Lectures 8: Two-, Three-, and Four-Way ANOVA

Junshu Bao

University of Pittsburgh
Table of contents

Introduction

Example: Treating Hypertension
   2-Way ANOVA Model
   3-Way ANOVA Model
   log Transformation

Example: School Attendance among Australian Children
Two-Way ANOVA

- Two categorical variables and one continuous outcome variable:
  - Independent variable # 1: A
  - Independent variable # 2: B
- Two-way ANOVA model:
  \[ Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \]
  - \( Y_{ijk} \) is the \( k^{th} \) outcome in the \( i^{th} \) level of A and \( j^{th} \) level of B
  - \( \mu \) is the overall mean
  - \( \alpha_i \) is the main effect of the \( i^{th} \) level of A
  - \( \beta_j \) is the main effect of the \( j^{th} \) level of B
  - \( (\alpha\beta)_{ij} \) is the first-order interaction between A and B
  - \( \epsilon_{ijk} \sim N(0, \sigma^2) \) is the error term
Assumptions

2-way and 3-way ANOVA assumptions

- Observations are independent
- Observations in each cell are normally distributed.
- Observations in each cell have the same variance.
Example: Treating Hypertension

Maxwell and Delaney (2003) describe a study investigating three possible treatments for hypertension.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Description</th>
<th>Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug</td>
<td>Medication</td>
<td>Drug X, Drug Y, Drug Z</td>
</tr>
<tr>
<td>Biofeed</td>
<td>Psychological feedback</td>
<td>Present, Absent</td>
</tr>
<tr>
<td>Diet</td>
<td>Special diet</td>
<td>Present, Absent</td>
</tr>
</tbody>
</table>

- There are 12 possible combinations of the 3 treatments: $3 \times 2 \times 2$.
- 72 subjects suffering from hypertension were recruited for the study, with 6 being randomly allocated to each of the 12 treatment combinations.
- Outcome variable: blood pressure reading (after treatment)
Example (cont.)

The number of subjects in each of the treatment combinations:

<table>
<thead>
<tr>
<th>Biofeed</th>
<th>Drug</th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>X</td>
<td>170</td>
<td>173</td>
</tr>
<tr>
<td></td>
<td>Y</td>
<td>175</td>
<td>176</td>
</tr>
<tr>
<td></td>
<td>Z</td>
<td>165</td>
<td>166</td>
</tr>
<tr>
<td>No</td>
<td>X</td>
<td>180</td>
<td>181</td>
</tr>
<tr>
<td></td>
<td>Y</td>
<td>185</td>
<td>186</td>
</tr>
<tr>
<td></td>
<td>Z</td>
<td>170</td>
<td>171</td>
</tr>
</tbody>
</table>

Questions:

▶ Any difference in mean blood pressure for the different levels of the three treatments?

▶ Any significant interactions between the treatments?
Reading Data

- `_n_` in SAS: Automatic variable saved internally. Indicates which row of data is being processed.

- **array** statement:
  - Defines an array by specifying a name.
  - An array could be thought of as a vector, matrix, etc.
  - Specifies related variables, simplifies processing for repeat statements.

- **do** Loops:
  - Repeats SAS statements a fixed number of times.
  - Use an index variable that changes with each repetition.
  - *When using with an array; index starts with 1 and ends with number of variables in array.*

- **output** statement:
  - Writes an observation to the output dataset with the current values of all variables.
  - When included within a do loop, results in index # of obs.
Descriptive Statistics

```plaintext
proc tabulate data=hyper;
   class drug diet biofeed;
   var bp;
   table drug*diet*biofeed,
       bp*(mean std n);
   format diet $YN. biofeed $PA.;
run;
```

Note that in the `table` statement you first specify the rows (treatment combinations), `drug*diet*biofeed`, and then specify the column (outcome) and the statistics requested, `bp*(mean std n)`.
Test for Homogeneity of Variance

```
proc anova data=hyper;
  class cell;
  model bp=cell;
  means cell / hovtest;
run;
```

Recall that the “cell” variable was created to contain all the 12 combinations of the three treatments.

- Test statistic: $F = 1.01$
- p-value = 0.4452
- Fail to reject the null.
Example: Treating Hypertension

Two-Way ANOVA Models (1)

Before we consider the full three-way model, we will fit the two-way models.

```sas
proc anova data = hyper;
   class diet drug;
   model bp = diet drug diet*drug;
   format diet $YN.;
   means diet drug diet*drug;
   ods output means = twowayDIET_DRUG;
run;
```

- The `anova` procedure is specifically for balanced designs.
- The `model` statement specified the model: \( Y = x_1 \ x_2 \ x_1 * x_2 \). A shorthand way: \( Y = x_1 | x_2 \)
- The `means` statement generates a table of cell means
- The `ods` output statement saved the means in a SAS data set.
Lectures 8: Two-, Three-, and Four-Way ANOVA

Example: Treating Hypertension

2-Way ANOVA Model

SAS Output of Interest

- Model p-value (model utility test)
  - Simultaneous effects
  - \( H_0: \alpha_i = \beta_j = (\alpha\beta)_{ij} = 0 \) for all \( i \) and \( j \).

- Source p-values
  - Main effect of \( A \): \( \alpha_i \)'s
  - Main effect of \( B \): \( \beta_j \)'s
  - Interaction of \( A \) and \( B \): \( (\alpha\beta)_{ij} \)'s

- Notes:
  - Order of variables is specified in the model statement.
  - Source p-value quantifies how significant the corresponding effect is.
  - If the interaction effect is not significant, we can re-fit a smaller model with only main effects (‘Main Effect model’).
  - If the interaction IS significant, to interpret the interaction, we draw an interaction plot.
Test Results

- Simultaneous effects:
  - Test statistic $F = 10.07$
  - p-value $< 0.0001$

- Source p-values
  - Main effect of diet: p-value $< 0.0001$
  - Main effect of drug: p-value $= 0.0002$
  - Interaction diet*drug: p-value $= 0.1057$

* The interaction is not significant.
Interaction Plot

```sas
proc sgplot data=twowayDIET_DRUG;
    series y=mean_bp x=diet / group=drug;
run;
```

Observations from the interaction plot:

- Drug X is significantly different with (smaller than) Drug Y and Drug Z no matter special diet is present or absent.
- For each drug, having special diet reduces the blood pressure.
Two-Way ANOVA Models (2)

Now, consider another 2-way model:

```plaintext
proc anova data = hyper;
   class drug biofeed;
   model bp = drug|biofeed;
   format diet $YN.;
   means biofeed*drug;
   ods output means = twowaybiofeed_DRUG;
run;
```

- Simultaneous effects:
  - Test statistic $F = 4.75$
  - p-value = 0.0009

- Source p-values
  - Main effect of drug: p-value < 0.0014
  - Main effect of biofeed: p-value = 0.0058
  - Interaction diet*biofeed: p-value = 0.6002
Three-Way ANOVA

- Three categorical variables and one continuous outcome variable:
  - Independent variable # 1: A
  - Independent variable # 2: B
  - Independent variable # 3: C

- Three-way ANOVA full model:

\[ Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k \]
\[ + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} \]
\[ + (\alpha\beta\gamma)_{ijk} \]
\[ + \epsilon_{ijkl} \]

where \((\alpha\beta\gamma)_{ijk}\) is the second-order interaction between the three variables.
Three-Way ANOVA Model

```sas
proc anova data=hyper;
   class diet drug biofeed;
   model bp=diet|drug|biofeed;
   format diet $YN. biofeed $PA.;
   means diet*drug*biofeed;
   ods output means=outmeans;
run;
```
Three-Way ANOVA Model: Test Results

- Simultaneous effects:
  - Test statistic $F = 7.66$
  - p-value $< 0.0001$

- Source p-values
  - Main effects:
    - drug: p-value $< 0.0001$
    - diet: p-value $< 0.0001$
    - biofeed: p-value $= 0.0006$
  - First-order interactions:
    - diet*drug: p-value $= 0.0638$
    - diet*biofeed: p-value $= 0.6529$
    - drug*biofeed: p-value $= 0.4425$
  - Second-order interaction:
    - diet*drug*biofeed: p-value $= 0.0388$

- Note that the second-order interaction is significant, though none of the first-order interactions is significant.
Interpretation of Interactions

What does a significant second-order interaction mean?

- The first-order interaction between two of the variables differs in form or magnitude in different levels of the remaining variable.

- The presence of a significant second-order interaction means that there is little point in drawing conclusions about either the non-significant first-order interactions or the significant main effects.

The interpretation of main effects may be misleading.
Interaction Plots

To better understand the second-order interaction, we may create the interaction plot.

```r
proc sgpanel data=outmeans;
  panelby drug / rows = 1 ;
  series y=mean_bp x=biofeed / group=diet;
run;
```

Observations:

- **Drug X:** diet has a negligible effect when biofeedback is present, but substantially reduces blood pressure when biofeedback is absent.
- **Drug Y:** the situation is the reverse of drug X.
- **Drug Z:** the blood pressure drop when the diet is given and when it is not is approximately equal for both levels of biofeedback.
Log-Transformation

A significant high-order interaction may make interpretation of the results from a factorial analysis of variance difficult. In such cases, a transformation of the data may help.

```plaintext
data hyper;
   set hyper;
   logbp=log(bp);
run;
proc anova data=hyper;
   class diet drug biofeed;
   model logbp=diet|drug|biofeed;
   format diet $YN. biofeed $PA.;
   means diet*drug*biofeed;
run;
```

Now the second-order interaction is only marginally significant (p-value = 0.0447). We can fit a main effect only model to the log-transformed blood pressures.
Main Effect Model for Log(BP)

```r
proc anova data=hyper;
   class diet drug biofeed;
   model logbp=diet drug biofeed;
   means drug / scheffe cldiff lines;
run;
```

- Simultaneous test: p-value < 0.0001.
- Source p-values:
  - diet: p-value < 0.0001
  - drug: p-value = 0.0001
  - biofeed: p-value = 0.0009
- Pairwise comparison:
  - Drug X is significantly different with Drugs Y and Z
  - Drug Y and Drug Z are not significantly different
Balanced versus Unbalanced Designs

- Balanced designs have the same number of observations in each cell.
  - Can use proc anova or proc glm
- Unbalanced designs have different numbers of subjects in each cell.
  - Should use proc glm

Notes: proc anova is used for the analysis of balanced data only, with some exceptions including one-way ANOVA.
Sum of Squares

Balanced versus unbalanced designs:

- For balanced designs, it is possible to partition the total variation (SST) in the response variable into non-overlapping or orthogonal sums of squares representing factor main effects and factor interactions.
- For unbalanced designs, there is no unique way of finding sum of squares for each effect since these effects are no longer independent of each other.
  * Order matters! The sum of squares that can be attributed to a factor depends on which factors have already been allocated a sum of squares.
- There are different ways sum of squares are calculated which matter significantly if you have unbalanced designs.
  - Type I Sum of Squares
  - Type III Sum of Squares:
Type I Sum of Squares

- Sequential, forward.
- If A, B, AB order:
  - Effect of A is estimated given no effects in model.
  - Effect of B is estimated given A is in the model.
  - Effect of AB is estimated given A and B in model.
- Order is important. Preferred for unbalanced designs
  - Principle of parsimony (begin with simplest model)
  - Significance of interactions without main effects makes little sense.
Type III Sum of Squares

- Order of input does not matter.
  - Effect of A is estimated given B and AB in model
  - Effect of B is estimated given A and AB in model
  - Effect of AB given A and B in model.

- ‘Given all other effects are present’, p-values test whether the effect of a factor is significant.

- When balanced data, Type I = Type III.

Notes: Nelder (1977) and Aitkin (1978) are strongly critical of “correcting” main effects sums of squares for an interaction term involving the corresponding main effect and recommend to use Type I sums of squares.
Example: School Attendance among Australian Children

- Unbalanced design: different number of students within each cell.
- A sociological study of 154 Aboriginal and non-aboriginal children reported by Quine (1975)
  - Independent variables:
    1. Cultural origin (aboriginal, non-aboriginal)
    2. Four grade levels (F0, F1, F2, F3)
    3. Type of learner (SL ‘slow learner’, AL ‘average learner’)
    4. Gender (female, male)
- Dependent variable: number of days absent from school
- Design: 2 x 4 x 2 x 2 factorial (4-way ANOVA).
Four-Way ANOVA Model

The usual model for $Y_{ijklm}$, the number of days absent for the $i^{th}$ child in the $j^{th}$ sex group, the $k^{th}$ age group, the $l^{th}$ cultural group and the $m^{th}$ learning group is

$$Y_{ijklm} = \mu + \alpha_j + \beta_k + \gamma_l + \delta_m$$

$$+ (\alpha \beta)_{jk} + (\alpha \gamma)_{jl} + (\alpha \delta)_{jm}$$

$$+ (\beta \gamma)_{kl} + (\beta \delta)_{km} + (\gamma \delta)_{lm}$$

$$+ (\alpha \beta \gamma)_{jkl} + (\alpha \beta \delta)_{jkm} + (\alpha \gamma \delta)_{jlm} + (\beta \gamma \delta)_{klm}$$

$$+ (\alpha \beta \gamma \delta)_{jkml}$$

$$+ \epsilon_{ijklm}$$

where $\epsilon_{ijklm} \sim N(0, \sigma^2)$. 

Example: School Attendance among Australian Children
### Data structure

<table>
<thead>
<tr>
<th>Cell</th>
<th>Origin</th>
<th>Sex</th>
<th>Grade</th>
<th>Type</th>
<th>Days of Absence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>M</td>
<td>F0</td>
<td>SL</td>
<td>2,11,14</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>M</td>
<td>F0</td>
<td>AL</td>
<td>5,5,13,20,22</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>M</td>
<td>F1</td>
<td>SL</td>
<td>6,6,15</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>M</td>
<td>F1</td>
<td>AL</td>
<td>7,14</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>M</td>
<td>F2</td>
<td>SL</td>
<td>6,32,53,57</td>
</tr>
<tr>
<td>30</td>
<td>N</td>
<td>F</td>
<td>F2</td>
<td>AL</td>
<td>1</td>
</tr>
<tr>
<td>31</td>
<td>N</td>
<td>F</td>
<td>F3</td>
<td>SL</td>
<td>8</td>
</tr>
<tr>
<td>32</td>
<td>N</td>
<td>F</td>
<td>F3</td>
<td>AL</td>
<td>1,9,22,3,3,5,15,18,22,37</td>
</tr>
</tbody>
</table>
Reading Data

data ozkids;
  infile 'ozkids.dat' dlm=’,’ expandtabs missover;
  input cell origin $ gender $ grade $ type $ days @;
    do until (days=.);
      output;
      input days @;
    end;
  run;

▶ The **expandtabs** option converts tabs to spaces so that the list input can be used to read the tab-separated values.
▶ The **dlm=’,’** option specifies that both spaces and commas are delimiters by including a space and a comma in the quotes.
▶ The **do** loop is used to output an observation for each value of days of absence. Read the textbook for more details.
Fitting Main Effect Only Models (1)

For unbalanced designs, \texttt{proc glm} should be used rather than \texttt{proc anova}. We begin by fitting one main effects only model.

\begin{verbatim}
proc glm data=ozkids;
  class origin gender grade type;
  model days=origin gender grade type /ss1 ss3;
run;
\end{verbatim}

- Both Type I and Type III sums of squares are requested.
- When dealing with a main effects only model, the Type III sums of squares can be used to identify the most important effects. (Here: ‘origin’ and ‘grade’)

---

Lectures 8: Two-, Three-, and Four-Way ANOVA
Example: School Attendance among Australian Children
Fitting Main Effect Only Models (2)

Now we will fit more main effects only models for different orders of the main effects.

```plaintext
proc glm data=ozkids;
   class origin gender grade type;
   model days=grade gender type origin /ss1;
run;
proc glm data=ozkids;
   class origin gender grade type;
   model days=type gender origin grade /ss1;
run;
proc glm data=ozkids;
   class origin gender grade type;
   model days=gender origin type grade /ss1;
run;
```

Since the Type III sums of squares are invariant to the order only Type I sums of squares are requested.
Fitting a Full Model

Next we fit a full factorial model as follows:

```plaintext
proc glm data=ozkids;
  class origin gender grade type;
  model days=origin gender grade type origin|gender|grade|type /ss1 ss3;
run;
```

We specify the main effects explicitly so that they are entered before any interaction terms when calculating Type I sums of squares.