ANOVA

Analysis **o**f **Va**riance

Origin of ANOVA

- Fisher, R. A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. Transactions of the Royal Society of Edinburgh 52: 399-433.
- Fisher, R. A. 1925. **Statistical Methods for Research Workers**. Oliver and Boyd, Edinburgh.
- Fisher, R. A. 1935. **The Design of Experiments**. Oliver and Boyd, Edinburgh.

Sir Ronald Aylmer Fisher

17 February 1890 – 29 July 1962

One way ANOVA

One way ANOVA

Rodent weight at different sites

Site1	Site2	Site3	Site4	Site5	Site6	Site7
9.4	16.8	27	21	24.3	17.7	16.5
8.7	30.8	28.9	23.4	29.7	19.7	20.7
13.3	33.6	32	27.5	19.9	21.5	23.5
13.6	40.5	32.7	27.5		27.9	26.4
15	48.9	35.5	30.5		34.8	26.7
15.2		45.6	31.9		40.2	29.5
17.7			32.5			29.8
18.6			33.8			31.9
22.2			33.8			35.5

Why not using t test?

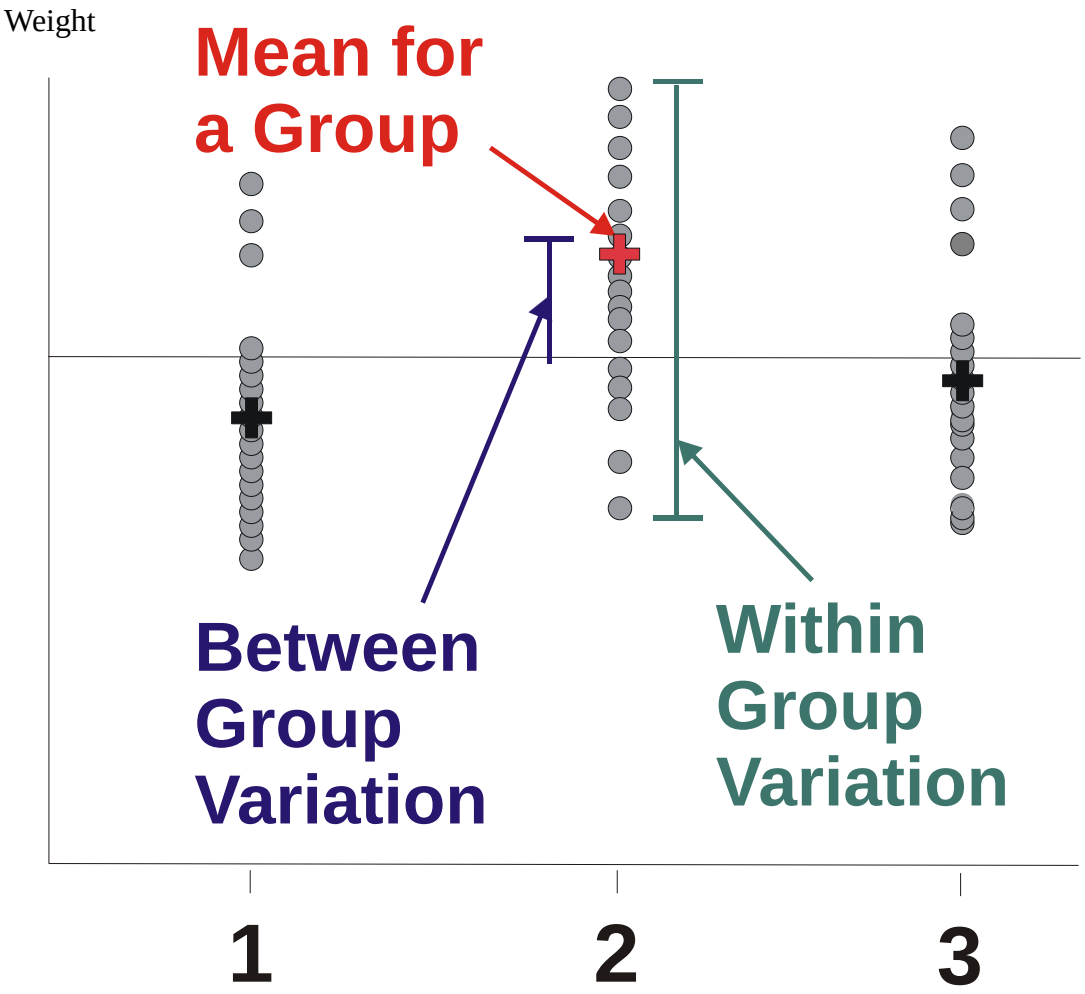
For a rodent species at 7 sites, we compare their mean weight at each site.

Number of T test needed: C(7, 2) = 21

Probability of no difference for all pairs at 95% confidence interval:
$0.95^{21} = 0.34$

Type I error is 1 - 0.34 = 0.66

Sources of variance



$\overline{X}$

$$\text{total SS} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$$

$$\text{among-groups SS} = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2$$

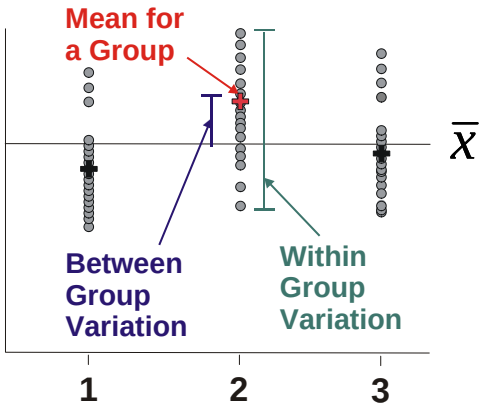
$$\text{winthin-groups SS} = \sum_{i=1}^{k}\left[\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2\right]$$

Sums of squares (SS)

$$\text{total SS} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 \qquad \text{total DF} = N - 1$$

$$\text{among-groups SS} = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2 \qquad \text{among-groups DF} = k - 1$$

$$\text{winthin-groups SS} = \sum_{i=1}^{k}\left[\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2\right] \qquad \text{winthin-groups DF} = \sum_{i=1}^{k}(n_i - 1) = N - k$$



Sums of squares (SS) and degrees of freedom (DF) are additive

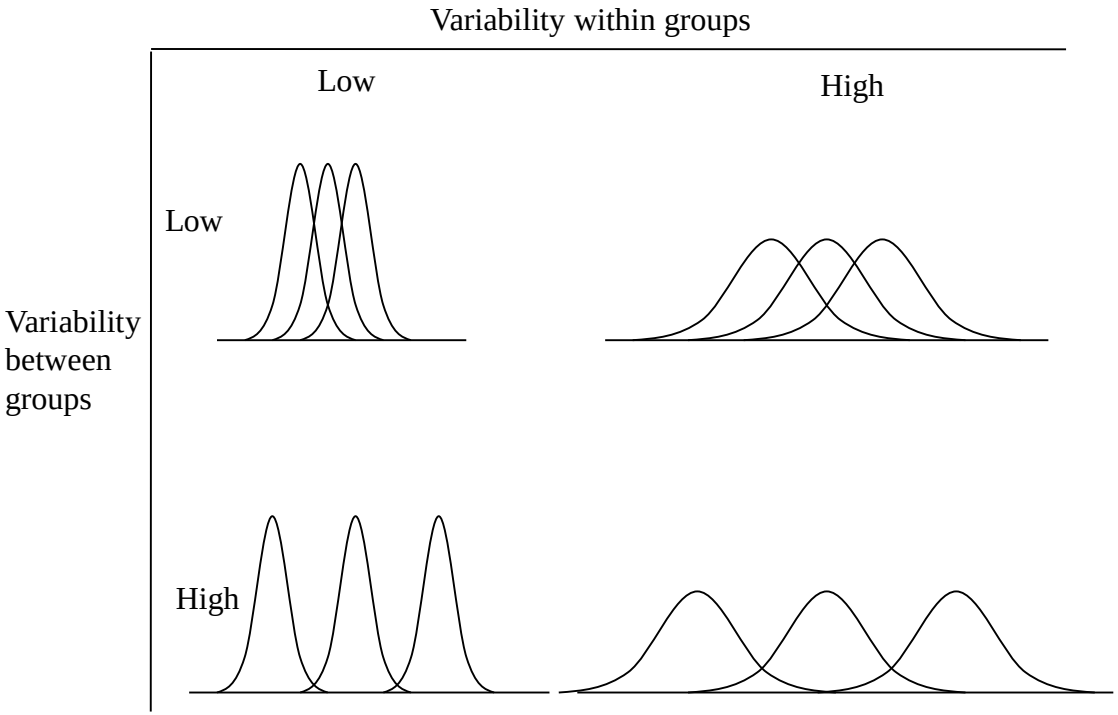
- Total SS = among-group SS + error SS
- Total DF = among-group DF + error DF
- MS = SS/DF

Partitioning the variance

$$SS_{\text{Total}} = SS_{\text{Within}} + SS_{\text{Between}}$$

SS_{Between}: Variability due to treatments

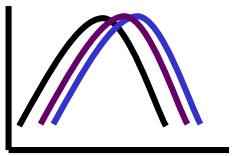
SS_{Within} : Variability due to other factors plus error variability



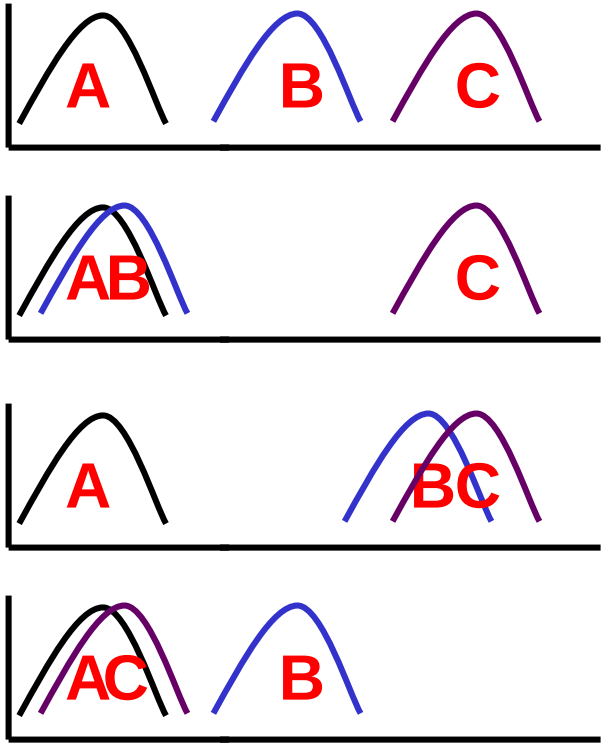
Null hypothesis

$H_0: \quad \mu_1 = \mu_2 = \mu_3$

$H_A: \quad$ not all the μ_j are equal



Ho



Ha

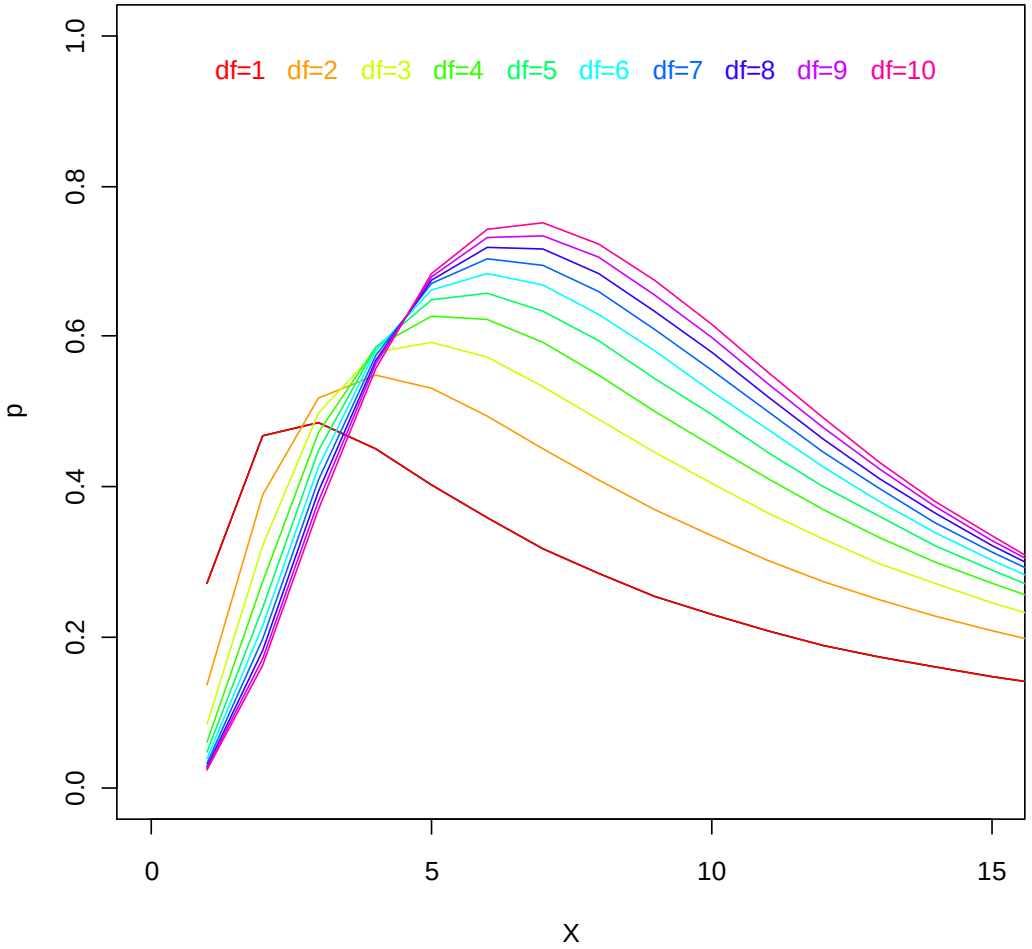
Testing the null hypothesis using *F* distribution

$$F = \frac{\text{among - group MS}}{\text{error MS}}$$

Mean square (MS) = Sum of square (SS) / DF

F-distribution

- F-test is always a one-tailed test.




```
# F distribution
df.1 <- 10 # fixed in the following plot
df.2 <- 1
f.d <- df (seq (0.1, 15, by=0.1), df.1, df.2)
plot (f.d, xlab='X', ylab='p', type='l',
      xlim=c(0,15), ylim=c(0,1))
for (df.2 in 1:10){
   f.d <- df (seq (0.1, 15, by=0.1), 10, df.2)
   lines(f.d, type='l', col=rainbow(10)[df.2])
   legend(df.2*1.3-1, 1, paste('df=', df.2, sep=''),
   text.col = rainbow(10)[df.2], box.lty=0, cex=1)
}
```

Distributions: *F*, *z*, *t*, and chi square

$$Z = f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$z^2 = \chi^2_{(1)}$$

$$t^2_{[v]} = F_{[1,v]}$$

$$(X^2_{[n]}/n)/(X^2_{[m]}/m) = F_{[n,m]}$$

$$X^2_{[v]}/v = F_{[v,\infty]}$$

mean(rnorm(10000)^2) ~1

Table of content – ANOVA

- One-way ANOVA
- Random Blocked Design
- Paired Comparison
- Two-way ANOVA
- Repeated measures ANOVA
- Hierarchical ANOVA
- Three way ANOVA
- Latin Square Design
- Split Plot Design
- Mixed-effects models

One way ANOVA

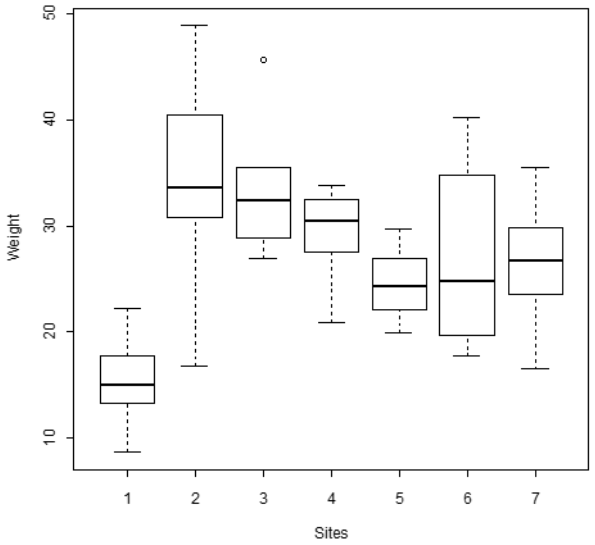
Rodent weight at different sites

Site1	Site2	Site3	Site4	Site5	Site6	Site7
9.4	16.8	27	21	24.3	17.7	16.5
8.7	30.8	28.9	23.4	29.7	19.7	20.7
13.3	33.6	32	27.5	19.9	21.5	23.5
13.6	40.5	32.7	27.5		27.9	26.4
15	48.9	35.5	30.5		34.8	26.7
15.2		45.6	31.9		40.2	29.5
17.7			32.5			29.8
18.6			33.8			31.9
22.2			33.8			35.5

Input data and check data

```
# One way ANOVA
# Rodent weight at 7 sites

site1 <- c(9.4, 8.7, 13.3, 13.6, 15, 15.2, 17.7, 18.6, 22.2)
site2 <- c(16.8, 30.8, 33.6, 40.5, 48.9)
site3 <- c(27.0, 28.9, 32, 32.7, 35.5, 45.6)
site4 <- c(21.0, 23.4, 27.5, 27.5, 30.5, 31.9, 32.5, 33.8, 33.8)
site5 <- c(24.3, 29.7, 19.9)
site6 <- c(17.7, 19.7, 21.5, 27.9, 34.8, 40.2)
site7 <- c(16.5, 20.7, 23.5, 26.4, 26.7, 29.5, 29.8, 31.9, 35.5)
```




```
rodent.survey <- data.frame(weight=c(site1,site2,site3,site4,site5,site6,site7),
      site=factor(c(rep("1",9),rep("2",5),rep("3",6),rep("4",9),rep("5",3),
      rep("6",6),rep("7",9))))

options(digits=3) # default value = 7
tapply(rodent.survey$weight, rodent.survey$site, mean)
tapply(rodent.survey$weight, rodent.survey$site, var)
tapply(rodent.survey$weight, rodent.survey$site, sum)

boxplot(weight~site, data=rodent.survey, xlab='Sites', ylab='Weight')
```

Input data and check data

rodent.survey

	weight	site
1	9.4	1
2	8.7	1
3	13.3	1
4	13.6	1
5	15.0	1
6	15.2	1
7	17.7	1
8	18.6	1
9	22.2	1
10	16.8	2
11	30.8	2
12	33.6	2
13	40.5	2

tapply(rodent.survey$weight, rodent.survey$site, mean)

1	2	3	4	5	6	7
14.9	34.1	33.6	29.1	24.6	27.0	26.7

tapply(rodent.survey$weight, rodent.survey$site, var)

1	2	3	4	5	6	7
18.5	142.6	43.3	21.1	24.1	81.2	34.0

tapply(rodent.survey$weight, rodent.survey$site, sum)

1	2	3	4	5	6	7
133.7	170.6	201.7	261.9	73.9	161.8	240.5

Model and results

```
# One Way Anova (Completely Randomized Design)
fit <- aov (weight~site, data=rodent.survey)
summary(fit)
# fit=lm(weight~site, data=rodent.survey); anova(fit)
```

Analysis of Variance Table

Response: weight

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
site	6	1888	314.6	6.88	4.6e-05 ***
Residuals	40	1830	45.8		

```
# report the means and the number of subjects/cell
print(model.tables(fit,"means"), digits=3)
```

site

	1	2	3	4	5	6	7
	14.9	34.1	33.6	29.1	24.6	27.0	26.7
rep	9.0	5.0	6.0	9.0	3.0	6.0	9.0

Predicted values and residuals

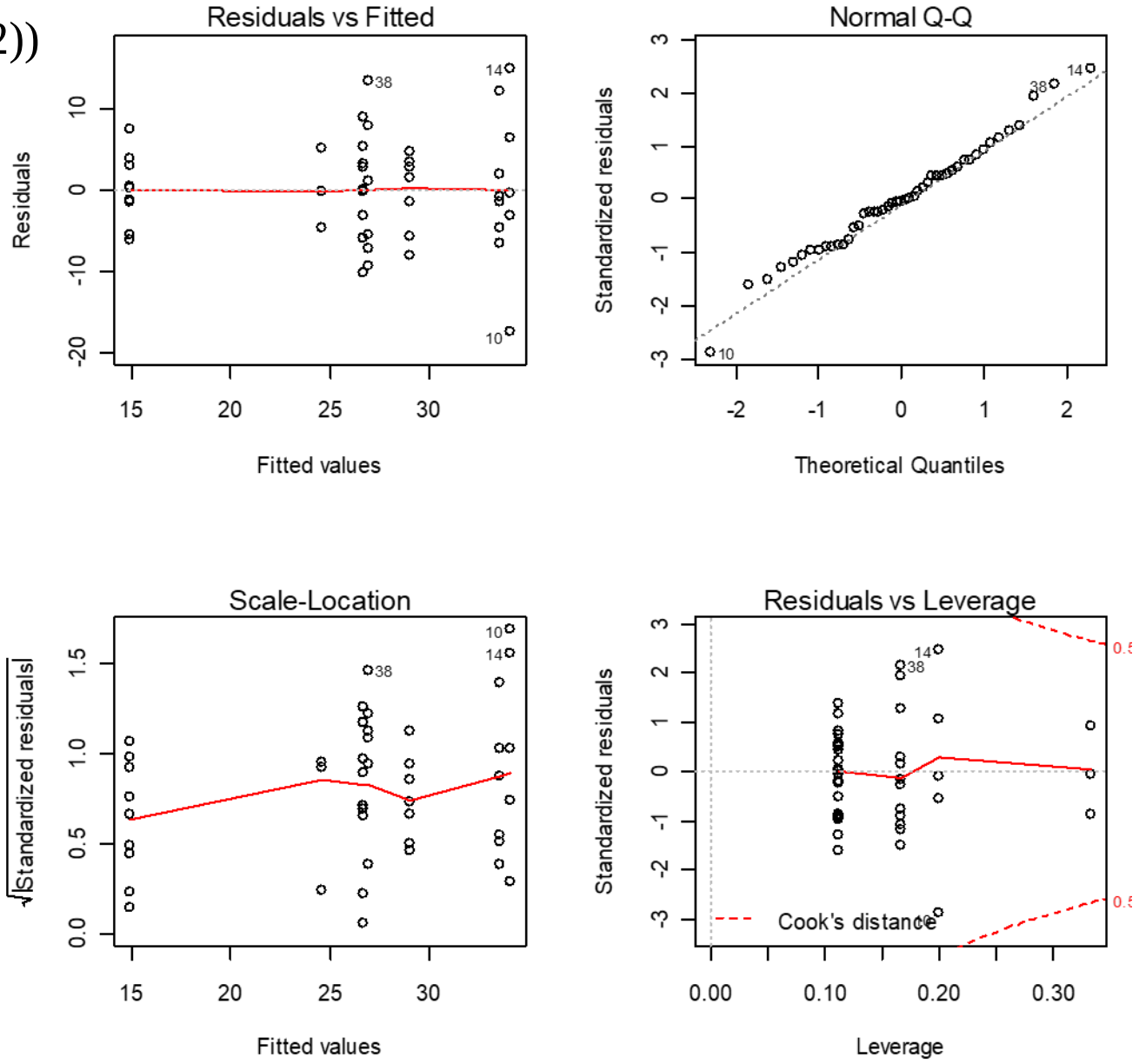
	weight	site	predicted	residual
1	9.4	1	14.86	−5.46
2	8.7	1	14.86	−6.16
3	13.3	1	14.86	−1.56
4	13.6	1	14.86	−1.26
5	15	1	14.86	0.14
6	15.2	1	14.86	0.34
7	17.7	1	14.86	2.84
8	18.6	1	14.86	3.74
9	22.2	1	14.86	7.34
10	16.8	2	34.12	−17.32
11	30.8	2	34.12	−3.32
12	33.6	2	34.12	−0.52
13	40.5	2	34.12	6.38
14	48.9	2	34.12	14.78
15	27	3	33.62	−6.62
16	28.9	3	33.62	−4.72
17	32	3	33.62	−1.62
18	32.7	3	33.62	−0.92
19	35.5	3	33.62	1.88
20	45.6	3	33.62	11.98

	weight	site	predicted	residual
21	21	4	29.1	−8.10
22	23.4	4	29.1	−5.70
23	27.5	4	29.1	−1.60
24	27.5	4	29.1	−1.60
25	30.5	4	29.1	1.40
26	31.9	4	29.1	2.80
27	32.5	4	29.1	3.40
28	33.8	4	29.1	4.70
29	33.8	4	29.1	4.70
30	24.3	5	24.63	−0.33
31	29.7	5	24.63	5.07
32	19.9	5	24.63	−4.73
33	17.7	6	26.97	−9.27
34	19.7	6	26.97	−7.27
35	21.5	6	26.97	−5.47
36	27.9	6	26.97	0.93
37	34.8	6	26.97	7.83
38	40.2	6	26.97	13.23
39	16.5	7	26.72	−10.22
40	20.7	7	26.72	−6.02
41	23.5	7	26.72	−3.22
42	26.4	7	26.72	−0.32
43	26.7	7	26.72	−0.02
44	29.5	7	26.72	2.78
45	29.8	7	26.72	3.08
46	31.9	7	26.72	5.18
47	35.5	7	26.72	8.78

```
# ANOVA fit
fit[[1]] # coefficients
fit[[2]] # residuals
fit[[5]] # predicted (fitted.values)
all <- cbind (rodent.survey,
      predicted=fit[[5]], residual=fit[[2]])
head(all)
```

Model performance

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



Leverage

Identifying those observations (y_i) that are far away from average predictor values.

In linear regression model, the leverage score (self-sensitivity or self-influence) for the data unit *i* is defined as:

$$h_{ii} = \frac{\partial \hat{y}_i}{\partial y_i}$$

Residuals normal?

shapiro.test (residuals) # *fit[[2]]*

```
 Shapiro-Wilk normality test

data:  rodent.survey$res
W = 0.9872, p-value = 0.88
```

Hypothesis of normality can not be rejected.

Normality satisfied

Assumptions of one-way ANOVA

- Observations are independent of each other
- Scores in groups are normally distributed
  - ANOVA is robust to violations of the normality assumption
- Variances in groups are homogeneous
  - ANOVA is robust if $n1 = n2 = \ldots = nJ$

Homogeneity of variance - Bartlett's test

The most common method employed to test for homogeneity of variances is *Bartlett's test* (Bartlett,1937a, 1937b; based on a principle of Neyman and Pearson. 1931). In this procedure, the test statistic is

$$B = (\ln s_p^2)(\sum_{i=1}^{k} \nu_i) - \sum_{i=1}^{k} \nu_i \ln s_i^2$$

s_i^2	9.39	8.56	7.66	8.39
ν_i	4	4	3	4

Where $\nu_1 = n_1 - 1$ and n_1 is the size of sample i . The pooled variance, s_p^2 , is calculated as before as $\sum_{i=1}^{k} SS_i / \sum_{i=1}^{k} \nu_i$. Many biologists prefer to operate with common logarithms (base 10), rather than with natural logarithms (base *e*); so Equation 10.44 may be written as

$$B = 2.30259\left[(\log s_p^2)\left(\sum_{i=1}^{k} \nu_i\right) - \sum_{i=1}^{k} \nu_i \log s_i^2\right]$$

The distribution of *B* is approximated by the chi-square distribution, with *k* -1 degrees of freedom, but a more accurate chi-square approximation is obtained by computing a correction factor.

$$C = 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k} \frac{1}{\nu_i} - \frac{1}{\sum_{i=1}^{k} \nu_i}\right) \qquad Bc = B/C$$

Bartlett's test - example

EXAMPL 10.13 Bartlett's test for homogeneity of variances (Zar 1999).

Nineteen pigs were divided into four groups, and each group was given a different feed. The data are weights in kilograms. And we wish to test whether the variance of weights is the same for pigs on all four feeds.

$H_0 : \sigma_1 = \sigma_2 = \sigma_3 = \sigma_4$ $\alpha=0.05$

H_A : The four population variances are heterogeneous (i.e. , are not all equal).

	Feed1	***Feed2***	***Feed3***	***Feed4***	
	60.8	68.7	102.6	87.9	
	57.0	67.7	102.1	84.2	
	65.0	74.0	100.2	83.1	
	58.6	66.3	96.5	85.7	
	61.7	69.8		90.3	
SS_i	37.57	34.26	22.97	33.55	$\sum SS_i = 128.35$
ν_i	4	4	3	4	$\sum \nu_i = 15$
s_i^2	9.39	8.56	7.66	8.39	
$\log s_i^2$	0.9727	0.9325	0.8842	0.9238	
$\nu_i \log s_i^2$	3.8908	3.7300	2.6526	3.6952	$\sum \nu_i \log s_i^2 = 13.9686$
$\frac{1}{\nu_i}$	0.250	0.250	0.333	0.250	$\sum \frac{1}{\nu_i} = 1.083$

Bartlett's test - example

$$s_p^2 = \frac{\sum SS_i}{\sum \nu_i} = \frac{128.35}{15} = 8.56$$

$$\log s_p^2 = 0.9325$$

$$B = 2.30259\left[\left(\log s_p^2\right)\left(\sum \nu_i\right) - \sum \nu_i \log s_i^2\right]$$
$$= 2.30259\left[(0.9325)(15) - 13.9686\right]$$
$$= 2.30259(0.0189)$$
$$= 0.0435$$

$$C = 1 + \frac{1}{3(k-1)}\left(\sum \frac{1}{\nu_i} - \frac{1}{\sum \nu_i}\right)$$
$$= 1 + \frac{1}{3(3)}\left(1.083 - \frac{1}{15}\right)$$
$$= 1.113$$

$$B_c = \frac{B}{C} = \frac{0.0435}{1.113} = 0.0391$$

$$\chi^2_{0.05,3} = 7.815$$

P=0.998

Do not reject H_0.

```
# Bartlett Test of Homogeneity of Variances (parametric)
bartlett.test (weight~site, data=rodent.survey)

# Figner-Killeen Test of Homogeneity of Variances (non-parametric)
fligner.test (weight~site, data=rodent.survey)
```

post-ANOVA comparisons

Tukey's multiple comparisons tests

The "Honestly Significantly Different" (HSD) test proposed by the statistician John Tukey is based on what is called the "studentized range distribution." To test all pairwise comparisons among means using the Tukey HSD, compute t for each pair of means using the formula:

$$t_s = \frac{M_i - M_j}{\sqrt{\frac{MSE}{n_h}}}$$

where M_i - M_j is the difference between the ith and jth means, MSE is the Mean Square Error, and n_h is the harmonic mean* of the sample sizes of groups i and j.

*The harmonic mean is the reciprocal of the arithmetic mean of the reciprocals.

Post-ANOVA comparisons

Tukey's multiple comparisons tests (Honestly Significantly Different (HSD) test)

TukeyHSD(fit)$site

	diff	lwr	upr	p
2-1	19.498	7.5	31.5	0.0002
3-1	18.994	7.7	30.3	0.00012
4-1	14.478	4.4	24.6	0.00129
5-1	12.378	-4.4	29.2	0.27443
6-1	12.344	1	23.7	0.02488
7-1	12.178	2.1	22.3	0.00983
3-2	-0.503	-13.5	12.5	1
4-2	-5.02	-17	7	0.84701
5-2	-7.12	-25.1	10.9	0.87775
6-2	-7.153	-20.2	5.9	0.61462
7-2	-7.32	-19.3	4.7	0.49364
4-3	-4.517	-15.8	6.8	0.8742
5-3	-6.617	-24.2	10.9	0.90033
6-3	-6.65	-19.1	5.8	0.6418
7-3	-6.817	-18.1	4.5	0.51096
5-4	-2.1	-18.9	14.7	0.9997
6-4	-2.133	-13.5	9.2	0.9969
7-4	-2.3	-12.4	7.8	0.99147
6-5	-0.033	-17.6	17.5	1
7-5	-0.2	-17	16.6	1
7-6	-0.167	-11.5	11.2	1

Generic recipe for data analysis with general linear model

1. State population, and conditions for taking sample.

2. Construct the model:

   (a) state the response variable;

   (b) state the explanatory variable(s);

   (c) state type of measurement scale for each of these;

   (d) write model relating response to explanatory variable.

3. State H_A/H_0 about terms in model, (and about parameters in model if appropriate). State α, the tolerance for Type 1 error.

4. Execute analysis: place data in model format, code model statement, obtain fitted values and residuals.

5. If regression line is used, examine plot of residuals against fitted values. If bowl or arch is evident, revise the form of the model (back to step 2).

6. Partition df and SS = df * var (Response) according to model, table SS, df, MS, F (by computer usually).

Courtesy of David Schneider

Generic recipe for data analysis with general linear model

7. Calculate Type 1 error (the p value) from density function (F or t distribution, etc.).
8. Check assumption for use of p-value from density function.
   (a) Residuals independent? (plot residuals versus residuals at lag 1)
   (b) Residuals homogeneous? (residual versus fit plot)
   (c) Residuals normal? (histogram of residuals, quantile plot or normal score).
9. If assumption are met then step 10. If not, decide whether to recompute p-value. Recompute better p-value by randomization or bootstrap if sample small ($n < 30$), or p near α.
10. Declare decision about model terms: if $p < \alpha$ then reject H_0 and accept H_A, if $p > \alpha$ then hold H_0 and reject H_A. Report conclusion with evidence: F-ratio, df1, df2, and p-value (not α) for each term.
11. Examine parameters of interest. Report conclusions with parameter estimates (means, slopes) and one measure of uncertainty (st. error, st. dev., or conf. intervals) .

Courtesy of David Schneider

The generic recipe

1. Population: weight (site)

2. $\text{Weight} = \beta_0 + \beta_{site} \times \text{Site} + \text{error}$

   df: 45 6 39

3. $H_0: \beta_{site} = 0$; $H_A: \beta_{site} \neq 0$; $\alpha = 0.05$

Example

Width of scutum of tick larvae, in microns, from 4 hosts (rabbits)

Width	Host	Width	Host
380	1	354	3
376	1	360	3
360	1	362	3
368	1	352	3
372	1	366	3
366	1	372	3
374	1	362	3
382	1	344	3
350	2	342	3
356	2	358	3
358	2	351	3
376	2	348	3
338	2	348	3
342	2	376	4
366	2	344	4
350	2	342	4
344	2	372	4
364	2	374	4
		360	4

Model – with Tukey test

```
# R code
fit <- aov (Width ~ Host, data=tick)

TukeyHSD(fit)$Host
```

Model

$$\text{Width} = \beta_0 + \beta_{\text{Host}} \times \text{Host} + \text{error}$$

Source	DF	SS	MS	F	P
Host	3	1808	602.6	5.26	0.004
Error	33	3778	114.5		
Total	36	5586	155.2		

Tukey test results

Tukey's Studentized Range (HSD) Test for width

Alpha	0.05
Error Degrees of Freedom	33
Error Mean Square	114.4849
Critical Value of Studentized Range	3.82537

Comparisons significant at the 0.05 level are indicated by ***.

host Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
1 - 4	10.917	-4.714	26.547	
1 - 3	16.942	3.937	29.948	***
1 - 2	17.850	4.121	31.579	***
4 - 1	-10.917	-26.547	4.714	
4 - 3	6.026	-8.259	20.310	
4 - 2	6.933	-8.012	21.879	
3 - 1	-16.942	-29.948	-3.937	***
3 - 4	-6.026	-20.310	8.259	
3 - 2	0.908	-11.266	13.081	
2 - 1	-17.850	-31.579	-4.121	***
2 - 4	-6.933	-21.879	8.012	
2 - 3	-0.908	-13.081	11.266	

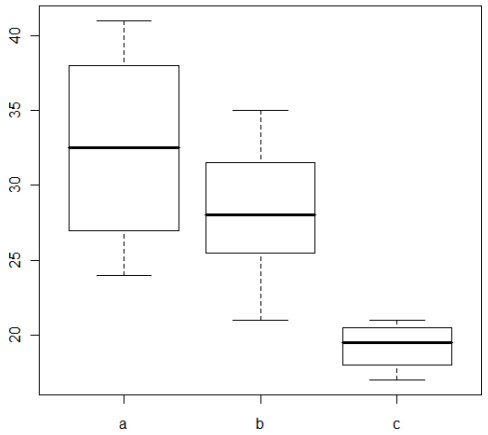
Width = fitted values + residuals

ROW	width	fits	res	fits = $\beta_0 + \beta_{host} \times$ host
1	380	372.250	7.7500	372.250 = 359.703 + 12.5473
2	376	372.250	3.7500	
7	374	372.250	1.7500	
8	382	372.250	9.7500	
9	350	354.400	-4.4000	354.400 = 359.703 - 5.3027
10	356	354.400	1.6000	
17	344	354.400	-10.4000	
18	364	354.400	9.6000	
19	354	355.308	-1.3077	355.308 = 359.703 - 4.3950
20	360	355.308	4.6923	
21	362	355.308	6.6923	
30	348	355.308	-7.3077	
31	348	355.308	-7.3077	
32	376	361.333	14.6667	361.333 = 359.703 + 1.6306
33	344	361.333	-17.3333	
34	342	361.333	-19.3333	
35	372	361.333	10.6667	
36	374	361.333	12.6667	
37	360	361.333	-1.3333	

R script

```
Datafilename <- "http://personality-project.org/R/datasets/R.appendix1.data" #tell where the data come from
data.ex1 <- read.table(datafilename, header = T)   #read the data into a table
aov.ex1 <- aov(Alertness ~ Dosage, data = data.ex1)  #do the analysis of variance
print(model.tables(aov.ex1,"means"),digits = 3)  #report the means and the number of subjects/cell
boxplot(Alertness ~ Dosage, data = data.ex1)       #graphical summary
summary(aov.ex1) #show the results
```

	Dosage	Alertness
1	a	30
2	a	38
3	a	35
4	a	41
5	a	27
6	a	24
7	b	32
8	b	26
9	b	31
10	b	29
11	b	27
12	b	35
13	b	21
14	b	25
15	c	17
16	c	21
17	c	20
18	c	19




```
Tables of means
Grand mean
27.66667

Dosage
        a    b    c
     32.5 28.3 19.3
rep   6.0  8.0  4.0
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Dosage	2	426.25	213.13	8.7887	0.002977 **
Residuals	15	363.75	24.25		

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

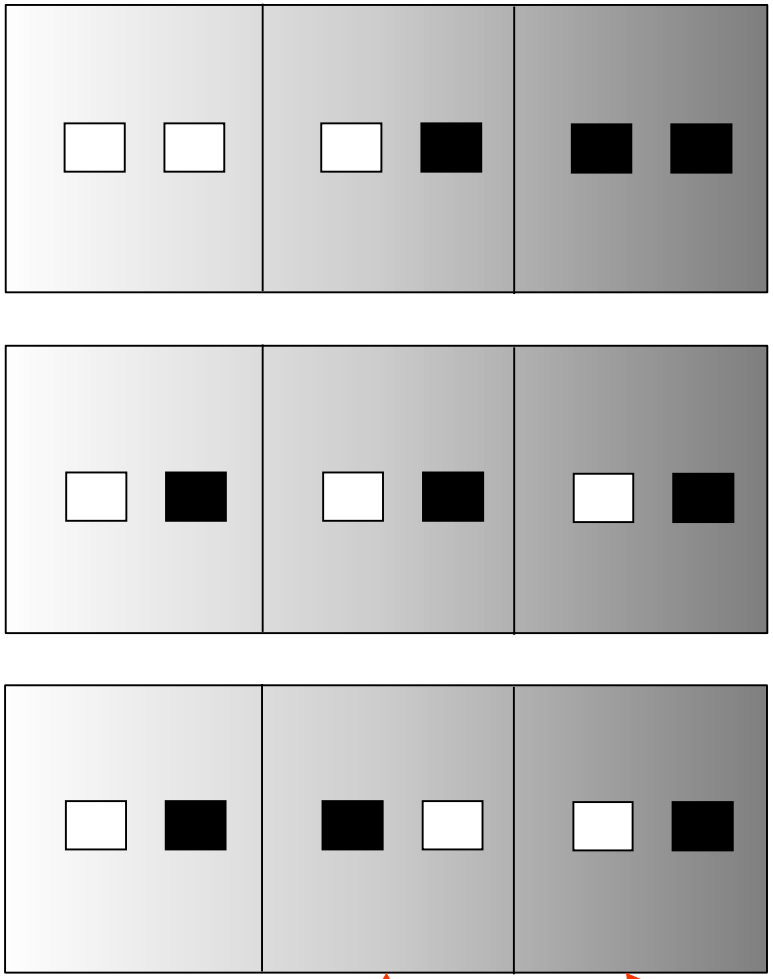
Randomized block designs

Example: randomized block design

CO_2 density

	Treat			
Incubator	A	B	C	D
1	5.27	5.27	5.94	5.53
2	5.27	5.22	4.88	4.96
3	5.88	5.83	5.38	5.53
4	5.44	5.38	5.27	5.32
5	5.66	5.44	5.38	4.88
6	6.22	6.22	5.61	5.92
7	5.83	5.72	5.38	4.88
8	5.27	5.11	5.12	4.44

Three options in assigning treatments:



- Randomly assign

Pros: Statistically robust
Cons: With small n, chance of all in a bad patch

- Systematic

Pros: No clumping possible
Cons: Violates random assumption of statistics, yet it usually well represents the population.

- Randomized block

The model for a randomized block design

$$y_{ij} = \beta_0 + \beta_i \times treatment + \beta_j \times block + \varepsilon_{ij}$$

$$i = 1,2,\ldots, t \qquad j = 1,2,\ldots, b$$

y_{ij} = the observation in the jth block receiving the ith treatment

β_0 = overall mean

β_i = the effect of the ith treatment

β_j = the effect of the jth block

ε_{ij} = random error

The ANOVA Table for a randomized Block Experiment

Source	S.S.	d.f.	M.S.	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		

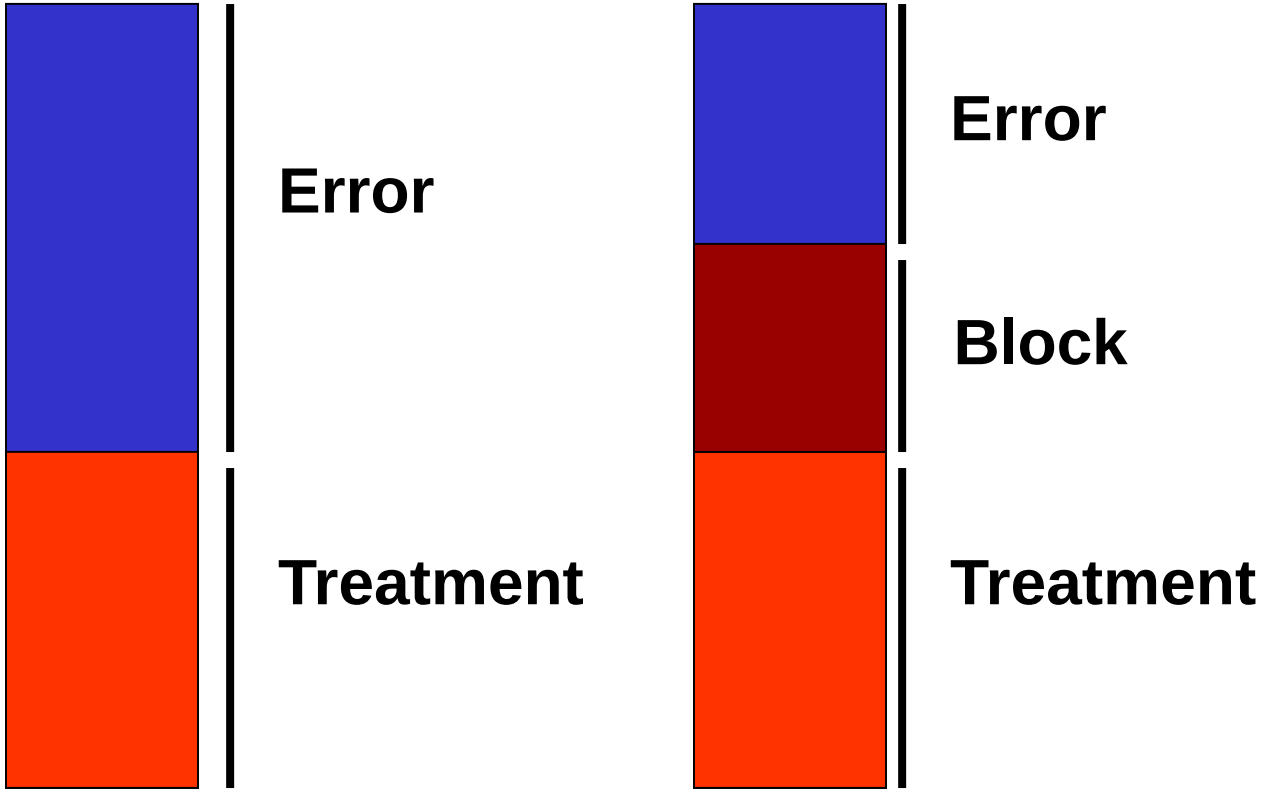
Notes

- A randomized block experiment is assumed to be a two-factor experiment.
- The factors are blocks and treatments.
- It is assumed that there is no interaction between blocks and treatments.
- The degrees of freedom for the interaction is used to estimate error.

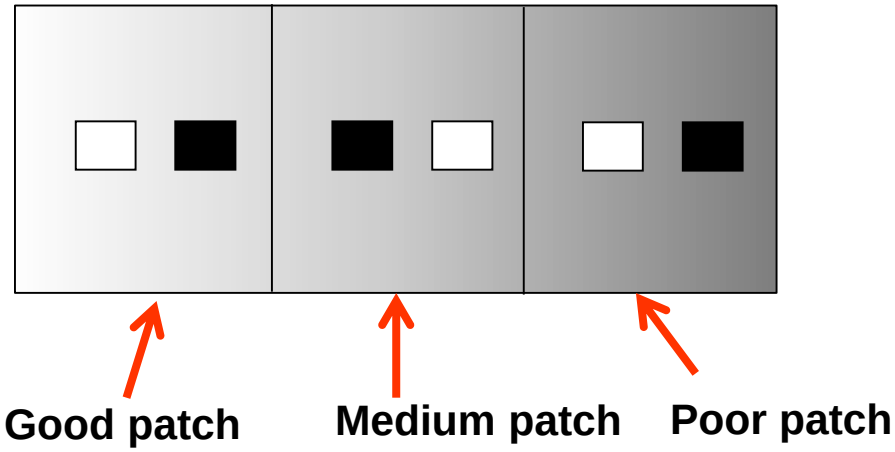
Can remove between-block variance from error variance:

increase power of test

$$F = \frac{\text{Treatment SS/treatment df}}{\text{Error SS/error df}}$$



Note - Randomized block



1. **Do not have to know if patches differ in quality**
2. **Must have all treatment combinations represented in each block**

Randomized block *F* test assumptions

1. Normality

   Populations are normally distributed

2. Homogeneity of Variance

   Populations have equal variances

3. Independence of Errors

   Independent random samples are drawn

4. No Interaction Between Blocks & Treatments

Example: randomized block design

CO_2 density

	Treat			
Incubator	A	B	C	D
1	5.27	5.27	5.94	5.53
2	5.27	5.22	4.88	4.96
3	5.88	5.83	5.38	5.53
4	5.44	5.38	5.27	5.32
5	5.66	5.44	5.38	4.88
6	6.22	6.22	5.61	5.92
7	5.83	5.72	5.38	4.88
8	5.27	5.11	5.12	4.44

Models

One way ANOVA:

$$\text{Density} = \beta_0 + \beta_{\text{treat}} \times \text{treat} + \text{error}$$

Randomized block designs:

$$\begin{aligned}\text{Density} = {} & \beta_0 + \beta_{\text{treat}} \times \text{treat} \\ & + \beta_{\text{incubator}} \times \text{incubator} \\ & + \text{error}\end{aligned}$$

R script - randomized block design

```
#Randomized Block Design
#Carbon dioxygen density at 8 incubators and 4 treatments
CO2 <- data.frame(ID=1:32, group=NA, treat=NA, density=NA)
n <- 0
for(i in 1:8){
            for(j in c('A','B','C','D')){
            n <- n+1
            CO2$group[n] = i
            CO2$treat[n] = j
}}
CO2$group   <-  factor(CO2$group)
CO2$treat     <- factor(CO2$treat)
CO2$density <- c(5.27,5.27,5.94,5.53,5.27,5.22,4.88,4.96,5.88,5.83,
    5.38,5.53,5.44, 5.38,5.27,5.32,5.66, 5.44,5.38,4.88,6.22,
    6.22,5.61,5.92,5.83,5.72,5.38,4.88,5.27,5.11,5.12,4.44)
fit1 <- aov(density ~ treat, data = CO2) # one way ANOVA
fit2 <- aov(density ~ group + treat, data = CO2) # Randomized Block Design

summary(fit1)
summary(fit2)
par(mfrow=c(2,2))
plot(fit2)
```

R results

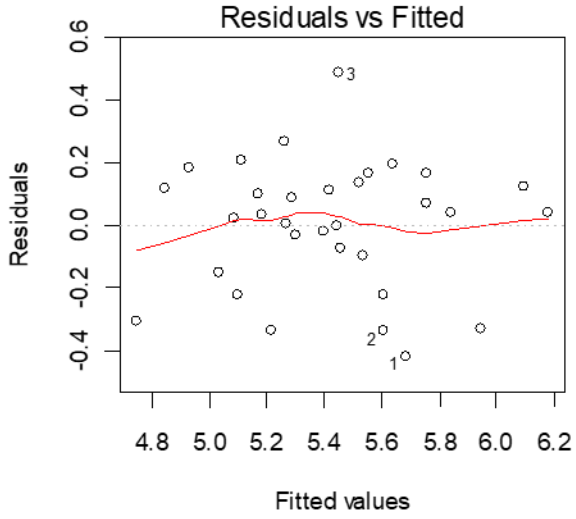
> summary(fit1)

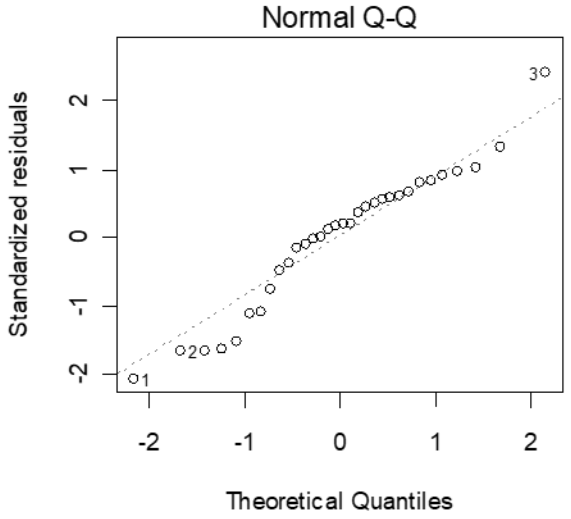
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
treat	3	0.83	0.277	1.88	0.16
Residuals	28	4.13	0.147		

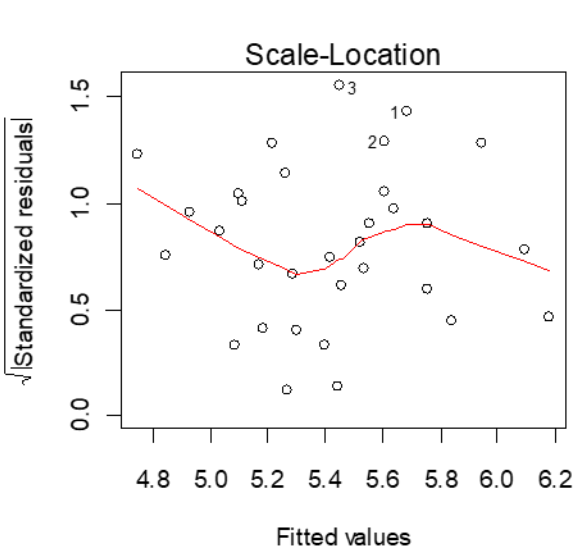
> summary(fit2)

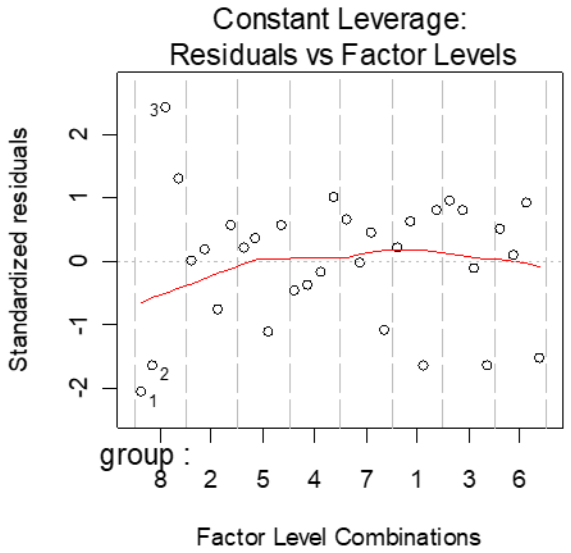
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	7	2.820	0.403	6.47	0.00039 ***
treat	3	0.831	0.277	4.45	0.01434 *
Residuals	21	1.308	0.062		

R plot(fit2)


Residuals vs Fitted
Residuals
Fitted values
3
2
1
-0.4
-0.2
0.0
0.2
0.4
0.6
4.8
5.0
5.2
5.4
5.6
5.8
6.0
6.2


Normal Q-Q
Standardized residuals
Theoretical Quantiles
3
2
1
-2
-1
0
1
2


Scale-Location
√|Standardized residuals|
Fitted values
3
1
2
0.0
0.5
1.0
1.5
4.8
5.0
5.2
5.4
5.6
5.8
6.0
6.2


Constant Leverage:
Residuals vs Factor Levels
Standardized residuals
Factor Level Combinations
group :
8 2 5 4 7 1 3 6
3
2
1
-2
-1
0
1
2

Reshape data

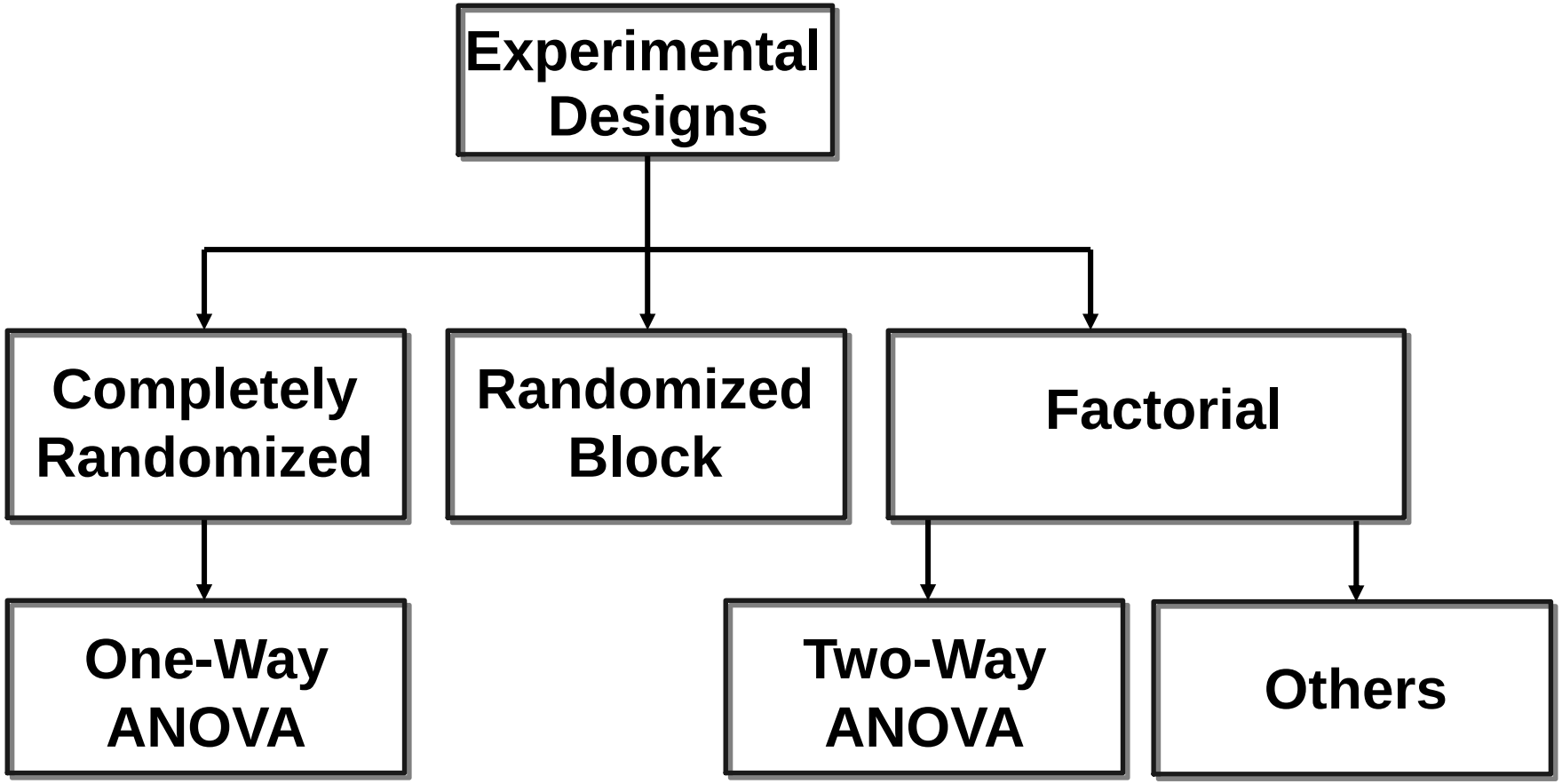
```
head(CO2)
```

ID	group	treat	density
1	1	A	5.27
2	1	B	5.27
3	1	C	5.94
4	1	D	5.53
5	2	A	5.27
6	2	B	5.22

```
library(reshape2)
Tab = melt(CO2, id=c("group", "treat"), na.rm=TRUE)
Mat = acast(Tab, group ~ treat ~ variable)
dim(Mat); Mat[ , , 2]
```

	A	B	C	D
1	5.27	5.27	5.94	5.53
2	5.27	5.22	4.88	4.96
3	5.88	5.83	5.38	5.53
4	5.44	5.38	5.27	5.32
5	5.66	5.44	5.38	4.88
6	6.22	6.22	5.61	5.92
7	5.83	5.72	5.38	4.88
8	5.27	5.11	5.12	4.44

Types of ANOVA



Two way ANOVA

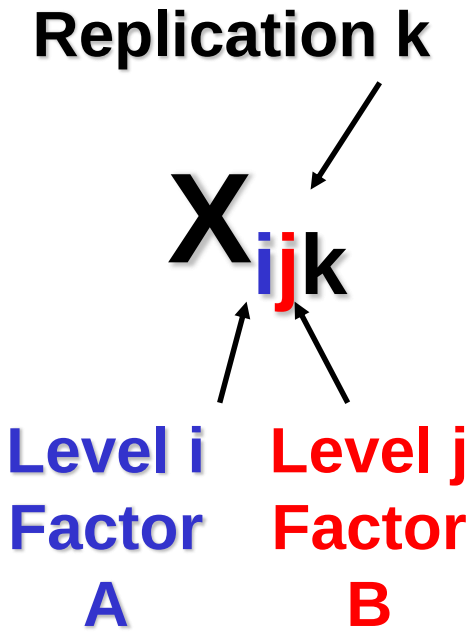
Two way ANOVA

Weight gain under different density (amount) and food (type)

Amount	Type	Gain									
high	beef	73	102	118	104	81	107	100	87	117	111
high	cereal	98	74	56	111	95	88	82	77	86	92
high	pork	94	79	96	98	102	102	108	91	120	105
low	beef	90	76	90	64	86	51	72	90	95	78
low	cereal	107	95	97	80	98	74	74	67	89	58
low	pork	49	82	73	86	81	97	106	70	61	82

Two-way ANOVA data table

Factor A	Factor B			
	1	2	...	b
1	X_{111}	X_{121}	...	X_{1b1}
	X_{112}	X_{122}	...	X_{1b2}
2	X_{211}	X_{221}	...	X_{2b1}
	X_{212}	X_{222}	...	X_{2b2}
:	:	:	:	:
a	X_{a11}	X_{a21}	...	X_{ab1}
	X_{a12}	X_{a22}	...	X_{ab2}



Two-way ANOVA null hypotheses

1. No Difference in Means Due to Factor A

   H_0: $\mu_{1..} = \mu_{2..} = \ldots = \mu_{a..}$

2. No Difference in Means Due to Factor B

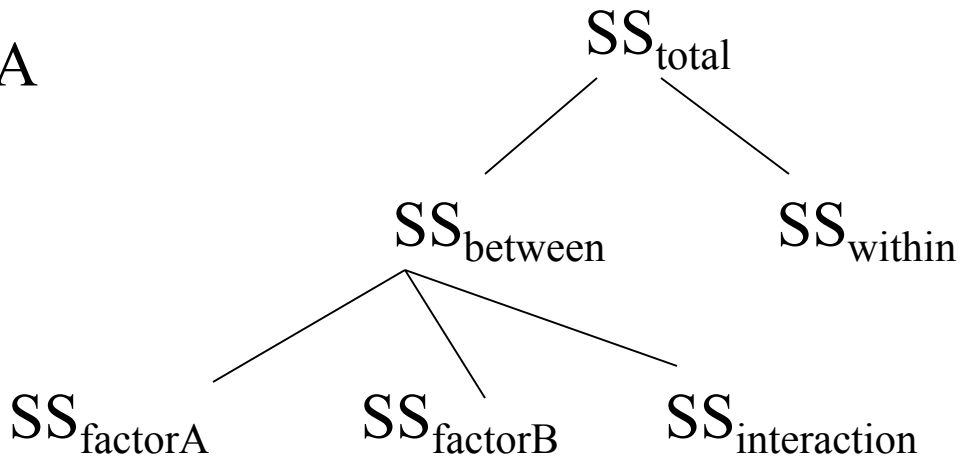
   H_0: $\mu_{.1.} = \mu_{.2.} = \ldots = \mu_{.b.}$

3. No Interaction of Factors A & B

   H_0: $AB_{ij} = 0$

Partitioning variance of two way ANOVA

Two Way ANOVA



Main Effect 1

$$F = \frac{s^2_{factorA}}{s^2_w}$$

Main Effect 2

$$F = \frac{s^2_{factorB}}{s^2_w}$$

Interaction

$$F = \frac{s^2_{interaction}}{s^2_w}$$

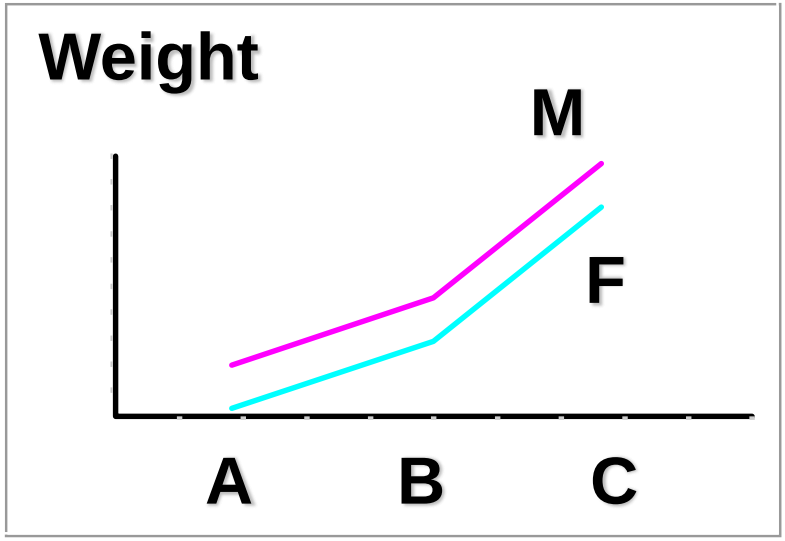
Interaction

- Occurs when effects of one factor vary according to levels of other factor
- When significant, interpretation of main effects (A & B) is complicated
- Can be detected
  - in data table, pattern of cell means in one row differs from another row
  - in graph of cell means, lines cross

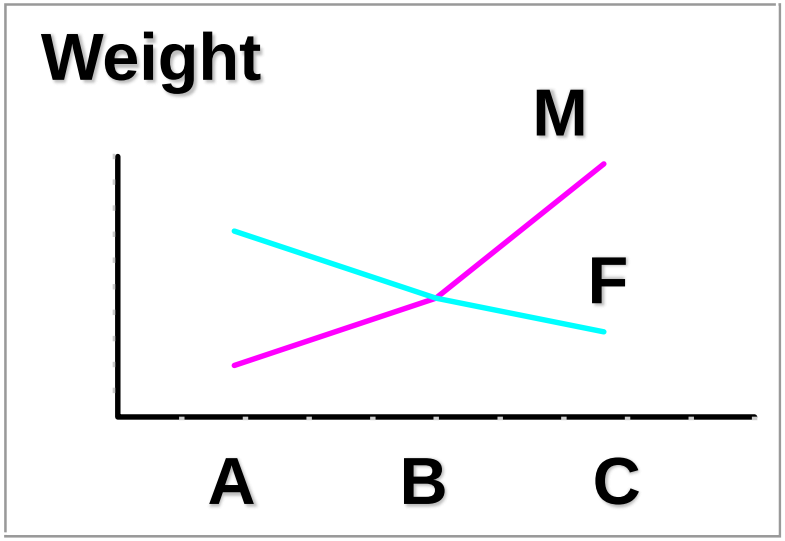
Graphs of interaction

Effects of gender (male and female) & location (countries A, B, and C) on body weight

No Interaction



Interaction



Graphs of interaction

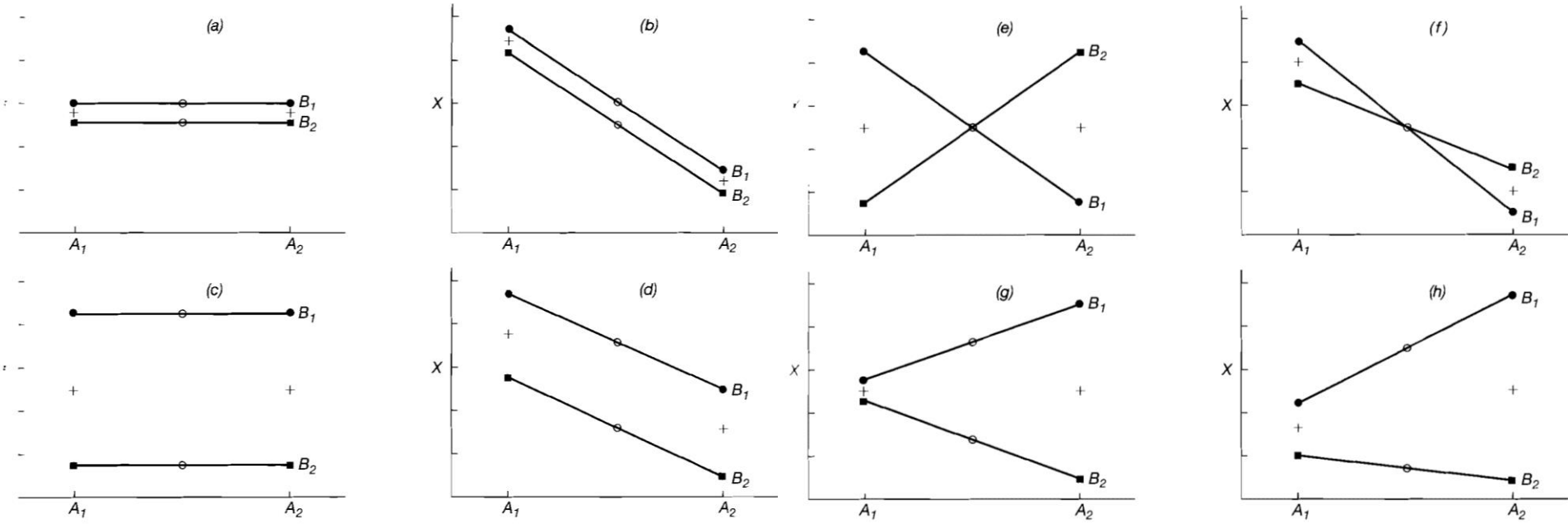
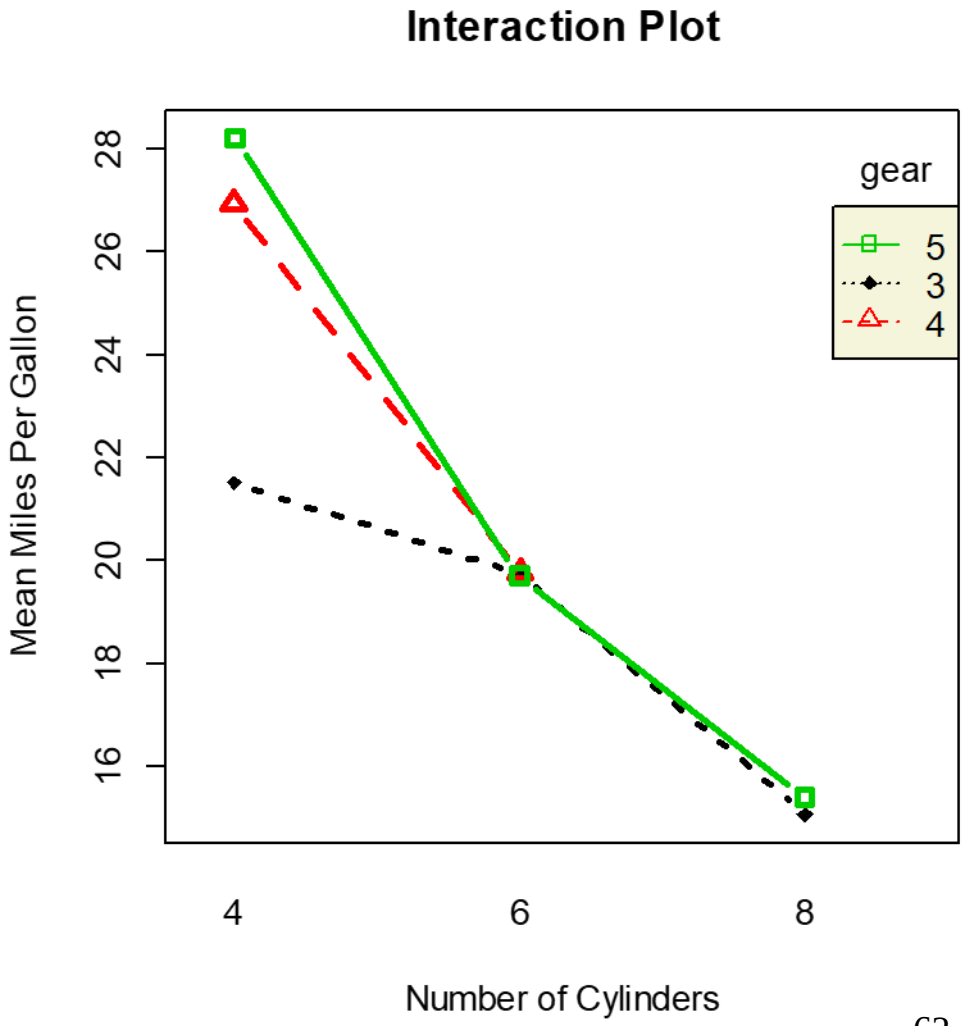


Figure 12.2 Means in a two-factor ANOVA, showing various effects of the two factors and their interaction. (a) No effect of factor *A*, small effect of factor *B* (and if there were no effect of *B* the two lines would coincide), and no interaction of *A* and *B*. (b) Large effect of factor *A*, small effect of factor *B*, and no interaction (which is the situation in Fig. 12.1). (c) No effect of *A*, large effect of *B*, and no interaction. (d) Large effect of *A*, large effect of *B*, and no interaction. (e) No effect of *A*, no effect of *B*, but interaction between *A* and *B*. (f) Large effect of *A*, no effect of *B*, with slight interaction. (g) No effect of *A*, large effect of *B*, with large interaction. (h) Effect of *A*, large effect of *B*, with large interaction.

(Zar 1999)

R script for two-way interaction plot

```
# Two-way Interaction Plot
attach(mtcars)
gears <- factor(gear)
cyl <- factor(cyl)
interaction.plot(cyl, gear, mpg,
          type="b", col=c(1:3),
          leg.bty="o", leg.bg="beige",
          lwd=2,   pch=c(18,24,22),
          xlab="Number of Cylinders",
          ylab="Mean Miles Per Gallon",
          main="Interaction Plot")
```




```
leg.bty: legend boundary type (o, n, l)
Leg.bg:  legend background
```

Variance partition of two way ANOVA

$$SS_{tot} = \sum\sum\sum (X - \bar{\bar{X}})^2$$

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

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

$$SS_{within} = \sum (X - \bar{X}_{AB})^2$$

$$SS_{interaction} = SS_{tot} - SS_{within} - SS_A - SS_B$$

Steps for two way ANOVA

Seed Germination = time + temperature + time × temperature

	Lower.T	Median.T	Higher.T	
Morning	1 2 3 $\bar{X}=2$	4 5 6 $\bar{X}=5$	3 1 2 $\bar{X}=2$	$\bar{X}=3$
Afternoon	1 4 1 $\bar{X}=2$	0 1 2 $\bar{X}=1$	0 0 0 $\bar{X}=0$	$\bar{X}=1$
	$\bar{X}=2$	$\bar{X}=3$	$\bar{X}=1$	$\bar{\bar{X}}=2$

Get the cell means, row means, column means, and grand mean

Steps for two way ANOVA

	Lower.T	Median.T	Higher.T
Morning	1 2 3 $\bar{X} = 2$	4 5 6 $\bar{X} = 5$	3 1 2 $\bar{X} = 2$
Afternoon	1 4 1 $\bar{X} = 2$	0 1 2 $\bar{X} = 1$	0 0 0 $\bar{X} = 0$

$\bar{\bar{X}} = 2$

Step 1: Get the SS_{total}

$$SS_{tot} = \sum (X - \bar{\bar{X}})^2$$

= 56

Steps for two way ANOVA

	Lower.T	Median.T	Higher.T
Morning	1 2 3 $\bar{X} = 2$	4 5 6 $\bar{X} = 5$	3 1 2 $\bar{X} = 2$
Afternoon	1 4 1 $\bar{X} = 2$	0 1 2 $\bar{X} = 1$	0 0 0 $\bar{X} = 0$

$\bar{\bar{X}} = 2$

Step 2: Get the SS_{within}

$$SS_w = \sum (X - \bar{X}_{cell})^2$$

= 14

Steps for two way ANOVA

	Lower.T	Median.T	Higher.T	
Morning	1 2 3 $\bar{X}=2$	4 5 6 $\bar{X}=5$	3 1 2 $\bar{X}=2$	$\bar{X}=3$
Afternoon	1 4 1 $\bar{X}=2$	0 1 2 $\bar{X}=1$	0 0 0 $\bar{X}=0$	$\bar{X}=1$
				$\bar{\bar{X}}=2$

Step 3: Get the SS_{time}

$$SS_{rowfactor} = \sum n_{row}(\bar{X}_{row} - \bar{\bar{X}})^2$$

= 18

Steps for two way ANOVA

	Lower.T	Median.T	Higher.T
Morning	1 2 3 $\bar{X}=2$	4 5 6 $\bar{X}=5$	3 1 2 $\bar{X}=2$
Afternoon	1 4 1 $\bar{X}=2$	0 1 2 $\bar{X}=1$	0 0 0 $\bar{X}=0$
	$\bar{X}=2$	$\bar{X}=3$	$\bar{X}=1$	$\bar{\bar{X}}=2$

Step 4: Get the SS_{temp}

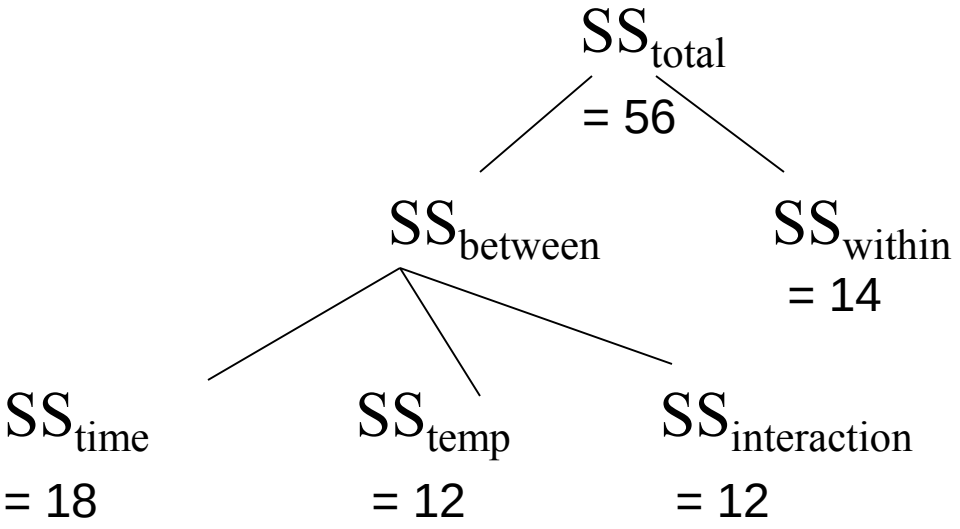
$$SS_{colfactor} = \sum n_{col}(\bar{X}_{col} - \bar{\bar{X}})^2$$

= 12

Steps for two way ANOVA

	Lower.T	Median.T	Higher.T
Morning	1 2 3 $\overline{X}=2$	4 5 6 $\overline{X}=5$	3 1 2 $\overline{X}=2$
Afternoon	1 4 1 $\overline{X}=2$	0 1 2 $\overline{X}=1$	0 0 0 $\overline{X}=0$

Step 5: Get the $SS_{interaction}$



Steps for two way ANOVA

	Lower.T	Median.T	Higher.T
Morning	1 2 3	4 5 6	3 1 2
Afternoon	1 4 1	0 1 2	0 0 0

Step 6: Finish the table

Source	SS	df	s^2	F	
Between					
SS_{time}	18	1	18	15.38	p < .01
SS_{temp}	12	2	6	5.12	p < .05
$SS_{time\ X\ temp}$	12	2	6	5.12	p < .05
Within	14	12	1.17		
Total	56	17			

Steps for two way ANOVA

If interaction exists (significant), split the two-way ANOVA to two one-way ANOVA.

Since there are only two levels of time, you know they are significantly different from one another

Since there are three levels of temperature, you will need to use post-hoc multiple comparisons (e.g. Tukey's HSD) to determine which levels differ from one another.

Assumptions for the two factor ANOVA

1. Observations within each sample are independent.
2. Populations are normally distributed.
3. Populations from which the samples are selected must have equal variances (homogeneity of variance).

1 2 3	4 5 6	3 1 2
1 4 1	0 1 2	0 0 0

Two way ANOVA example

Weight gain under different density (amount) and food (type)

Amount	Type	Gain									
high	beef	73	102	118	104	81	107	100	87	117	111
high	cereal	98	74	56	111	95	88	82	77	86	92
high	pork	94	79	96	98	102	102	108	91	120	105
low	beef	90	76	90	64	86	51	72	90	95	78
low	cereal	107	95	97	80	98	74	74	67	89	58
low	pork	49	82	73	86	81	97	106	70	61	82

Model

$$\begin{aligned} \text{Gain} = {} & \beta_0 + \beta_{\text{amount}} \times \text{amount} \\ & + \beta_{\text{type}} \times \text{type} \\ & + \beta_{\text{type} \times \text{amount}} \times \text{amount} \times \text{type} \\ & + \text{error} \end{aligned}$$

R – two way ANOVA

```
# 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)
```

R results

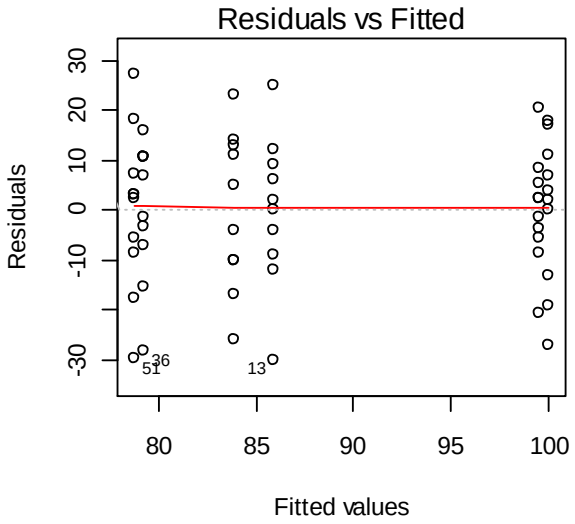
> summary(fit)

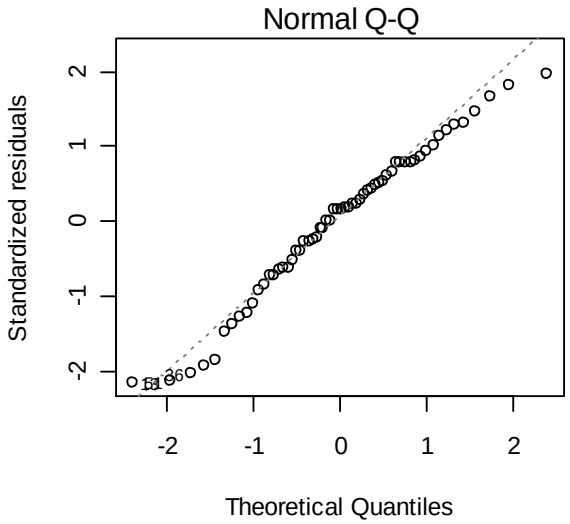
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
amount	1	3168	3168	14.77	0.00032 ***
food	2	267	133	0.62	0.54113
amount:food	2	1178	589	2.75	0.07319 .
Residuals	54	11586	215		

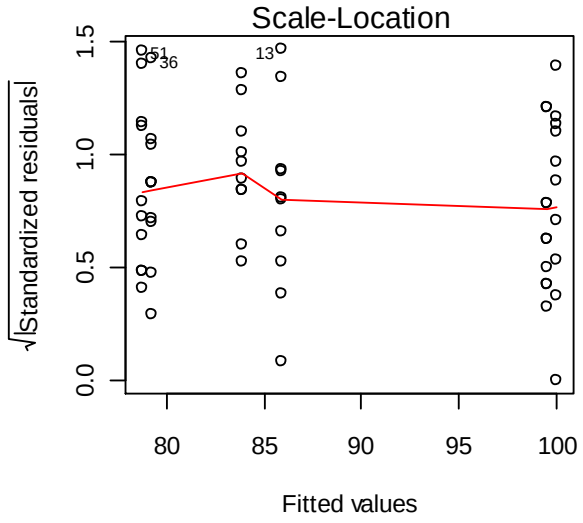
 # Because the interaction term is not significant, the final model is:

fit <- aov(gain ~ amount + food, data = weight.gain)

R – model performance







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
```

Assignment

General objectives: learn one way ANOVA.

- Prepare your own data,
- 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.

R script

```
# Input data
site1 <- c(9.4, 8.7, 13.3, 13.6, 15, 15.2, 17.7, 18.6, 22.2)
site2 <- c(16.8, 30.8, 33.6, 40.5, 48.9)
site3 <- c(27.0, 28.9, 32, 32.7, 35.5, 45.6)

rodent.survey <- data.frame(weight=c(site1,site2,site3),
      site=factor(c(rep("1",9),rep("2",5),rep("3",6))))

# Check data
options(digits=3) # default value = 7
tapply(rodent.survey$weight, rodent.survey$site,mean)
tapply(rodent.survey$weight, rodent.survey$site,var)

boxplot(weight~site, data=rodent.survey, xlab='Sites', ylab='Weight')

# Bartlett Test of Homogeneity of Variances (parametric)
bartlett.test (weight~site, data=rodent.survey)

# One Way ANOVA
fit <- aov (weight ~ site, data = rodent.survey)
summary(fit)
plot(fit)
```