Topic 21 - Two Factor ANOVA

STAT 525 - Fall 2013

Outline

- Data
- Model
- Parameter Estimates
  - Equal Sample Size
  - One replicate per cell
  - Unequal Sample size

Overview

- Now have two factors (A and B)
- Suppose each factor has two levels
- Could analyze as one factor with 4 levels
  - Trt 1: A high, B high
  - Trt 2: A high, B low
  - Trt 3: A low, B high
  - Trt 4: A low, B low
- Use contrasts to test for main effects an interaction

$$A \text{ main effect } = \frac{\text{Trt1} + \text{Trt2}}{2} - \frac{\text{Trt3} + \text{Trt4}}{2}$$

Example

An experiment is conducted to study the effect of hormones injected into test rats. There are two distinct hormones (A,B) each with two distinct levels. For purposes here, we will consider this to be four different treatments labeled {A,a,B,b}. Each treatment is applied to six rats with the response being the amount of glycogen (in mg) in the liver.

Treatment	Responses					
A	106	101	120	86	132	97
a	51	98	85	50	111	72
B	103	84	100	83	110	91
b	50	66	61	72	85	60

Example

Three contrasts are of interest. They are:

Comparison	A	a	B	b
Hormone A vs Hormone B	1	1	-1	-1
Low level vs High level	1	-1	1	-1
Equivalence of level effect	1	-1	-1	1

Can we reanalyze the experiment in such a way that these sum of squares are already separated?

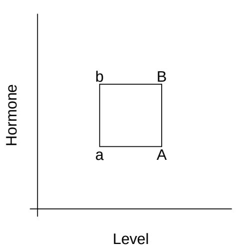
Two-way Factorial

- Break up the treatments into the two factors
- Example also known as a 2^2 factorial
- Investigates <u>all</u> combos of factor levels
- Single "replicate" involves ab trials

Two-way Factorial Layout

- Often presented as table or plot

	Level	
Hormone	Low	High
A	xxxxxx	xxxxxx
B	xxxxxx	xxxxxx



Data for Two Factor ANOVA

- Y is the response variable
- Factor A has levels $i = 1, 2, ..., a$
- Factor B has levels $j = 1, 2, ..., b$
- Y_{ijk} is the k^{th} observation from cell (i, j)
- Chapter 19 assumes $n_{ij} = n$
- Chapter 20 assumes $n_{ij} = 1$
- Chapter 23 allows n_{ij} to vary

Example Page 833

- Castle Bakery supplies wrapped Italian bread to a large number of supermarkets
- Bakery interested in the set up of their store display
  - Height of display shelf (top, middle, bottom)
  - Width of display shelf (regular, wide)
- Twelve stores equal in sales volume were selected
- Randomly assigned equally to each of 6 combinations
- Y is the sales of the bread
  - $i = 1, 2, 3$ and $j = 1, 2$
  - $n_{ij} = n = 2$

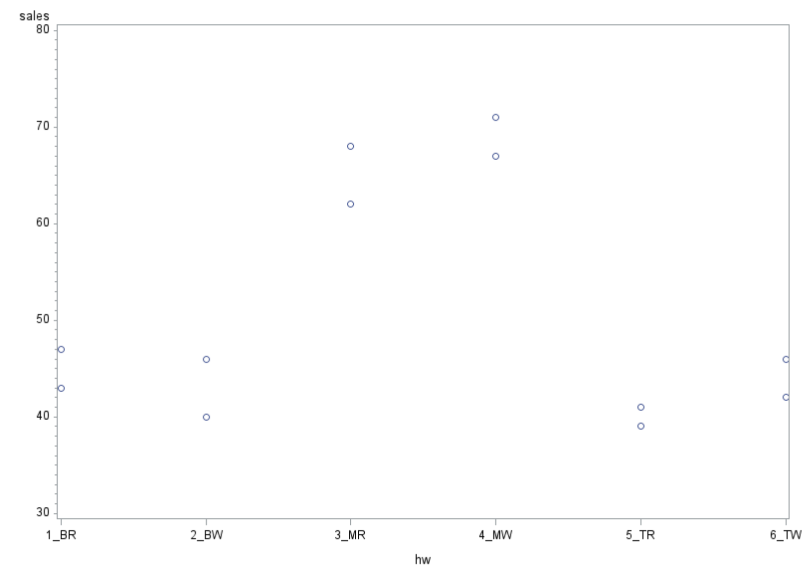
SAS Commands

```
data a1; infile 'u:\.www\datasets525\CH19TA07.txt';
   input sales height width;

proc print;
run;

data a1; set a1;
   if height eq 1 and width eq 1 then hw='1_BR';
   if height eq 1 and width eq 2 then hw='2_BW';
   if height eq 2 and width eq 1 then hw='3_MR';
   if height eq 2 and width eq 2 then hw='4_MW';
   if height eq 3 and width eq 1 then hw='5_TR';
   if height eq 3 and width eq 2 then hw='6_TW';

symbol1 v=circle i=none;
proc gplot data=a1;
   plot sales*hw/frame;
```


Sales Data

Obs	sales	height	width
1	47	1	1
2	43	1	1
3	46	1	2
4	40	1	2
5	62	2	1
6	68	2	1
7	67	2	2
8	71	2	2
9	41	3	1
10	39	3	1
11	42	3	2
12	46	3	2

Scatterplot



The Model

- All observations assumed independent
- All observations normally distributed with
  - a mean that may depend on levels of factors A and B
  - constant variance
- Often presented in terms of cell means or factor effects

The Cell Means Model

- Expressed mathematically

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

  where μ_{ij} is the theoretical mean or expected value of all observations in cell (i, j)
- The ε_{ijk} are iid $N(0, \sigma^2)$ which implies the Y_{ijk} are independent $N(\mu_{ij}, \sigma^2)$
- Parameters
  - $\{\mu_{ij}\}$, $i = 1, 2, .., a$, $j = 1, 2, ..., b$
  - σ^2

Estimates / Inference

- Estimate μ_{ij} by the sample mean of the observations in cell (i, j)

$$\hat{\mu}_{ij} = \overline{Y}_{ij.}$$

- Can also estimate variance using observations in cell (i, j)

$$s_{ij}^2 = \sum (Y_{ijk} - \overline{Y}_{ij.})^2/(n-1)$$

- These s_{ij}^2 are combined for single estimate of σ^2

ANOVA Table : $n_{ij} = n$

- Similar ANOVA table construction ($\overline{Y}_{ij.}$ is fitted value)

Source of Variation	df	SS
Model	$ab-1$	$n\sum\sum(\overline{Y}_{ij.} - \overline{Y}_{...})^2$
Error	$ab(n-1)$	$\sum\sum\sum(Y_{ijk} - \overline{Y}_{ij.})^2$
Total	$abn-1$	$\sum\sum\sum(Y_{ijk} - \overline{Y}_{...})^2$

$$\overline{Y}_{...} = \sum\sum\sum Y_{ijk}/abn$$
$$\overline{Y}_{ij.} = \sum Y_{ij.}/n$$

- Can further break down into Factor A, Factor B and interaction effects using contrasts

Factor Effects Model

- Breaks down cell means

  $\mu = \sum_i \sum_j \mu_{ij}/(ab)$

  $\mu_{i.} = \sum_j \mu_{ij}/b$ and $\mu_{.j} = \sum_i \mu_{ij}/a$

  $\alpha_i = \mu_{i.} - \mu$ and $\beta_j = \mu_{.j} - \mu$

  $(\alpha\beta)_{ij} = \mu_{ij} - (\mu + \alpha_i + \beta_j)$

- Interaction effect is the difference between the cell mean and the additive or main effects model. Explains variation not explained by main effects.

Factor Effects Model

- Statistical model is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \quad \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, b \\ k = 1, 2, \ldots, n \end{cases}$$

  μ - grand mean

  α_i - ith level effect of factor A (ignores B)

  β_j - jth level effect of factor B (ignores A)

  $(\alpha\beta)_{ij}$ - interaction effect of combination ij

- Like one-way model this is over-parameterized.
- Must include $a + b + 1$ model constraints.

  $\sum_i \alpha_i = 0 \quad \sum_j \beta_j = 0 \quad \sum_i (\alpha\beta)_{ij} = 0 \quad \sum_j (\alpha\beta)_{ij} = 0$

Factor Effects Estimates

- Constraints on previous page result in

  $\widehat{\mu} = \overline{Y}_{...}$

  $\widehat{\alpha}_i = \overline{Y}_{i..} - \overline{Y}_{...}$

  $\widehat{\beta}_j = \overline{Y}_{.j.} - \overline{Y}_{...}$

  $\widehat{(\alpha\beta)}_{ij} = \overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...}$

- The predicted value and residual are

  $\widehat{Y}_{ijk} = \overline{Y}_{ij.}$

  $e_{ijk} = Y_{ijk} - \overline{Y}_{ij.}$

Questions about our Example

- Does the height of the display affect sales?
  - If yes, will need to do pairwise comparisons
- Does the width of the display affect sales?
  - If yes, will need to do pairwise comparisons
- Does the effect of height depend on the width?
- Does the effect of width depend on the height?
  - If yes to either of these last two, we have an interaction

Partitioning the Sum of Squares

$$\begin{aligned} Y_{ijk} - \overline{Y}_{...} &= (\overline{Y}_{i..} - \overline{Y}_{...}) + (\overline{Y}_{.j.} - \overline{Y}_{...}) + \\ &\quad (\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...}) + (Y_{ijk} - \overline{Y}_{ij.}) \end{aligned}$$

- Consider $\sum\sum\sum (Y_{ijk} - \overline{Y}_{...})^2$
- Right hand side simplifies to

$$\begin{aligned} & bn \sum_i (\overline{Y}_{i..} - \overline{Y}_{...})^2 + \\ & an \sum_j (\overline{Y}_{.j.} - \overline{Y}_{...})^2 + \\ & n \sum_i \sum_j (\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...})^2 + \\ & \sum_i \sum_j \sum_k (Y_{ijk} - \overline{Y}_{ij.})^2 \end{aligned}$$

Partitioning the Sum of Squares

- Can be written as

$$\text{SSTO} = \text{SSA} + \text{SSB} + \text{SSAB} + \text{SSE}$$

- Degrees of freedom also broken down
- Under normality, all SS/σ^2 independent

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square
Factor A	SSA	$a-1$	MSA
Factor B	SSB	$b-1$	MSB
Interaction	SSAB	$(a-1)(b-1)$	MSAB
Error	SSE	$ab(n-1)$	MSE
Total	SSTO	$abn-1$	

Hypothesis Testing

- Can show: Fixed Case

E(MSE)=σ^2
E(MSA) $= \sigma^2 + bn\sum \alpha_i^2/(a-1)$
E(MSB) $= \sigma^2 + an\sum \beta_j^2/(b-1)$
E(MSAB) $= \sigma^2 + n\sum (\alpha\beta)_{ij}^2/(a-1)(b-1)$

- Use F-test to test for A, B, and AB effects

$$F^* = \frac{\text{SSA/(a}-1)}{\text{SSE/(ab(n}-1))}$$

$$F^* = \frac{\text{SSB/(b}-1)}{\text{SSE/(ab(n}-1))}$$

$$F^* = \frac{\text{SSAB/(a}-1)(\text{b}-1)}{\text{SSE/(ab(n}-1))}$$

SAS Commands

```
proc glm data=a1;
   class height width;
   model sales=height width height*width;
   means height width height*width;

proc means data=a1;
   var sales; by height width;
   output out=a2 mean=avsales;

symbol1 v=square i=join c=black;
symbol2 v=diamond i=join c=black;
proc gplot data=a2;
   plot avsales*height=width/frame;
run;
```

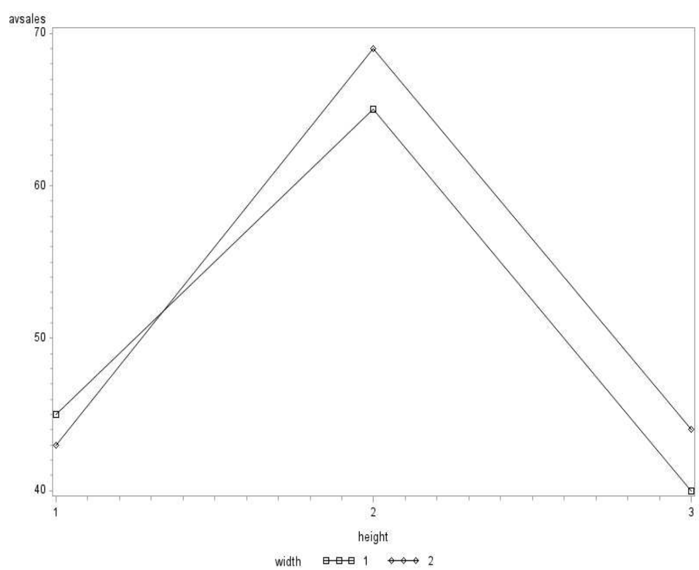
Output

```
The GLM Procedure
   Class Level Information
Class         Levels    Values
height             3    1 2 3
width              2    1 2

Number of observations    12

                            Sum of
Source            DF       Squares   Mean Square  F Value  Pr > F
Model              5   1580.000000    316.000000    30.58  0.0003
Error              6     62.000000     10.333333
Corrected Total   11   1642.000000

R-Square     Coeff Var      Root MSE    sales Mean
0.962241      6.303040      3.214550      51.00000

Source            DF     Type I SS   Mean Square  F Value  Pr > F
height             2   1544.000000    772.000000    74.71  <.0001
width              1     12.000000     12.000000     1.16  0.3226
height*width       2     24.000000     12.000000     1.16  0.3747

Source            DF   Type III SS   Mean Square  F Value  Pr > F
height             2   1544.000000    772.000000    74.71  <.0001
width              1     12.000000     12.000000     1.16  0.3226
height*width       2     24.000000     12.000000     1.16  0.3747
```


Output

```
Level of          ------------sales------------
height       N            Mean          Std Dev
1            4      44.0000000       3.16227766
2            4      67.0000000       3.74165739
3            4      42.0000000       2.94392029

Level of          ------------sales------------
width        N            Mean          Std Dev
1            6      50.0000000       12.0664825
2            6      52.0000000       13.4313067

Level of     Level of          ------------sales------------
height       width        N            Mean          Std Dev
1            1            2      45.0000000       2.82842712
1            2            2      43.0000000       4.24264069
2            1            2      65.0000000       4.24264069
2            2            2      69.0000000       2.82842712
3            1            2      40.0000000       1.41421356
3            2            2      44.0000000       2.82842712
```


Interaction Plot



SAS Commands

```
proc glm data=a1;
   class height width;
   model sales=height|width;
   means height / tukey lines;

proc glm data=a1;
   class height width;
   model sales=height width;
   means height / tukey lines;
run;
```

Results

- There appears to be no interaction between height and width (P=0.37) → The effect of width (or height) is the same regardless of height (or width). Because of this, we can focus on the main effects (averages out the other effect).
- The main effect for width is not statistically significant (P=0.32) → Width does not affect sales of bread
- The main effect for height is statistically significant ($P < 0.0001$). From the scatterplot and interaction plot, it appears the middle location is better than the top and bottom. Pairwise testing (adjusting for multiple comparisons) can confirm this.

Pooling Insignificant Terms

- Some argue that an insignificant interaction should be dropped from the model (i.e., pooled with error)
- See last GLM call

  $\mathrm{SSE}^* = \mathrm{SSE} + \mathrm{SSAB}$

  $\mathrm{df}^*_{\mathrm{E}} = ab(n-1) + (a-1)(b-1)$
- This increases DF but could inflate $\hat{\sigma}^2$
- Possibly result in a Type II error
- Rule of thumb: Only pool when dfe small (e.g., < 5) and P-value of the interaction is large (e.g., > 0.25)

Output

```
Tukey's Studentized Range (HSD) Test for sales

Error Degrees of Freedom                   6
Error Mean Square                   10.33333
Critical Value of Studentized Range  4.33902
Minimum Significant Difference         6.974

          Mean      N    height
A       67.000      4    2

B       44.000      4    1
B
B       42.000      4    3

**** POOLING ****
Error Degrees of Freedom                   8
Error Mean Square                      10.75
Critical Value of Studentized Range  4.04101
Minimum Significant Difference        6.6247

          Mean      N    height
A       67.000      4    2

B       44.000      4    1
B
B       42.000      4    3
```


Background Reading

- KNNL Chapter 19
- knnl833.sas