STAT 525 FALL 2018

Chapter 23
Two-Factor Studies with Unequal Sample Sizes

Professor Min Zhang

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)
- Now $k = 1, 2, ..., n_{ij}$

Example (Page 954)

- Synthetic growth horomone administered to growth horomone deficient pre-pubescent children
- Interested in two factors
  - Gender ($a = 2$)
  - Bone development level ($b = 3$)
- Y is the difference between growth rate during treatment and prior to treatment
- Set up as balanced design ($n = 3$) but four children were unable to complete the study
  - $i = 1, 2$ and $j = 1, 2, 3$
  - $n_{ij} = 3, 2, 2, 1, 3, 3$

Recall: General Plan of Two-Factor ANOVA

- Construct scatterplot / interaction plot
- Run full model
- Check assumptions
  - Residual plots
  - Histogram / QQplot
  - Ordered residuals plot
- Check significance of interaction

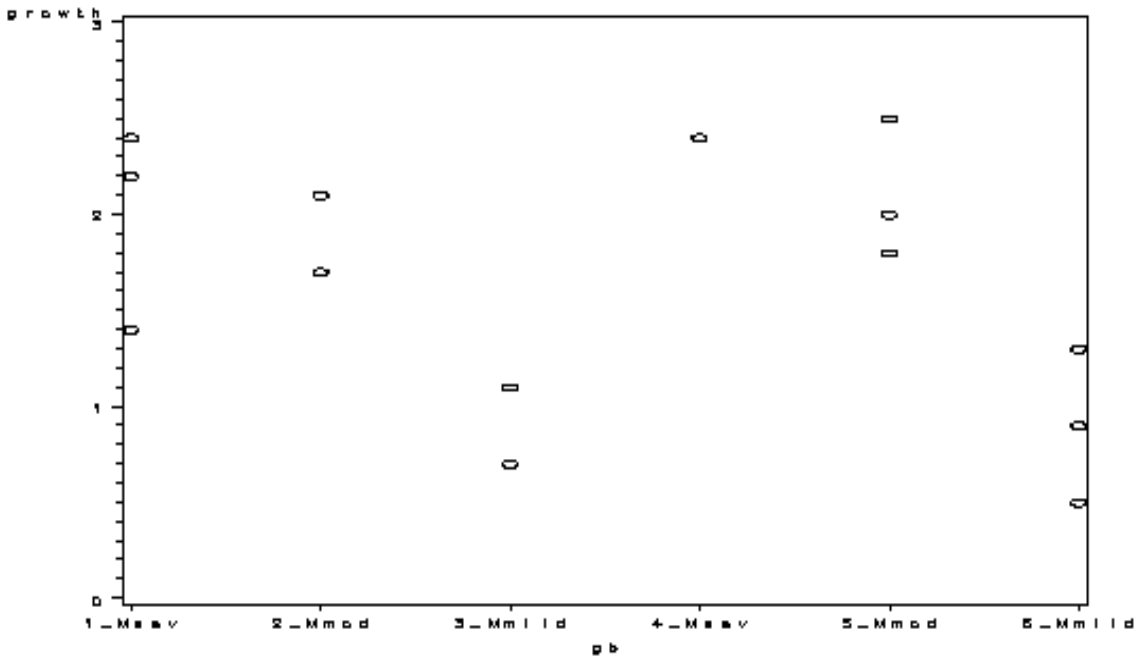
```
options nocenter;
data a1; infile 'u:\.www\datasets525\CH23TA01.txt';
    input growth gender bone;

proc print data=a1; run; quit;
```

Obs	growth	gender	bone
1	1.4	1	1
2	2.4	1	1
3	2.2	1	1
4	2.1	1	2
5	1.7	1	2
6	0.7	1	3
7	1.1	1	3
8	2.4	2	1
9	2.5	2	2
10	1.8	2	2
11	2.0	2	2
12	0.5	2	3
13	0.9	2	3
14	1.3	2	3

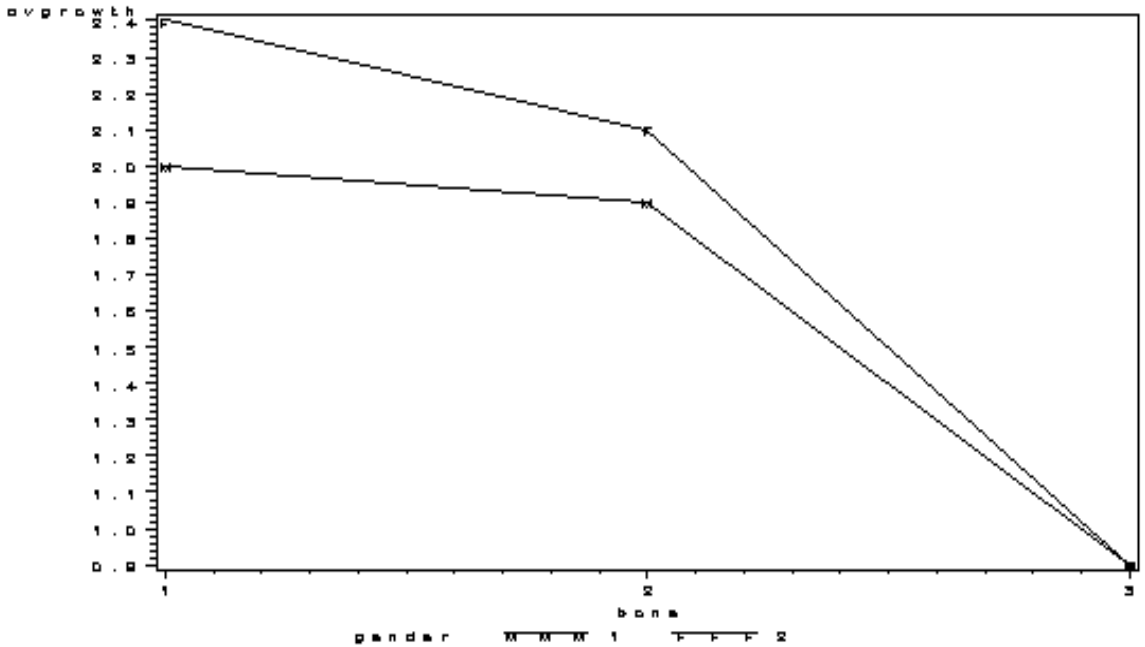
```
data a1; set a1;
   if (gender eq 1)*(bone eq 1) then gb='1_Msev ';
   if (gender eq 1)*(bone eq 2) then gb='2_Mmod ';
   if (gender eq 1)*(bone eq 3) then gb='3_Mmild';
   if (gender eq 2)*(bone eq 1) then gb='4_Msev ';
   if (gender eq 2)*(bone eq 2) then gb='5_Mmod ';
   if (gender eq 2)*(bone eq 3) then gb='6_Mmild';

/*----------- Scatterplot -----------*/
symbol1 v=circle i=none c=black;
proc gplot data=a1;
   plot growth*gb/frame;
run; quit;
```



```
proc means data=a1;
    output out=a2 mean=avgrowth;
    by gender bone;

/*------------ Interaction Plot ------------*/
symbol1 v='M' i=join c=black;
symbol2 v='F' i=join c=black;
proc gplot data=a2;
    plot avgrowth*bone=gender/frame;
run; quit;
```



The Cell Means Model

- Expressed numerically

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

  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

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

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

- For each cell (i, j), also estimate of the variance

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

- These s^2_{ij} are pooled to estimate σ^2

Example (Page 954)

```
proc glm data=a1;
    class gender bone;
    model growth=gender|bone/solution;
run; quit;
```

- The `solution` option gives parameter estimates for the factor effects model under the GLM constraints (aka SAS approach)
  - $\alpha_2 = 0$
  - $\beta_3 = 0$
  - $(\alpha\beta)_{13} = (\alpha\beta)_{23} = (\alpha\beta)_{21} = (\alpha\beta)_{22} = 0$
- Produces the cell means in the usual way

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	4.47428571	0.89485714	5.51	0.0172
Error	8	1.30000000	0.16250000		
Corrected Total	13	5.77428571			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
gender	1	0.00285714	0.00285714	0.02	0.8978
bone	2	4.39600000	2.19800000	13.53	0.0027
gender*bone	2	0.07542857	0.03771429	0.23	0.7980

Source	DF	Type III SS	Mean Square	F Value	Pr > F
gender	1	0.12000000	0.12000000	0.74	0.4152
bone	2	4.18971429	2.09485714	12.89	0.0031
gender*bone	2	0.07542857	0.03771429	0.23	0.7980

Parameter			Estimate	Std Error	t Value	Pr > |t|
Intercept			0.9000 B	0.23273733	3.87	0.0048
gender	1		-0.0000 B	0.36799004	-0.00	1.0000
gender	2		0.0000 B	.	.	.
bone	1		1.5000 B	0.46547467	3.22	0.0122
bone	2		1.2000 B	0.32914029	3.65	0.0065
bone	3		0.0000 B	.	.	.
gender*bone	1	1	-0.4000 B	0.59336610	-0.67	0.5192
gender*bone	1	2	-0.2000 B	0.52041650	-0.38	0.7108
gender*bone	1	3	0.0000 B	.	.	.
gender*bone	2	1	0.0000 B	.	.	.
gender*bone	2	2	0.0000 B	.	.	.
gender*bone	2	3	0.0000 B	.	.	.

Comments

- For studies with equal sample sizes, Type I and Type III SS were identical, due to the fact that the sample sizes were all the same made the variables completely orthogonal.
- When sample sizes are unequal, the SS do not break down in the usual way. The various SS that we can calculate will not necessarily add up to the SSM.
  - Type I: $0.003 + 4.396 + 0.075 = 4.474 =$ Model SS
  - Type III: $0.120 + 4.190 + 0.075 = 4.385 \neq$ Model SS
- SAS actually calculates four types of SS (I, II, III, IV). It does `ss1` and `ss3` by default but you can also ask for `ss2` and `ss4`.
- We will focus on Type I and Type III in ANOVA.
  - Type I and Type III hypotheses are different
  - Most prefer Type III analysis
  - Can be misleading if n_{ij} widely different
  - Use contrasts to understand the difference

Type I

- Type I SS refer to the difference in SS when variables are added sequentially in the model, i.e. $SS(A)$, $SS(B|A)$, $SS(A\times B|A,B)$.
- Type I weights each observation equally, with the result that the treatments are weighted in proportion to their $n_{i,j}$.

Type II

- Type II SS referred to the difference in SSM when a variable is included last in the model or not (i.e $SS(A|B,A\times B)$, $SS(B|A,A\times B)$, $SS(A\times B|A,B)$.
- Type II also weights each observation equally, with the result that the treatments are weighted in proportion to their $n_{i,j}$.

Type III

- As Type II SS, Type III SS referred to the difference in SSM when a variable is included last in the model or not (i.e $SS(A|B, A \times B)$, $SS(B|A, A \times B)$, $SS(A \times B|A, B)$
- Type III SS adjust for the cells having different $n_{i,j}$, by weighting each treatment equally, so that the observations are weighted differently.
- When the sample sizes are unequal, Type III SS are more informative about the treatments than Type I.
- The Type III SS are calculated using regression with indicator variables to do the ANOVA, and to calculate the SSM for the full and reduced models.
- In Sections 23.2-3, KNNL are discussing Type III SS (they don't call them that; the type numbers are a SAS convention).

Type IV

- Type IV SS are like Type III, except that Type IV additionally take into account possibly empty cells ($n_{i,j} = 0$).
- If there are empty cells, then Type IV SS are preferred. See KNNL Section 23.4 about empty cells.

Contrast for $A * B$

- Same for Type I and Type III
- Null Hypothesis is that the mean profiles are parallel (recall interaction plot)
- Null hypothesis can be expressed

$$\mu_{12} - \mu_{11} = \mu_{22} - \mu_{21}$$

and

$$\mu_{13} - \mu_{12} = \mu_{23} - \mu_{22}$$

```
contrast 'gender*bone Type I&III'
  gender*bone 1 -1 0 -1 1 0,  gender*bone 0 1 -1 0 -1 1;
```

Type III Contrast for A

- Null Hypothesis is that the marginal gender means are the same
- Null hypothesis can be expressed

$$\begin{aligned}
1 \times \mu_{11} &= 1(\mu + \alpha_1 + \beta_1 + (\alpha\beta)_{11}) \\
1 \times \mu_{12} &= 1(\mu + \alpha_1 + \beta_2 + (\alpha\beta)_{12}) \\
1 \times \mu_{13} &= 1(\mu + \alpha_1 + \beta_3 + (\alpha\beta)_{13}) \\
-1 \times \mu_{21} &= -1(\mu + \alpha_2 + \beta_1 + (\alpha\beta)_{21}) \\
-1 \times \mu_{22} &= -1(\mu + \alpha_2 + \beta_2 + (\alpha\beta)_{22}) \\
-1 \times \mu_{23} &= -1(\mu + \alpha_2 + \beta_3 + (\alpha\beta)_{23}) \\
\hline
&= 3\alpha_1 - 3\alpha_2 + (\alpha\beta)_{1.} - (\alpha\beta)_{2.}
\end{aligned}$$

```
contrast 'gender Type III'
  gender 3 -3  gender*bone 1 1 1 -1 -1 -1;
estimate 'gender Type III'
  gender 3 -3  gender*bone 1 1 1 -1 -1 -1;
```

Type I Contrast for A

- Null Hypothesis is that the "weighted" marginal gender means are the same
- Null hypothesis can be expressed

$$
\begin{aligned}
3 \times \mu_{11} &= 3(\mu + \alpha_1 + \beta_1 + (\alpha\beta)_{11}) \\
2 \times \mu_{12} &= 2(\mu + \alpha_1 + \beta_2 + (\alpha\beta)_{12}) \\
2 \times \mu_{13} &= 2(\mu + \alpha_1 + \beta_3 + (\alpha\beta)_{13}) \\
-1 \times \mu_{21} &= -1(\mu + \alpha_2 + \beta_1 + (\alpha\beta)_{21}) \\
-3 \times \mu_{22} &= -3(\mu + \alpha_2 + \beta_2 + (\alpha\beta)_{22}) \\
-3 \times \mu_{23} &= -3(\mu + \alpha_2 + \beta_3 + (\alpha\beta)_{23}) \\
\hline
&= 7\alpha_1 - 7\alpha_2 + 2\beta_1 - \beta_2 - \beta_3 + \\
&\quad 3(\alpha\beta)_{11} + 2(\alpha\beta)_{12} + 2(\alpha\beta)_{13} - \\
&\quad (\alpha\beta)_{21} - 3(\alpha\beta)_{22} - 3(\alpha\beta)_{23}
\end{aligned}
$$

```
contrast 'gender Type I'
  gender 7 -7  bone 2 -1 -1  gender*bone 3 2 2 -1 -3 -3;
estimate 'gender Type I'
  gender 7 -7  bone 2 -1 -1  gender*bone 3 2 2 -1 -3 -3;
```

```
proc glm data=a1;
    class gender bone;
    model growth=gender|bone;
    contrast 'gender*bone Type I and III'
      gender*bone 1 -1 0 -1 1 0,  gender*bone 0 1 -1 0 -1 1;
    contrast 'gender Type III'
      gender 3 -3  gender*bone 1 1 1 -1 -1 -1;
    estimate 'gender Type III'
      gender 3 -3  gender*bone 1 1 1 -1 -1 -1;
    contrast 'gender Type I'
      gender 7 -7  bone 2 -1 -1  gender*bone 3 2 2 -1 -3 -3;
    estimate 'gender Type I'
      gender 7 -7  bone 2 -1 -1  gender*bone 3 2 2 -1 -3 -3;
    contrast 'bone Type III'
      bone 2 -2 0  gender*bone 1 -1 0 1 -1 0,
      bone 0 2 -2  gender*bone 0 1 -1 0 1 -1;
run; quit;
```

	DF	Contrast SS	Mean Square	F Value	Pr > F
gender*bone Type I&III	2	0.07542857	0.03771429	0.23	0.7980
gender Type III	1	0.12000000	0.12000000	0.74	0.4152
gender Type I	1	0.00285714	0.00285714	0.02	0.8978
bone Type III	2	4.18971429	2.09485714	12.89	0.0031

Parameter	Estimate	Standard Error	t Value	Pr > |t|
gender Type III	-0.60000000	0.69821200	-0.86	0.4152
gender Type I	0.20000000	1.50831031	0.13	0.8978

Recall: If Interaction Not Significant

- Determine whether pooling is beneficial
  - If yes, rerun analysis without interaction
- Check significance of main effects
  - If factor insignificant, determine whether pooling is beneficial
    - If yes, rerun analysis as one-way ANOVA
  - If statistically significant factor has more than two levels, use multiple comparison procedure to assess differences
    - Contrasts and linear combinations can also be used

```
proc glm data=a1;
    class gender bone;
    model growth=gender bone/solution;
run; quit;
```

```
                          Sum of
Source                DF     Squares  Mean Square  F Value  Pr > F
Model                  3  4.39885714   1.46628571    10.66  0.0019
Error                 10  1.37542857   0.13754286
Corrected Total       13  5.77428571

R-Square     Coeff Var      Root MSE    growth Mean
0.761801      22.57456      0.370868       1.642857

Source                DF   Type I SS  Mean Square  F Value  Pr > F
gender                 1  0.00285714   0.00285714     0.02  0.8883
bone                   2  4.39600000   2.19800000    15.98  0.0008

Source                DF  Type III SS  Mean Square  F Value  Pr > F
gender                 1  0.09257143   0.09257143     0.67  0.4311
bone                   2  4.39600000   2.19800000    15.98  0.0008

                                 Standard
Parameter       Estimate             Error  t Value    Pr > |t|
Intercept    0.968571429 B   0.18572796     5.22      0.0004
gender    1 -0.171428571 B   0.20896028    -0.82      0.4311
gender    2  0.000000000 B    .              .         .
bone      1  1.260000000 B   0.25931289     4.86      0.0007
bone      2  1.120000000 B   0.23455733     4.77      0.0008
bone      3  0.000000000 B    .              .         .
```

```
proc glm data=a1;
    class gender bone;
    model growth=gender bone;
    means gender bone/ tukey lines;
run; quit;
```

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

Alpha                                      0.05
Error Degrees of Freedom                     10
Error Mean Square                      0.137543
Critical Value of Studentized Range     3.15106
Minimum Significant Difference           0.4417

           Mean      N    gender
A        1.6571      7    1
A
A        1.6286      7    2


Alpha                                      0.05
Error Degrees of Freedom                     10
Error Mean Square                      0.137543
Critical Value of Studentized Range     3.87676
Minimum Significant Difference           0.6692
Harmonic Mean of Cell Sizes            4.615385

           Mean      N    bone
A        2.1000      4    1
A
A        2.0200      5    2

B        0.9000      5    3
```

```
proc glm data=a1;
    class gender bone;
    model growth=gender bone;
    lsmeans gender bone/ adjust=tukey pdiff;
run; quit;
```

```
Adjustment for Multiple Comparisons: Tukey-Kramer

                          H0:LSMean1=
                growth      LSMean2
gender          LSMEAN       Pr > |t|
1           1.59047619         0.4311
2           1.76190476

               growth      LSMEAN
bone           LSMEAN      Number
1         2.14285714           1
2         2.00285714           2
3         0.88285714           3

     Least Squares Means for effect bone
     Pr > |t| for H0: LSMean(i)=LSMean(j)

          Dependent Variable: growth
i/j               1             2             3
   1                        0.8538        0.0017
   2         0.8538                       0.0020
   3         0.0017        0.0020
```

Comments

- The `means` and `lsmeans` are not the same
  - `means` : raw sample mean - similar to a weighted average of cell means (Type I)
  - `lsmeans`: uses parameter estimates - similar to Type III approach
- `lsmeans` most commonly used

Chapter Review

- Two-factor ANOVA with unequal sample sizes
  - Data
  - Model
  - Parameter Estimates
  - Inference