Two Way Analysis of Variance

Professor: Hammou El Barmi
Columbia University

Introduction

- Goal: compare the means of a single variable at different levels of two factors A and B in scientific experiments .
- Suppose factor A has a levels and factor B has b levels
- In total we have ab treatments
- We assume that each treatment level, we have n_{ij} experimental units and let Y_{ijk} be the kth observation when $A = i$ and $B = j$
- We will assume first that $n_{ij} \equiv n$ for all (i,j) (we say that the design is balanced)
- We assume that the Y_{ijk} are independent and that

$$Y_{ijk} \sim N(\mu_{ij}, \sigma^2)$$

Let

$$\mu_{i\bullet} = \frac{\sum_{j=1}^{b} \mu_{ij}}{b}, \quad \mu_{\bullet j} = \frac{\sum_{i=1}^{a} \mu_{ij}}{a}, \quad \mu_{\bullet\bullet} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b} \mu_{ij}}{ab}$$

we can represent the means as follows

		Factor B				
		B_1	B_2	$\ldots$	B_b	
5*Factor A	A_1	μ_{11}	μ_{12}	$\ldots$	μ_{1b}	$\mu_{1\bullet}$
	A_2	μ_{21}	μ_{22}	$\ldots$	μ_{2b}	$\mu_{2\bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	A_a	μ_{a1}	μ_{a2}	$\ldots$	μ_{ab}	$\mu_{2\bullet}$
		$\mu_{\bullet 1}$	$\mu_{\bullet 2}$	$\ldots$	$\mu_{\bullet b}$	$\mu_{\bullet\bullet}$

- Let

$$\bar{Y}_{ij\bullet} = \frac{\sum_{k=1}^{n} Y_{ijk}}{n} \quad \bar{Y}_{\bullet j\bullet} = \frac{\sum_{i=1}^{a}\sum_{k=1}^{n} Y_{ijk}}{an}$$

$$\bar{Y}_{i\bullet\bullet} = \frac{\sum_{j=1}^{b}\sum_{k=1}^{n} Y_{ijk}}{kn} \quad \bar{Y}_{\bullet\bullet\bullet} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} Y_{ijk}}{abn}$$

- Let also

$$SST = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk} - \bar{Y}_{\bullet\bullet\bullet})^2 \quad SSA = nb\sum_{i=1}^{k}(\bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet\bullet\bullet})^2$$

$$SSB = na\sum_{i=1}^{k}(\bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet\bullet\bullet})^2 \quad SSAB = n\sum_{i=1}^{k}(\bar{Y}_{ij\bullet} - \bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet j\bullet} + \bar{Y}_{\bullet\bullet\bullet})^2$$

$$SSE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk} - \bar{Y}_{ij\bullet})^2$$

Then

$$SST = SSA + SSB + SSAB + SSE$$

ANOVA table

Source	SS	df	MS	E(MS)
Factor A	SSA	a-1	$\frac{SSA}{a-1}$	$\sigma^2 + \frac{bn\sum_{i=1}^{a}(\mu_{i\bullet}-\mu_{\bullet\bullet})^2}{a-1}$
Factor B	SSB	b-1	$\frac{SSB}{b-1}$	$\sigma^2 + \frac{an\sum_{i=1}^{a}(\mu_{\bullet j}-\mu_{\bullet\bullet})^2}{b-1}$
Factor AB	SSAB	(a-1)(b-1)	$\frac{SSAB}{(a-1)(b-1)}$	$\sigma^2 + \frac{n\sum_{i=1}^{a}\sum_{j=1}^{b}(\mu_{ij}-\mu_{i\bullet}-\mu_{\bullet j}+\mu_{\bullet\bullet})^2}{(a-1)(b-1)}$
Error	SSE	ab(n-1)	$\frac{SSE}{ab(n-1)}$	σ^2
Total	SST	abn-1		

- Notice that

$$\frac{E(MSA)}{E(MSE)} = 1 \quad \Leftrightarrow \quad \sum_{i=1}^{a} (\mu_{i\bullet} - \mu_{\bullet\bullet})^2 = 0$$
$$\Leftrightarrow \quad \mu_{1\bullet} = \mu_{2\bullet} = \ldots = \mu_{a\bullet}$$

- We should always start with testing the interaction. If interaction is present, then we ignore main effects look at it as one factor with ab levels
- if no interaction, compare levels of A ignoring B and compare levels of B ignoring A

F-test

- Test for interaction

$$H_0 : \text{No interation and} H_a : \text{Yes interaction}$$

- The test statistic is

$$F = \frac{MSAB}{MSE}$$

- Reject H_0 if

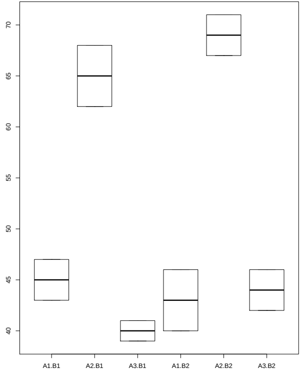
$$F > F(1-\alpha, (a-1)(b-1), ab(n-1))$$

  or if p-value less than α

Example

```
> response<-c(47,43, 62,68, 41,39, 46,40,67,71,42,46)
> A<-c(rep(c(rep("A1",2),rep("A2",2),rep("A3",2)) ,2 ))
> B<-c(rep("B1",6),rep("B2",6))
>boxplot(response~A*B)
```

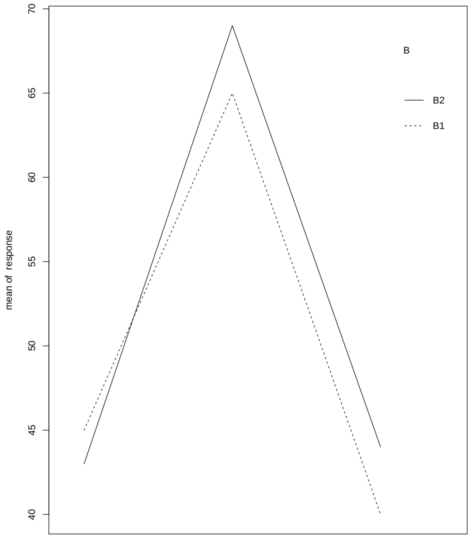
Figure: Box Plots



Example

```
> tapply(response, list(A),mean)
A1 A2 A3
44 67 42
> tapply(response, list(B),mean)
B1 B2
50 52
> interaction.plot(A,B,response)
   B1 B2
A1 45 43
A2 65 69
A3 40 44
```

Figure: Interaction Plot



Example

```
> summary(aov(response~A*B))
            Df Sum Sq Mean Sq F value   Pr(>F)
A            2   1544   772.0  74.710 5.75e-05
B            1     12    12.0   1.161    0.323
A:B          2     24    12.0   1.161    0.375
Residuals    6     62    10.3

A           ***
B
A:B
```

We have $SSA =, 1544 SSB = 12, SSAB = 24, SSE = 62$. their degrees of freedom are, 2, 1, 2 and 6, respectively
The p-value of the test for interaction is 0.375. We reject H_0 and conclude that there is no interaction.

```
Model without interaction:
> summary(aov(response~A+B))
            Df Sum Sq Mean Sq F value   Pr(>F)
A            2   1544   772.0  71.814 7.75e-06
B            1     12    12.0   1.116    0.322
Residuals    8     86    10.8
```

From this output we can conclude that A level means (p-value¡0.05) are different but B level means (p-value=0.322) are not (There an A effect but no B effect) We drop B and refit the model

```
> summary(aov(response~A))
            Df Sum Sq Mean Sq F value  Pr(>F)
A            2   1544   772.0    70.9 3.1e-06
Residuals    9     98    10.9
```

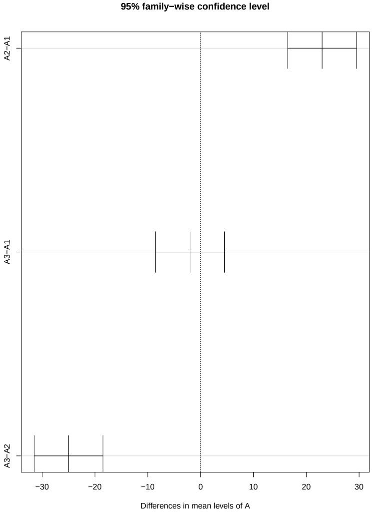
Factor A is significant at $\alpha = 0.01$

```
> fit<-aov(response~A)
> tk<-TukeyHSD(fit, "A")
> plot(tk)
> tk
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = response ~ A)

$A
      diff       lwr       upr     p adj
A2-A1   23  16.48532  29.51468 0.0000108
A3-A1   -2  -8.51468   4.51468 0.6789461
A3-A2  -25 -31.51468 -18.48532 0.0000054
```

Figure: Tukey method based confidence intervals



```
data example;
input   response A $   B $ ;
datalines;
47     A1    B1
43     A1    B1
62     A1    B2
68     A1    B2
41     A2    B1
39     A2    B1
46     A2    B2
40     A2    B2
67     A3    B1
71     A3    B1
42     A3    B2
46     A3    B2
;
proc glm;
class A B;
model response= A B A*B;
run;
```

- We have

$$\begin{aligned} Y_{ijk} &= \mu_{ij} + \epsilon_{ijk} \\ &= \mu_{\bullet\bullet} + \mu_{i\bullet} - \mu_{\bullet\bullet} + \mu_{\bullet j} - \mu_{\bullet\bullet} + \mu_{ij} - \mu_{i\bullet} - \mu_{\bullet j} + \mu_{\bullet\bullet} + \epsilon_{ijk} \\ &= \mu_{\bullet\bullet} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \end{aligned}$$

  where

$$\alpha_i = \mu_{i\bullet} - \mu_{\bullet\bullet}, \quad \beta_j = \mu_{\bullet j} - \mu_{\bullet\bullet} \quad \text{and} \quad (\alpha\beta)_{ij} = \mu_{ij} - \mu_{i\bullet} - \mu_{\bullet j} + \mu_{\bullet\bullet}$$

  and

$$\sum_{i=1}^{a} \alpha_i = 0, \quad \sum_{j=1}^{b} \beta_j = 0, \quad \text{and} \sum_{i=1}^{a} (\alpha\beta)_{ij} = \sum_{j=1}^{b} (\alpha\beta)_{ij} = 0$$

- The αs measure the effect of factor A, the βs measure the effect of factor B and the $(\alpha\beta)_{ij}$s are the interaction terms

Suppose $a = b = 2$.

$$\begin{array}{ll} \mu_{11} = \mu_{\bullet\bullet} + \alpha_1 + \beta_1 + (\alpha\beta)_{11} & \mu_{12} = \mu_{\bullet\bullet} + \alpha_1 + \beta_2 + (\alpha\beta)_{12} \\ \mu_{21} = \mu_{\bullet\bullet} + \alpha_2 + \beta_1 + (\alpha\beta)_{21} & \mu_{22} = \mu_{\bullet\bullet} + \alpha_2 + \beta_2 + (\alpha\beta)_{22} \end{array}$$

Note that since
$\alpha_1+\alpha_2 = 0, \beta_1+\beta_2 = 0, (\alpha\beta)_{11}+(\alpha\beta)_{12} = 0, (\alpha\beta)_{21}+(\alpha\beta)_{22} = 0, (\alpha\beta)_{11}+(\alpha\beta)_{21} = 0$ and $(\alpha\beta)_{12} + (\alpha\beta)_{22} = 0$ we have

Factor Effects Model

$$\begin{array}{ll} \mu_{11} = \mu_{\bullet\bullet} + \alpha_1 + \beta_1 + (\alpha\beta)_{11} & \mu_{12} = \mu_{\bullet\bullet} + \alpha_1 - \beta_1 - (\alpha\beta)_{11} \\ \mu_{21} = \mu_{\bullet\bullet} - \alpha_1 + \beta_1 - (\alpha\beta)_{11} & \mu_{22} = \mu_{\bullet\bullet} - \alpha_1 - \beta_1 + (\alpha\beta)_{11} \end{array}$$

from the model we have

$$\begin{aligned} \mu_{11} - \mu_{21} &= (\mu_{\bullet\bullet} + \alpha_1 + \beta_1 + (\alpha\beta)_{11}) - (\mu_{\bullet\bullet} - \alpha_1 + \beta_1 - (\alpha\beta)_{11}) \\ &= 2\alpha_1 + 2(\alpha\beta)_{11} \end{aligned}$$

and

$$\begin{aligned} \mu_{12} - \mu_{22} &= (\mu_{\bullet\bullet} + \alpha_1 - \beta_1 - (\alpha\beta)_{11}) - (\mu_{\bullet\bullet} - \alpha_1 - \beta_1 + (\alpha\beta)_{11}) \\ &= 2\alpha_1 - 2(\alpha\beta)_{11} \end{aligned}$$

So when

$$2\alpha_1 + 2(\alpha\beta)_{11} = 2\alpha_1 - 2(\alpha\beta)_{11}$$

we have

$$\begin{aligned} \mu_{11} - \mu_{21} = \mu_{12} - \mu_{22} &\Leftrightarrow 2\alpha_1 + 2(\alpha\beta)_{11} = 2\alpha_1 - 2(\alpha\beta)_{11} \\ &\Leftrightarrow (\alpha\beta)_{11} = 0 \end{aligned}$$

If you fix the level of one factor and look at the means, difference does not depend on that level if there is no inteaction

Factor Effects Model

In general

$$\mu_{ij} - \mu_{i'j} = \mu_{ij'} - \mu_{i'j'} \quad \text{for all } i, i', j, j' \quad \Leftrightarrow \quad (\alpha\beta)_{ij} = 0 \quad \text{for all} \quad (i,j)$$

		A1	A2	B	A1B	A2B
A1	B1	1	0	1	1	0
A1	B1	1	0	1	1	0
A2	B1	0	1	1	0	1
A2	B1	0	1	1	0	1
A3	B1	-1	-1	1	-1	-1
A3	B1	-1	-1	1	-1	-1
A1	B2	1	0	-1	-1	0
A1	B2	1	0	-1	-1	0
A2	B2	0	1	-1	0	-1
A2	B2	0	1	-1	0	-1
A3	B2	-1	-1	-1	1	1
A3	B2	-1	-1	-1	1	1

Use regression techniques and test

$$H_0 : \text{reduced model versus} \qquad H_a : \text{full model}$$

use partial F-test

```
Fit Full Model
> fit1<-lm(response~A1+A2+B+A1*B+A2*B)
> summary(fit1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   51.000      0.928  54.959 2.44e-09 ***
A1            -7.000      1.312  -5.334  0.00177 **
A2            16.000      1.312  12.192 1.85e-05 ***
B             -1.000      0.928  -1.078  0.32261
A1:B           2.000      1.312   1.524  0.17835
A2:B          -1.000      1.312  -0.762  0.47494
---
Residual standard error: 3.215 on 6 degrees of freedom
Multiple R-squared:  0.9622,Adjusted R-squared:  0.9308
F-statistic: 30.58 on 5 and 6 DF,  p-value: 0.0003384
```

```
Fit Reduced Model
> fit2<-lm(response~A1+A2+B)
> summary(fit2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  51.0000     0.9465  53.884 1.56e-11 ***
A1           -7.0000     1.3385  -5.230 0.000793 ***
A2           16.0000     1.3385  11.953 2.21e-06 ***
B            -1.0000     0.9465  -1.057 0.321579

Residual standard error: 3.279 on 8 degrees of freedom
Multiple R-squared:  0.9476,Adjusted R-squared:  0.928
F-statistic: 48.25 on 3 and 8 DF,  p-value: 1.813e-05
```

```
> anova(fit2,fit1)
Analysis of Variance Table

Model 1: response ~ A1 + A2 + B
Model 2: response ~ A1 + A2 + B + A1 * B + A2 * B
  Res.Df RSS Df Sum of Sq      F Pr(>F)
1      8  86
2      6  62  2        24 1.1613 0.3747
```