>
<td>11.2500000</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>14.25</td>
<td>14.2500000</td>
<td>4</td>
</tr>
</tbody>
</table>
### Output from LSMEANS

Least Squares Means for effect A*B

Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: SF

<table>
<thead>
<tr>
<th>i/j</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.9956</td>
<td>0.3349</td>
<td>0.0083</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.9956</td>
<td>0.4372</td>
<td>0.0112</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.3349</td>
<td>0.4372</td>
<td>0.0991</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.0083</td>
<td>0.0112</td>
<td>0.0991</td>
<td></td>
</tr>
</tbody>
</table>
Analysis of Covariance

• The analysis of variance is concerned with the problem of determining whether a particular quantitative variable is related to one or more (usually qualitative) factors.
• Sometimes the response variable is also related to some additional quantitative variable (called a covariate).
• The Analysis of Covariance (ANCOVA) is used to estimate factor effects over and above the effect of the covariate.
• ANCOVA can be performed using PROC GLM or PROC REG. In addition to the usual assumptions for ANOVA, it is also necessary that there be no interaction effect between the covariate and the factor.
In a study to determine the effect of weaning conditions on the weight of 9-week-old pigs (NWWT), weaning weights and nine-week weights were recorded for pigs from 3 litters. One of these litters was weaned at approximately 21 days (EARLY), the next at 28 days (MEDIUM) and the third at 35 days (LATE). Perform an ANCOVA using NWWT as the response variable, weaning time as the factor and weaning weight as the covariate. Are any assumptions violated?
Ancova Example: Pigs

Reading in the file:

/* pigs.sas */
DATA PIGS;
  INFILE 'pigs.txt';
  INPUT WEANTIME $ WEANWT1 $ WWT1-WWT7 WEANWT2 $ NWWT1-NWWT7;
  WWT = WWT1; NWWT = NWWT1; OUTPUT;
  WWT = WWT2; NWWT = NWWT2; OUTPUT;
  WWT = WWT3; NWWT = NWWT3; OUTPUT;
  WWT = WWT4; NWWT = NWWT4; OUTPUT;
  WWT = WWT5; NWWT = NWWT5; OUTPUT;
  WWT = WWT6; NWWT = NWWT6; OUTPUT;
  WWT = WWT7; NWWT = NWWT7; OUTPUT;
  KEEP WEANTIME WWT NWWT;
RUN;
Interaction effects

- Check for interaction effects between WWT and WEANTIME. If there is a significant effect here, then this means that the covariate is not independent of the levels of the factor.

  PROC GLM;
  CLASS WEANTIME;
  MODEL NWWT = WEANTIME WWT WEANTIME*WWT;
  RUN;

* The large p-value (.519) indicates that the interaction effect is not significant.
## Pigs data ANCOVA output

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>WEANTIME</td>
<td>2</td>
<td>77.2380952</td>
<td>38.6190476</td>
<td>3.23</td>
<td>0.0681</td>
</tr>
<tr>
<td>WWT</td>
<td>1</td>
<td>394.0805861</td>
<td>394.0805861</td>
<td>32.98</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>WWT*WEANTIME</td>
<td>2</td>
<td>16.3788118</td>
<td>8.1894059</td>
<td>0.69</td>
<td>0.5190</td>
</tr>
</tbody>
</table>
**ANCOVA**

We can proceed to check for main effects:

```
PROC GLM;
  CLASS WEANTIME;
  MODEL NWWT = WEANTIME WWT;
  LSMEANS WEANTIME / PDIFF ADJUST = TUKEY;
RUN; QUIT;
```

From the ANCOVA output, we see strong evidence that the covariate \( WWT \) is related to \( NWWT \), and that there is moderate evidence to suggest that \( WEANTIME \) is related to \( NWWT \) (\( p\)-value = .0591).
## Pigs data ANCOVA output

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>WEANTIME</td>
<td>2</td>
<td>77.2380952</td>
<td>38.6190476</td>
<td>3.36</td>
<td>0.0591</td>
</tr>
<tr>
<td>WWT</td>
<td>1</td>
<td>394.0805861</td>
<td>394.0805861</td>
<td>34.24</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>
Pigs data ANCOVA output

Least Squares Means for effect WEANTIME

Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: NWWT

<table>
<thead>
<tr>
<th>i/j</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>0.0005</td>
<td>0.3462</td>
</tr>
<tr>
<td>2</td>
<td>0.0005</td>
<td></td>
<td>0.0029</td>
</tr>
<tr>
<td>3</td>
<td>0.3462</td>
<td>0.0029</td>
<td></td>
</tr>
</tbody>
</table>
**Using Proc Reg for ANCOVA**

**ANCOVA** can also be done with **PROC REG**. In order to do this, one must construct ‘dummy’ variables.

```sas
DATA PIGS2;
   SET PIGS;
   IF WEANTIME = 'EARLY' THEN DO;
      WT1 = 0;
      WT2 = 1;
   END;
   IF WEANTIME = 'MEDIUM' THEN DO;
      WT1 = 1;
      WT2 = 0;
   END;
```

IF WEANTIME = ’LATE’ THEN DO;
   WT1 = 1;
   WT2 = 1;
END;
PROC REG;
   MODEL NWWT = WWT WT1 WT2;
RUN; QUIT;

The output is in the form of a regression model relating NWWT to WWT, WT1 and WT2.
Summary

- The analysis of variance is used to test for differences in the means among 3 or more populations.
- **PROC GLM** must be used if the samples coming from the different populations or treatment groups are of different sizes. Otherwise, **PROC ANOVA** can be used.
- If the data are clearly not normally distributed, and the treatment groups are defined by the different levels of a single factor (i.e. a 1-way layout), then **PROC NPAR1WAY** is to be used.