ANOVA

- Two way ANOVA without replication
- Paired comparison
- Comparing means with a control
- Three way ANOVA
- Latin square design
- Hierarchical (nested) ANOVA

Review

One-way ANOVA

Randomized block designs

Two-way ANOVA

Quick R (http://www.statmethods.net/)

```
# One Way Anova (Completely Randomized Design)
attach(mtcars); head(mtcars)

fit1 <- aov(mpg ~ cyl, data=mtcars)

B=cyl; A=gear; x=wt

# Randomized Block Design (B is the blocking factor)
fit2 <- aov(mpg ~ A + B, data=mtcars)

# Two Way Factorial Design
fit3 <- aov(mpg ~ A + B + A:B, data=mtcars)
fit4 <- aov(mpg ~ A*B, data=mtcars) # same thing

# Analysis of Covariance
fit5 <- aov(mpg ~ A + x, data=mtcars)

summary(fit1) # display Type I ANOVA table
drop1(fit1,~.,test="F") # type III SS and F Tests
```

The table for one way ANOVA

Source	SS	d.f.	MS	F	p-value
Treat	SS_T	t-1	MS_T	MS_T/MS_E	
Error	SS_E	(n-t)	MS_E		

Randomized block experiment

Source	SS	d.f.	MS	F	p-value
Treat	SS_T	t-1	MS_T	MS_T/MS_E	
Block	SS_B	b-1	MS_B	MS_B/MS_E	
Error	SS_E	(t-1)(b-1)	MS_E		

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21	6	160	110	3.9	2.62	16.46	0	1	4	4
Mazda RX4 Wag	21	6	160	110	3.9	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.44	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.46	20.22	1	0	3	1
Duster 360	14.3	8	360	245	3.21	3.57	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.19	20	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.15	22.9	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.44	18.3	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.44	18.9	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.07	17.4	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.73	17.6	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.78	18	0	0	3	3
Cadillac Fleetwood	10.4	8	472	205	2.93	5.25	17.98	0	0	3	4
Lincoln Continental	10.4	8	460	215	3	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.2	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.9	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.7	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318	150	2.76	3.52	16.87	0	0	3	2
AMC Javelin	15.2	8	304	150	3.15	3.435	17.3	0	0	3	2

Quiz

```
data(mtcars)
nrow(mtcars) # 32
mtcars$cyl = as.factor(mtcars$cyl)
levels(mtcars$cyl) # “4” “6” “8”

model = aov(mpg~cyl, data = mtcars)
summary(model)
```

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
cyl					4.98 E-09
Resid-uals		301.3			
Total		1126.1			

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21	6	160	110	3.9	2.62	16.46	0	1	4	4
Mazda RX4 Wag	21	6	160	110	3.9	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.44	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.46	20.22	1	0	3	1
Duster 360	14.3	8	360	245	3.21	3.57	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.19	20	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.15	22.9	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.44	18.3	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.44	18.9	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.07	17.4	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.73	17.6	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.78	18	0	0	3	3
Cadillac Fleetwood	10.4	8	472	205	2.93	5.25	17.98	0	0	3	4
Lincoln Continental	10.4	8	460	215	3	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.2	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.9	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.7	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318	150	2.76	3.52	16.87	0	0	3	2
AMC Javelin	15.2	8	304	150	3.15	3.435	17.3	0	0	3	2

Quiz

```
data(mtcars)
nrow(mtcars) # 32
mtcars$cyl = as.factor(mtcars$cyl)
levels(mtcars$cyl) # “4” “6” “8”

model = aov(mpg~cyl, data = mtcars)
summary(model)
```

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
cyl	2	824.8	412.4	39.7	4.98 E-09
Resid-uals	29	301.3	10.4		
Total	31	1126.1			

Assumptions of ANOVA

– Observations are independent of each other

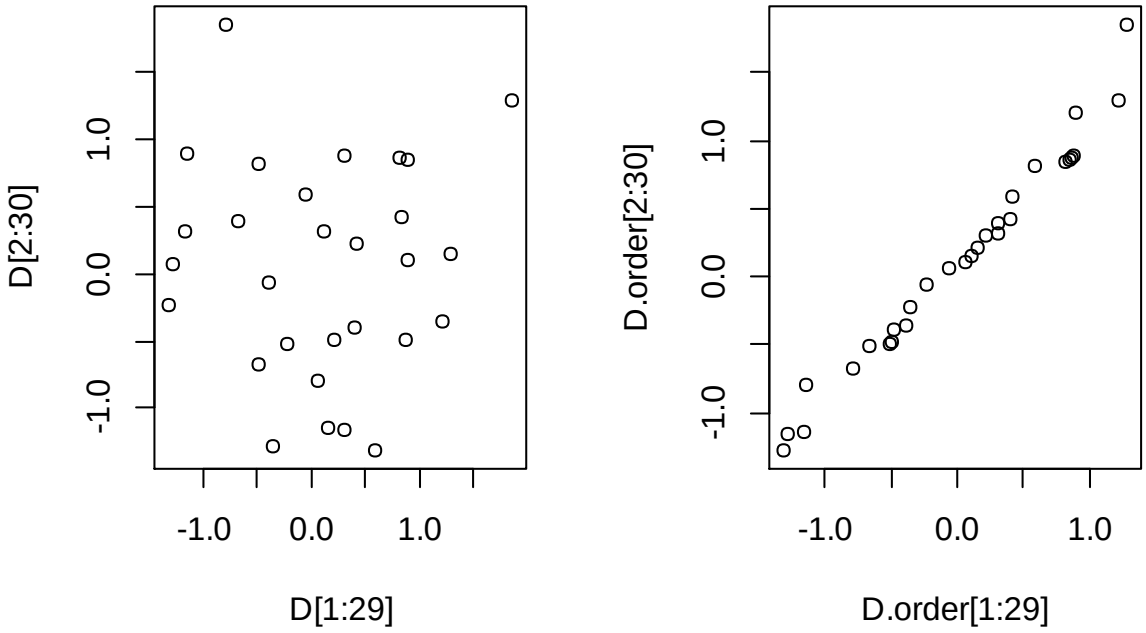
– Residuals are normally distributed

– Variances in groups are homogeneous

– Residuals are homogeneous

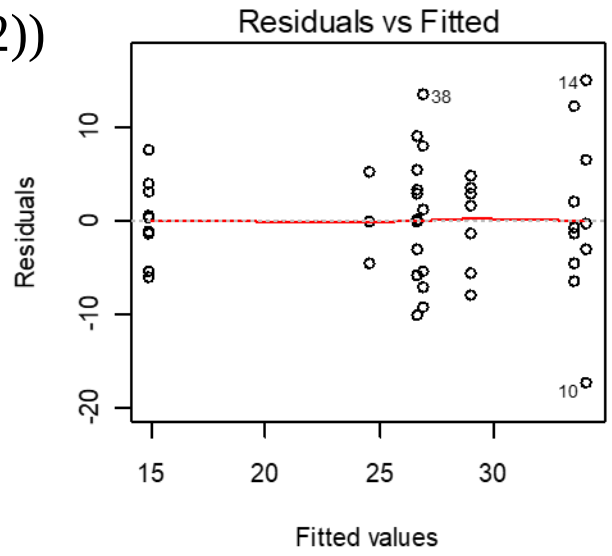
How to check independence

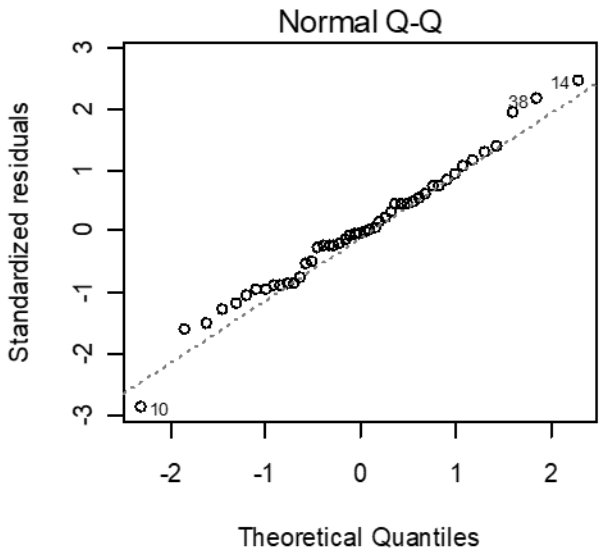
```
par(mfrow=c(1,2))
D = rnorm(30)
plot(D[1:29],D[2:30])
D.order = sort(D)
plot(D.order[1:29],D.order[2:30])
```

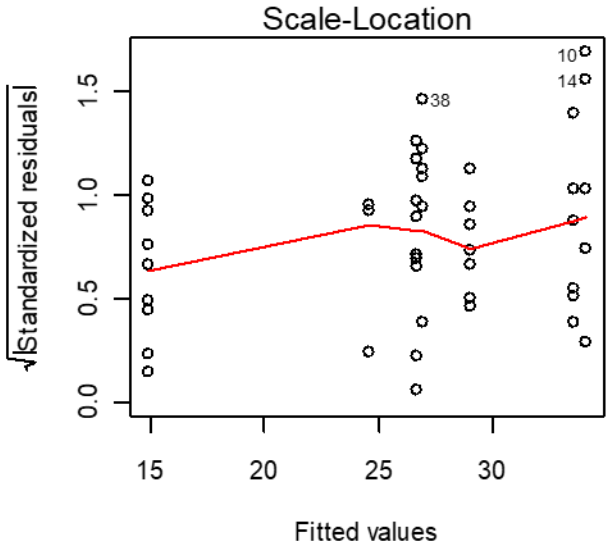


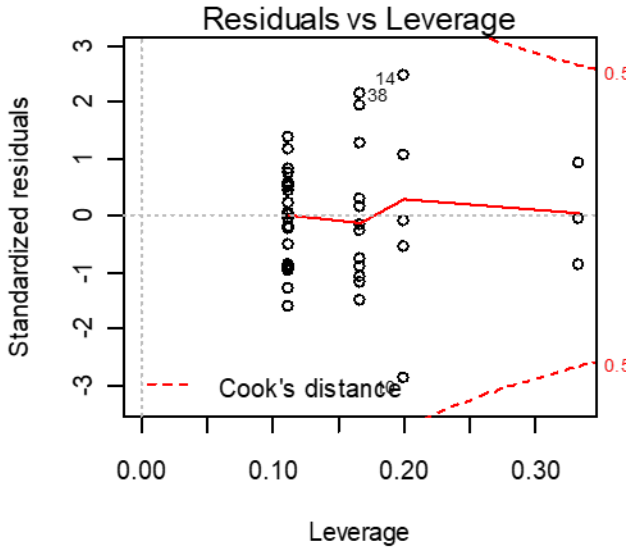
Model performance

```
par(mfrow=c(2,2))
plot(fit)
```









Two way ANOVA without replication – interaction can't be checked

Source of variance	Sum of squares (SS)	Degrees of freedom (DF)	Mean square (MS)
Total	$\sum_{i=1}^{a}\sum_{j=1}^{b} X_{ij}^2 - C$	N-1	
Factor A	$\frac{\sum_{i=1}^{a}\left(\sum_{j=1}^{b} X_{ij}\right)^2}{b} - C$	a-1	$\frac{\text{factor A SS}}{\text{factor A DF}}$
Factor B	$\frac{\sum_{j=1}^{b}\left(\sum_{i=1}^{a} X_{ij}\right)^2}{a} - C$	b-1	$\frac{\text{factor B SS}}{\text{factor B DF}}$
Remainder	total SS - factor *A* SS - factor *B* SS	(a-1)(b-1)	$\frac{\text{remainder SS}}{\text{remainder DF}}$

Here, a is the number of levels in factor *A*, and *b* is the number of levels in factor *B*.

$$C = \frac{\left(\sum_{i=1}^{a}\sum_{j=1}^{b} X_{ij}\right)^2}{N}, \text{and } \mathrm{N} = ab$$

Two way ANOVA without replication

No enough degree of freedom for quantifying the interaction term

fit1 <- aov(W ~ plot + type, data = mydata)

W	plot	type
7.33	1	1
7.49	1	2
7.27	1	3
7.18	1	4
7.56	1	5
7.81	1	6
7.46	1	7
7.84	1	8
7.29	2	1
7.64	2	2
7.25	2	3
7.67	2	4
7.04	2	5
7.1	2	6
7.74	2	7
7.43	2	8
7.7	3	1
7.49	3	2
7.27	3	3
7.65	3	4
7.83	3	5
7.81	3	6
7.46	3	7
7.14	3	8

Paired comparison

Facial width

Individual	Age (year) 0	1
1	7.33	7.49
2	7.11	7.27
3	7.27	7.93
4	7.63	7.56
5	7.56	7.81
6	7.81	7.46
.	.	.
15	6.94	7.49

fw	age	ind
7.33	0	1
7.49	0	2
7.27	0	3
7.18	0	4
7.56	0	5
7.81	0	6
7.46	0	7
7.84	0	8
7.29	0	9
7.64	0	10
7.25	0	11
7.67	0	12
7.04	0	13
7.1	0	14
7.74	0	15
7.43	1	1
7.7	1	2
7.49	1	3
7.27	1	4
7.65	1	5
7.83	1	6
7.81	1	7
7.46	1	8
7.14	1	9
7.84	1	10
7.44	1	11
7.95	1	12
7.71	1	13
7.8	1	14
7.89	1	15

Model

$$fw = \beta_0 + \beta_{age} \times age + error$$

$$fw = \beta_0 + \beta_{age} \times age + \beta_{ind} \times ind + error$$

Comparing means with a control

Group	Weight.DNA (mg/g)						
Normal	12.3	13.2	13.7	15.2	15.4	15.8	16.9
Exp1	10.8	11.6	12.3	12.7	13.5	13.5	14.8
Exp2	9.8	10.3	11.1	11.7	11.7	12.0	12.3

Dunnett's test

Compute a t-test between each experimental group and the control group using the formula:

$$t_d = \frac{M_i - M_c}{\sqrt{\frac{2MSE}{n_h}}}$$

where M_i is the mean of the ith experimental group, M_c is the mean of the control group, MSE is the mean square error as computed from the analysis of variance, and n_h is the harmonic mean of the sample sizes of the experimental group and the control group.

The degrees of freedom (df) for the test are equal to N-a where N is the total number of subjects in all groups and "a" is the number of groups (including the control).

Dunnett's test

```
Group <- factor(c("A","A","B","B","B","C","C","C","D","D","D","E","E","F","F","F"))
Value <- c(5,5.09,4.63,4.58,4.72,5,5.08,4.24,5.09,5.19,4.58,6.16,6.85,7.68,7.07,6.48)
data <- data.frame(Group, Value)
aov <- aov(Value ~ Group, data)
library(multcomp)
summary(glht(aov, linfct=mcp(Group="Dunnett")))
```

Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts

Fit: aov(formula = Value ~ Group, data = data)

Linear Hypotheses:

	Estimate	Std. Error	t value	Pr(>|t|)
B - A == 0	-0.40167	0.36699	-1.094	0.6880
C - A == 0	-0.27167	0.36699	-0.740	0.8939
D - A == 0	-0.09167	0.36699	-0.250	0.9988
E - A == 0	1.46000	0.40202	3.632	0.0168 *
F - A == 0	2.03167	0.36699	5.536	<0.001 ***

Three way ANOVA

Species	Temp	Sex	Rate
1	M	F	2.6
1	H	M	2.9
1	H	M	2.8
1	H	M	3.4
.	.	.	.
1	H	F	3
1	H	F	2.7
2	L	M	2.1
2	L	M	2.2
2	L	F	2.3
2	L	F	2

species temp sex rate
1 L M 1.9
1 L M 1.8
1 L M 1.6
1 L M 1.4
1 L F 1.8
1 L F 1.7
1 L F 1.4
1 L F 1.5
1 M M 2.3
1 M M 2.1
1 M M 2.0
1 M M 2.6
1 M F 2.4
1 M F 2.7
1 M F 2.4
1 M F 2.6
1 H M 2.9
1 H M 2.8
1 H M 3.4
1 H M 3.2
1 H F 3.0
1 H F 3.1
1 H F 3.0
1 H F 2.7
2 L M 2.1
2 L M 2.0
2 L M 1.8
2 L M 2.2
2 L F 2.3
2 L F 2.0
2 L F 1.9
2 L F 1.7
2 M M 2.4
2 M M 2.6
2 M M 2.7
2 M M 2.3
2 M F 2.0
2 M F 2.3
2 M F 2.1
2 M F 2.4
2 H M 3.6
2 H M 3.1
2 H M 3.4
2 H M 3.2
2 H F 3.1
2 H F 3.0
2 H F 2.8
2 H F 3.2
3 L M 1.1
3 L M 1.2
3 L M 1.0
3 L M 1.4
3 L F 1.4
3 L F 1.0
3 L F 1.3
3 L F 1.2
3 M M 2.0
3 M M 2.1
3 M M 1.9
3 M M 2.2
3 M F 2.4
3 M F 2.6
3 M F 2.3
3 M F 2.2
3 H M 2.9
3 H M 2.8
3 H M 3.0
3 H M 3.1
3 H F 3.2
3 H F 2.9
3 H F 2.8
3 H F 2.9

Model

$$\text{Rate} = \beta_0 + \beta_{species} \times \text{species} + \beta_{temp} \times \text{temp} + \beta_{sex} \times \text{sex} + \beta_{temp \times species} \times \text{temp} \times \text{species} + \beta_{sex \times species} \times \text{sex} \times \text{species} + \beta_{temp \times sex} \times \text{temp} \times \text{sex} + \beta_{temp \times species \times sex} \times \text{temp} \times \text{species} \times \text{sex} + \text{error}$$

R script

```
# Three way ANOVA
Dat = read.table('d:/ioz/statistics/2015/3way.ANOVA.txt',
        sep=' ', header=T)
Dat$species <- as.factor(Dat$species)
model <- aov(rate ~ species * temp* sex, data=Dat)

summary(model)
summary.lm(model)
```

Species	Temp	Sex	Rate
1	M	F	2.6
1	H	M	2.9
1	H	M	2.8
1	H	M	3.4
.	.	.	.
1	H	F	3
1	H	F	2.7
2	L	M	2.1
2	L	M	2.2
2	L	F	2.3
2	L	F	2

species temp sex rate
1 L M 1.9
1 L M 1.8
1 L M 1.6
1 L M 1.4
1 L F 1.8
1 L F 1.7
1 L F 1.4
1 L F 1.5
1 M M 2.3
1 M M 2.1
1 M M 2.0
1 M M 2.6
1 M F 2.4
1 M F 2.7
1 M F 2.4
1 M F 2.6
1 H M 2.9
1 H M 2.8
1 H M 3.4
1 H M 3.2
1 H F 3.0
1 H F 3.1
1 H F 3.0
1 H F 2.7
2 L M 2.1
2 L M 2.0
2 L M 1.8
2 L M 2.2
2 L F 2.3
2 L F 2.0
2 L F 1.9
2 L F 1.7
2 M M 2.4
2 M M 2.6
2 M M 2.7
2 M M 2.3
2 M F 2.0
2 M F 2.3
2 M F 2.1
2 M F 2.4
2 H M 3.6
2 H M 3.1
2 H M 3.4
2 H M 3.2
2 H F 3.1
2 H F 3.0
2 H F 2.8
2 H F 3.2
3 L M 1.1
3 L M 1.2
3 L M 1.0
3 L M 1.4
3 L F 1.4
3 L F 1.0
3 L F 1.3
3 L F 1.2
3 M M 2.0
3 M M 2.1
3 M M 1.9
3 M M 2.2
3 M F 2.4
3 M F 2.6
3 M F 2.3
3 M F 2.2
3 H M 2.9
3 H M 2.8
3 H M 3.0
3 H M 3.1
3 H F 3.2
3 H F 2.9
3 H F 2.8
3 H F 2.9

Results

summary(model)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
species	2	1.8175	0.9088	24.4751	2.72E-08***
temp	2	24.6558	12.3279	332.0237	2.20E-16***
sex	1	0.0089	0.0089	0.2394	0.6266
species:temp	4	1.1017	0.2754	7.4177	7.75E-05***
species:sex	2	0.3703	0.1851	4.9863	0.0103*
temp:sex	2	0.1753	0.0876	2.3603	0.1041
species:temp:sex	4	0.2206	0.0551	1.485	0.2196
Residuals	54	2.005	0.0371		

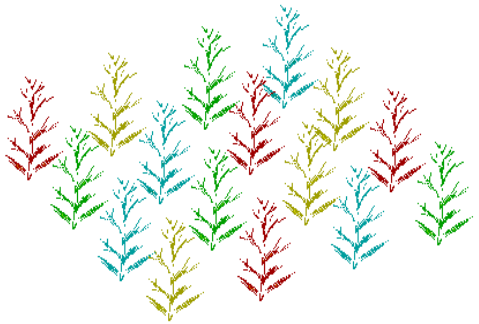
```
model <- aov(rate ~ species * temp* sex
                    - species : temp: sex,
                    data=Dat)
```

Results

summary.lm(model)

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.95E+00	9.64E-02	30.619	2.00E-16	***
species2	7.50E-02	1.36E-01	0.55	0.58428	
species3	-6.99E-17	1.36E-01	0	1	
tempL	-1.35E+00	1.36E-01	-9.908	9.47E-14	***
tempM	-4.25E-01	1.36E-01	-3.119	0.00291	**
sexM	1.25E-01	1.36E-01	0.917	0.36301	
species2:tempL	3.00E-01	1.93E-01	1.557	0.12534	
species3:tempL	-3.75E-01	1.93E-01	-1.946	0.05685	.
species2:tempM	-4.00E-01	1.93E-01	-2.076	0.04268	*
species3:tempM	-1.50E-01	1.93E-01	-0.778	0.4397	
species2:sexM	1.75E-01	1.93E-01	0.908	0.36781	
species3:sexM	-1.25E-01	1.93E-01	-0.649	0.51928	
tempL:sexM	-5.00E-02	1.93E-01	-0.259	0.79625	
tempM:sexM	-4.00E-01	1.93E-01	-2.076	0.04268	*
species2:tempL:sexM	-2.00E-01	2.73E-01	-0.734	0.46617	
species3:tempL:sexM	-7.07E-17	2.73E-01	0	1	
species2:tempM:sexM	4.00E-01	2.73E-01	1.468	0.14794	
species3:tempM:sexM	7.50E-02	2.73E-01	0.275	0.78419	

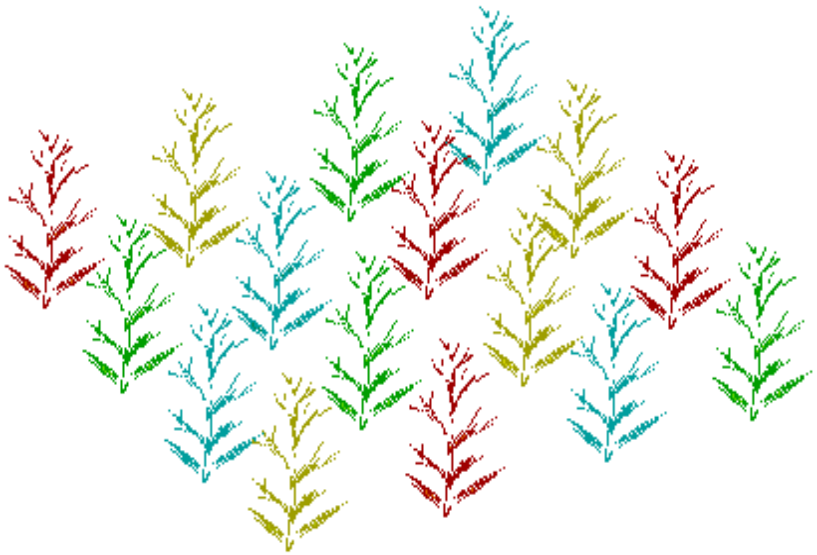
Latin square design



- Treatments are assigned within rows and columns, with each treatment once per row and once per column.
- There are equal numbers of rows, columns, and treatments (orthogonally designed).
- Useful where the experimenter desires to control variation in two different directions.

4×4 Latin square design

Different colors represent different treatments. There are 4 treatments (A-D) assigned to 4 rows (I-IV, e.g. independent days) and 4 columns (1-4, e.g. different species).



Latin square design solutions

(digits indicate treatments)

$$\begin{bmatrix} 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \\ 3 & 4 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 1 & 3 \\ 3 & 1 & 4 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 5 & 1 & 4 \\ 3 & 5 & 4 & 2 & 1 \\ 4 & 1 & 2 & 5 & 3 \\ 5 & 4 & 1 & 3 & 2 \end{bmatrix} \quad \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 4 & 1 & 5 & 3 \\ 3 & 5 & 4 & 2 & 1 \\ 4 & 1 & 5 & 3 & 2 \\ 5 & 3 & 2 & 1 & 4 \end{bmatrix}$$

Variance partation

Source of variation	Degrees of freedom[a]	Sums of squares (SS)	Mean square (MS)	F
Rows (*R*)	r-1	SS_R	$SS_R/(r-1)$	MS_R/MS_E
Columns (*C*)	r-1	SS_C	$SS_C/(r-1)$	MS_C/MS_E
Treatments (*Tr*)	r-1	SS_{Tr}	$SS_{Tr}/(r-1)$	MS_{Tr}/MS_E
Error (*E*)	(r-1)(r-2)	SS_E	$SS_E/((r-1)(r-2))$	
Total (*Tot*)	r^2-1	SS_{Tot}		

where r=number of treatments, rows, and columns.

Example

Change in Blood Sugar Levels in Mice:
Four individuals of mice, four days, and four treatments are arranged in a Latin square design. The response is the mean change in blood sugar for 4 animals. The treatments are levels of insulin coded as follows:

Level A for 150 micro units
B 300
C 600
D 1200

```
1 I   B -4.5
1 II  D 92.33
1 III C 59.83
1 IV  A -45.
2 I   C 91.83
2 II  A -48.33
2 III D 168.99
2 IV  B 89.
3 I   D 86.16
3 II  B -78.16
3 III A -24.17
3 IV  C 101.0
4 I   A -.17
4 II  C 68.83
4 III B 25.17
4 IV  D 177.17
```

Mouse	Day							
	1		2		3		4	
I	B	-4.5	C	91.83	D	86.16	A	-0.17
II	D	92.33	A	-48.33	B	-78.16	C	68.83
III	C	59.83	D	168.99	A	-24.17	B	25.17
IV	A	-45	B	89	C	101	D	177.17

Model

$$\text{Sugar} = \beta_0 + \beta_{\text{insulin}} \times \text{insulin} + \beta_{\text{day}} \times \text{day} + \beta_{\text{group}} \times \text{group} + \text{error}$$

df?

R script for another example

```
# Book "Linear Models with R" by Faraway
library(faraway)
data(abrasion)
lines <-
"id run position material wear
1 1 1 C 235
2 1 2 D 236
3 1 3 B 218
4 1 4 A 268
5 2 1 A 251
6 2 2 B 241
7 2 3 D 227
8 2 4 C 229
9 3 1 D 234
10 3 2 C 273
11 3 3 A 274
12 3 4 B 226
13 4 1 B 195
14 4 2 A 270
15 4 3 C 230
16 4 4 D 225"
abrasion.data <- read.table(con <-
   textConnection(lines), header=TRUE)
close(con)

matrix(abrasion.data$material, 4, 4)
abrasion.data$run = as.factor(abrasion.data$run)
abrasion.data$position = as.factor(abrasion.data$position)
fit1 = aov(wear ~ run + position + material, abrasion.data)
fit2 = lm  (wear ~ run + position + material, abrasion.data)
summary(fit1)
summary(fit2)
```

R results

```
matrix(abrasion.data$material, 4, 4)
     [,1]  [,2]  [,3]  [,4]
[1,] "C"   "A"   "D"   "B"
[2,] "D"   "B"   "C"   "A"
[3,] "B"   "D"   "A"   "C"
[4,] "A"   "C"   "B"   "D"
```

id	run	position	material	wear
1	1	1	C	235
2	1	2	D	236
3	1	3	B	218
4	1	4	A	268
5	2	1	A	251
6	2	2	B	241
7	2	3	D	227
8	2	4	C	229
9	3	1	D	234
10	3	2	C	273
11	3	3	A	274
12	3	4	B	226
13	4	1	B	195
14	4	2	A	270
15	4	3	C	230
16	4	4	D	225

```
summary(fit2)
Coefficients:
             Estimate  Std.Error t value Pr(>|t|)
(Intercept)  254.750    6.187    41.174   1.37e-08 ***
run2          -2.250    5.534    -0.407   0.698423
run3          12.500    5.534     2.259   0.064657 .
run4          -9.250    5.534    -1.671    0.145658
position2     26.250    5.534     4.743   0.003180 **
position3      8.500    5.534     1.536   0.175454
position4      8.250    5.534     1.491   0.186608
materialB    -45.750    5.534    -8.267   0.000169 ***
materialC    -24.000    5.534    -4.337   0.004892 **
materialD    -35.250    5.534    -6.370   0.000703 ***
```

Hierarchical (nested) ANOVA

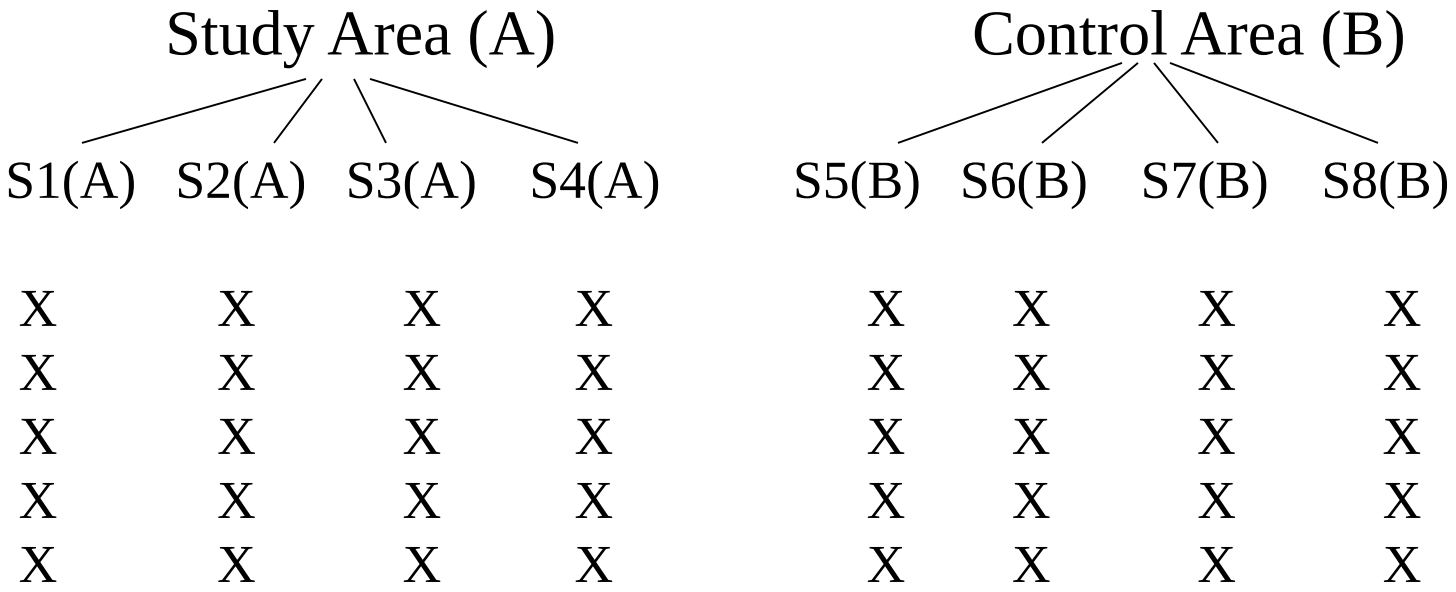
Hierarchical (nested) ANOVA

- In some two-factor experiments the level of one factor , say B, is not “cross” or “cross classified” with the other factor, say A, but is “NESTED” with it.

- The levels of B are different for different levels of A.
  - For example: 2 Areas (Study vs Control)
    - 4 sites per area, each with 5 replicates.
    - There is no link from any sites on one area to any sites on another area.

Example

- There are 8 sites, not 2, not 4 either.



X = replications

Number of sites (S)/replications need not be equal with each sites.

Analysis is carried out using a nested ANOVA not a two-way ANOVA.

Nested ANOVA vs. two-way ANOVA

- A Nested design is not the same as a two-way ANOVA which is represented by:

	A1	A2	A3
B1	X X X X X	X X X X X	X X X X X
B2	X X X X X	X X X X X	X X X X X
B3	X X X X X	X X X X X	X X X X X

Nested, or hierarchical designs are very common in environmental effects monitoring studies. There are several "Study" and several "Control" Areas.

Objectives

- The nested design tests two things: (1) difference between "Study" and "Control" areas, and (2) the variability of the sites within areas.
- If we fail to find a significant variability among the sites within areas, then a significant difference between areas would suggest that there is an treatment effect.
- In other words, the variability is due to differences between areas (treatment) and not to variability among the sites.

Treatment

Area 1	Area 2
Site1	Site3
Site2	Site4

Notes

- In this kind of situation, however, it is highly likely that we will find variability among the sites.

- Even if it should be significant, however, we can still test to see whether the difference between the areas is significantly larger than the variability among the sites with areas.

Statistical Model

$$Y_{ijk} = \mu + A_i + B_{(i)j} + \varepsilon_{(ij)k}$$

i indexes "A" (often called the "major factor")

(i)j indexes "B" within "A" (B is often called the "minor factor")

(ij)k indexes replication

$i = 1, 2, \ldots, M$

$j = 1, 2, \ldots, m$

$k = 1, 2, \ldots, n$

Model (continued)

$$Y_{ijk} = \bar{Y}_{\cdots} + \left(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdots}\right) + \left(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot}\right) + \left(Y_{ijk} - \bar{Y}_{ij\cdot}\right)$$

and

$$\sum_i \sum_j \sum_k \left(Y_{ijk} - \bar{Y}_{\cdots}\right)^2 = \sum_i \sum_j \sum_k \left(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdots}\right)^2 + \sum_i \sum_j \sum_k \left(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot}\right)^2 + \sum_i \sum_j \sum_k \left(Y_{ijk} - \bar{Y}_{ij\cdot}\right)^2$$

Model (continued)

M_1				M_2				M_3				Areas	i
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	Sites	j
10	12	8	13	11	13	9	10	13	14	7	10		
14	8	10	12	14	11	10	9	10	13	9	7	Repl.	k
9	10	12	11	8	9	8	8	16	12	5	4		
11	10	10	12	11	11	9	9	13	13	7	7		
10.75				10.0				10.0					
10.25													

Or,

$$TSS = SS_A + SS_{(A)B} + SSW_{error}$$

$$= m.n\sum_{i=1}^{M}\left(\bar{Y}_{i..} - \bar{Y}_{...}\right)^2 + n\sum_{i=1}^{M}\sum_{j=1}^{m}\left(\bar{Y}_{ij.} - \bar{Y}_{i..}\right)^2 + \sum_{i=1}^{3}\sum_{j=1}^{4}\sum_{k=1}^{3}\left(Y_{ijk} - \bar{Y}_{ij.}\right)^2$$

Degrees of freedom:

M.m.n -1 = (M-1) + M(m-1) + Mm(n-1)

In lecture 5, Two-way ANOVA:

$$SS_A = \sum n_B n_{AB}(\bar{X}_A - \bar{\bar{X}})^2$$

$$SS_B = \sum n_A n_{AB}(\bar{X}_B - \bar{\bar{X}})^2$$

Example

M=3, m=4, n=3; 3 Areas, 4 sites within each area, 3 replications per site, total of (M.m.n = 36) data points

M_1				M_2				M_3				Areas
1	2	3	4	5	6	7	8	9	10	11	12	Sites
10	12	8	13	11	13	9	10	13	14	7	10	
14	8	10	12	14	11	10	9	10	13	9	7	Repl.
9	10	12	11	8	9	8	8	16	12	5	4	
11	10	10	12	11	11	9	9	13	13	7	7	— $\bar{Y}_{ij.}$
	10.75				10.0				10.0			$\bar{Y}_{i..}$
						10.25						$\bar{Y}_{...}$

Example (continue)

$SS_A = 4 \times 3 [(10.75-10.25)^2 + (10.0-10.25)^2 + (10.0-10.25)^2]$

$= 12 (0.25 + 0.0625 + 0.625) = 4.5$

$SS_{(A)B} = 3 [(11-10.75)^2 + (10-10.75)^2 + (10-10.75)^2 + (12-10.75)^2 +$

$(11-10)^2 + (11-10)^2 + (9-10)^2 + (9-10)^2 +$

$(13-10)^2 + (13-10)^2 + (7-10)^2 + (7-10)^2]$

$= 3 (42.75) = 128.25$

$TSS = 240.75$

$SSW_{error} = 108.0$

	M_1				M_2				M_3				**Areas**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**Sites**
	10	12	8	13	11	13	9	10	13	14	7	10	
	14	8	10	12	14	11	10	9	10	13	9	7	**Repl.**
	9	10	12	11	8	9	8	8	16	12	5	4	
$\bar{Y}_{ij.}$	11	10	10	12	11	11	9	9	13	13	7	7	
$\bar{Y}_{i..}$		10.75				10.0				10.0			
$\bar{Y}_{...}$						10.25							

ANOVA Table for Example

Nested ANOVA: Observations versus Area, Sites

Source	DF	SS	MS	F	P
Area	2	4.50	2.25	0.158	0.856
Sites (A)B	9	128.25	14.25	3.167	0.012**
Error	24	108.00	4.50		
Total	35	240.75			

$= MS_A/MS_{(A)B}$

$= MS_{(A)B}/MSW_{error}$

Summary

- Nested designs are very common in environmental monitoring
- It is a refinement of the one-way ANOVA
- All assumptions of ANOVA hold: normality of residuals, constant variance, etc.
- Can be easily computed using R, SAS, MINITAB, etc.
- Need to be careful about the proper ratio of the Mean squares.
- Always use graphical methods e.g. boxplots and normal plots as visual aids to aid analysis.

Example: Hierarchical (nested) ANOVA

Length	Mosquito	Cage
58.5	1	1
59.5	1	1
77.8	2	1
80.9	2	1
84.0	3	1
83.6	3	1
70.1	4	1
68.3	4	1
69.8	1	2
69.8	1	2
56.0	2	2
54.5	2	2
50.7	3	2
49.3	3	2
63.8	4	2
65.8	4	2
56.6	1	3
57.5	1	3
77.8	2	3
79.2	2	3
69.9	3	3
69.2	3	3
62.1	4	3
64.5	4	3

Model

$$\text{Length} = \beta_0 + \beta_{\text{cage}} \times \text{cage} + \beta_{\text{mosquito(cage)}} \times \text{mosquito (cage)} + \text{error}$$

df?

```
lines <-
"Length   Mosquito   Cage
58.5        1          1
59.5        1          1
77.8        2          1
80.9        2          1
84.0        3          1
83.6        3          1
70.1        4          1
68.3        4          1
69.8        1          2
69.8        1          2
56.0        2          2
54.5        2          2
50.7        3          2
49.3        3          2
63.8        4          2
65.8        4          2
56.6        1          3
57.5        1          3
77.8        2          3
79.2        2          3
69.9        3          3
69.2        3          3
62.1        4          3
64.5        4          3 "
dat <- read.table(con <-
    textConnection(lines),
    header=TRUE); close(con)
```

R code

```
dat$Mosquito = as.factor(dat$Mosquito)
dat$Cage = as.factor(dat$Cage)

# two way ANOVA
summary(aov(Length ~ Cage * Mosquito, dat))
# nested ANOVA
summary(aov(Length ~ Cage / Mosquito, dat))
```

```
> summary(aov(Length ~ Cage * Mosquito, dat))
              Df  Sum Sq  Mean Sq  F value   Pr(>F)
Cage           2  665.7   332.8    255.70  1.45e-10 ***
Mosquito       3  260.2    86.7     66.63  9.44e-08 ***
Cage:Mosquito  6 1460.5   243.4    187.00  3.86e-11 ***
Residuals     12   15.6     1.3
> summary(aov(Length ~ Cage / Mosquito, dat))
              Df  Sum Sq  Mean Sq  F value   Pr(>F)
Cage           2  665.7   332.8    255.7   1.45e-10 ***
Cage:Mosquito  9 1720.7   191.2    146.9   6.98e-11 ***
Residuals     12   15.6     1.3
```

R results

R does not compute the correct F-statistics, because it uses the residual MS for the denominator in all calculations, which is not applicable for nested ANOVA.

The interaction sum of squares in the nested case is the sum of the subject (nested in) effect and interaction in the crossed model case, as are the degrees of freedom.

```
# Manually computation of the F-statistic and p value is needed.
fit = summary(aov(Length ~ Cage / Mosquito, dat))
F.value = fit[[1]] [1, 3] / fit[[1]] [2, 3]
p = pf(F.value, fit[[1]] [1, 1], fit[[1]] [2, 1], lower=FALSE); p # 0.23
```

Or, lme() or lmer() are preferred.

```
# Another version: nested ANOVA
summary(aov(Length ~ Cage +
            Error(Cage / Mosquito),
            dat))
```

Error: Cage

	Df	Sum Sq	Mean Sq
Cage	2	665.7	332.8

Error: Cage:Mosquito

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Residuals	9	1721	191.2		

Error: Within

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Residuals	12	15.62	1.302		

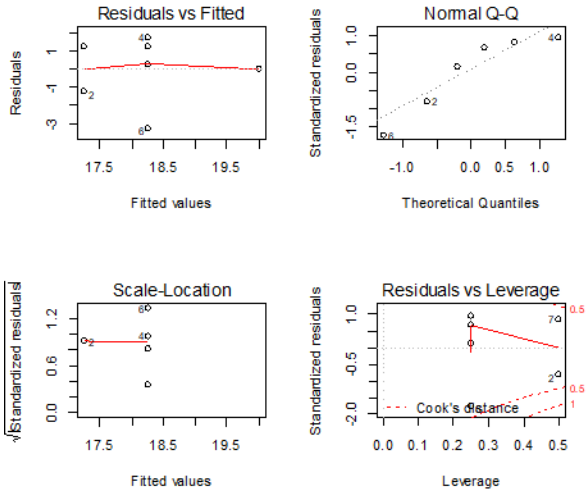
Assignment scores so far

```
score = c(19.5, 16, 20, 20, 18.5, 15, 18.5)
dept  = c('UCAS_Bio', 'UCAS_Math', 'HUST', 'UCAS_Bio', 'UCAS_Bio',
          'UCAS_Bio', 'UCAS_Math')
```

tapply(score, dept, mean); tapply(score, dept, length)

```
HUST    UCAS_Bio UCAS_Math
20.00   18.25    17.25
```

```
HUST    UCAS_Bio UCAS_Math
1       4        2
```




```
fit = aov(score ~ dept)
summary(fit)
```

```
            Df Sum Sq  Mean Sq F value Pr(>F)
dept        2  5.054   2.527   0.55    0.615
Residuals 4  18.375  4.594
```

```
TukeyHSD(fit)
plot(fit)
```

Assignment

General objectives: learn two-way ANOVA.

- Generate your own data (In order to better understand an interaction in a 2-way ANOVA, you create data that has an interaction with a P <.05. Please describe the experiment and data for interpretation purposes.)
- Provide a brief introduction to the data set,
- Formally state the hypotheses that you are going to test (Ho's and Ha's),
- Satisfy assumptions of normality of residuals, homogeneity of variances, and independency of residuals, homogeneous of residuals
- Provide a print out's of the data set, programs and their output.
- Indicate in your results and discussion section what you found, i.e. did you reject your null, and the conclusions that you have drawn from the analysis. Since there will be an interaction (P<0.05) then you must break apart your factors and do 1-way ANOVA.

R script

```
# Bartlett Test of Homogeneity of Variances (parametric)
bartlett.test(split(Dat$y, list(Dat$x1, Dat$x2)))
bartlett.test(Dat$y ~ Dat$x1 * Dat$x2)


# Two Way ANOVA
weight.gain=data.frame(ID=1:60, amount=NA, food=NA, gain=NA)
n=0
for(i in c('high','low')){
                for(j in c('beef','cereal','port')){
                                for(k in 1:10){
                                n=n+1
                                weight.gain$amount[n]=i
                                weight.gain$food[n]=j
}               }               }
weight.gain$gain=c(73,102,118,104,81,107,100,87,117,111,
98,74,56,111,95,88,82,77,86,92, 94,79,96,98,102,102,108,91,120,105,
90,76,90,64,86,51,72,90,95,78, 107,95,97,80,98,74,74,67,89,58,
49,82,73,86,81,97,106,70,61,82)

fit =  aov(gain ~ amount + food + amount:food, data=weight.gain)
fit <- aov(gain ~ amount * food, data=weight.gain) # same thing

summary(fit)

par(mfrow=c(2,2)); plot(fit)
```