McMaster University | Integrated Biomedical Engineering & Health Sciences Program

IBEHS - 4QZ3
Modelling of Biological Systems

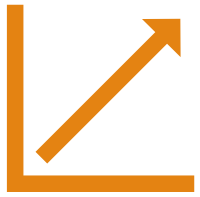
Lecture 3

TAYLOR DEVET MASC.

PHD. CANDIDATE BIOLOGICAL AND BIOMEDICAL ENGINEERING
MCGILL UNIVERSITY

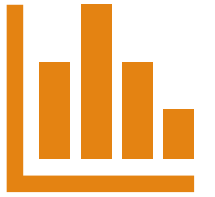
SHRINERS HOSPITAL FOR CHILDREN

Todays Aims…



Correlation

Complete Randomized Design



Post Hoc Testing

Let’s go back to the cardiac data…….

Cardiac Example

Null Hypothesis: None of the independent variables are of any value in explaining the variation in cardiac force output.

i.e. $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

Alternative, H_1 : not all β's are zero.

Source	df	SS	MS	F_c	$F_{i,j,\alpha}$
regression	4	197382.43	49458.11	34.48	2.84
residual	21	30121.92	1434.38	--------	--------
TOTAL	25	227954.35	--------	--------	--------

$F_C > F_{i,j,\alpha}$. Therefore, reject H_0; at least one β is not zero.

SSCP Matrix

Sum of Squares and Cross Product Matrix

Main diagonal are sum of squares for each column

Off main diagonals are sums of cross products

Inverse of this times Stdev gives us a matrix with variances and covariances

26.0000	49.2900	6.5150	19.8100	23.0700
49.2900	97.9781	12.7980	39.0012	45.9843
6.5150	12.7980	1.7540	5.2688	6.1600
19.8100	39.0012	5.2688	16.6387	18.9306
23.0700	45.9843	6.1600	18.9306	23.1015

$(X'X)$

0.9186	-0.3780	-0.9663	-0.1060	0.1796
-0.3780	0.4209	-0.6846	-0.0550	-0.2328
-0.9663	-0.6846	20.1257	-2.3701	-1.0967
-0.1060	-0.0550	-2.3701	1.5028	-0.3842
0.1796	-0.2328	-1.0967	-0.3842	0.9346

$(X'X)^{-1}$

Variance Covariance Matrix

$$\begin{vmatrix} Var(b_0) & Cov(b_0,b_1) & Cov(b_0,b_2) & Cov(b_0,b_3) & Cov(b_0,b_4) \\ Cov(b_1,b_0) & Var(b_1) & Cov(b_1,b_2) & Cov(b_1,b_3) & Cov(b_1,b_4) \\ Cov(b_2,b_0) & Cov(b_2,b_1) & Var(b_2) & Cov(b_2,b_3) & Cov(b_2,b_4) \\ Cov(b_3,b_0) & Cov(b_3,b_1) & Cov(b_3,b_2) & Var(b_3) & Cov(b_3,b_4) \\ Cov(b_4,b_0) & Cov(b_4,b_1) & Cov(b_4,b_2) & Cov(b_4,b_3) & Var(b_4) \end{vmatrix} = s^2 \times (X'X)^{-1}$$

$$= \begin{vmatrix} \mathbf{1317.55} & -542.151 & -1386.08 & -152.066 & 257.6247 \\ -542.151 & \mathbf{603.774} & -981.925 & -78.8711 & -333.958 \\ -1386.08 & -981.925 & \mathbf{28867.82} & -3399.59 & -1573.08 \\ -152.066 & -78.8711 & -3399.59 & \mathbf{2155.64} & -551.089 \\ 257.6247 & -333.958 & -1573.08 & -551.089 & \mathbf{1340.625} \end{vmatrix}$$

Cardiac Example cont

$(X'X)^{-1}$

0.9186	-0.3780	-0.9663	-0.1060	0.1796
-0.3780	0.4209	-0.6846	-0.0550	-0.2328
-0.9663	-0.6846	20.1257	-2.3701	-1.0967
-0.1060	-0.0550	-2.3701	1.5028	-0.3842
0.1796	-0.2328	-1.0967	-0.3842	0.9346

With rejection of HO, one can test for specific factors.

e.g. $H_0 : \beta_4 = c$

where c is any number specified (often zero).

1317.55	-542.151	-1386.08	-152.066	257.6247
-542.151	**603.774**	-981.925	-78.8711	-333.958
-1386.08	-981.925	**28867.82**	-3399.59	-1573.08
-152.066	-78.8711	-3399.59	**2155.64**	-551.089
257.6247	-333.958	-1573.08	-551.089	**1340.625**

$$t = \frac{b_4 - c}{\text{StdErr}(b_4)} = \frac{b_4 - c}{\sqrt{s^2 \cdot (X'X)^{-1}}} = \frac{40.31 - 0}{\sqrt{1434.38 \cdot 0.9346}}$$

$$= 1.1$$

$$\beta = \begin{bmatrix} -185.33 \\ 97.76 \\ 256.97 \\ 126.57 \\ 40.28 \end{bmatrix}$$

- s^2 is the residual (MS) from the full model (ANOVA) table.

Therefore, there is little evidence to reject $H_0 : \beta_4 = 0$

Cardiac Example cont

A 95% confidence interval may be calculated for β_4:

$$b_4 \pm t_{\alpha,df} \times stderr(b_4) = 40.3 \pm (2.080 \times 36.61)$$

$$stderr(b_4) = \sqrt{s^2 \times (X'X)^{-1}_{5,5}}$$

Where: $t_{\alpha,df} = t_{0.05,26-5} = 2.080$

$t = 1.1$

Therefore, $-35.8 \leq b_4 \leq 116.4$

Reject Ho? $\beta_4 = 0$ Based on the confidence interval it could very well be zero.

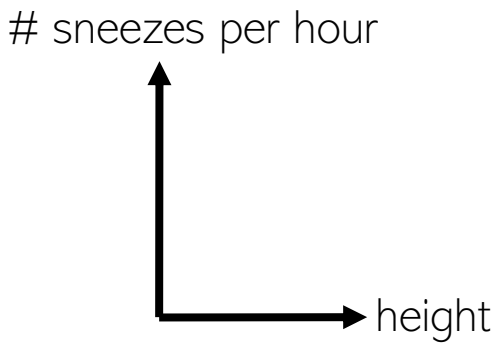
Correlated Variables

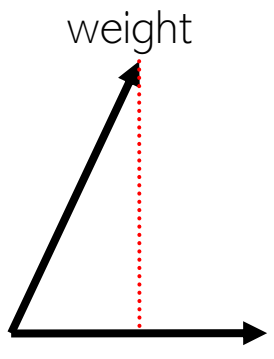
Need to understand whether independent variables are correlated.

If independent variables are not correlated among themselves they are said to be orthogonal.

When correlated they are non-orthogonal.

Non-correlated independent variables = Orthogonal





- weight and height are correlated

Multicollinearity

A lot of these calculations pend on the X'X matrix being invertible.

This problem is referred to as the multicollinearity problem

Multicollinearity occurs when there are high correlations between two or more predictor variables.

i.e. one predictor variable can be used to predict the other giving redundancy and skewing the model.

Multicollinearity

Examples of multicollinear predictors:

1) a person's height and weight

2) years of education and annual income.

3) Height in metres and height in feet

An easy way to detect:

Calculate correlation coefficients (r) for all pairs of predictor variables. If r is exactly +1 or -1, this is called perfect multicollinearity and one of the variables should be removed from the model

Main Causes of Multicollinearity:

1) Data-based
- poorly designed experiments
- data that is 100% observational, or data collection methods that cannot be manipulated.
- In some cases, variables may be highly correlated (usually due to collecting data from purely observational studies) and there is no error on the researcher's part. Can test this in advance!!

2) Structural multicollinearity
- caused by the researcher, poorly creating new predictor variables

There is a fix!! Plan the experiment!

Observational Example:

Estimate 4 slopes and 1 intercept of zero (i.e. not estimating)

```
X =

     1     0     0     0
     0     1     0     1
     0     0     0     1
     1     0     0     0
     0     1     0     1
     0     0     0     1
     1     0     0     0
     0     1     0     1
     0     0     0     1
     1     0     0     0
     0     1     0     1
     0     0     0     1
```

```
>> X'*X

ans =

     4     0     0     0
     0     4     0     4
     0     0     0     0
     0     4     0     8

>> (X'*X)^-1
Warning: Matrix is singular to working precision.

ans =

   Inf   Inf   Inf   Inf
   Inf   Inf   Inf   Inf
   Inf   Inf   Inf   Inf
   Inf   Inf   Inf   Inf
```

Observational Example:

Estimate 4 slopes and 1 intercept as b

```
X =

     1     1     0     0     0
     1     0     1     0     0
     1     0     0     1     0
     1     0     0     0     1
     1     1     0     0     0
     1     0     1     0     0
     1     0     0     1     0
     1     0     0     0     1
     1     1     0     0     0
     1     0     1     0     0
     1     0     0     1     0
     1     0     0     0     1
     1     1     0     0     0
     1     0     1     0     0
     1     0     0     1     0
     1     0     0     0     1
```

```
>> X'*X

ans =

    16     4     4     4     4
     4     4     0     0     0
     4     0     4     0     0
     4     0     0     4     0
     4     0     0     0     4
```

$$(X'X)^{-1} = 1e+15 \cdot \begin{bmatrix} 4.50 & -4.50 & -4.50 & -4.50 & -4.50 \\ -4.50 & 4.50 & 4.50 & 4.50 & 4.50 \\ -4.50 & 4.50 & 4.50 & 4.50 & 4.50 \\ -4.50 & 4.50 & 4.50 & 4.50 & 4.50 \\ -4.50 & 4.50 & 4.50 & 4.50 & 4.50 \end{bmatrix}$$

Essentially this is infinite!!

Correlated Variables

In the heart example the independent variables are correlated.

i.e. force is related to a combination of ions!

- What about sodium, β4 ? We tested and accepted H0: β4 = 0

The general approach to the analysis is the comparison of two competing models:

1) The full model $$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i$$

2) The reduced model $$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

Model Reduction

Fit the full model (done already) and reduced model to the data and for each obtain the model SS and residual SS.

- The appropriate SS for testing H0:$\beta 4 = 0$ is:

$$R(\beta_4 \mid \beta_1, \beta_2, \beta_3) = \text{SS model (full) - SS model (reduced)}$$

$$= \text{SS residual (reduced) - SS residual (full)}$$

- this 'R' notation helps to differentiate the different SS.

- 'R' means 'Reduction in residual SS'

Model Reduction

The full model (already calculated) is:

$$R(\beta_1, \beta_2, \beta_3, \beta_4) = \text{SS (full model)} = 197832.43$$

The reduced model is:

$$R(\beta_1, \beta_2, \beta_3) = \text{SS (reduced model)} = 196096.77$$

Therefore,

$$R(\beta_4 \mid \beta_1, \beta_2, \beta_3) = R(\beta_1, \beta_2, \beta_3, \beta_4) - R(\beta_1, \beta_2, \beta_3)$$

$$= 197832.43 - 196096.77$$

$$= 31857.58 - 30121.92$$

$$= 1735.66$$

NOTE: red #s are from reduced model - you should verify these !!

Model Reduction

The reduced model is fitted in the same way as the full model !!

$$\hat{\mu}_{Y.123} = -193.07 + 107.80X_1 + 304.24X_2 + 143.13X_3$$

Source	df	SS	MS
Reduced model	3 ← 4	196096.77 197k...	65365.59
residual	22 21	31857.58 30121	1448.07
TOTAL	25 25	227954.35	---------

NOTE: you should verify the ANOVA and reduced b results !!

Model Reduction

- the df for $\beta 4 = 1$ (i.e. 4 - 3)

Calculation of the F statistic to test H0:$\beta 4 = 0$

For the denominator MS of the F-test the residual MS from the FULL MODEL is used!

$$F = \frac{R(\beta_4 \mid \beta_1, \beta_2, \beta_3)/df_{\beta_4}}{Error(MS)_{FULL}} = \frac{1735.66/1}{1434.38} = 1.21$$

$$F_{1,21,\alpha=0.05} = 4.32 \text{ (table)}$$

Since $F_c < F_{table}$ there is little evidence to suggest $\beta_4 \neq 0$

Which model do we want?

Use F statistic

$$F = \frac{\frac{SSE\ reduced - SSE\ Full}{df\ reduced - df\ full}}{\frac{SSE\ full}{df\ full}}$$

$$F = \frac{\frac{31857.58 - 30121.92}{22 - 21}}{\frac{30121.9}{21}}$$

F = 1735.66/1434.37 = 1.21

$$\frac{df\ Reduced - df\ Full}{df\ Full} =$$

$F(alpha\ 1,21\ ) = 4.32$

Model Reduction

Therefore, it can be concluded that Na is not related to force after allowing for the influence of Cl^-, PO4 and K^+.

Note: If we had only done simple linear regression of Y on X4, ignoring X1, X2, and X3:

- Would have been fitting the model $Yi=\beta 0 + \beta 4X4 + \varepsilon i$)
- Testing the significance of the regression (i.e. H0:$\beta 4 = 0$) would have resulted in Fc = 34.4
- Testing two VERY different hypotheses

→ Therefore careful experimental planning are essential !

Estimating Y for Given Values of X's

Example. Estimate the contractile force with Cl-=2, PO4=0.3, and K+=0.8 in the buffer. (note X4 removed, based on previous test)

Therefore the estimate of force is:

$$\hat{\mu}_{Y.123} = -193.07 + 107.80(2) + 304.24(0.3) + 143.13(0.8) = 228.3\ N$$

(β_1, β_2, β_3)

In other words, we estimate the force output will be 228.3 N.

Standard Error of Estimate:

Recall, the b's are not independent.

Consequently the standard error will involve the covariances among b's as well as their variances.

$$\hat{\mu}_{Y.123} = x'_0 \times b$$

Note:
- Where x0' = [1, 2, 0.3, 0.8] and

$$var(\hat{\mu}_{Y.123}) = x'_0 \cdot (X'X)^{-1} \cdot x_0 \cdot \sigma^2$$

↑ ~~5×5~~ 4×4 → MS(E) FULL MODEL

Standard Error of Estimate

$$x_0'(X'X)^{-1}x_0 = \begin{vmatrix} 1 & 2 & 0.3 & 0.8 \end{vmatrix} \times \begin{vmatrix} 0.8840 & -0.3332 & -0.7556 & -0.0322 \\ -0.3332 & 0.3629 & -0.9578 & -0.1507 \\ -0.7556 & -0.9578 & 18.8388 & -2.8209 \\ -0.0322 & -0.1507 & -2.8209 & 1.3449 \end{vmatrix} \times \begin{vmatrix} 1 \\ 2 \\ 0.3 \\ 0.8 \end{vmatrix} = 0.06869$$

Therefore: $SE = \sqrt{Var(\mu_{123}) \cdot \sigma^2} = \sqrt{0.06869 \cdot 1448.07} = 9.973$

Hence, a 95% confidence interval can be constructed for the true contractility with Cl^-=2, PO_4=0.3, and K^+=0.8 as :

228.3 ± (2.074)(9.973) = (207.6, 249.0) (Student's-t with df = 22)

An Introduction To Statistical Design: Terminology

ANOVA = analysis of variance

rmANOVA = repeated measures ANOVA

CRD = completely randomized design

RCBD = randomized complete block design

ANCOVA = analysis of covariance

SPD = split plot design

FACTORIAL designs

MANOVA = multivariate analysis of variance

etc. etc. etc.....

ANOVA Part 1: One way

testing to see whether many means come from the same population

Goal

- Determine likely values of measure if samples in each group are from the same Population
- Develop a measure for the difference between experimental groups based on the means and using the estimate of variability for scaling of the difference

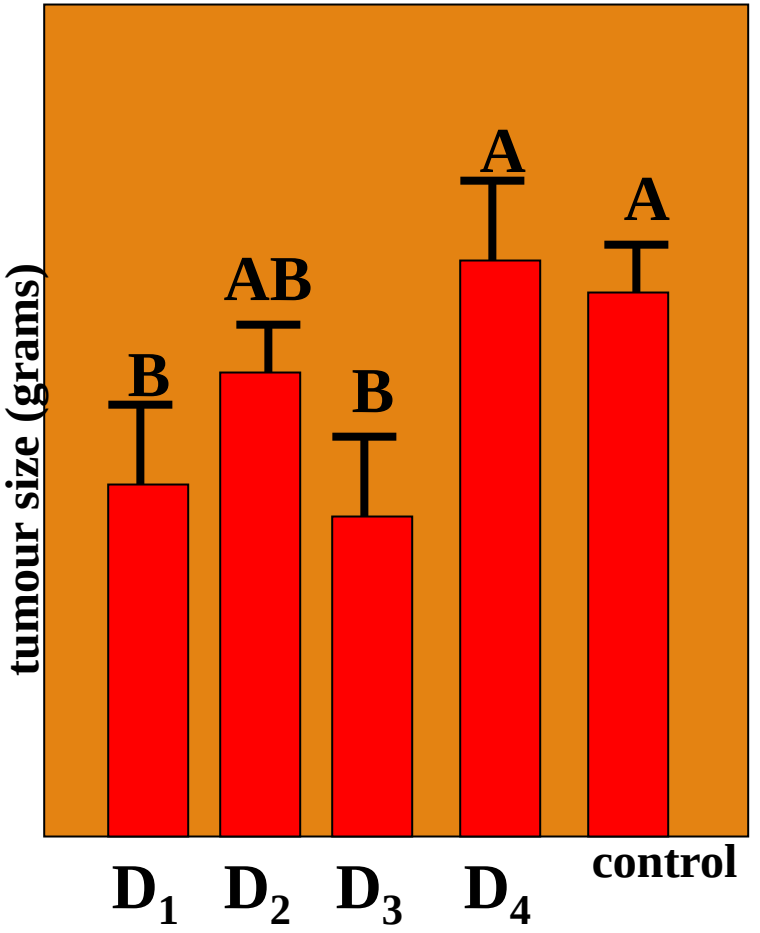
Example:

- 4 new anticancer drugs compared against a control

Null Hypothesis: all treatments the same

Alternative: atleast one is different.

- letters denote treatments that are significantly different from one another.



Use F test for Multiple Groups

Is there a difference between m groups of n samples each? (also sometimes i and j, or t and r).

To verify if all groups are from the same population one guesses that they are actually identical and validates or invalidates the statement.

If the hypothesis is true:

- The average variance of the individual groups should be smaller or equal to the variance of a given population.

Completely Randomized Design (CRD)

- treatments are randomly assigned, completely at random to the experimental units, which are assumed to be homogeneous.
- Model 1: Fixed Effects Case (examine effects of treatments)
- Model 2: Random Effects Case (identify sources of variation)

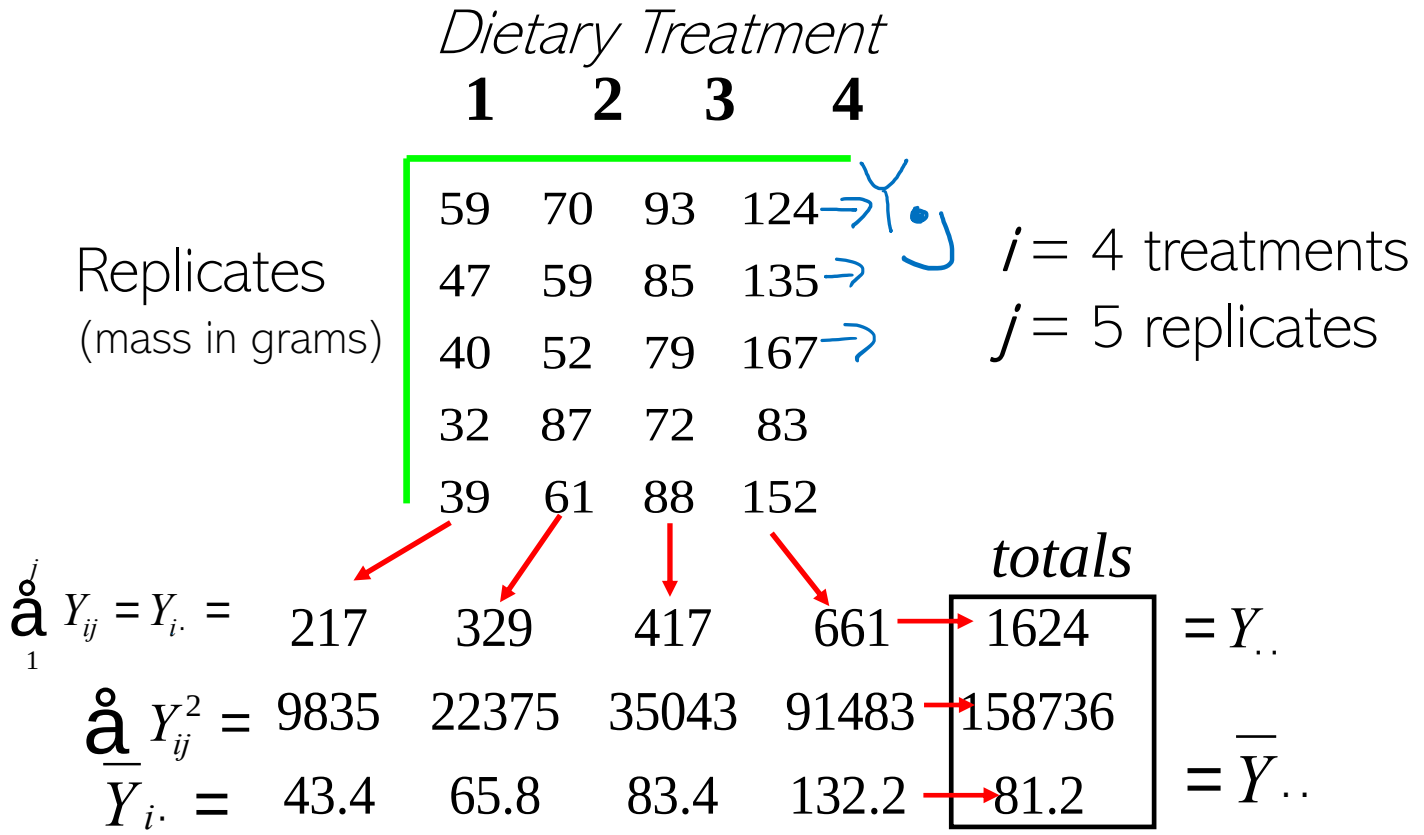
Treatments

Replicates	T_1	T_2	T_3	T_4	T_5
	1.1	3.2	5.6	8.2	8.3
	1.7	3.4	4.8	7.6	8.0
	1.5	3.0	5.1	7.9	8.4

matrix notation goes from larger to smaller groups e.g. block, treatment, replicate, etc.

Model I: Fixed Model

Example: Want to examine the effect of 4 diets on the growth of rainbow trout fry.



Model I: Fixed Model

Compute Sums of Squares:

$$Total(SS) = \sum Y_{ij}^2 - \frac{(Y_{..})^2}{i \times j} = 158736 - \boxed{\frac{1624^2}{4 \cdot 5}} = 26867.2$$

SS (Treatments) =

$$SS(T) = \frac{\sum (Y_{i\cdot})^2}{j} - \frac{(Y_{..})^2}{i \times j} = \frac{217^2 + 329^2 + 417^2 + 661^2}{5} - 21359.2$$

SS(error) = SS(E) = Total(SS) - SS(T) $\approx$ 5000

Model I: Fixed Model

Source	df	SS	MS	F_c	$F_{i,j,\alpha}$
treatment	3	21359.2	7119.73	20.68	3.24
error	16	5508.0	344.25	--------	--------
TOTAL	19	26867.2	--------	--------	--------

$df_{treatment}$ = (# treatments - 1) = (i - 1)
df_{TOTAL} = (i x j) - 1
df_{error} = df_{TOTAL} - $df_{treatment}$

MS = SS / df
F_C = T(MS) / E(MS)

$F_{i,j,\alpha}$ = from table
e.g. $F_{3,16,0.05}$ = 3.24

Model I: Fixed Model

The Bottom line…..

Since FC > Ftable, reject H0.

Therefore:

- there is a difference between one or more treatments
- Following ANOVA we would need to explore the differences (more on this soon…)

Model II: Random Effects Model

- not interested in specific treatments, but rather on sources of variation.

e.g. In order to study the sources of variation in synthesis of protein gumbycin, a sample of 5 cell cultures was selected at random from an incubator by a chemical engineer. A total of 4 western blots for the protein were made on each of the 5 randomly selected cultures.

protein content ($ng/10^6$ cells)

1	2	3	4	5
85	62	46	67	54
81	67	52	57	72
83	61	55	65	68
76	58	41	54	45

= culture number

$i = 5$

$j = 4$

$Y_{\cdot\cdot} = 1249$

$\sum Y_{ij}^2 = 81023$

Model II: Random Effects Model

Null Hypothesis:

- there is no variation between cultures i.e. HO: $\sigma t^2 = 0$

Calculation of ANOVA table is exactly the same:

Source	df	SS	MS	F_c	$F_{i,j,\alpha}$
Between	4	2233.7	558.425	10.61	3.06
within	15	789.25	52.617	--------	--------
TOTAL	19	3022.95	--------	--------	--------

Model I vs Model II?

What are we estimating?

- the MS(T) is an estimate of :
- Model 1: Fixed Effects Case (examine effects of treatments)
- Model 2: Random Effects Case (identify sources of variation)

	Model I	Model II
treatment (between)	$s^2 + \frac{j}{i-1}\sum t_i^2$	$s^2 + js_t^2$
error (within)	s^2	s^2

Model I vs Model II?

- in the protein case FC > Ftable.

- therefore σt^2 is significantly different from zero, and hence there is significant variability between cell cultures with respect to protein gumbycin synthesis.

- can σt^2 be estimated (i.e. variance between cultures) ? YES :

$$s_t^2 = \boldsymbol{s}_t^2 = \frac{MS(T) - MS(E)}{j} = \frac{558.425 - 52.617}{4}$$

$$= 126.45$$

Model I vs Model II?

- in this protein experiment the ratio of σt^2 to σ^2 = 126.452 : 52.616 or about 2.4 to 1

What percentage of the total variation does the variation between (σt^2) cultures account for ?

$$\text{total var} = \sigma_t^2 + \sigma^2 = 126.452 + 52.616 = 179.066$$

$$\frac{100 \times 126.452}{179.066} = 70.6 = 71\%$$

Real life experiments

- what can be done when there is unequal replication ?

- for example, consider an experiment to assess anti-carcinogens.
  - Rats were pre-medicated with one of 5 anticarcinogens prior to being given a single dose of benzo[a]pyrene.
  - The next day you want to assess mass of feed eaten.
  - However, for whatever reason many of the rats have died !!

	1	2	3	4	5	= treatment
mass of feed eaten (g) for each rat	8.4	6.5	7.2	7.2	7.9	
	7.6	8.1	7.4	7.5	9.6	
	8.2	7.7	6.2	-	9.9	
	7.4	-	6.6	-	-	
	8.2	-	-	-	-	

Real Life Experiments

Source	df	SS	MS	F_c	$F_{i,j,\alpha}$
treatment	4	9.8208	2.4552	5.435	3.26
error	12	5.4203	0.4517	---------	---------
TOTAL	16	15.2412	---------	---------	---------

except….

$df_{TOTAL} = n - 1$ (i.e. total samples - 1)

$df_{error} = n - i$ (i.e. total samples - # treatments)

$df_{treatment} = i - 1$ (i.e. total treatments - 1)

	Model I	Model II
treatment BALANCED	$s^2 + \frac{j}{i-1}\sum t_i^2$	$s^2 + js_t^2$
error	s^2	s^2
treatment UNEQUAL	$s^2 + \frac{\sum j_i (t_i)^2}{i-1}$	$s^2 + \frac{n - (\sum j^2)/n}{i-1} s_t^2$
error	s^2	s^2

Multiple Comparisons: Post hoc Procedures

What happens if you reject H0 ?

- need to explore where the differences lie and their magnitudes.

Methods will vary in conservatism

Repeated t-tests can result in errors so we need other methods

Statistical Testing of Means

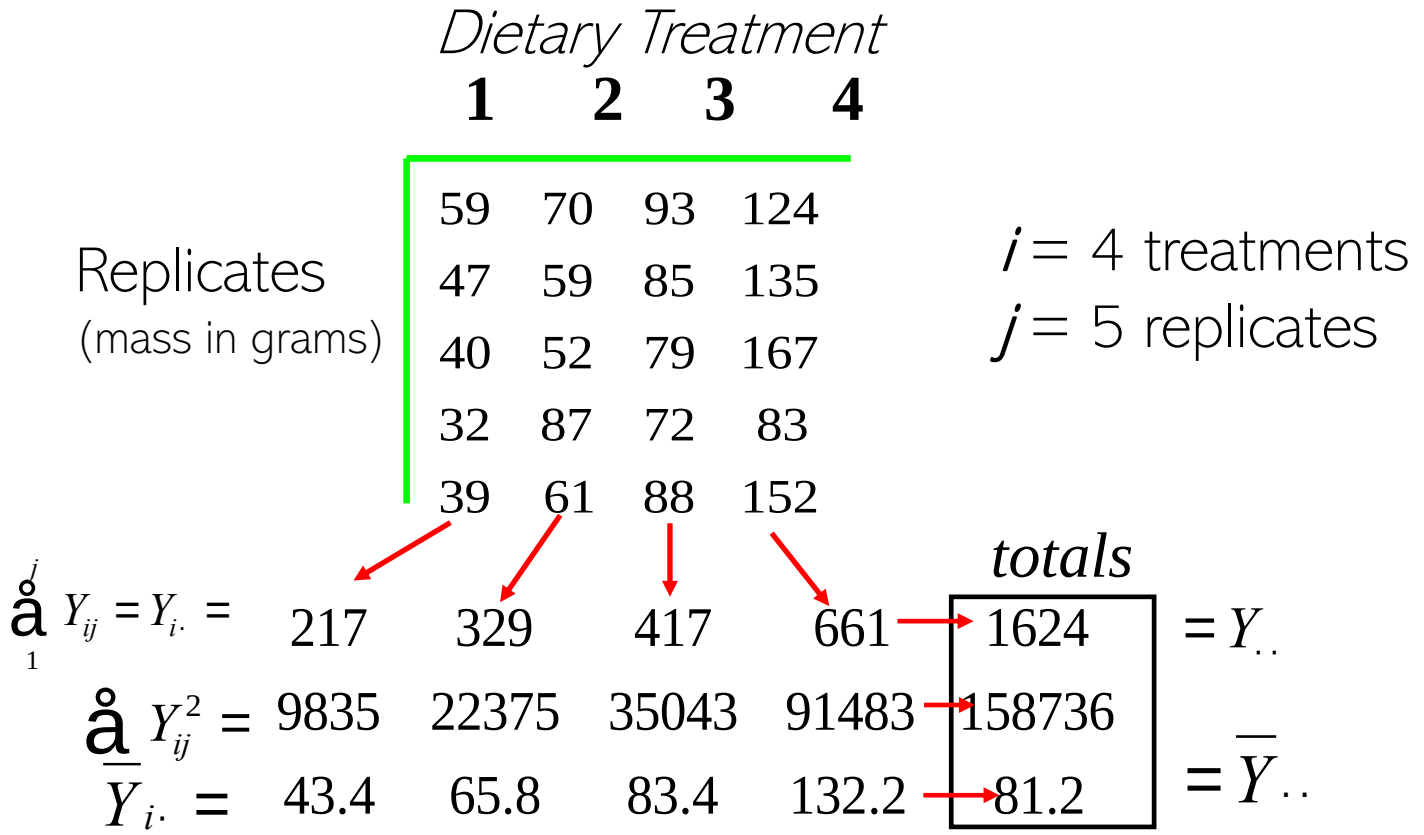
1) Student's t-test (2 means)
2) Least Significant Difference (lsd)
3) Duncan's new multiple range test
4) Contrast analysis
5) Scheffé Test

There are many others you can investigate on your own time.

e.g. SNK (Student-Neumann Keuls) test

Model I: Fixed Model

Example: Want to examine the effect of 4 diets on the growth of rainbow trout fry.



Model I: Fixed Model

Compute Sums of Squares:

$$Total(SS) = \sum Y_{ij}^2 - \frac{(Y_{..})^2}{i \times j} = 158736 - \frac{1624^2}{4 \times 5} = 26867.2$$

SS (Treatments) =

$$SS(T) = \frac{\sum (Y_{i.})^2}{j} - \frac{(Y_{..})^2}{i \times j} = \frac{217^2 + 329^2 + 417^2 + 661^2}{5} - \frac{1624^2}{4 \times 5} = 21359.2$$

SS(error) = SS(E) = Total(SS) - SS(T)

Model I: Fixed Model

Source	df	SS	MS	F_c	$F_{i,j,\alpha}$
treatment	3	21359.2	7119.73	20.68 >	3.24
error	16	5508.0	344.25	--------	--------
TOTAL	19	26867.2	--------	--------	--------

$df_{treatment}$ = (# treatments - 1) = (i - 1)
df_{TOTAL} = (i x j) - 1
df_{error} = df_{TOTAL} - $df_{treatment}$

MS = SS / df
F_C = T(MS) / E(MS)

$F_{i,j,\alpha}$ = from table
e.g. $F_{3,16,0.05}$ = 3.24

Least Significant Difference (lsd)

- examines differences between means

- ideally this is used for planned comparisons (i.e. specify in advance of getting the data.)

The equation for the standard error of the difference between 2 means is:

$$SE = \left(\overline{Y}_{i_a\cdot} - \overline{Y}_{i_b\cdot}\right) = \sqrt{E(MS) \times \frac{2}{j}} = \sqrt{344.25 \cdot \frac{2}{5}} = 11.73$$

$$lsd = t_{n,a/2} \times \sqrt{E(MS) \times \frac{2}{j}} = 2.120 \cdot 11.73 = 24.876$$

- where $\nu = df_{E(MS)} = i(j-1) = 4(5-1) = 16$

Least Significant Difference (lsd)

If lsd < (difference between 2 means), then reject H_0

- i.e. the means are significantly different (= sig. different)*
NOTE: 4/6 possible combinations were declared as significant different (i.e. 66%).

Difference Table: (24.877)

Diets	2 (65.8)	3 (83.4)	4 (132.2)
1 (43.4)	22.4	40.0*	88.8*
2 (65.8)	-----	17.6	66.4*
3 (83.4)	-----	-----	48.8*

Least Significant Difference (lsd)

Underscore Representation

- Underline pairs of means that are NOT significantly different.

1(43.4) 2(65.8) 3(83.4) 4(132.2)

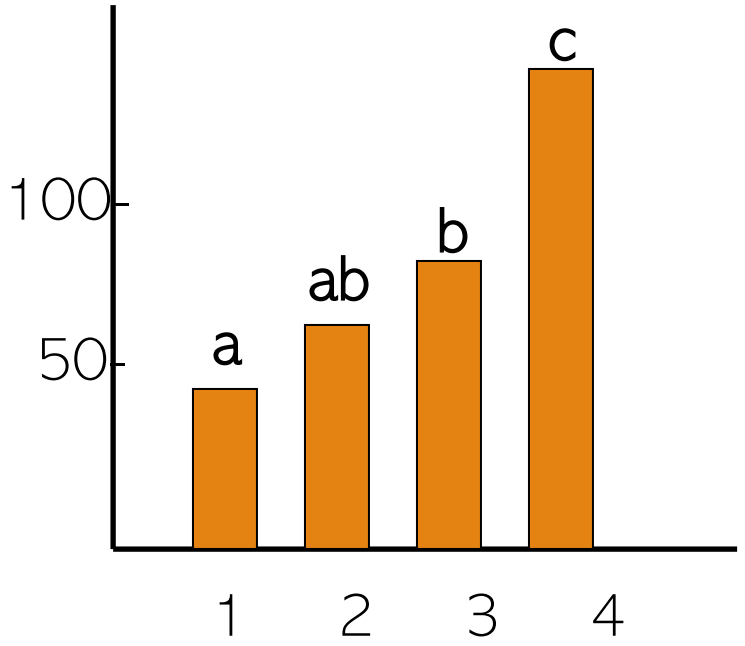
lsd results *:*

Diets 1 & 2, and diets 2&3 are not significantly different from each other.

However, diets 1&3, 1&4, 2&4, and 3&4 are significantly different.

Graphical Representation

same letter = *not* significantly different



Some Notes About the Results:

1 thumb
2 ☝
3 clap
4 smile

- Which diet would you use if you wished to raise the largest fish in the shortest period of time ?

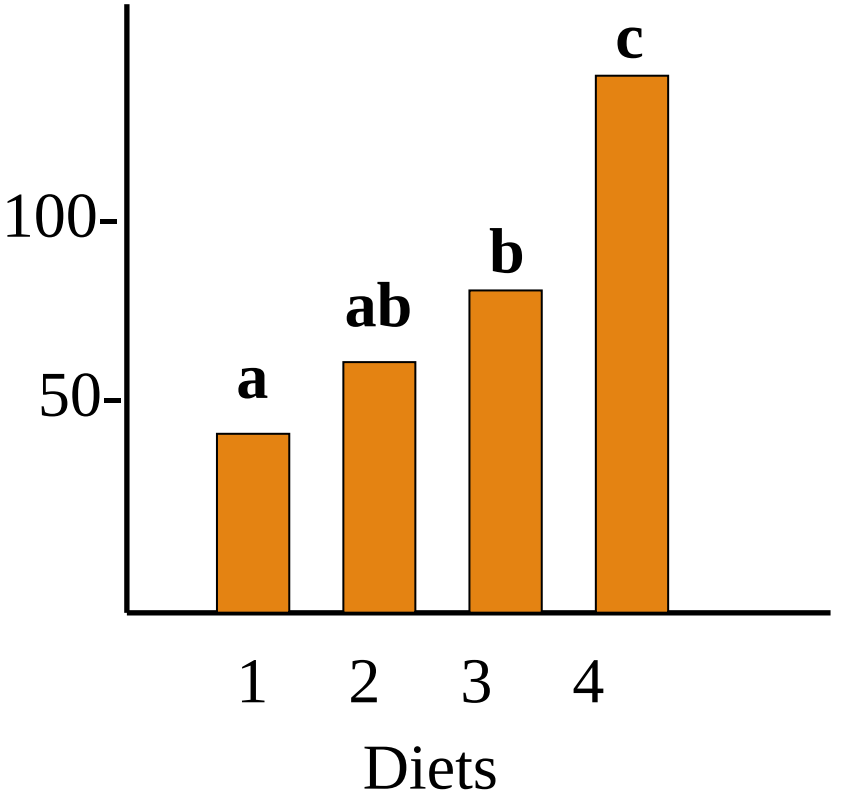
4

- If diet 3 cost much more than diet 2, which yields a greater weight gain per unit dollar of expenditure ?

2

- Suppose Diets 1 and 2 are equally priced. If they are the only 2 available diets which should be used ?

either.



Duncan's Multiple Range Method

- differs from lsd method which has a single 'least significant difference' with which to compare treatment effects.

- the Duncan method employs test criteria which vary in magnitude, depending on the number of means involved in the test.

First compute the standard error of a sample mean:

$$SE\left(\bar{Y}\right) = \sqrt{\frac{MS(E)}{j}} = \sqrt{\frac{344.25}{5}} = 8.29759$$

Duncan's Multiple Range Method

- consult table to determine the values of Studentized range: $q_a(p,fe)$
  - q_a = significance level (e.g. $\alpha = 0.05$)
  - p = number of means (i.e. treatments) being tested
  - f_e = the number of degrees of freedom of experimental error
    - (i.e. i(j-1) = 16, in this example)

Decision Rule:

- if the difference in means is greater than the calculated significant ranges parameter, Rp, then it is declared significant.

$$R_p = q_a(p, f_e) \times \sqrt{\frac{MS(E)}{j}}$$

Duncan's Multiple Range Method

SE(Y) = 8.29759

$$R_p = q_a(p, f_e) \times \sqrt{\frac{MS(E)}{j}}$$

p	2	3	4
$q_a(p,f_e)$	$q_{0.05}(2,16) = 3.00$	$q_{0.05}(3,16) = 3.15$	$q_{0.05}(4,16) = 3.23$
R_p	3.00(8.29759) = 24.893	3.15(8.29759) = 26.137	3.23(8.29759) = 26.801

Difference Table

Diets	2(65.8)	3(83.4)	4(132.2)
1(43.4)	22.4	40.0*	88.8*
2(65.8)	-----	17.6	66.4*
3(83.4)	-----	-----	48.8*

Duncan's Multiple Range Method

Notice, in this particular circumstance, Duncan's and lsd result in the same conclusions for the data.

- This can differ !!

If the difference in means is particularly close then the lsd method will result in significance more often (i.e. is less conservative than Duncan's).

Linear Combinations of Means (i.e. contrasts)

- suppose there are a number of ways to improve the length of time of musculoskeletal repair after injury:

X1 = blueberry and kale smoothies .

X2 = supplements of branched chain AAs + proline .

X3 = static magnetic field (B0)

X4 = pulsed magnetic field (B0)

X5 = Therapeutic Ultrasound (US)

Example of a linear combination (2 sample case):

$L_1 : \overline{X}_1 - \overline{X}_2$ - estimates $\mu_1 - \mu_2$ or the difference between dietary treatments

Linear Combinations of Means

Another possible linear combination:

$$L_2 : \overline{X}_3 - \overline{X}_4$$

- measures mean difference in time of healing for magnetic field (B_0) methods.

While:

$$L_3 : \frac{1}{2}\overline{X}_3 + \frac{1}{2}\overline{X}_4$$

- i.e. an estimate of ½(μ_3+μ_4)
- measures the average time to achieve full repair due to magnets.

Linear Combinations of Means

The Question:

- does US take less time, on average, than magnetic fields (ignoring any difference between pulsed vs. static B0)?

Can be written mathematically:

$$m_5 < \frac{1}{2}(m_3 + m_4)?$$

US B

Alternatively:

$$m_5 - \frac{1}{2}(m_3 + m_4) < 0?$$

To estimate this difference among mean values we would use:

$$L_4 : \overline{X}_5 - \frac{1}{2}(\overline{X}_3 + \overline{X}_4)$$

Then test whether L_4 was significantly different from 0.

Linear Combinations of Means

More specifically in this case we would be doing a 1-sided test, since we are only looking at whether there is evidence that μ_5 *is less* than the average of μ_3 and μ_4.

To compare diet with the average of 'engineering' methods we would ask whether the average repair times due to diet, estimated by:

$$\frac{1}{2}\overline{X}_1 + \frac{1}{2}\overline{X}_2$$

differs from the average times taken when using tech, estimated by:

$$\frac{1}{3}\overline{X}_3 + \frac{1}{3}\overline{X}_4 + \frac{1}{3}\overline{X}_5$$

Linear Combinations of Means

The difference between these 2 estimates:

$$L_5 : \frac{1}{2}\overline{X}_1 + \frac{1}{2}\overline{X}_2 - \frac{1}{3}\overline{X}_3 - \frac{1}{3}\overline{X}_4 - \frac{1}{3}\overline{X}_5$$

estimates:

$$\frac{1}{2}m_1 + \frac{1}{2}m_2 - \frac{1}{3}m_3 - \frac{1}{3}m_4 - \frac{1}{3}m_5$$

Linear Combinations of Means

To reiterate, here are the linear combinations again:

$$L_1 : \overline{X}_1 - \overline{X}_2$$

$$L_2 : \overline{X}_3 - \overline{X}_4$$

$$L_3 : \frac{1}{2}\overline{X}_3 + \frac{1}{2}\overline{X}_4$$

$$L_4 : \overline{X}_5 - \frac{1}{2}\left(\overline{X}_3 + \overline{X}_4\right)$$

$$L_5 : \frac{1}{2}\overline{X}_1 + \frac{1}{2}\overline{X}_2 - \frac{1}{3}\overline{X}_3 - \frac{1}{3}\overline{X}_4 - \frac{1}{3}\overline{X}_5$$

Linear Combinations of Means

Now here are the same linear contrasts in tabular form:

Means	$\overline{X}_1$	$\overline{X}_2$	$\overline{X}_3$	$\overline{X}_4$	$\overline{X}_5$	
Coefficients:	λ_1	λ_2	λ_3	λ_4	λ_5	**$\Sigma\lambda$**
Combination						
L_1	+1	-1	0	0	0	0
L_2	0	0	+1	-1	0	0
L_3	0	0	+1/2	+1/2	0	1
L_4	0	0	-1/2	-1/2	+1	0
L_5	+1/2	+1/2	-1/3	-1/3	-1/3	0

Linear Combinations of Means

Definition:

- Linear combinations with $\Sigma\lambda = 0$ are called contrasts.

- A sample contrast, denoted L, is an estimator of the population contrast.

- The Standard Error of this estimate is:

$$SE(L) = SE\left(\sum l_i \overline{X}_i\right) = \sqrt{s^2 \times \frac{\left(\sum l_i^2\right)}{n_i}}$$

Where s^2 is the MS(E) from the ANOVA table, $\Sigma\lambda_i^2$ is the coefficient sum of squares, and n_i is the number of samples in the i^{th} group.

Linear Combinations of Means

If, for example, each $n_i = 5$, then the standard errors for the different linear combinations are:

$$SE(L) = \sqrt{s^2 \cdot \frac{\left(\sum \lambda_i^2\right)}{n_i}}$$

Combination	$\Sigma\lambda^2$	SE(L)
L1	2.000	$s \times \sqrt{2.000/5}$
L2	2.000	$s \times \sqrt{2.000/5}$
L3	0.500	$s \times \sqrt{0.500/5}$
L4	1.500	$s \times \sqrt{1.500/5}$
L5	0.833	$s \times \sqrt{0.833/5}$

Hypothesis Testing Using Contrasts:

Consider L4, a comparison of ultrasound with the average of magnetic field induced repair times:

$$L_4 : \overline{X}_5 - \frac{1}{2}\left(\overline{X}_3 + \overline{X}_4\right)$$

The $\Sigma\lambda^2 = 1.500$. If all 3 means are based on samples of 5 times each, then:

$$SE(L) = \sqrt{s^2 \times \frac{\left(\text{å } l_i^2\right)}{n_i}} = S\sqrt{\frac{1.5}{5}} = S \cdot 0.5477$$

Hypothesis Testing Using Contrasts:

If s = pooled estimate of the population standard deviation (σ), based on 5 pooled variances s_1^2, s_2^2, s_3^2, s_4^2, s_5^2, each with 4 df

- then s will have 5 x 4 (i.e. 20) degrees of freedom.

Using a Student's t-table: $t_{\alpha=0.05,20} = 1.725$; $t_{\alpha=0.025,20} = 2.086$

The null hypothesis to be tested: $H_0 : \mu_5 - \frac{1}{2}(\mu_3 + \mu_4) = 0$

Which can also be re-written as: $H_0 : \mu_5 = \frac{1}{2}(\mu_3 + \mu_4)$

Hypothesis Testing Using Contrasts:

Test H_0 against one of the alternative hypothesis:

$$H_A : \mu_5 \neq \tfrac{1}{2}(\mu_3 + \mu_4)$$ (2-sided alternative)

$$H_A : \mu_5 > \tfrac{1}{2}(\mu_3 + \mu_4)$$ (1-sided alternative)

$$H_A : \mu_5 < \tfrac{1}{2}(\mu_3 + \mu_4)$$ (another 1-sided alternative)

Hypothesis Testing Using Contrasts:

A t-test can be used here:

$$t = \frac{\text{Estimated Value} - \text{Hypothesized True Value}}{\text{Standard Error of Estimated Value}} \qquad t = \frac{\left(\overline{X}_5 - \frac{1}{2}\left(\overline{X}_3 + \overline{X}_4\right)\right) - 0}{0.5477 \times s}$$

Depending on whether we have chosen a 1 or 2 sided test we would reject H0 in favour of the selected alternative hypothesis:

- if $|t| > 2.086$ (i.e. $t>2.086$ or $t<2.086$) for the 2 sided test.
- if $t > +1.725$ for the 1 sided test for the alternative which predicted $\mu 5 > 1/2(\mu 3+\mu 4)$.
- if $t < -1.725$ for the 1 sided test for the alternative which predicted $\mu 5 < 1/2(\mu 3+\mu 4)$.

Hypothesis Testing Using Contrasts:

The 95% confidence interval for a linear combination:

(Estimated value – $t_{\alpha,df}$, x stderr of estimated value) < Linear combination of true values < (Estimated value + $t_{\alpha,df}$, x stderr of estimated value)

In this case:

$$\left(\overline{X}_5 - \frac{1}{2}\left(\overline{X}_3 + \overline{X}_4\right)\right) - \left(2.086 \cdot 0.5477s\right) \le \left(m_5 - \frac{1}{2}\left(m_3 + m_4\right)\right) \le \left(\overline{X}_5 - \frac{1}{2}\left(\overline{X}_3 + \overline{X}_4\right)\right) + \left(2.086 \cdot 0.5477s\right)$$

Orthogonal Contrasts:

Means	$\overline{X}_1$	$\overline{X}_2$	$\overline{X}_3$	$\overline{X}_4$	$\overline{X}_5$	
Coefficients:	λ_1	λ_2	λ_3	λ_4	λ_5	$\Sigma\lambda$
Combination						
L_1	+1	-1	0	0	0	0
L_2	0	0	+1	-1	0	0
L_4	0	0	-1/2	-1/2	+1	0
L_5	+1/2	+1/2	-1/3	-1/3	-1/3	0

- Notice each contrast looks at a different characteristic of the data
- Not all contrasts look at genuinely different characteristics !!

Hypothesis Testing Using Contrasts:

For example, the contrasts:

$$\overline{X}_1 - \overline{X}_2 \qquad \overline{X}_1 - \overline{X}_3 \qquad \overline{X}_2 - \overline{X}_3$$

compare 1 with 2, 1 with 3, and 2 with 3. The third contrast, however, really tells us nothing we couldn't have figured out with the other two, since:

$$\overline{X}_2 - \overline{X}_3 = \overline{X}_1 - \overline{X}_3 - \left(\overline{X}_1 - \overline{X}_2\right)$$

A way to ensure that contrasts are looking at completely different aspects of the data is to require that all contrasts be **orthogonal**.

Hypothesis Testing Using Contrasts:

The numerical verification that 2 contrasts are orthogonal is that the sum of the products of their corresponding coefficients is zero.

For example, L_1 and L_2 are orthogonal:

L1	+1	-1	0	0	0	
L2	0	0	+1	-1	0	
Products	0	0	0	0	0	Sum = 0

Also, L4 and L5 are orthogonal:

L4	0	0	-1/2	-1/2	1	
L5	+1/2	+1/2	-1/3	-1/3	-1/3	
Products	0	0	1/6	1/6	-1/3	Sum = 0

Hypothesis Testing Using Contrasts:

However, X_1 - X_2 and X_1 - X_3 are not orthogonal:

X1-X2	+1	-1	0	0	0	
X1-X3	+1	0	-1	0	0	
Products	+1	0	0	0	0	Sum = +1

i.e. if 2 contrasts, with q number of coefficients:

$$\sum_{i=1}^{q} l_{A_i} l_{B_i} = 0$$

Then contrasts A and B are orthogonal.

Non-orthogonal contrasts do not provide any extra information !!

Hypothesis Testing Using Contrasts:

It's a bit tricky at first to come up with orthogonal contrasts. The best thing to do is think up contrasts which address specific and distinct questions- Then check for orthogonality.

e.g. How do 1 and 2 compare ?

1 0 0 1

KS β_1

How do 3 and 4 compare ?

How do 3 and 4 compare with 5 ?

How do 1 and 2 together compare with 3, 4, and 5 together ?

→ These questions led to the orthogonal contrasts L1, L2, L3, and L5

Hypothesis Testing Using Contrasts:

NOTES:
- if we have *i* treatments then there exactly *i-1* possible orthogonal contrasts.
- the *i-1* is exactly equal to the **df** for measuring variability among the treatment means.
- These orthogonal contrasts correspond to a decomposing of this variability !!
- each contrast has **1 df** associated with it.

Sum of Squares of Contrasts:

- a measure of a size of a contrast is it's sum of squares i.e. SS(L)

$$SS(L) = \frac{n \times (\text{estimated value of } L)^2}{\Sigma \lambda_i^2}$$

- the estimated value of L is calculated usually using mean values.

- Totals can also be used (and are equivalent)

$$L_1^{Totalbased} : \overline{X}_{1\cdot} + \overline{X}_{2\cdot}$$

Where $\overline{X}_{1\cdot}$ is the total of all observations taken for the first treatment

$$SS(L) = \frac{(\text{estimated value of } L^{\text{Totalbased}})^2}{n \times \Sigma \lambda_i^2}$$

Back to the fish example

We'd like to test the following:

for L_1 $\quad H_0 : \mu_4 - \mu_2$

for L_2 $\quad H_0 : \mu_4 - 1/2(\mu_2 + \mu_3)$

Note- try it yourself !! Check that L_1 and L_2 are orthogonal !!

Fish Example Contrasts

Contrast L_1: $\hat{L}_1 = \overline{X}_4 - \overline{X}_2 = 132.2 - 65.8 = 66.4$

diet 4 mean diet 2

The coefficients are: $\lambda_1 = 0; \lambda_2 = -1; \lambda_3 = 0; \lambda_4 = +1$

$1^2 + 1^2$

The sum of squares of $\Sigma\lambda$

$$SE(L_1) = \sqrt{s^2 \times \frac{(\text{å } l^2)}{n_i}} = \sqrt{344.25 \cdot \frac{2}{5}} = 11.73 \qquad t = \frac{\hat{L}_1 - 0}{11.73} = \frac{66.4}{11.73} = 5.66$$

From t-table, with df = 16, t = 2.120 (2 tailed)

$t_{table} = 2.12$

Since $t_{calc} > t_{table}$ reject H_0 i.e. diets 4 and 2 are significantly different

Fish Example Contrasts

Similarly, Contrast L_2:

$$\hat{L}_2 = \overline{X}_4 - \frac{1}{2}\left(\overline{X}_2 - \overline{X}_3\right) = 132.2 - \frac{1}{2}\left(65.8 - 83.4\right)$$

The coefficients are: $\lambda_1 = 0$; $\lambda_2 = -1/2$; $\lambda_3 = -1/2$; $\lambda_4 = +1$

The sum of squares of $\Sigma\lambda$ (i.e. $\Sigma\lambda^2$) = 1.5

$$SE(L_1) = \sqrt{s^2 \times \frac{\left(\text{å } l^2\right)}{n_i}} = \sqrt{344.25 \times \frac{1.5}{5}} = 10.16$$

t_{calc}=6.535

More Experimental Designs

- CRD with subsampling
- Randomized Complete Block Design (RCBD)
- Analysis of Covariance (ANCOVA)

Subsampling

- the term used to describe the situation in which more than one observation is taken per experimental unit.
- such observations are made on sampling units.
- when subsampling is performed, the linear model and the ANOVA must be expanded to take into account the variation among samples (the source of sampling error)

CRD with Subsampling: Model I (Treatment Effects)

Consider the following experiment:

- A new drug phelphodyne-HCl, was thought to enhance liver cyt-P450 in people with late-stage cirrhosis

- when given with alcohol the effect was thought to be diminished.

- a total of 6 randomly assigned patients were used to test this drug:

    - 2 controls (no drug)

    - 2 phelphodyne-HCl

    - 2 phelphodyne-HCl + ethanol

- After an appropriate time and dose, 4 liver biopsies were taken under ultrasound/MRI (co-registered) guidance from each patient and cyt-P450 was measured in piece.


sentinelle
MEDICAL

(Source: http://www.skills.uct.ac.za/activities.htm)

Experimental Objectives:

1) To determine if there was a significant difference among the three treatments
2) To estimate the 2 variance components:
   - variation among measurements within patients
     (i.e. sampling error).
   - variation among patients within a given treatment
     (i.e. experimental error).

CRD with Subsampling: Model I

The data ($Y_{i,j,k}$) from 6 randomly chosen/assigned patients and 4 randomly selected pieces of liver from each.

	Control		Phelphodyne		Phelphodyne +EtOH	
Patients	**1**	**2**	**1**	**2**	**1**	**2**
cyt-P450 readings	131	148	157	152	124	140
	130	143	153	155	125	138
	125	150	154	162	136	138
	131	150	149	161	130	139
	517	**591**	**613**	**630**	**515**	**555**
	1108		**1243**		**1070**	

$Y_{ij\bullet}$

$Y_{i\bullet\bullet}$

$\Sigma Y_{ij\bullet}^2 = 1962489$

$\Sigma Y_{i\bullet\bullet}^2 = 3917613$

$Y_{\bullet\bullet\bullet} = 3421$

CRD with Subsampling: Model I

Notes:

- an experimental unit here is a patient
- a sampling unit is a piece of patient liver

Degrees of Freedom

- i = 3 (treatments)
- j = 2 (replicates)
- k = 4 (subsamples)
- Total df = (3x2x4)-1 = 24-1 = 23

i j k

There are 2 hypothesis that can be tested:

- 1) H0:$\tau i = 0$, for all i vs. HA:$\tau i \neq 0$ (i.e. treatment effect)
- 2) H0:$\sigma_e^2 = 0$ vs. HA:$\sigma_e^2 \neq 0$ (i.e. error variance)

Compute Sums of Squares:

$$Total(SS) = \sum Y_{ijk}^2 - \frac{(Y_{...})^2}{i \times j \times k} = 490875 - \frac{3421^2}{24} = 3239.96$$

$$\text{SS(Treatments)} = SS(T) = \frac{\sum (Y_{i..})^2}{j \times k} - \frac{(Y_{...})^2}{i \times j \times k} = 2066.5833$$

$$\text{SS(Subsamples)} = SS(SS) = \frac{\sum Y_{ij.}^2}{k} - \frac{(Y_{...})^2}{i \times j \times k} = 2987.2083$$

SS(sampling error) = SS(SE) = Total(SS) - SS(SS) = 252.75

SS(experimental error) = SS(EE) = SS(SS) - SS(T) = 920.625

ANOVA Table

Source	df	SS	MS	Fc	Fi,j,α
subsamples (AMONG PATIENTS)	5	2987.2083	-----	-----	-----
Treatment	2	2066.5833	1033.2417	3.367	9.55
Exp. Error	3	920.625	306.875	21.85	3.16
Samp. Error	18	252.75	14.042	-----	-----
TOTAL	23	3239.9583	-----	-----	-----

*Note treatment df + EE df = subsamp df

$df_{TOTAL} = (i \cdot j \cdot k) - 1 = (3x2x4)-1 = 23$

$df_{SS} = (i \cdot j) - 1 = (3x2) - 1 = 5$

$df_{SE} = ij(k-1) = 3x2(4-1) = 18$

$df_{T} = i - 1 = 3 - 1 = 2$

$df_{EE} = i(j-1) = 3(2-1) = 3$

ANOVA Analysis

F_C (among treatments) = 1033.2417 / 306.875 = 3.367

Since $F_{0.05,2,3} = 9.55$

we Accept $H_0: \tau_i = 0$. Therefore there is NO effect of treatment.

F_C (experimental error) = 306.875 / 14.042 = 21.854

Since $F_{0.05,3,18} = 3.16$

We REJECT $H_0: \sigma_e^2 = 0$. Therefore there is A significant source of error between patients.

REMEMBER CRD what MS is estimating !!

	Model I	Model II
treatment BALANCED	$\sigma^2 + \frac{j}{i-1}\sum t_i^2$	$\sigma^2 + j\sigma_t^2$
error	σ^2	σ^2
treatment UNEQUAL	$\sigma^2 + \frac{\sum j_i (t_i)^2}{i-1}$	$\sigma^2 + \frac{n - \left(\sum j^2\right)/n}{i-1}\sigma_t^2$
error	σ^2	σ^2

CRD with Subampling

For Fixed (Type I) Models, what does the mean (MS) estimate?

ANOVA Table

$$E[MS(T)] = s^2 + ks_e^2 + \frac{j \times k}{i-1}\sum t_i^2$$

$$E[MS(EE)] = s^2 + ks_e^2$$

$$E[MS(SE)] = s^2$$

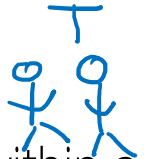
Estimation of variance components:

$$\hat{s}_e^2 = s_e^2 = (306.875 - 14.042)/4 = 73.208$$

$$\hat{s}^2 = s^2 = 14.042 \qquad s^2 + s_e^2 = 87.250$$

Where is variance?

σ^2 = variation among cyt-P_{450} within patients
s_ε^2 = variation in patients within treatments
$\sigma^2 + s_\varepsilon^2$ = total variance (within and among patients within a treatment)



The variation within patients (i.e. among liver samples) accounts for 100 x (14.042 / 87.250) = 16.1% of the estimated total variance.

The variation among patients of a given treatment accounts for 100 x (73.208 / 87.250) = 83.9% of the estimated total variance.

i.e. There is approximately 5.1x as much variation among patients as there is within patients.

CRB SS NOTES:

- The critical F Value (FC) value for treatments was noted to be quite small. This is due to the large denominator

- The large MS(EE) indicates high variability among the patients for any one preparation.

- Although no differences in treatments were detected, any differences may actually have been hidden by the variation among experimental units.

- Future experiments should take into account heterogeneity of experimental units.

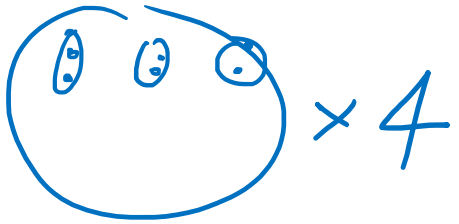
Also could performing a block design in which 2 blocks of 3 patients are examined too.

CRD with Subsampling (Model II- Random Effects)

Consider this experiment:

- an experiment was conducted to assess the precision to which EEG could be measured for application in a brain-computer interface.

  - 4 people were randomly selected
  - 3 brain regions were randomly chosen from each person
  - 2 samples taken per region for signal power analysis



Primary Objectives:

1) Estimate EEG signal power
2) Find the EEG signal power standard error
3) Isolate and estimate the sources of variation.

CRD with SS (Model II-Random Effects)

From this analysis:
- recommendation can be made with regard to optimizing future brain EEG sampling strategies.
- This would allow the researcher to reduce the standard error of their estimate in future studies.

Person	1			2			3			4		
Brian Region	1	2	3	1	2	3	1	2	3	1	2	3
Subsample	3.48	3.72	3.03	2.66	2.07	2.39	2.97	3.94	2.75	3.98	4.27	3.51
	3.29	3.68	3	2.64	2.12	2.39	2.86	3.64	2.75	4.07	4.32	3.51
	6.77	**7.4**	**6.03**	**5.3**	**4.19**	**4.78**	**5.83**	**7.58**	**5.5**	**8.05**	**8.59**	**7.02**
		20.2			**14.27**			**18.91**			**23.66**	

$Y_{ij\cdot}$

$Y_{i\cdot\cdot}$

$Y_{\cdot\cdot\cdot} = 77.09$

i = 4 pat.
j = 3 region
k = 2 sub samp
total df = (ijk) - 1 = 23

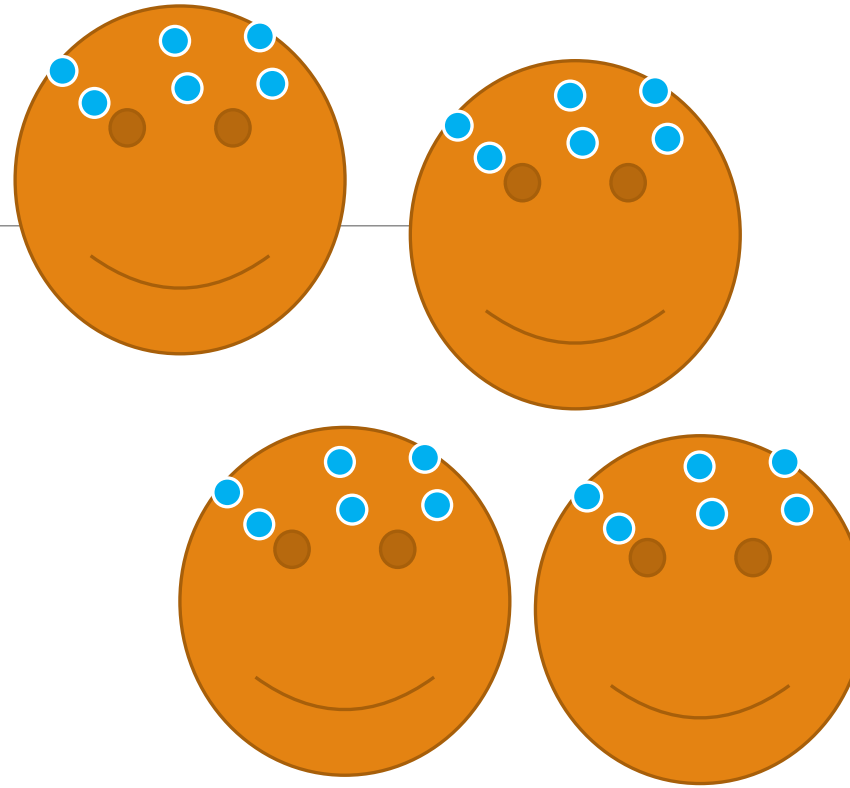
CRD II SS Model Equation:

$$Y_{ijk} = m + t_i + e_{ij} + d_{ijk}$$

Treatment Effect

Experimental Error

Sampling Error



Necessary assumptions:

- τ_i are $N(0, \sigma_\tau^2)$, where σ_τ^2 is the variation between people
- ε_{ij} are $N(0, \sigma_\varepsilon^2)$, where σ_ε^2 is the variation between brain areas
- δ_{ijk} are $N(0, \sigma_\delta^2)$, and σ_δ^2 is variation among samples between brain areas.

CRD II SS ANOVA

Source	df	SS	MS	Fc	Fi,j,α
subsamples (AMONG AREAS)	11	10.19055	-----	-----	-----
Treatment (AMONG PEOPLE)	3	7.56035	2.5201167	7.665 1τ	4.07
Exp. Error	8	2.6302	0.320775	49.41 2ε	2.85
Samp. Error	12	0.07985	0.0066542	-----	-----
TOTAL	23	10.2704	-----	-----	-----

*Note treatment df + EE df = subsamp df

Testable hypothesis

H_0: $\sigma_\tau^2 = 0$ (i.e. no significant difference between people)

H_0: $\sigma_\varepsilon^2 = 0$ (i.e. no significant difference between brain areas)

CRD II SS ANOVA Results:

- variance component due to brain areas within people is significantly different from 0?

49.41 | 72.85

reject H_0 EEG power varies between areas.

H_0 ~~$\sigma_\varepsilon^2 = 0$~~

- variance component from person to person is significantly different from 0?

7.665 | 4.07

reject H_0 EEG power varies b/w people

H

In CRD II with Subsampling

Random Effects (Type II) Models: what does the mean (MS) estimate?

people

$$E[MS(T)] = s_d^2 + k s_e^2 + jk s_t^2$$

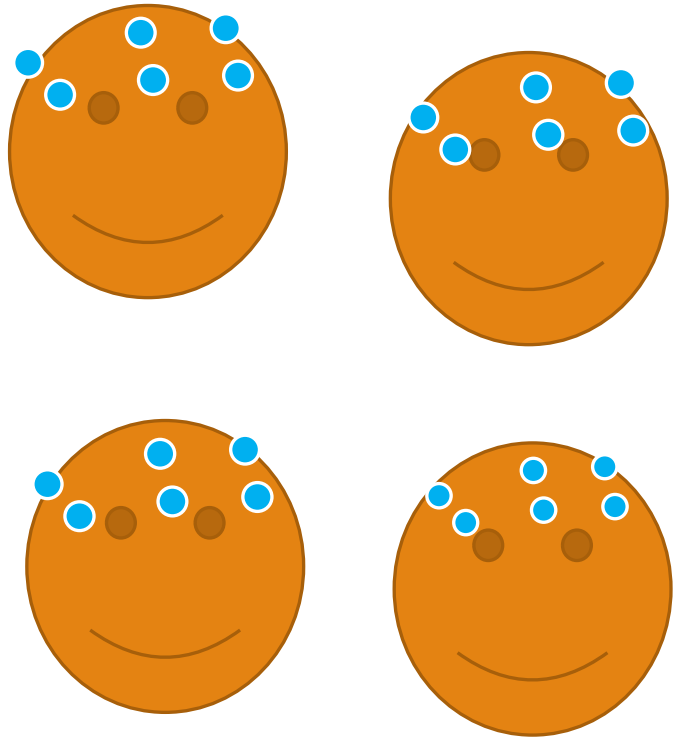
$$E[MS(EE)] = s_d^2 + k s_e^2$$

$$E[MS(SE)] = s_d^2$$

people.

among areas

within areas



Estimates of the 3 variance components:

within areas

$$\hat{s}_d^2 = s^2 = MS(SE) = 0.0066$$

among areas

$$\hat{s}_e^2 = s_e^2 = \frac{MS(EE) - MS(SE)}{k} = \frac{0.328775 - 0.0066542}{2} = 0.1611$$

among ppl

$$\hat{s}_\tau^2 = s_\tau^2 = \frac{MS(T) - MS(EE)}{jk} = \frac{2.5201167 - 0.320775}{6} = 0.365232$$

$$Total = s^2 + s_e^2 + s_t^2 = 0.0066 + 0.1611 + 0.3652 = 0.5329$$

CRD II with Subsampling Conclusions

Variation within brain areas represents = 100*(0.0066 / 0.5329) = 1.24%

Variation among brain areas represents = 100*(0.1611 / 0.5329) = 30.22%

Variation among people represents = 100*(0.3652 / 0.5329) = 68.54%

Estimate the total mean EEG power = 3.212 ± 0.01665

Standard Error = $SE = \sqrt{\frac{MS(EE)}{n}} = \sqrt{\frac{0.0066542}{24}} = 0.01665$

more ppl

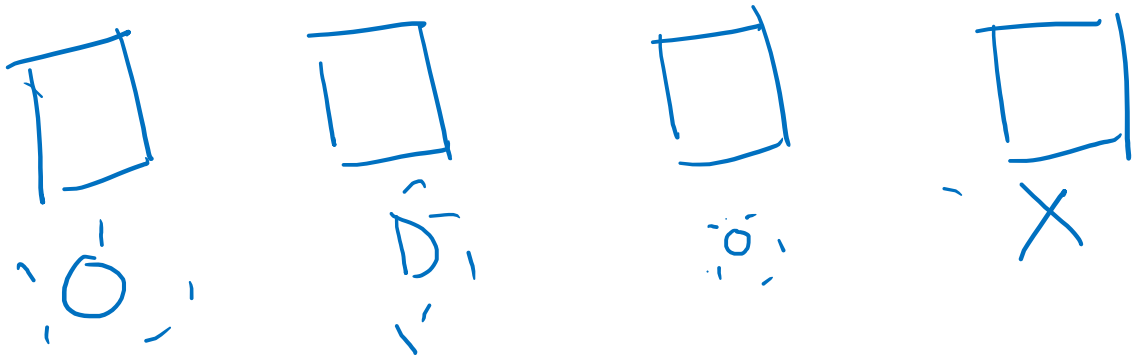
Randomized Complete Block Design

- removes source of variation

- if it is known in advance that the experimental units are NOT homogeneous then the CRD is no longer appropriate.

- the RCBD is used to remove sources of heterogeneity among experimental units.

- here experimental units are allocated to blocks such that those assigned to the same block should be similar in response to their treatment (i.e. homogeneous as possible).

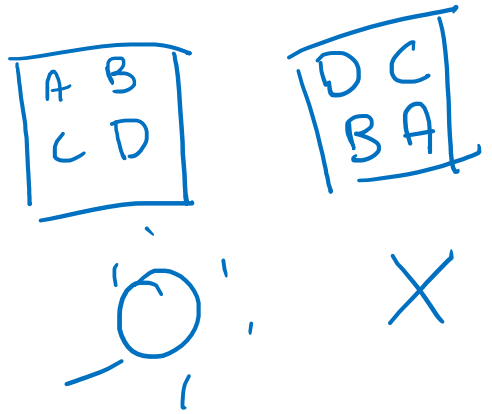


RCBD

- treatments are then allocated to the experimental units of each block, by a separate randomization that is carried out within each block.

Some blocking factors could include:

- DATE of experiment
- cage battery for animal housing
- plot of land
- incubator or oxygen chamber
- individual hospital



RCBD Fruit Fly Example

Consider the following genetics experiment.

- 5 people in the lab were all assigned to a project where they were to assess protein levels of CuZnSOD (superoxide dismutase) in Drosophila melanogaster (fruit fly) that had been transfected with human CuZnSOD. The 'boss' wanted to know which cross (i.e. homozygote (hom), heterozygote (het), or wild type (wld)) had higher CuZnSOD.

	PERSON (j)					$\overline{Y}$	$Y_{i\cdot}$	$\Sigma Y_{i\cdot}^2$
type(i)	1	2	3	4	5			
HOM	39.87	38.16	42.08	40.84	39.44	40.078	200.39	8039.948
HET	42.51	40.82	44.17	43.46	42.99	42.79	213.95	9161.273
WLD	45.76	43.14	46.29	44.95	45.89	45.206	226.03	10224.2
	128.14	122.12	132.54	129.25	128.32		640.37	27425.42
	5490.695	4983.518	5864.479	5577.18	5509.546			

$Y_{\cdot j}$

$\Sigma Y_{\cdot j}^2$

$Y_{\cdot\cdot}$

ΣY_{ij}^2

RCBD Fruit Fly Example

For this RCBD design:

A block is an individual person

A treatment is a genotype

An experimental unit is a fruit fly

A model equation for the RCBD design:

$$Y_{ij} = \mu + \tau_i + B_j + \varepsilon_{ij}$$

Null hypothesis: H_0: all $\tau_i = 0$
Alternative: H_A: all $\tau_i \neq 0$

Null hypothesis: H_0: all $B_j = 0$
Alternative: H_A: all $B_j \neq 0$

RCBD Fruit Fly Example

$df_{block} = j - 1 = 5 - 1 = 4$
$df_{treat} = i - 1 = 3 - 1 = 2$
$df_{total} = (i\,j) - 1 = 3x5-1 = 14$
$df_{error} = total - (block + treat) = (i-1)(j-1) = 8$
$i = 3$ (treatments); $j = 5$ (blocks)

$$Total(SS) = \sum Y_{\bullet j}^{\,2} - \frac{Y_{\bullet\bullet}^2}{ij} = 27425.417 - \frac{(640.37)^2}{5 \cdot 3} = 87.168$$

$$SS(Blocks) = \frac{\sum Y_{\bullet j}^2}{i} - \frac{Y_{\bullet\bullet}^2}{ij} = \frac{(128.14)^2 + \ldots\ldots (128.32)^2}{3} - \quad = 18.948$$

RCBD Fruit Fly Example

df_{block} = j - 1 = 5 - 1 = 4
df_{treat} = i -1 = 3 -1 = 2
df_{total} = (i j) - 1 = 3x5-1 = 14
df_{error} = total - (block + treat) = (i-1)(j-1) = 8
i = 3 (treatments); j = 5 (blocks)

$$SS(Treat) = \frac{\sum Y_{i\bullet}^{\ 2}}{j} - \frac{Y_{\bullet\bullet}^{2}}{ij} = \frac{200.39^2 + 213.95^2 + 226.03^2}{5} - \frac{(640.3)^2}{3.5}$$

$$= 65.814$$

SS(Error) = Total(SS) - [SS(blocks) + SS(Treat)]

$$= 87.168 - [18.948 + 65.814]$$

RCBD Fruit Fly Example

Source	df	SS	MS	F_c	$F_{i,j,\alpha}$
$Block_{person}$	4	18.947706	4.73926	15.76	3.84
$treatment_{type}$	2	65.81397	32.906987	109.4	4.46
Exp. Error	8	2.406194	0.3007743	-----	-----
TOTAL	14	87.167873	-----	-----	-----

Here $F_{i,j,\alpha} = F_{2,8,0.05} = 4.46$

Conclusion: Reject H_0, (i.e. $F_C > F_{table}$)
i.e. treatments are different !

Here $F_{i,j,\alpha} = F_{4,8,0.05} = 3.84$

Conclusion: Reject H_0, (i.e. $F_C > F_{table}$)
i.e. Blocks are NOT homogeneous !

RCBD Fruit Fly Example

If did not block on person, we could potentially contaminate the real source of differences in the data with differences between the people's lab techniques.

So, where's the source of differences ?

- can use lsd, Scheffe, etc.

e.g. critical value for lsd:

$$lsd = t_{\nu,\alpha/2} \times \sqrt{E(MS) \times \frac{2}{j}} = 2.306 \times \sqrt{0.3007743 \times \frac{2}{5}} = 0.799852$$

Where $t_{\nu,\ \alpha/2} = t_{8,\ 0.025} = 2.306$

ν (nu) = df for E(MS) = (i-1)(j-1) = 8

RCBD Fruit Fly Example

	HET (42.79)	WLD (45.206)
HOM (40.078)	2.712*	5.128*
HET (42.79)	-------	2.416 *

* = significantly different

Final Statements

- the 'boss' had hoped CuZnSOD transfection would work. Obviously it didn't !!

If anything the resultant flies had less.

- the lab has 5 people with significantly differing technical skills.

Differences between CRD and RCBD

- if we didn't block on person E(MS) would have been equal to 21.3539/12 = 1.7795

- therefore the Fc = 18.49 (re: Ftable = F2,12,0.05 = 3.89).

- the CRD doesn't partition the Error.

Let's revisit those trout fry:
Before, we had 4 diets, 5 fish/diet. Now let's suppose we obtained 4 fish (1/diet) from each of 5 hatcheries.

- this time, 4 fish had been randomized to 4 diets at each fish hatchery.

- fish are homogeneous in their response to treatment, but the hatchery may be a source of heterogeneity we wish to remove.

Fish Example: CRD

Example: Want to examine the effect of 4 diets on the growth of rainbow trout fry.

Replicates (mass in grams)

Dietary Treatment 1	2	3	4
59	70	93	124
47	59	85	135
40	52	79	167
32	87	72	83
39	61	88	152
217	329	417	661

i = 4 treatments

j = 5 replicates

totals

$$\sum_{1}^{j} Y_{ij} = Y_{i\bullet} = \quad 217 \quad 329 \quad 417 \quad 661 \rightarrow 1624 = Y_{\bullet\bullet}$$

$$\sum Y_{ij}^2 = \quad 9835 \quad 22375 \quad 35043 \quad 91483 \rightarrow 158736$$

$$\bar{Y}_{i\bullet} = \quad 43.4 \quad 65.8 \quad 83.4 \quad 132.2 \rightarrow 81.2 = \bar{Y}_{\bullet\bullet}$$

Fish Example: RCBD

Example: Want to examine the effect of 4 diets on the growth of rainbow trout fry in n 4 different hatcheries

i = 4 treatments

j = 5 blocks

Hatchery

Dietary Treatment

Replicates (mass in grams)

	1	2	3	4	$\sum_1^i Y_{ij} = Y_{\bullet j}$
A	59	70	93	124	346
B	47	59	85	135	326
C	40	52	79	167	338
D	32	87	72	83	274
E	39	61	88	152	340
	217	329	417	661	totals

$\sum_1^j Y_{ij} = Y_{i\bullet} =$	217	329	417	661	1624	$= Y_{\bullet\bullet}$
$\sum Y_{ij}^2 =$	9835	22375	35043	91483	158736	
$\overline{Y}_{i\bullet} =$	43.4	65.8	83.4	132.2	81.2	$= \overline{Y}_{\bullet\bullet}$

Compute Sums of Squares CRD:

$$Total(SS) = \sum Y_{ij}^2 - \frac{(Y_{\bullet\bullet})^2}{i \times j} = 158736 - \frac{1624^2}{4 \times 5} = 26867.2$$

SS(Treatments)=

$$SS(T) = \frac{\sum (Y_{i\bullet})^2}{j} - \frac{(Y_{\bullet\bullet})^2}{i \times j} = \frac{217^2 + 329^2 + 417^2 + 661^2}{5} - \frac{1624^2}{4 \times 5} = 21359.2$$

SS(error) = SS(E) = Total(SS) - SS(T)

ANOVA: CRD

Source	df	SS	MS	F_c	$F_{i,j,\alpha}$
treatment	3	21359.2	7119.73	20.68	3.24
error	16	5508.0	344.25	-----	-----
Total	19	26867.2	-----	-----	-----

$df_{treatment}$ = (# treatments - 1) = (i - 1)

df_{TOTAL} = (i × j) - 1

df_{error} = df_{TOTAL} - $df_{treatment}$

MS = SS / df

F_C = T(MS) / E(MS)

$F_{i,j,\alpha}$ = from table

e.g. $F_{3,16,0.05}$ = 3.24

ANOVA: RCBD

Source	df	SS	MS	F_c	$F_{i,j,\alpha}$
$Block_{hatchery}$	4	859.2	214.8	0.554	3.26
$treatment_{diet}$	3	21359.20	7119.733	18.38	3.49
Exp. Error	12	4648.8	387.4		
Total	19	26867.2	-----	-----	-----

Here $F_{i,j,\alpha} = F_{3,12,0.05} = 3.49$ Conclusion: Reject H_0, (i.e. $F_C > F_{table}$)

The difference?

- Error SS from CRD gets divided up into Error(SS) & Block(SS) in the RCBD design.
- The RCBD design removes some of the experimental error as error due to block effect.

CRD

Source	df	SS	MS	F_c	$F_{i,j,\alpha}$
treatment	3	21359.2	7119.73	20.68	3.24
error	16	5508.0	344.25	-----	-----
Total	19	26867.2	-----	-----	-----

RCBD

Source	df	SS	MS	F_c	$F_{i,j,\alpha}$
$Block_{hatchery}$	4	859.2	214.8	0.554	3.26
$treatment_{diet}$	3	21359.20	7119.733	18.38	3.49
Exp. Error	12	4648.8	387.4		
Total	19	26867.2	-----	-----	-----

Assessing the Efficiency of Blocking

$$\hat{\sigma}^2_{RCBD} = MS(E) = 387.4$$

$$\hat{s}^2_{CRD} = \frac{(j-1)s^2_{block} + j(i-1)s^2}{(i-1)(j-1)} = \frac{(5-1)214.8 + 5(4-1)387.4}{(4-1)(5-1)} = 344.25$$

$$\frac{\hat{\sigma}^2_{RCBD}}{\hat{\sigma}^2_{CRD}} = \frac{387.4}{344.25} = 1.125$$

- if this ratio > 1.0 then the RCBD is not any more efficient.

Fruit Fly Example

$$\hat{\sigma}^2_{RCBD} = MS(E) = 0.3007743$$

$$\hat{\sigma}^2_{CRD} = \frac{(j-1)s^2_{block} + j(i-1)s^2}{(i-1)(j-1)} = 1.7795$$

$$\frac{\hat{\sigma}^2_{RCBD}}{\hat{\sigma}^2_{CRD}} = \frac{0.3007743}{1.7795} = 0.169 = 16.9\%$$

Interpretation:

- a CRD design with, say, 100 experimental units not assembled into blocks will give answers that are about as precise as those for a RCBD with about 17 experimental units !!

RCBD with Subsampling

weanling rats fed 5 diets for 2 weeks

A = ZnDF

B = PEM

C = ZnPF

D = ZnAL

E = +ZnAL

- measured final weight (all started at exactly the same weight)

	DIET						
BLOCK	A	B	C	D	E	Totals	
1	72	82	110	117	138	519	56741
	61	87	105	103	116	472	46380
	58	79	99	110	113	459	44275
subtotal	191	248	314	330	367	1450	
2	54	82	106	117	127	486	50694
	55	76	97	108	119	455	44035
	61	80	102	114	131	488	50682
subtotal	170	238	305	339	377	1429	
3	65	83	110	122	139	519	57419
	53	80	99	104	117	453	43515
	50	75	98	125	125	473	48979
subtotal	168	238	307	351	381	1445	442720
TOTALS	529	724	926	1020	1125	43[illegible]4	

$Y_{...}$ ΣY_{ij}^2

RCBD with Subsampling Model Equation

Assumptions:

1) ε_{ij} are $N(0, s\varepsilon^2)$
2) $\Sigma\beta_j = 0$
3) $\Sigma\tau_i = 0$
4) δ_{ijk} are $N(0, s\delta^2)$

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij} + \delta_{ijk}$$

Block

SS

treat

How many parameters does the model try to fit ?

= (# treatments) + (# blocks) + 3 = 11

(i.e. $\tau 1$, $\tau 2$, $\tau 3$, $\tau 4$, $\tau 5$, $\beta 1$, $\beta 2$, $\beta 3$, μ, $\sigma\delta^2$, $\sigma\varepsilon^2$)

RCBD with Subsampling :Calculations

i = 5 (treatments) = diet
j = 3 (blocks) = cage battery
k = 3 (subsamples) = rat

$$Total(SS) = \sum Y_{ijk}^2 - \frac{Y_{\bullet\bullet\bullet}^2}{ijk} = \frac{4324^2}{(5)(3)(3)} = 442720\text{-}415488.36 = 27231.65$$

$$SS(AmongRats) = \frac{\sum Y_{ij\bullet}^2}{k} - \frac{Y_{\bullet\bullet\bullet}^2}{ijk} = \frac{191^2 + 248^2 + \ldots + 381^2}{3} - \frac{4324^2}{(5)(3)(3)} = 25600.98$$

$$SS(Error) = Total(SS) - SS(AmongRats) = 27231.65 - 25600.98 = 1630.67$$

$$SS(Blocks) = \frac{\sum Y_{\bullet j\bullet}^2}{ik} - \frac{Y_{\bullet\bullet\bullet}^2}{ijk} = \frac{1450^2 + 1429^2 + 1445^2}{(5)(3)} - \frac{4324^2}{(5)(3)(3)} = 16.04$$

$$SS(T) = \frac{\sum Y_{i\bullet\bullet}^2}{jk} - \frac{Y_{\bullet\bullet\bullet}^2}{ijk} = \frac{529^2 + 724^2 + \ldots + 1175^2}{(3)(3)} - \frac{4324^2}{(5)(3)(3)} = 25346.98$$

$$SS(ExpError) = SS(AmongRats) - [SS(Blocks) + SS(T)] = 25600.98 - [16.04 + 25346.98] = 237.96$$

RCBD with SS ANOVA Table:

Source	df	SS	MS	F_c	$F_{i,j,\alpha}$
Among $ExpUnits_{Rats}$	14	25600.98	------	------	------
$Block_{battery}$	2	16.04	4.01	------	------
$treatment_{diet}$	4	25346.98	6336.74	213.07	3.84
Exp. Error	8	237.96	29.745	0.5479	2.27
ERROR	30	1630.67	54.36	------	------
TOTAL	44	27231.65	------	------	------

$df_{expUnits} = ij-1 = (5)(3)-1 = 14$
$df_{Block} = j-1 = 2$
$df_{treatment} = i-1 = 4$

$df_{ExpError} = (i-1)(j-1) = 4x2 = 8$
$df_{Error} = ij(k-1) = (5)(3)(3-1) = 30$
$df_{Total} = ijk-1 = (5)(3)(3)-1 = 44$

RCBD with Subsampling: ANOVA

$$F_C(treatment) = \frac{MS(T)}{MS(EE)} = \frac{6336.74}{29.745} = 213.07$$

$$F_C(ExpError) = \frac{MS(EE)}{MS(E)} = \frac{29.745}{54.36} = 0.5427$$

- To test the null hypothesis of no differences between treatments one should use MS(T)/MS(EE) as above.
- However, if MS(EE) $\leq$ MS(E) it is recommended that you should use the pooled error:

$$MS(PooledError) = \frac{SS(EE) + SS(E)}{df_{EE} + df_E} = \frac{237.96 + 1630.67}{8 + 40} = 49.2$$

RCBD with Subsampling: ANOVA

$$F_C(treatment_{pooled}) = \frac{MS(T)}{MS(PE)} = \frac{6336.74}{49.2} = 128.86$$

This parameter would be used to test the null hypothesis concerning treatment effects.

- this situation could happen when the variation among experimental units is insignificant, and error is only within experimental units, as measured by the sampling error (i.e. MS(E)).
- in other words MS(E) and MS(EE) are essentially measuring the same thing (i.e. the quantity σ^2)

RCBD with Subsampling: MS

$$MS(B) = \sigma^2 + k\sigma_\varepsilon^2 + \frac{ik}{(j-1)}\sum \beta_j^2 = \sigma^2 + 3\sigma_\varepsilon^2 + 7.5\sum \beta_j^2$$

$$MS(T) = \sigma^2 + k\sigma_\varepsilon^2 + \frac{jk}{(i-1)}\sum \beta_j^2 = \sigma^2 + 3\sigma_\varepsilon^2 + 2.25\sum \beta_j^2$$

$$MS(EE) = \sigma^2 + k\sigma_\varepsilon^2$$

$$MS(E) = \sigma^2$$

RCBD with Subsampling: Results

The best estimate of $\sigma\varepsilon^2 = 0$

→ This is because we accepted the null hypothesis that experimental error was not significant.

(i.e. Fc < Ftable = 0.5479 < 2.27)

The best estimate of σ^2 is the pooled error 49.2

Is there a significant difference among diets ?

→ YES reject H_0 since FC > Ftable (i.e. 213.1 > 3.84)

RCBD with Subsampling: Results

So, there is a significant diet effect. Where is(are) the differences?

To evaluate the nature of these differences one can use the lsd method:

$$lsd = t_{\nu,\alpha/2} \times \sqrt{MS(EE) \times \frac{2}{jk}} = 2.306 \times \sqrt{29.74 \times \frac{2}{(3)(3)}} = 5.9282$$

Note that k is added, only in subsampling

Note: other uses of the lsd method uses E(MS), the error mean SS. For subsampling the MS(EE) is used instead.

RCBD with Subsampling: Post Hoc

Diet	B_{PEM}	C_{ZnPF}	D_{ZnAL}	E_{+ZnAL}
A_{ZnDF}	21.666*	44.111*	54.555*	66.222*
B_{PEM}		22.445*	32.889*	44.556*
C_{ZnDF}			10.444*	22.111*
D_{ZnAL}				11.667*

* = significant

Where,

A_{ZnDF} = 58.778 g

B_{PEM} = 80.444 g

C_{ZnPF} = 102.889 g

D_{ZnAL} = 113.333 g

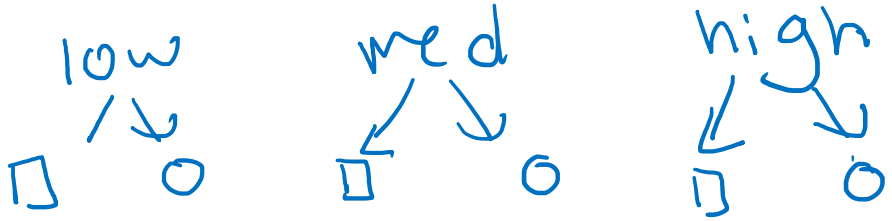
E_{+ZnAL} = 125.0 g

Other Important Designs

1). Factorial Design

- to this point only one factor has been investigated (e.g. diet on weight gain)

- what about treatment combinations that are somehow jointly responsible for the response



2). Analysis of Covariance (ANCOVA)

- any of the other models. However, include a covariate term (e.g. initial age, initial weight, scalp/skin impedance, SNR, etc.)