Lecture 5

Single factor design and analysis

Completely randomized design (CRD)

Completely randomized design

- In the design of experiments, completely randomized designs are for studying the effects of one primary factor without the need to take other nuisance variables into account
- The experiment compares the values of a response variable based on the different levels of that primary factor. For completely randomized designs, the levels of the primary factor are randomly assigned to the experimental units.

Completely randomized design

- A study design with only one independent factor (e.g. category) of treatment in which the factor is manipulated at multiple levels. Often used in experimental design to determine the effect of a certain treatment or intervention.
- May be contrasted with factorial design, which evaluates the effects of two or more factors simultaneously.

Completely randomized design

- In the experiment, only one factor A, and a levels A_1, A_2,…, A_a. In each level A_i, there are r_i replications, i=1, 2, 3, …, a
- If r_1= r_2=…=r_a, the design is balanced. Otherwise, it is an unbalanced design.
- y_{ij} is the result of ith level and jth replication.

Example

- For an unbalanced design, A1 has 7 samples, A2 has 5 samples, A3 has 6 samples, and A4 has 6 samples. In total, there are 24 samples.

Levels of factor A	Number of experiment units
A1	1, 2, 3, 4, 5, 6, 7 (7)
A2	8, 9, 10, 11, 12 (5)
A3	13, 14, 15, 16, 17, 18 (6)
A4	19, 20, 21, 22, 23, 24 (6)

Example

- Can we arrange the 24 experiment units in the order of the four levels? No.
- The attention and skill proficiency of manipulators may change during the experiments. And the light intensity may also be different. The observations may not be independent!
- So we should use random design to solve this problem.

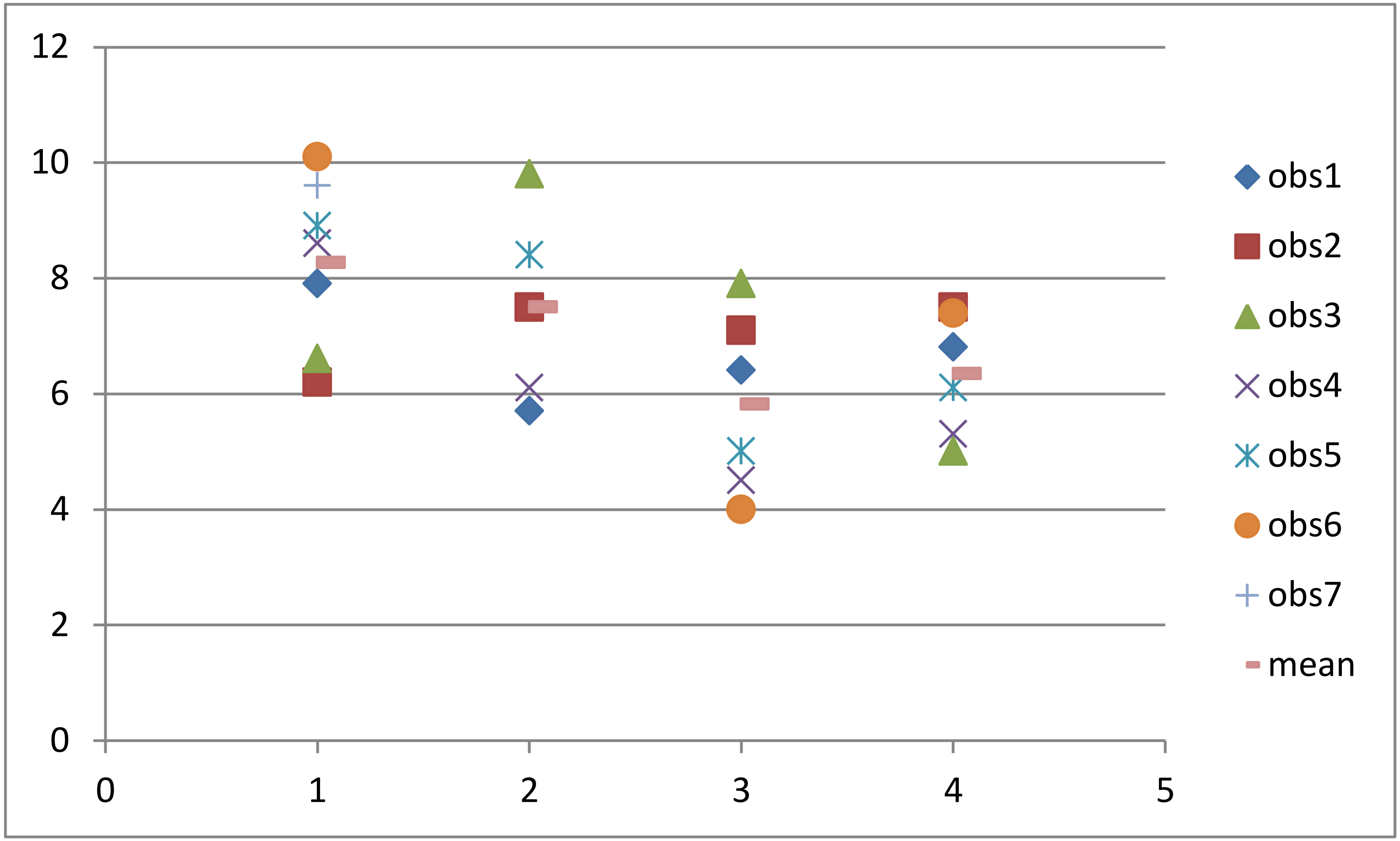
	Randomiza	RandNum
Plot1	A1	0.016941
Plot2	A3	0.040356
Plot3	A4	0.04234
Plot4	A3	0.063423
Plot5	A1	0.125257
Plot6	A3	0.126621
Plot7	A4	0.208301
Plot8	A4	0.221024
Plot9	A1	0.255363
Plot10	A2	0.334577
Plot11	A4	0.366835
Plot12	A2	0.418322
Plot13	A4	0.441407
Plot14	A2	0.635086
Plot15	A3	0.762573
Plot16	A2	0.799705
Plot17	A3	0.802349
Plot18	A1	0.829444
Plot19	A4	0.838652
Plot20	A3	0.848814
Plot21	A2	0.853357
Plot22	A1	0.860988
Plot23	A1	0.964504
Plot24	A1	0.9751

Example

- 24 experiments for Folic acid content in green tea

Levels of factor A	Observed data (mg)	Sample mean
A1	7.9, 6.2, 6.6, 8.6, 8.9, 10.1, 9.6	8.27
A2	5.7, 7.5, 9.8, 6.1, 8.4	7.50
A3	6.4, 7.1, 7.9, 4.5, 5.0, 4.0	5.82
A4	6.8, 7.5, 5.0, 5.3, 6.1, 7.4	6.35

Dot-plot



- Are the differences caused by chance or not? We use analysis of variance (ANOVA) for further analysis.

General Data

Levels of factor A	Data	Sum	Mean
A_1	$y_{11}\ y_{12}\ \cdots\ y_{1r_1}$	$T_1 = y_{11} + y_{12} + \cdots + y_{1r_1}$	$\bar{y}_1 = T_1/r_1$
A_2	$y_{21}\ y_{22}\ \cdots\ y_{2r_2}$	$T_2 = y_{21} + y_{22} + \cdots + y_{2r_2}$	$\bar{y}_2 = T_2/r_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_a	$y_{a1}\ y_{a2}\ \cdots\ y_{ar_a}$	$T_a = y_{a1} + y_{a2} + \cdots + y_{ar_a}$	$\bar{y}_a = T_a/r_a$

Basic assumptions

- 1. Normality: samples y_{i1}, y_{i2}, …, y_{ir_i} under level A_i have the Normal distribution

$$N(\mu_i, \sigma_i^2), i = 1, 2, \cdots, a$$

- 2. Homogeneity of Variance: the variances of the *a* Normal distributions are the same, i.e. $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_a^2 = \sigma_\varepsilon^2$

- 3. Randomness. All data y_{ij} are independent.

Targets

- 1. Are the means of the *a* levels $\mu_1, \mu_2, \cdots, \mu_a$ the same? (using One-way ANOVA)

- If the means are not the same, which difference between means is significant? (using multiple comparison)

The linear model

- The model is

$$y_{ij} = \mu_i + \varepsilon_{ij}, i = 1, 2, \cdots, a;\ j = 1, 2, \cdots, r_i$$

- ε_{ij} is the experimental error of the *i*th level and *j*th experiment. $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$, i.i.d
- Theory 1: y_{ij} are sum of a constant μ_i and random error ε_{ij}
- Theory 2: $E(\varepsilon_{ij}) = 0, Var(\varepsilon_{ij}) = \sigma_\varepsilon^2$, so

$$E(y_{ij}) = \mu_i, Var(y_{ij}) = \sigma_\varepsilon^2$$

The linear model

- Theory 3: $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$, so $y_{ij} \sim N(\mu_i, \sigma_\varepsilon^2)$
- Theory 4: The random errors are independent, so all y_{ij} are also independent.

Least square estimation

- Minimize

$$SS_{\varepsilon} = \sum_{i=1}^{a} \sum_{j=1}^{r_i} (y_{ij} - \mu_i)^2$$

$$= \sum_{j=1}^{r_1} (y_{1j} - \mu_1)^2 + \sum_{j=1}^{r_2} (y_{2j} - \mu_2)^2 + \cdots + \sum_{j=1}^{r_a} (y_{aj} - \mu_a)^2$$

- The least square estimator of μ_i is

$$\hat{\mu}_i = \bar{y}_i, i = 1, 2, \cdots, a$$

- In previous example,

$$\hat{\mu}_1 = 8.27, \hat{\mu}_2 = 7.50, \hat{\mu}_3 = 5.82, \hat{\mu}_4 = 6.35$$

One-way ANOVA

Hypothesis in one-way ANOVA

- The one-way analysis of variance is used to test the claim that more than 2 population means are equal
- This is an extension of the two independent samples *t*-test
- H_0: $\mu_1 = \mu_2 = \cdots = \mu_a$
- H_A: $\mu_1, \mu_2, \cdots, \mu_a$ are not equal.
- If we reject H_0 under the significance level α, then factor A is significant under the level α. Otherwise, factor A is not significant.

Sum of squares

- Definition

$$\bar{y} = (y_1 + y_2 + \cdots + y_a)/a$$

$$Q = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_a - \bar{y})^2 = \sum_{i=1}^{a} (y_i - \bar{y})^2$$

- Because $\sum_{i=1}^{a} (y_i - \bar{y}) = 0$

- There are only *a*-1 independent deviations in *Q*, we call the number of independent deviations in sum of squares as degree of freedom for sum of squares which is often denoted as *f*.

Distribution of sum of square (*Q*)

- Theorem: assume $y_1, y_2, \ldots, y_a$ is a sample from *a* normal distribution N(μ, σ^2). Then
- 1. Sample mean

$$\bar{y} \sim N(\mu, \sigma^2/a)$$

- 2. Ratio of sum of squares to σ^2 is

$$Q/\sigma^2 \sim \chi^2(a-1)$$

- 3. $\bar{y}$ and *Q* are independent

Decomposing the sum of squares

- Mean of all data y_{ij} is

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{r_i} y_{ij} = \frac{1}{n}\sum_{i=1}^{a} r_i \bar{y}_i$$

- Define $SS_{\mathrm{T}} = \sum_{i=1}^{a}\sum_{j=1}^{r_i}(y_{ij} - \bar{y})^2$, degree of freedom $f_T = n - 1$

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{r_i}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{a}\sum_{j=1}^{r_i}(y_{ij} - \bar{y}_i + \bar{y}_i - \bar{y})^2$$

$$= \sum_{i=1}^{a}\sum_{j=1}^{r_i}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{a} r_i(\bar{y}_i - \bar{y})^2$$

$$\overset{Define}{=} SS_{\varepsilon} + SS_A$$

Sum of Squares

- SS_A is the sum between groups, with degree of freedom *a*-1; SS_ε is the sum within groups (i.e. sum squares of errors), with degree of freedom *n*-*a*

Sum of Squares

- Then

$$SS_{\mathrm{T}} = \sum_{i=1}^{a}\sum_{j=1}^{r_i}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{a}\sum_{j=1}^{r_i} y_{ij}^2 - n\bar{y}^2,\ f_T = n-1$$

$$SS_{\mathrm{A}} = \sum_{i=1}^{a} r_i(\bar{y}_i - \bar{\mathrm{y}})^2 = \sum_{i=1}^{a} r_i \bar{y}_i^2 - n\bar{y}^2, f_A = a-1$$

$$SS_{\varepsilon} = SS_T - SS_A,\ f_{\varepsilon} = n-a$$

Example

- The data and sum of squares

Level	Data	Rep	Mean
A_1	7.9, 6.2, 6.6, 8.6, 8.9, 10.1, 9.6	r_1=7	8.27
A_2	5.7, 7.5, 9.8, 6.1, 8.4	r_2=5	7.5
A_3	6.4, 7.1, 7.9, 4.5, 5.0, 4.0	r_3=6	5.82
A_4	6.8, 7.5, 5.0, 5.3, 6.1, 7.4	r_4=6	6.35
		n=24	7.02

Example

- So

$$SS_{\mathrm{A}} = 7 \times 8.27^2 + 5 \times 7.5^2 + 6 \times 5.82^2 + 6 \times 6.35^2 - 24 \times 7.02^2 = 23.50$$

$$SS_{\mathrm{T}} = (7.9^2 + 6.2^2 + \cdots + 6.1^2 + 7.4^2) - 24 \times 7.02^2 = 65.27$$

$$SS_{\varepsilon} = 65.27 - 23.50 = 41.77$$

Mean square

- Mean square is sum of squares divided by its degree of freedom

$$MS_{\varepsilon} = \frac{SS_{\varepsilon}}{n-a} \qquad MS_{A} = \frac{SS_{A}}{a-1}$$

- Theory: Under the basic assumption of single factor design, we have:

$$E(SS_{\varepsilon}) = (n-a)\sigma_{\varepsilon}^{2} \quad E(SS_{A}) = (a-1)\sigma_{\varepsilon}^{2} + \sum_{i=1}^{a} r_i(\mu_i - \mu)^2$$

- Where $\mu = \frac{1}{n}\sum_{i=1}^{a} r_i\mu_i = E(\bar{\mathrm{y}})$

Distributions under H_0

- It is proved, under H_0,

$$SS_A / \sigma_\varepsilon^2 \sim \chi^2(a-1)$$

$$SS_\varepsilon / \sigma_\varepsilon^2 \sim \chi^2(n-a)$$

SS_A and SS_ε are independent.

- Then $\dfrac{SS_A / \sigma_\varepsilon^2 / (a-1)}{SS_\varepsilon / \sigma_\varepsilon^2 / (n-a)} = \dfrac{MS_A}{MS_\varepsilon}$ ~ $F(a\text{-}1,n\text{-}a)$

i.e. $F = \dfrac{MS_A}{MS_\varepsilon}$ ~ $F(a\text{-}1,n\text{-}a)$

The analysis of variance table for the single factor

Source	Degrees of freedom	Sum of squares	Mean square	F ratio
Factor A	$f_A=a-1$	$SS_A=\sum_{i=1}^{a} r_i(\bar{y}_i-\bar{y})^2$	$MS_A=\frac{SS_A}{a-1}$	$F=\frac{MS_A}{MS_\varepsilon}$
Error	$f_\varepsilon=n-a$	$SS_\varepsilon=\sum_{i=1}^{a}\sum_{j=1}^{r_i}(y_{ij}-\bar{y}_i)^2$	$MS_\varepsilon=\frac{SS_\varepsilon}{n-a}$	
Total T	$f_T=n-1$	$SS_T=\sum_{i=1}^{a}\sum_{j=1}^{r_i}(y_{ij}-\bar{y})^2$		

- Given significance level α , find the $F_{1-\alpha}(a-1,n-a)$, then
- If $F>F_{1-\alpha}(a-1,n-a)$, reject H_0
- If $F\le F_{1-\alpha}(a-1,n-a)$, accept H_0

Example (continued)

- We have calculated the sum of squares. The table of ANOVA is

Source	Degrees of freedom (DF)	Sum of squares (SS)	Mean square (MS)	F value
Factor A	3	23.50	7.83	3.75*
Error	20	41.77	2.09	
Total T	23	65.27		

- $\alpha = 0.05, F_{0.95}(3,20) = 3.10, F > 3.10$, reject H_0, i.e. the four classes have significant difference.

Example (continued)

- Meanwhile, we can get the unbiased estimation of σ^2: $\hat{\sigma}_\varepsilon^2 = 2.09$
- Estimation of means are

$$\hat{\mu}_1 = 8.27, \hat{\mu}_2 = 7.50, \hat{\mu}_3 = 5.82, \hat{\mu}_4 = 3.5$$

- The mean under A_1 is the largest.

$$\alpha = 0.05, t_{1-\alpha/2}(n-a) = t_{0.975}(20) = 2.0860, r_1 = 7, \hat{\sigma}_\varepsilon = 1.45$$

$$\text{So } \bar{y}_1 \pm t_{1-\alpha/2}(n-a)\hat{\sigma}_\varepsilon / \sqrt{r_1} = 8.27 \pm 2.0860 \times 1.45 / \sqrt{7} = 8.27 \pm 1.14$$

- The interval estimation of μ_1 is [7.13, 9.41]

Balanced experiment

- If the experiment has the same number of replications in every level, the design is a balanced experiment.
- Advantages:
  - Exclude the impact of different replications
  - The equations for calculation are simpler.

ANOVA of balanced experiment

$$r_1 = r_2 = \cdots = r_a = r$$

$$SS_{\mathrm{T}} = \sum_{i=1}^{a}\sum_{j=1}^{r}(y_{ij} - \overline{y})^2 = \sum_{i=1}^{a}\sum_{j=1}^{r} y_{ij}^2 - ar\overline{y}^2,\ f_T = ar - 1$$

$$SS_A = r\sum_{i=1}^{a}(\overline{\mathrm{y}}_i - \overline{y})^2 = r\sum_{i=1}^{a}\overline{y}_i^2 - ar\overline{y}^2,\ f_A = a - 1$$

$$SS_\varepsilon = SS_T - SS_A,\ f_\varepsilon = (r-1)a$$

- Under H_0, $SS_{\mathrm{A}} / \sigma_\varepsilon^2 \ \sim\ \chi^2(a-1)$

$$SS_\varepsilon / \sigma_\varepsilon^2 \ \sim\ \chi^2((r-1)a)$$

ANOVA table

Source	Degrees of freedom	Sum of squares	Mean square	Expected MS	F ratio
Factor A	$f_A=a-1$	$SS_A = \sum_{i=1}^{a} r(\bar{y}_i - \bar{y})^2$	$MS_A = \frac{SS_A}{a-1}$	$\sigma_\varepsilon^2 + r\sigma_A^2$	$F = \frac{MS_A}{MS_\varepsilon}$
Error	$f_\varepsilon=(r-1)a$	$SS_\varepsilon = \sum_{i=1}^{a}\sum_{j=1}^{r}(y_{ij} - \bar{y}_i)^2$	$MS_\varepsilon = \frac{SS_\varepsilon}{(r-1)a}$	σ_ε^2	
Total T	$f_T=ra-1$	$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{r}(y_{ij} - \bar{y})^2$			

Let’s work on the previous example together on our computers

Levels of factor A	Observed data (mg)
A_1	7.9, 6.2, 6.6, 8.6, 8.9, 10.1, 9.6
A_2	5.7, 7.5, 9.8, 6.1, 8.4
A_3	6.4, 7.1, 7.9, 4.5, 5.0, 4.0
A_4	6.8, 7.5, 5.0, 5.3, 6.1, 7.4

- Do the randomization of the 24 plots
- Build the ANOVA table for the observed data

Randomly complete block design (RCBD)

Blocking to increase precision by grouping the experimental units into homogeneous blocks to compare treatments within a more uniform environment

Complete block design (CBD)

- If every treatment is used and replicated the same number of times in every block, the design is a complete block design (CBD).
- If each treatment is used once in every block, it is a randomly complete block design (RCBD).
- Here we consider a experiment with ***a*** treatments and ***b*** blocks (replications).

Statistical model of RCBD

i=1, 2, …, *a* for the *a* treatments; *j*=1, 2, …,*b* for the *b* replications

$$y_{ij} = \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..})$$

$$+ [y_{ij} - (\bar{y}_{i.} - \bar{y}_{..}) - (\bar{y}_{.j} - \bar{y}_{..}) - \bar{y}_{..}]$$

$$= \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

- μ : the general mean
- α_i : the treatment effect
- β_j : the block effect
- ε_{ij} : the experimental error

Field layout of 8 mutants and 3 blocks (i.e. 3 replications)

Block 1	RandN	Block 2	RandN	Block 3	RandN
B	0.31	D	0.3	G	0.07
A	0.33	E	0.37	F	0.21
E	0.38	H	0.43	B	0.39
F	0.4	G	0.45	A	0.56
C	0.45	A	0.64	C	0.78
D	0.68	B	0.68	H	0.79
H	0.73	C	0.87	D	0.82
G	0.96	F	0.99	E	0.94

Example: Observations of 8 mutants and 3 blocks (i.e. 3 replications)

Mutants	Observations			Mean across replications $\bar{y}_{i.}$	Mutant effects α_i
	Rep I	Rep II	Rep III		
A	10.9	9.1	12.2	10.7	-0.9
B	10.8	12.3	14.0	12.4	0.8
C	11.1	12.5	10.5	11.4	-0.2
D	9.1	10.7	10.1	10.0	-1.6
E	11.8	13.9	16.8	14.2	2.6
F	10.1	10.6	11.8	10.8	-0.8
G	10.0	11.5	14.1	11.9	0.3
H	9.3	10.4	14.4	11.4	-0.2
Mean across mutants $\bar{y}_{.j}$	10.4	11.4	13.0	11.6 ($\bar{y}$)	
Block effects β_j	-1.2	-0.2	1.4		

ANOVA of RCBD

$$SS_T = \sum_{i=1,\cdots,8;\, j=1,\cdots,3} (y_{ij} - \bar{y})^2 = \sum_{i=1,\cdots,8;\, j=1,\cdots,3} y_{ij}{}^2 - 24\bar{y}^2 = 84.61$$

$$SS_A = 3\sum_{i=1}^{8} \bar{y}_{i.}^2 - 24\bar{y}^2 = 3\sum_{i=1,\cdots,8} \alpha_i^2 = 34.08$$

$$SS_B = 8\sum_{j=1}^{3} \bar{y}_{.j}^2 - 24\bar{y}^2 = 8\sum_{j=1}^{3} \beta_j^2 = 27.56$$

ANOVA of RCBD

$$SS_T = \sum_{i;j} (y_{ij} - \mu)^2$$

$$= \sum_{i;j} [(\bar{y}_{i\cdot} - \mu) + (\bar{y}_{\cdot j} - \mu) + (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \mu)]^2$$

$$= r\sum_{i} (\bar{y}_{i\cdot} - \mu)^2 + n\sum_{j} (\bar{y}_{\cdot j} - \mu)^2 + \sum_{i;j} (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \mu)^2$$

$$= SS_A + SS_B + SS_\varepsilon$$

$$SS_\varepsilon = SS_T - SS_A - SS_B = 22.97$$

Table of ANOVA

Source of variation	Degree of freedom (df)	Sum of squares (SS)	Mean squares (MS)	F-test	Pr > F
Total	$a \times b$-1 = 23	84.61			
Mutants	a-1 = 7	34.08	4.87	2.97*	0.0395
Blocks	b-1 = 2	27.56	13.78	8.40**	0.004
Error	$(a$-1$) \times (b$-1$)$ = 14	22.97	1.64		

A sum of squares for blocks is partitioned out of the sum of squares of experimental error. The blocked design will markedly improve the precision on the estimates of treatment means if the reduction in SS_{ε} with blocking is substantial.

Confidence interval of treatment mean

- Standard error of treatment mean

$$s_{\bar{y}_i.} = \sqrt{\frac{MS_\varepsilon}{b}} = \sqrt{\frac{1.64}{3}} = 0.74$$

- The 95% confidence interval (CI)

$$CI_{\bar{y}_i.} = \bar{y}_i. \pm t_{0.975}(14) \times s_{\bar{y}_i.} = \bar{y}_i. \pm 2.14 \times s_{\bar{y}_i.} = \bar{y}_i. \pm 1.58$$

- Mutant A: (9.15, 12.31); B: (10.79, 13.95); C: (9.79, 12.95); D: (8.39, 11.55); E: (12.59, 15.75); F: (9.25, 12.41); G: (9.79, 12.95)

Test of hypothesis of treatment mean

- F statistic to test the null hypothesis of no yield difference among the eight mutants

$$F = \frac{MS_A}{MS_\varepsilon} = \frac{4.87}{1.64} = 2.97$$

- Critical value

$$F_{0.05}(7,14) = 2.76$$

- Observed significance level

$$P > F = F(2.76,7,14) = 0.0395$$

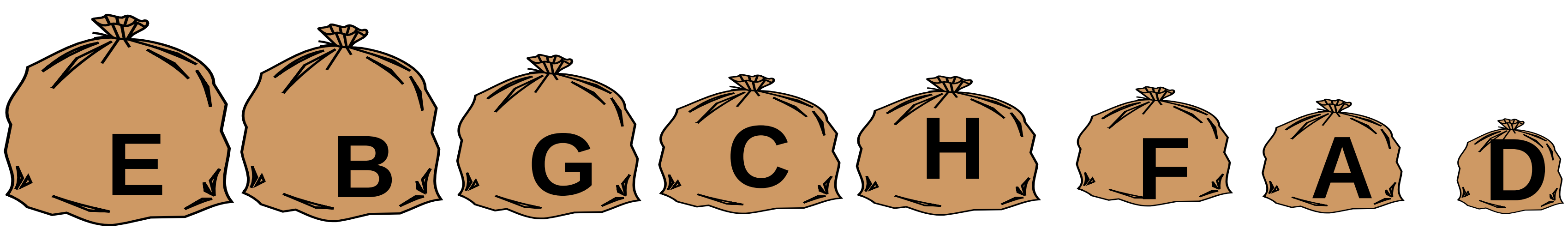
Estimation of variance component

Source	DF	MS	Expected MS
Total	$a \times b$-1 = 23		
Mutants	a-1 = 7	4.87	$\sigma_{\varepsilon}^{2} + b\sigma_{G}^{2}$
Blocks	b-1 = 2	13.78	
Error	(a-1)$\times$(b-1) = 14	1.64	σ_{ε}^{2}

- Error variance $\sigma_{\varepsilon}^{2} = 1.64$
- Genotypic variance $\sigma_{G}^{2} = (MS_{\mathrm{A}} - MS_{\varepsilon}) / b = 1.08$
- Repeatability (H) $H = \sigma_{G}^{2} / (\sigma_{G}^{2} + \sigma_{\varepsilon}^{2}) = 39.71\%$

Reporting the experiment

- Analysis of yield data indicates significant differences in yield among the eight wheat mutants
- Mutant E produces the highest yield
- Mutant D is clearly inferior to the others



E	B	G	C	H	F	A	D
14.2 ±1.58	12.4 ±1.58	11.9 ±1.58	11.4 ±1.58	11.4 ±1.58	10.8 ±1.58	10.7 ±1.58	10.0 ±1.58

Compared to One-way ANOVA

Source	Degrees of freedom	Sum of squares	Mean square	F ratio
Mutants	7	34.08	4.87	1.54
Error	16	50.53	3.16	
Total T	23	84.61		

- $\alpha = 0.05, F_{0.95}(7,16) = 2.66, F < 2.66$, we can't reject H_0, i.e. the eight mutants doesn't have significant difference.
- Here error variance=3.16>1.64 (error using the last analysis method)

Let's work on the previous example together on our computers

Mutants	Rep I	Rep II	Rep III
A	10.9	9.1	12.2
B	10.8	12.3	14.0
C	11.1	12.5	10.5
D	9.1	10.7	10.1
E	11.8	13.9	16.8
F	10.1	10.6	11.8
G	10.0	11.5	14.1
H	9.3	10.4	14.4

- Do the randomization of the three blocks
- Build the ANOVA table for the observed data