

Week 9--IES 612-STA 4-573-STA 4-576.doc

IES 612/STA 4-573/STA 4-576 Winter 2008

"ANOVA" MODELS for standard designs

- Completely Randomized Design (CRD)
- Random Complete Block Design (RCBD)
- Latin Squares (LS)

CRD with a single factor ...

Numeric data: independent random samples from "t" poplns/trmts obtained **OR** random samples randomly assigned to one of t treatments.

Assume $y_{ij} \sim$ independent $N(\mu_i, \sigma_\varepsilon^2)$, $i = 1, 2, \dots, t$ (poplns or trmts), $j = 1, 2, \dots, n_i$ (observations)

n_i = number of observations from the i^{th} population and $n_T = n_1 + n_2 + \dots + n_t$

CRD MODEL: $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, where $\varepsilon_{ij} =$ random error \sim independent $N(0, \sigma_\varepsilon^2)$
 μ = arbitrary reference or baseline point
 α_i = effect of i^{th} treatment

CRD ANOVA Table

Source	SS	df	MSQ	E{MSQ}	Fobs
Treatment or Model	SSTr	t-1	MSTr = SSTr/(t-1)	$\sigma_\varepsilon^2 + n \frac{\sum (\alpha_i - \bar{\alpha})^2}{t-1}$	MSTr/MSE
Error	SSE	$n_T - t$	MSE = SSE/($n_T - t$)	σ_ε^2	
Total	SSTot	$n_T - 1$			

NOTES/COMMENTS

1. Notational warning: book uses single summation for multiple sums $\sum_{i,j} y_{ij} = \sum_{i=1}^t \sum_{j=1}^{n_i} y_{ij}$

2. Why does an F-test work? Use Expected Mean Squares!

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_t \Leftrightarrow H_0: \alpha_1 = \alpha_2 = \alpha_3 = \dots = \alpha_t \Leftrightarrow \sum (\alpha_i - \bar{\alpha})^2 = 0$$

$$\Leftrightarrow E(\text{MSTr}) = E(\text{MSE})$$

3. Multiple Comparisons: Test all $H_0: \mu_i = \mu_j$, using Tukey's, Bonferroni, Scheffe.

4. CRD Advantages:

- i. easy to construct
- ii. easy to analyze
- iii. can be used for any number of treatments

5. CRD Disadvantages:

- i. Best suited for relatively few treatments
- ii. EUs must be as homogeneous as possible [may need more observations in a CRD to detect a particular effect size when compared to an RCBD or other designs]

5. Assumptions To Check

- i. Constant Variance of Errors (the ϵ_{ij})
- ii. Normality of Errors (the ϵ_{ij})
- iii. Independence of Errors (the ϵ_{ij})
- iv. No Outliers

Example Bacteria growth in meat under different packaging conditions (again!)

Using R

```
> MeatANOVA = lm(logCount ~ Condition , data=Meat)
> anova(MeatANOVA)
Analysis of Variance Table
Response: logCount
      Df Sum Sq Mean Sq F value    Pr(>F)
Condition  3  32.873   10.958   94.584 1.376e-06 ***
Residuals  8   0.927    0.116
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(aov(logCount ~ Condition, data=Meat))
Analysis of Variance Table
Response: logCount
      Df Sum Sq Mean Sq F value    Pr(>F)
Condition  3  32.873   10.958   94.584 1.376e-06 ***
Residuals  8   0.927    0.116
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> TukeyHSD(aov(logCount ~ Condition, data=Meat))
  Tukey multiple comparisons of means
 95% family-wise confidence level
Fit: aov(formula = logCount ~ Condition, data = Meat)
$Condition
      diff          lwr          upr         p adj
mixed-CO2    3.90    3.010038    4.789962 0.0000031
plastic-CO2    4.12    3.230038    5.009962 0.0000020
vacuum-CO2     2.14    1.250038    3.029962 0.0002639
plastic-mixed  0.22   -0.669962    1.109962 0.8563618
```

vacuum-mixed	-1.76	-2.649962	-0.870038	0.0010160
vacuum-plastic	-1.98	-2.869962	-1.090038	0.0004549

Using SAS

```
title "One-way ANOVA";
title2 "ANOVA Bacteria in meat data";
data meat;
  input condition $ logcount @@;
  cards;
plastic 7.66 plastic 6.98 plastic 7.80 vacuum 5.26 vacuum 5.44 vacuum 5.80
mixed 7.41 mixed 7.33 mixed 7.04 CO2 3.51 CO2 2.91 CO2 3.66
;
proc glm data=meat order=data;
  class condition;
  model logcount=condition;
run;
```

The GLM Procedure

Dependent Variable: logcount

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4.64640000	4.64640000	81.87	0.0008
Error	4	0.22700000	0.05675000		
Corrected Total	5	4.87340000			

R-Square	Coeff Var	Root MSE	logcount Mean
0.953421	3.733896	0.238223	6.380000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
condition	1	4.64640000	4.64640000	81.87	0.0008

Source	DF	Type III SS	Mean Square	F Value	Pr > F
condition	1	4.64640000	4.64640000	81.87	0.0008

RANDOMIZED COMPLETE BLOCK DESIGN (RCBD)

A single factor (treatment) of interest with "t" levels and a block "factor" with "b" levels AND we have observations on every treatment in every block. That is, we have an observation in every cell of the table below!

		Block			
		1	2	...	b
Treatment	1				
	2				
	...				
	t				

Recall that the Block Factor or BLOCKS are defined as a homogeneous unit formed in advance and treatments are randomly assigned within blocks (if "t" units in each block then RCBD).

RCBD MODEL: $y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$, where $\varepsilon_{ij} = \text{random error} \sim \text{independent } N(0, \sigma_\varepsilon^2)$
 $\mu = \text{arbitrary reference or baseline point}$
 $\alpha_i = \text{effect of } i^{\text{th}} \text{ treatment}$
 $\beta_j = j^{\text{th}} \text{ block effect}$

Assume $y_{ij} \sim \text{independent } N(\mu + \alpha_i + \beta_j, \sigma_\varepsilon^2)$, $i = 1, 2, \dots, t$ (trmts), $j = 1, 2, \dots, b$ (blocks)

Mean of $Y_{ij} = E(y_{ij})$		Block			
		1	2	...	b
Treatment	1	$\mu + \alpha_1 + \beta_1$	$\mu + \alpha_1 + \beta_2$...	$\mu + \alpha_1 + \beta_b$
	2	$\mu + \alpha_2 + \beta_1$	$\mu + \alpha_2 + \beta_2$...	$\mu + \alpha_2 + \beta_b$

	t	$\mu + \alpha_t + \beta_1$	$\mu + \alpha_t + \beta_2$...	$\mu + \alpha_t + \beta_b$

Notice: Difference of means in the same block differ only by the α 's.

RCBD ANOVA Table

Source	SS	df	MSQ	E{MSQ}	Fobs
Treatment	SSTr	t-1	MSTr = SSTr/(t-1)	$\sigma_{\epsilon}^2 + n \frac{\sum (\alpha_i - \bar{\alpha})^2}{t-1}$	MSTr/MSE
Block	SSB	b-1	MSB = SSB/(b-1)		
Error	SSE	(b-1)(t-1)	MSE = SSE/(bt-t)	σ_{ϵ}^2	
Total	SSTot	bt-1			

TESTS:

$H_0: \alpha_1 = \alpha_2 = \alpha_3 = \dots = \alpha_t$ vs $H_0: \alpha$'s not all equal

Test Statistic: $F_{obs} = MST_r/MSE$

RR: Reject H_0 if $F_{obs} > F_{\alpha, t-1, (b-1)(t-1)}$ or p-value: $\text{Prob}(F_{t-1, (b-1)(t-1)} > F_{obs})$

NOTES/COMMENTS

- Some argue that blocks should not be tested since no randomization basis for test.
- RCBD has $n_T = b \cdot t$ total observations b/c form "b" blocks with "t" units each.
- Spend "b-1" of error degrees of freedom on blocks [potential COST] in hopes of achieving a smaller residual error for testing treatment effects.
- RCBD Advantages
 - Useful for comparing "t" means in the presence of **ONE** extraneous source of variability
 - Easy analysis
 - Easy design to construct
 - Can accommodate any number of treatments including factorial treatment structure in any number of blocks.
- RCBD Disadvantages
 - Best suited for a relatively small number of treatments
 - Controls only one source of variability [LATIN SQUARES control for 2 sources of variability]
 - Treatment effect must be approximately the same from block to block (ie no interaction between Treatment effect and Block effect).

6. Efficiency of RCBD relative to CRD

In a balanced ($n_i = r$) CRD, $\hat{\text{var}}(\bar{y}_i) = \hat{\text{var}}_{\text{CRD}}(\bar{y}_i) = \frac{\text{MSE}_{\text{CRD}}}{r}$.

In a RCBD, $\hat{\text{var}}(\bar{y}_i) = \hat{\text{var}}_{\text{RCBD}}(\bar{y}_i) = \frac{\text{MSE}_{\text{RCBD}}}{b}$.

$\text{RE}(\text{RCB}, \text{CRD}) = \text{rel. efficiency}(\text{of RCBD to CRD}) = \frac{\hat{\text{var}}_{\text{RCBD}}(\bar{y}_i)}{\hat{\text{var}}_{\text{CRD}}(\bar{y}_i)}$.

IF $\text{RE}(\text{RCBD}, \text{CRD}) < 1$, then this implies that $\hat{\text{var}}_{\text{RCBD}}(\bar{y}_i) < \hat{\text{var}}_{\text{CRD}}(\bar{y}_i)$ and the RCBD is more efficient, "better," than the CRD.

Hence using a "smaller" number of observations, the RCBD produces estimates with the smaller variance!

7. Multiple Comparisons: Test all $H_0: \alpha_i = \alpha_j$, using Tukey's, Bonferroni, Scheffe.
Note that rarely would one perform a MC on the Block Levels!

8. Assumptions To Check

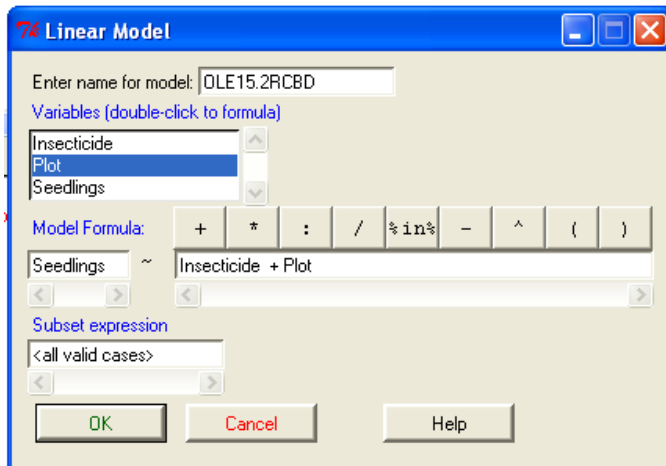
- i. Constant Variance of Errors (the ε_{ij})
- ii. Normality of Errors (the ε_{ij})
- iii. Independence of Errors (the ε_{ij})
- iv. No Outliers

Example OL 15.2

Using R

With *OLE15.2* the active data set, using RCommander:

Statistics > Fit Models > Linear model ...



With the following output:

```
> OLE15.2RCBD <- lm(Seedlings ~ Insecticide + Plot , data=OLE15.2)
> summary(OLE15.2RCBD)
Call:
lm(formula = Seedlings ~ Insecticide + Plot, data = OLE15.2)

Residuals:
    Min       1Q   Median       3Q      Max
-3.000e+00 -1.000e+00  1.226e-14  1.000e+00  2.000e+00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)      56.000      1.472  38.045 2.20e-08 ***
Insecticide[T.2]  29.000      1.472  19.702 1.11e-06 ***
Insecticide[T.3]  22.000      1.472  14.946 5.65e-06 ***
Plot[T.2]         -7.000      1.700  -4.118 0.00623 **
Plot[T.3]         8.000      1.700   4.707 0.00330 **
Plot[T.4]         7.000      1.700   4.118 0.00623 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.082 on 6 degrees of freedom
Multiple R-squared:  0.9887,    Adjusted R-squared:  0.9792
F-statistic: 104.8 on 5 and 6 DF,  p-value: 9.408e-06
```

```

> anova(OLE15.2RCBD)
Analysis of Variance Table

Response: Seedlings
      Df Sum Sq Mean Sq F value    Pr(>F)
Insecticide  2 1832.00   916.00 211.385 2.74e-06 ***
Plot         3   438.00   146.00  33.692 0.0003767 ***
Residuals    6    26.00    4.33
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(lm(Seedlings ~ Insecticide, data=OLE15.2))
Warning in model.matrix.default(mt, mf, contrasts) :
  variable 'Insecticide' converted to a factor
Analysis of Variance Table

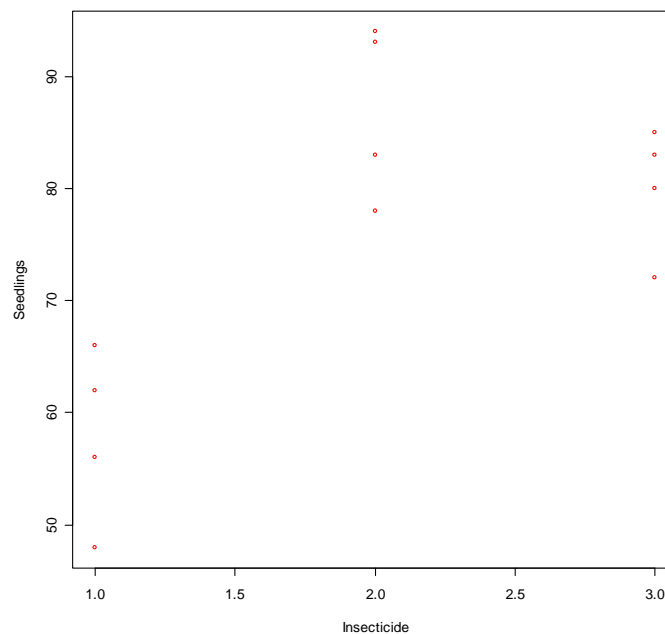
Response: Seedlings
      Df Sum Sq Mean Sq F value    Pr(>F)
Insecticide  2 1832.00   916.00 17.767 0.0007498 ***
Residuals    9   464.00    51.56
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

> scatterplot(Seedlings~Insecticide, reg.line=FALSE, smooth=FALSE, labels=FALSE,
boxplots=FALSE, span=0.5, pch=c(Plot), data=OLE15.2B)

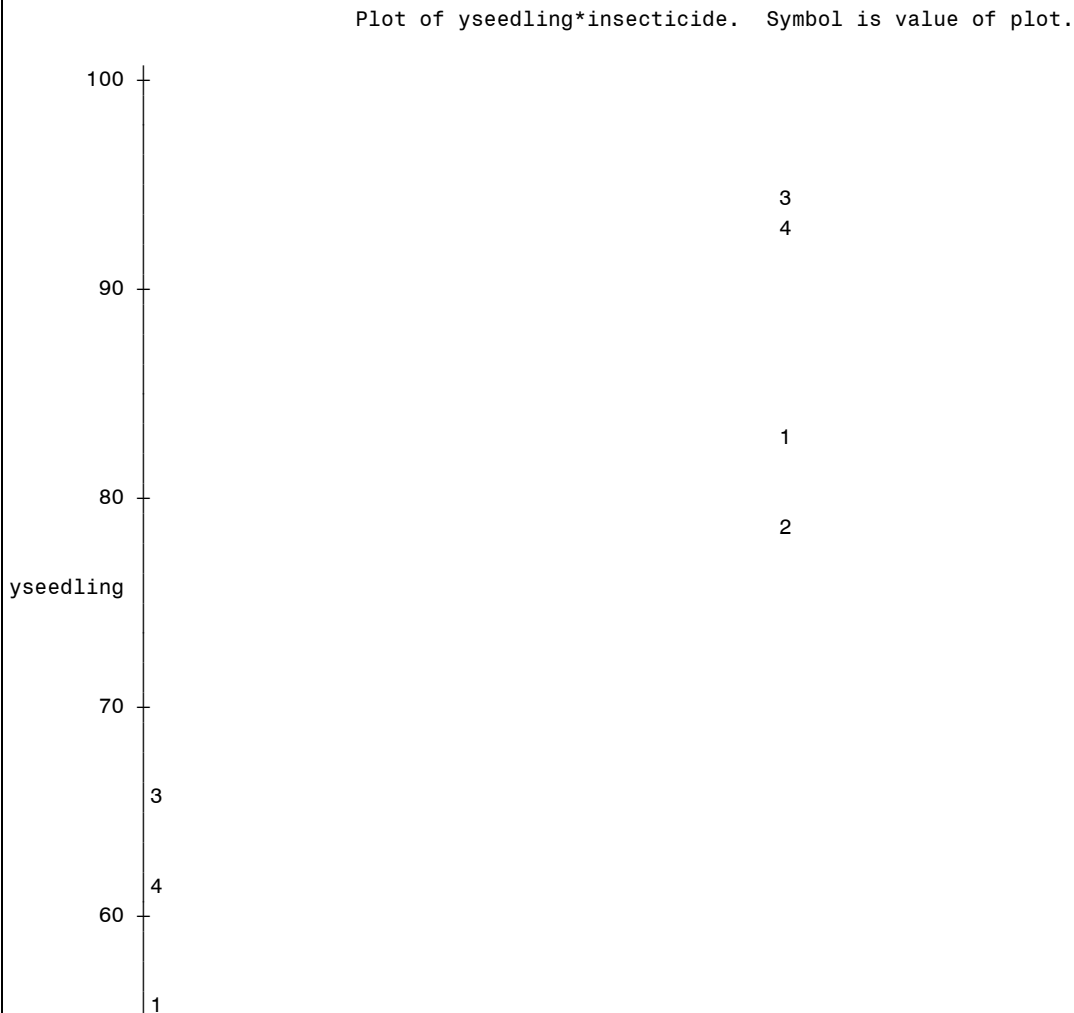
```

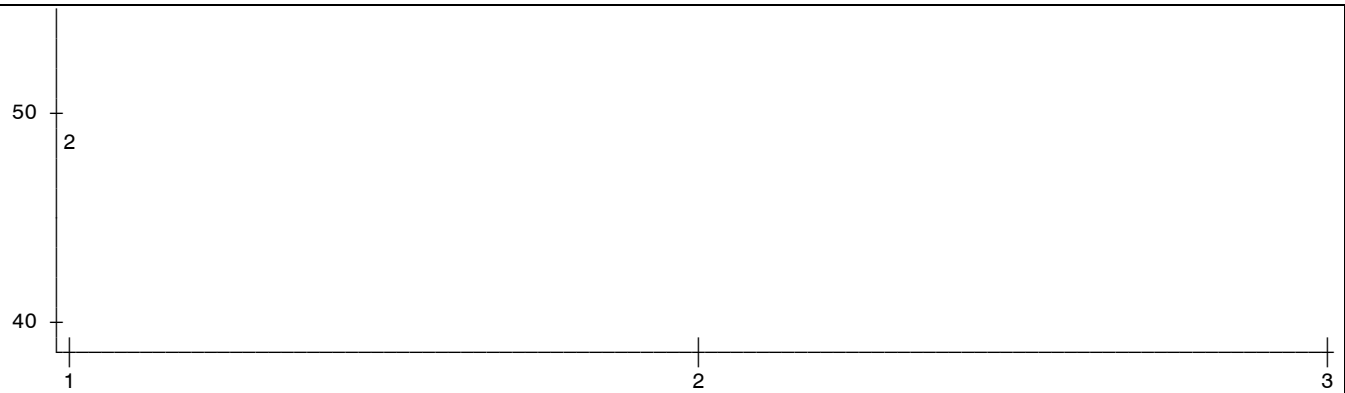


Using SAS

```
options ls=110 pageno=1 formdlim="+" nodate;
title "RCBD - block=plot trt=insecticide";
title2 "Ott/Longnecker p. 868 - example 15.2";
data drcbd;
  input insecticide plot yseedling @@;
  datalines;
1 1 56 1 2 48 1 3 66 1 4 62
2 1 83 2 2 78 2 3 94 2 4 93
3 1 80 3 2 72 3 3 83 3 4 85
;
proc plot;
  plot yseedling*insecticide=plot;
proc glm;
  class plot insecticide;
  model yseedling = plot insecticide;
  means insecticide / tukey;
proc glm;
  class insecticide;
  model yseedling = insecticide;
run;
```

RCBD - block=plot trt=insecticide
Ott/Longnecker p. 868 - example 15.2





insecticide

```

+++++
RCBD - block=plot trt=insecticide
Ott/Longnecker p. 868 - example 15.2
2

```

The GLM Procedure

Class Level Information

Class	Levels	Values
plot	4	1 2 3 4
insecticide	3	1 2 3

Number of Observations Read 12
 Number of Observations Used 12

```

+++++
RCBD - block=plot trt=insecticide
Ott/Longnecker p. 868 - example 15.2
3

```

The GLM Procedure

Dependent Variable: yseedling

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2270.000000	454.000000	104.77	<.0001
Error	6	26.000000	4.333333		
Corrected Total	11	2296.000000			

R-Square 0.988676
 Coeff Var 2.775555
 Root MSE 2.081666
 yseedling Mean 75.00000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
plot	3	438.000000	146.000000	33.69	0.0004
insecticide	2	1832.000000	916.000000	211.38	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
plot	3	438.000000	146.000000	33.69	0.0004
insecticide	2	1832.000000	916.000000	211.38	<.0001

```

+++++
RCBD - block=plot trt=insecticide
Ott/Longnecker p. 868 - example 15.2
4

The GLM Procedure

Tukey's Studentized Range (HSD) Test for yseedling

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error
rate than REGWQ.

Alpha                0.05
Error Degrees of Freedom      6
Error Mean Square          4.333333
Critical Value of Studentized Range  4.33902
Minimum Significant Difference    4.5162

Means with the same letter are not significantly different.

Tukey Grouping
      Mean      N  insecticide
A      87.000    4    2
B      80.000    4    3
C      58.000    4    1
+++++
RCBD - block=plot trt=insecticide
Ott/Longnecker p. 868 - example 15.2
5

The GLM Procedure

      Class Level Information

Class      Levels  Values
insecticide      3    1 2 3

Number of Observations Read      12
Number of Observations Used      12
+++++
RCBD - block=plot trt=insecticide
Ott/Longnecker p. 868 - example 15.2
6

The GLM Procedure

Dependent Variable: yseedling

Source              DF      Sum of
                   Squares    Mean Square    F Value    Pr > F
Model                2      1832.000000    916.000000    17.77    0.0007
Error                9       464.000000    51.555556
Corrected Total     11      2296.000000

R-Square      Coeff Var      Root MSE    yseedling Mean
0.797909      9.573626      7.180220      75.00000

Source              DF      Type I SS    Mean Square    F Value    Pr > F
insecticide        2      1832.000000    916.000000    17.77    0.0007

```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
insecticide	2	1832.000000	916.000000	17.77	0.0007

Latin Square Design

Suppose our Treatment has "t" levels.

We consider two sources of extraneous variation, call them Block Factors: A and B, BOTH with "t" levels!

A "t x t" Latin Square has "t" rows and "t" columns ("t" treatments are randomly assigned to EUs within rows and columns so that every treatment appears in every row and column)

e.g. t=3

	B ₁	B ₂	B ₃		B ₁	B ₂	B ₃		B ₁	B ₂	B ₃
A ₁	T ₁	T ₂	T ₃	A ₁	T ₂	T ₃	T ₁	A ₁	T ₃	T ₁	T ₂
A ₂	T ₃	T ₁	T ₂	A ₂	T ₁	T ₂	T ₃	A ₂	T ₂	T ₃	T ₁
A ₃	T ₂	T ₃	T ₁	A ₃	T ₃	T ₁	T ₂	A ₃	T ₁	T ₂	T ₃

LATIN SQUARE MODEL:

$y_{ij} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ij}$, where ε_{ij} = random error \sim independent $N(0, \sigma_\varepsilon^2)$
 μ = arbitrary reference or baseline point
 α_i = effect of i^{th} "A" Block
 β_j = effect of j^{th} "B" Block
 γ_j = effect of k^{th} Treatment

$y_{ij} = \mu + A_i + B_j + Tr_k + \varepsilon_{ij}$,

Assume $y_{ij} \sim$ independent $N(\mu + \alpha_i + \beta_j + \gamma_k, \sigma_\varepsilon^2)$

Mean of $Y_{ij} = E(y_{ij})$		B Block		
		1	2	3
A Block	1	$\mu + \alpha_1 + \beta_1 + \gamma_1$	$\mu + \alpha_1 + \beta_2 + \gamma_2$	$\mu + \alpha_1 + \beta_3 + \gamma_3$
	2	$\mu + \alpha_2 + \beta_1 + \gamma_3$	$\mu + \alpha_2 + \beta_2 + \gamma_1$	$\mu + \alpha_2 + \beta_3 + \gamma_2$
	3	$\mu + \alpha_3 + \beta_1 + \gamma_2$	$\mu + \alpha_3 + \beta_2 + \gamma_3$	$\mu + \alpha_3 + \beta_3 + \gamma_1$

Now compare the observations with

Treatment 1:

Treatment 2:

Treatment 3:

LATIN SQUARE ANOVA Table

Source	SS	df	MSQ	E{MSQ}	Fobs
Treatment	SSTr	t-1	MSTr = SSTr/(t-1)	$\sigma_{\epsilon}^2 + n \frac{\sum(\gamma_i - \bar{\gamma})^2}{t-1}$	MSTr/MSE
Block A	SSA	t-1	MSA = SSA/(t-1)		
Block B	SSB	t-1	MSB = SSB/(t-1)		
Error	SSE		MSE= SSE/(t ² -t)	σ_{ϵ}^2	
Total	SSTot	t ² -1			

TESTS:

H₀: $\gamma_1 = \gamma_2 = \gamma_3 = \dots = \gamma_t$ vs H₀: γ 's not all equal

Test Statistic: $F_{obs} = MST_r/MSE$

RR: Reject H₀ if p-value: $\text{Prob}(F_{t-1, (b-1)(t-1)} > F_{obs})$ small

NOTES/COMMENTS

1. Again we argue that neither block effect should not be tested.

2. Multiple Comparison of the Treatment Effects would be done as usual.

3. Latin Square Advantage

- i. Accounts for Two extraneous sources of variation.
- ii. Requires far fewer EU's than an RCBD with two Blocks!

4. RCBD Disadvantage

- i. The number of levels of both block factors AND the treatment MUST all be the same!

Factorial Designs

Most time more than one variable "makes up" or "defines the treatment factor. Suppose we have two factors of interest, all them Factor A and Factor B, with "a" and "b" levels, respectively.

Factorial MODEL:

$$y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \quad \text{where } \varepsilon_{ij} = \text{random error} \sim \text{independent } N(0, \sigma_\varepsilon^2)$$

μ = arbitrary reference or baseline point
 α_i = effect of the i^{th} level of the A Factor
 β_j = effect of the j^{th} level of the B Factor
 $(\alpha\beta)_{ij}$ = INTERACTION effect the i^{th} level of the A Factor with the j^{th} level of the B Factor
 $i = 1, \dots, a$ (Factor A levels)
 $j = 1, \dots, b$ (Factor B levels)
 $k = 1, \dots, n_{ij}$
 N = sum of all the n_{ij}

$$y_{ij} = \mu + A_i + B_j + A*B_{ij} + \varepsilon_{ij},$$

Assume $y_{ij} \sim \text{independent } N(\mu + \alpha_i + \beta_j + \gamma(\alpha\beta)_{ij}, \sigma_\varepsilon^2)$

E(y _{ijk})		Factor B			
		1	2	...	b
Factor A	1	$\mu + \alpha_1 + \beta_1 + (\alpha\beta)_{11}$	$\mu + \alpha_1 + \beta_2 + (\alpha\beta)_{12}$...	$\mu + \alpha_1 + \beta_b + (\alpha\beta)_{1b}$
	2	$\mu + \alpha_2 + \beta_1 + (\alpha\beta)_{21}$	$\mu + \alpha_2 + \beta_2 + (\alpha\beta)_{22}$...	$\mu + \alpha_2 + \beta_b + (\alpha\beta)_{2b}$

	a	$\mu + \alpha_a + \beta_1 + (\alpha\beta)_{a1}$	$\mu + \alpha_a + \beta_2 + (\alpha\beta)_{a2}$...	$\mu + \alpha_a + \beta_b + (\alpha\beta)_{ab}$

Notice: Difference of means in the same level of one factor differ by the α 's AND the interaction terms $(\alpha\beta)$'s.

	E(y _{ijk})	Factor B			
		1	2	...	b
Factor A	1	$\mu + \alpha_1 + \beta_1 + (\alpha\beta)_{11}$	$\mu + \alpha_1 + \beta_2 + (\alpha\beta)_{12}$...	$\mu + \alpha_1 + \beta_b + (\alpha\beta)_{1b}$
	2	$\mu + \alpha_2 + \beta_1 + (\alpha\beta)_{21}$	$\mu + \alpha_2 + \beta_2 + (\alpha\beta)_{22}$...	$\mu + \alpha_2 + \beta_b + (\alpha\beta)_{2b}$

Twoway Factorial ANOVA Table

Source	SS	df	MSQ	F _{obs}
Factor A	SSA	a-1	MSA = SSA/(a-1)	MSA/MSE
Factor B	SSB	b-1	MSB = SSB/(b-1)	MSB/MSE
Interaction	SSAB	(a-1)(b-1)	MSAB = SSAB/(a-1)(b-1)	MSAB/MSE
Error	SSE	N-ab	MSE=SSE/(N-ab)	
Total	TSS	N-1		

TESTS:

H₀: $\alpha\beta_{ij} = 0$ for all i,j (No interaction)

Test Statistic: $F_{obs} = MSAB/MSE$

RR: Reject H₀ if $F_{obs} > F_{\alpha, (a-1)(b-1), N-ab}$

P-value: $\text{Prob}(F_{(a-1)(b-1), N-ab} > F_{obs})$

-

H₀: $\alpha_1 = \alpha_2 = \alpha_3 = \dots = \alpha_a = 0$ (No A Main Effect)

Test Statistic: $F_{obs} = MSA/MSE$

RR: Reject H₀ if $F_{obs} > F_{\alpha, a-1, N-ab}$

P-value: $\text{Prob}(F_{a-1, N-ab} > F_{obs})$

H₀: $\beta_1 = \beta_2 = \beta_3 = \dots = \beta_b = 0$ (No B Main Effect)

Test Statistic: $F_{obs} = MSB/MSE$

RR: Reject H₀ if $F_{obs} > F_{\alpha, b-1, N-ab}$

P-value: $\text{Prob}(F_{b-1, N-ab} > F_{obs})$

$y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ij}$ where $\varepsilon_{ij} \sim \text{ind. } N(0, \quad)$

```

title "Factorial example: Factor A=pesticide Factor B=variety";
title2 "Ott/Longnecker p. 901 - example 15.8";
data dfact;
  input variety pesticide yield @@;
  datalines;
1 1 49 1 1 39 1 2 50 1 2 55 1 3 43 1 3 38 1 4 53 1 4 48
2 1 55 2 1 41 2 2 67 2 2 58 2 3 53 2 3 42 2 4 85 2 4 73
3 1 66 3 1 68 3 2 85 3 2 92 3 3 69 3 3 62 3 4 85 3 4 99
;
* Generate the Mean Profile plot;
proc sort; by variety pesticide;
proc means noprint; by variety pesticide;
  var yield;
  output out=factmean mean=ymean;
run;
proc print;
run;
proc plot data=factmean;
  plot ymean*pesticide=variety;
run;

```

```
* Test components of the Two-way anova model;  
proc glm data=dfact;  
  class pesticide variety;  
  model yield = variety pesticide variety*pesticide;  
  means variety / tukey;  
  means pesticide / tukey;  
run;
```

COMMENT: This plot suggests NO INTERACTION between pesticide and variety. It also suggests a difference in both PESTICIDE levels and VARIETY levels. Let's see if this is observed in the formal hypothesis tests.

COMMENT: We would fail to reject the (null) hypothesis of NO INTERACTION between pesticide and variety ($P=0.1817$). The main effects of VARIETY and PESTICIDE are both significant at P -values of $<.0001$ and $.0001$, respectively. Thus, we conclude that YIELD differs for both different varieties and pesticides; however, these factors do not interact.

COMMENT: TYPE III table = TYPE I table if the n_{ij} are the same in all factor level combinations (balanced data). TYPE I corresponds to sequential tests (test of term given all terms above it) while TYPE III corresponds to partial/adjusted tests (test of term given all other terms are in the model). It is usually recommended that you consider the TYPE III tests.

Comment: The TUKEY procedure is comparing means of VARIETY levels that are pooled across levels of the PESTICIDE factor. This makes sense if the factors do not interact.

COMMENT: If you have significant interactions present, then you may want to analyze the study as a one-way anova. In the variety-pesticide study, you have $3 \times 4 = 12$ unique factor level combinations that define the treatments. We can reanalyze these data using a one-way anova with 12 levels. ASIDE: This is mainly a pedagogical exercise since the FACTORS did not interact, there is no strong reason to do this unless you want to identify the variety-pesticide combination that leads to the maximal response.

COMMENT: Variety = 1, 2, 3 and Pesticide = 1, 2, 3, 4 so defining COMBO = $10 \times \text{variety} + 1 \times \text{pesticide}$ yields a treatment with levels 11, 12, 13, 14, 21, 22, 23, 24, 31, 32, 33, 34.

```
title "Factorial - Factor A=pesticide Factor B=variety";
title2 "Ott/Longnecker p. 901 - example 15.8";
title3 "redo as a one-way anova";
data dfact;
  input variety pesticide yield @@;
  combo = 10*variety + 1*pesticide;      * coding of combined treatment;
datalines;
1 1 49 1 1 39 1 2 50 1 2 55 1 3 43 1 3 38 1 4 53 1 4 48
2 1 55 2 1 41 2 2 67 2 2 58 2 3 53 2 3 42 2 4 85 2 4 73
3 1 66 3 1 68 3 2 85 3 2 92 3 3 69 3 3 62 3 4 85 3 4 99
;
proc glm;
  class combo;
```

```

model yield = combo;
means combo / tukey;
run;

```

Factorial - Factor A=pesticide Factor B=variety
 Ott/Longnecker p. 901 - example 15.8
 redo as a one-way anova

The GLM Procedure

Class Level Information		
Class	Levels	Values
combo	12	11 12 13 14 21 22 23 24 31 32 33 34

Number of Observations Read	24
Number of Observations Used	24

Dependent Variable: yield

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	6680.458333	607.314394	14.36	<.0001
Error	12	507.500000	42.291667		
Corrected Total	23	7187.958333			

R-Square	Coeff Var	Root MSE	yield Mean
0.929396	10.58149	6.503204	61.45833

Source	DF	Type I SS	Mean Square	F Value	Pr > F
combo	11	6680.458333	607.314394	14.36	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
combo	11	6680.458333	607.314394	14.36	<.0001

Tukey's Studentized Range (HSD) Test for yield

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	12
Error Mean Square	42.29167
Critical Value of Studentized Range	5.61464
Minimum Significant Difference	25.819

Means with the same letter are not significantly different.						
Tukey Grouping				Mean	N	combo
		A		92.000	2	34
		A				
B		A		88.500	2	32
B		A				
B		A	C	79.000	2	24
B		A	C			
B	D	A	C	67.000	2	31
B	D		C			
B	D	E	C	65.500	2	33
	D	E	C			
	D	E	C	62.500	2	22
	D	E				
	D	E		52.500	2	12
	D	E				
	D	E		50.500	2	14
	D	E				
	D	E		48.000	2	21
	D	E				

Means with the same letter are not significantly different.						
Tukey Grouping			Mean	N	combo	
	D	E	47.500	2	23	
	D	E				
	D	E	44.000	2	11	
		E				
		E	40.500	2	13	

Suppose your data are not balanced. This will often be the case even if the design starts out as balanced (beakers break, algal blooms kill all organisms in an aquarium, etc.) What will this do to the output of a factorial analysis? HINT: compare the TYPE I and TYPE III tables.

```

title "Factorial - Factor A=pesticide Factor B=variety";
title2 "Ott/Longnecker p. 901 - example 15.8";
title3 "what if missing data in a couple of cells";
data dfact;
  input variety pesticide yield @@;
  datalines;
1 1 49 1 1 39 1 2 . 1 2 55 1 3 43 1 3 38 1 4 53 1 4 48
2 1 55 2 1 41 2 2 67 2 2 58 2 3 53 2 3 42 2 4 85 2 4 .
3 1 . 3 1 68 3 2 85 3 2 92 3 3 69 3 3 62 3 4 85 3 4 99
;
proc glm;
  class pesticide variety;
  model yield = variety pesticide variety*pesticide;
run;

```

Factorial - Factor A=pesticide Factor B=variety
 Ott/Longnecker p. 901 - example 15.8
 what if missing data in a couple of cells

The GLM Procedure

Class Level Information		
Class	Levels	Values
pesticide	4	1 2 3 4
variety	3	1 2 3

Number of Observations Read	24
------------------------------------	----

Number of Observations Used	21
------------------------------------	----

Dependent Variable: yield

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	6480.809524	589.164502	12.59	0.0004
Error	9	421.000000	46.777778		
Corrected Total	20	6901.809524			

R-Square	Coeff Var	Root MSE	yield Mean
0.939002	11.16858	6.839428	61.23810

Source	DF	Type I SS	Mean Square	F Value	Pr > F
variety	2	4108.666667	2054.333333	43.92	<.0001
pesticide	3	1864.336975	621.445658	13.29	0.0012
pesticide*variety	6	507.805882	84.634314	1.81	0.2035

Source	DF	Type III SS	Mean Square	F Value	Pr > F
variety	2	3096.800000	1548.400000	33.10	<.0001
pesticide	3	2096.211538	698.737179	14.94	0.0008
pesticide*variety	6	507.805882	84.634314	1.81	0.2035

I showed you an ANCOVA analysis where the assumptions were violated (the slopes were not equal when comparing Tahoe Keys to Eagle lake with respect to log(DO) - depth relationships). The next example is one where the traditional ANCOVA assumption holds.

ANCOVA

$$y_{ij} = \mu + \alpha_i + \beta x_{ij} + \varepsilon_{ij} \text{ where } \varepsilon_{ij} \sim \text{ind. } N(0, \quad)$$

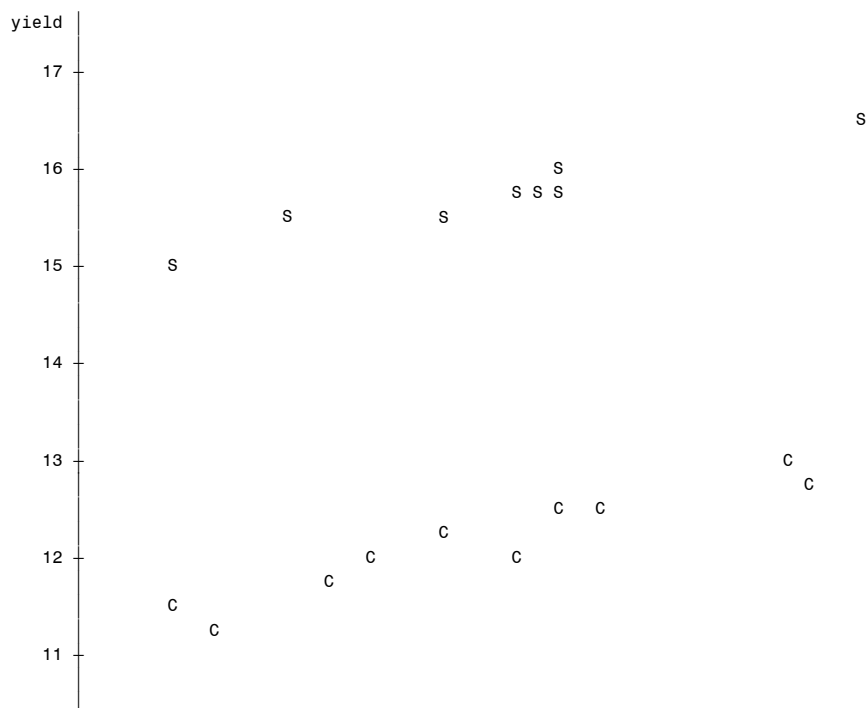
```
title "ANCOVA - Factor =Fertilizer Covariate=height";
title2 "Ott/Longnecker p. 947 - example 16.1";
data dancova;
  input fertilizer $ yield height @@;
  datalines;
C 12.2 45 C 12.4 52 C 11.9 42 C 11.3 35 C 11.8 40 C 12.1 48
C 13.1 60 C 12.7 61 C 12.4 50 C 11.4 33
S 16.6 63 S 15.8 50 S 16.5 63 S 15.0 33 S 15.4 38 S 15.6 45
S 15.8 50 S 15.8 48 S 16.0 50 S 15.8 49
F 9.5 52 F 9.5 54 F 9.6 58 F 8.8 45 F 9.5 57 F 9.8 62
F 9.1 52 F 10.3 67 F 9.5 55 F 8.5 40
;

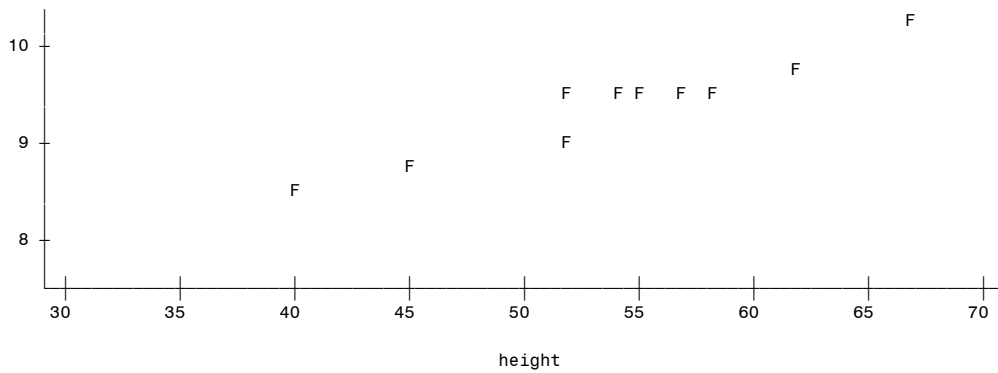
proc plot;
  plot yield*height=fertilizer;
  run;

proc glm;
  class fertilizer;
  model yield = height|fertilizer;
  run;

proc glm;
  class fertilizer;
  model yield = height fertilizer;
  lsmeans fertilizer / pdiff;
  run;
```

Plot of yield*height. Symbol is value of fertilizer.





NOTE: 2 obs hidden.

Notice: The yield is linearly related to the covariate (height) in each fertilizer group.

ANCOVA - Factor =Fertilizer Covariate=height
 Ott/Longnecker p. 947 - example 16.1

The GLM Procedure

Class Level Information		
Class	Levels	Values
fertilizer	3	C F S

Number of Observations Read	30
Number of Observations Used	30

Dependent Variable: yield

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	214.4372247	42.8874449	2887.70	<.0001
Error	24	0.3564420	0.0148517		
Corrected Total	29	214.7936667			

R-Square	Coeff Var	Root MSE	yield Mean
0.998341	0.978334	0.121868	12.45667

Source	DF	Type I SS	Mean Square	F Value	Pr > F
height	1	0.4721494	0.4721494	31.79	<.0001
fertilizer	2	213.9038045	106.9519022	7201.30	<.0001

Source	DF	Type I SS	Mean Square	F Value	Pr > F
height*fertilizer	2	0.0612708	0.0306354	2.06	0.1491

Source	DF	Type III SS	Mean Square	F Value	Pr > F
height	1	6.65321124	6.65321124	447.97	<.0001
fertilizer	2	6.69631934	3.34815967	225.44	<.0001
height*fertilizer	2	0.06127080	0.03063540	2.06	0.1491

ANCOVA - Factor =Fertilizer Covariate=height
Ott/Longnecker p. 947 - example 16.1

The GLM Procedure

Dependent Variable: yield

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	214.3759539	71.4586513	4447.85	<.0001
Error	26	0.4177128	0.0160659		
Corrected Total	29	214.7936667			

R-Square	Coeff Var	Root MSE	yield Mean
0.998055	1.017537	0.126751	12.45667

Source	DF	Type I SS	Mean Square	F Value	Pr > F
height	1	0.4721494	0.4721494	29.39	<.0001
fertilizer	2	213.9038045	106.9519022	6657.08	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
height	1	6.6932872	6.6932872	416.62	<.0001
fertilizer	2	213.9038045	106.9519022	6657.08	<.0001

The GLM Procedure
Least Squares Means

fertilizer	yield LSMEAN	LSMEAN Number
C	12.3141728	1
F	9.1700172	2
S	15.8858099	3

Comment: The LSMEANS compares the yields for the different fertilizer groups after adjusting for the covariate.

Least Squares Means for effect fertilizer Pr > t for H0: LSMean(i)=LSMean(j) Dependent Variable: yield			
i/j	1	2	3
1		<.0001	<.0001
2	<.0001		<.0001
3	<.0001	<.0001	

Note: To ensure overall protection level, only probabilities associated with pre-planned comparisons should be used.

Finally, suppose that the factor levels were not FIXED but where sampled from some population of factor levels. This naturally leads to a random (or mixed) effects model. Here is a simple illustration.

Random Effects Models

$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ where $\alpha_i \sim N(0, \quad)$ and $\varepsilon_{ij} \sim \text{ind. } N(0, \quad)$

```

title "Random effect";
title2 "Ott/Longnecker p. 981 - example 17.1";
data draneff;
  input station intensity @@;
  datalines;
1 20 1 1050 1 3200 1 5600 1 50
2 4300 2 70 2 2560 2 3650 2 80
3 100 3 7700 3 8500 3 2960 3 3340
;
proc glm;
  class station;

```

```

model intensity=station;
random station;
run;

ods html close;

```

Random effect
Ott/Longnecker p. 981 - example 17.1

The GLM Procedure

Class Level Information		
Class	Levels	Values
station	3	1 2 3

Number of Observations Read	15
Number of Observations Used	15

Random effect
Ott/Longnecker p. 981 - example 17.1

The GLM Procedure

Dependent Variable: intensity

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	20259573.3	10129786.7	1.38	0.2884
Error	12	87989600.0	7332466.7		
Corrected Total	14	108249173.3			

R-Square	Coeff Var	Root MSE	intensity Mean
0.187157	94.06622	2707.853	2878.667

Source	DF	Type I SS	Mean Square	F Value	Pr > F
station	2	20259573.33	10129786.67	1.38	0.2884

Source	DF	Type III SS	Mean Square	F Value	Pr > F
--------	----	-------------	-------------	---------	--------

Source	DF	Type III SS	Mean Square	F Value	Pr > F
station	2	20259573.33	10129786.67	1.38	0.2884

The GLM Procedure

Source	Type III Expected Mean Square
station	Var(Error) + 5 Var(station)