Lecture 10
The Analysis of Variance

Experimental Design

- The **sampling plan** or **experimental design** determines the way that a sample is selected.
- In an **observational study,** the experimenter observes data that already exist. The **sampling plan** is a plan for collecting this data.
- In a **designed experiment,** the experimenter imposes one or more experimental conditions on the experimental units and records the response.

Experimental Design

- Must design an experiment that will test your hypothesis.
- This experiment will allow you to change some conditions or variables to test your hypothesis.

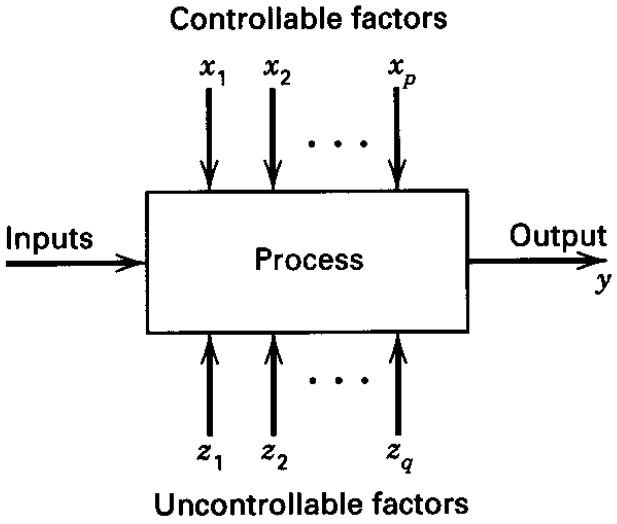


Figure 1-1 General model of a process or system.

Variables

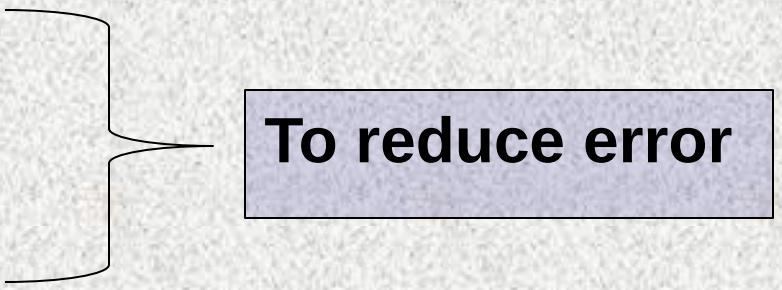
Variables are things that change.

The ***independent variable*** is the variable that is purposely changed. It is the manipulated variable.

The ***dependent variable*** changes in response to the independent variable. It is the responding variable.

The Basic Principles of Experimental Design

- Two aspects to any experimental problem:
  1. The design of the experiment
  2. Statistical analysis of the data
- **Three basic principles of experimental design**
  1. **Replication**
  2. **Randomization**
  3. **Blocking**



Definitions

- An **experimental unit** is the object on which a measurement or measurements) is taken.
- A **factor** is an independent variable whose values are controlled and varied by the experimenter.
- A **level** is the intensity setting of a factor.
- A **treatment** is a specific combination of factor levels.
- The **response** is the variable being measured by the experimenter.

Example

- A group of people is randomly divided into an experimental and a control group. The control group is given an aptitude test after having eaten a full breakfast. The experimental group is given the same test without having eaten any breakfast.

Experimental unit : person

Response : Score on test

Treatments : Breakfast or no breakfast

Factor = meal

Levels = Breakfast or no breakfast

Example

- The experimenter in the previous example also records the person's gender. Describe the factors, levels and treatments.

Experimental unit = person

Response = score

Factor #1 = meal

Factor #2 = gender

Levels = breakfast or no breakfast

Levels = male or female

Treatments:

male and breakfast, female and breakfast, male and no breakfast, female and no breakfast

The Analysis of Variance (ANOVA)

- All measurements exhibit **variability.**
- The total variation in the response measurements is broken into portions that can be attributed to various **factors**.
- These portions are used to judge the effect of the various factors on the experimental response.

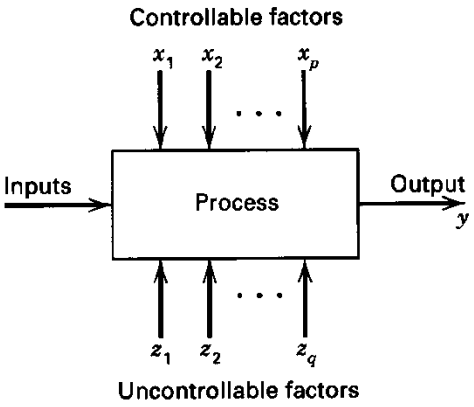
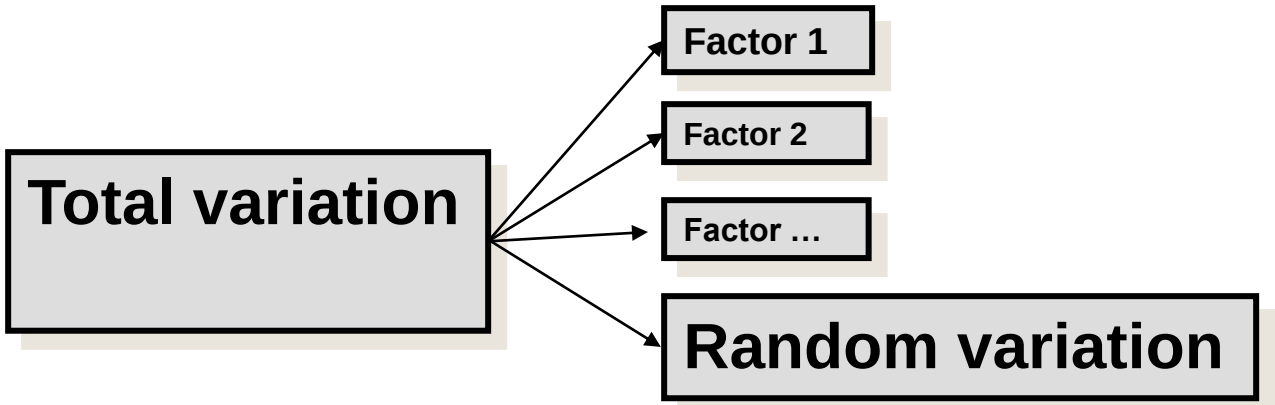


Figure 1-1 General model of a process or system.

The Analysis of Variance

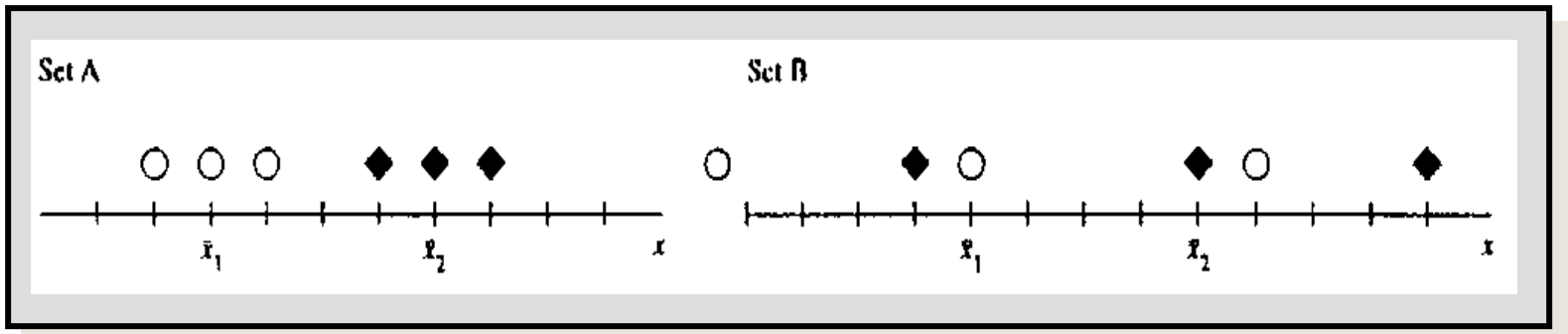
- If an experiment has been properly designed,



•We compare the variation due to any one factor to the typical random variation in the experiment.

The variation between the sample means is larger than the typical variation within the samples.

The variation between the sample means is about the same as the typical variation within the samples.



Assumptions

1. The observations within each population are normally distributed with a common variance σ^2.
2. Assumptions regarding the sampling procedures are specified for each design.

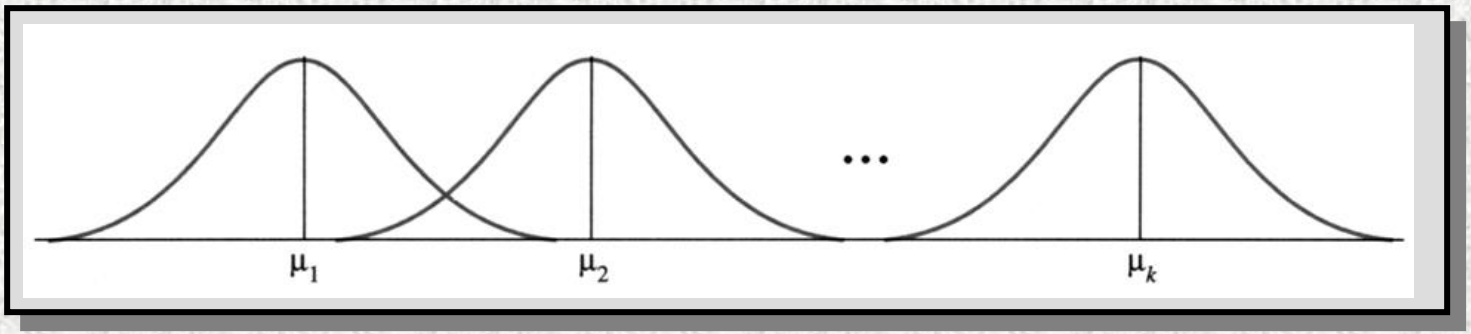
Analysis of variance procedures are fairly robust when sample sizes are equal and when the data are fairly mound-shaped.

Two Designs

1. **Completely Randomized Design (CRD)**
   an extension of the two independent sample *t*-test.
2. **Completely Randomized Block Design (CRBD)**
   an extension of the paired difference test.

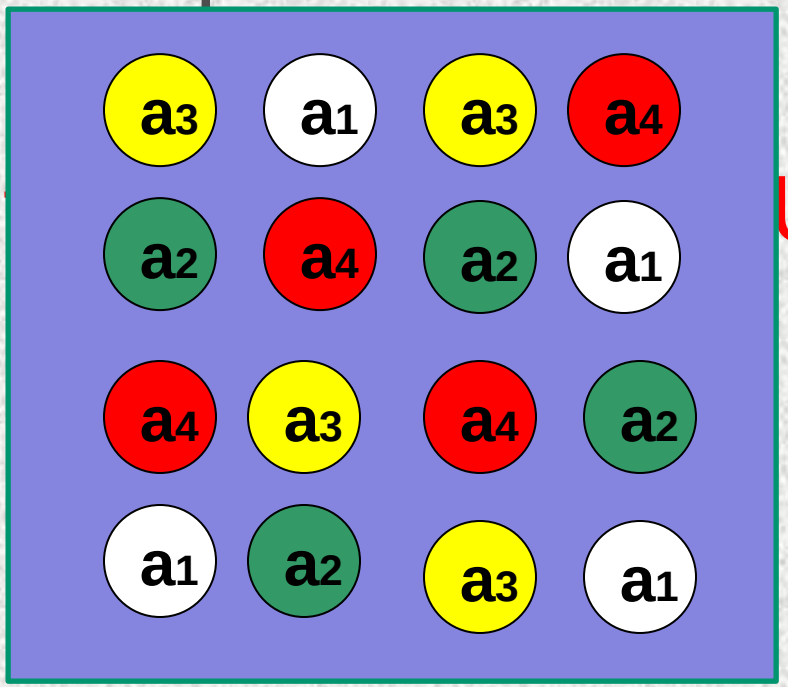
1. The Completely Randomized Design *(CRD)*

- A **one-way classification** in which one factor is set at *k* different levels.
- The *k* levels correspond to *k* different normal populations, which are the **treatments**.
- Are the *k* population means the same, or is at least one mean different from the others?



Randomization in CRD

- Factor : A
- A has 4 levels (a_1, a_2,a_3, and a_4)
- There are 4 replications each level (balance)
- So, there U



Assumptions

1. Randomness & Independence of Errors
   - Independent Random Samples are Drawn for each condition
2. Normality
   - Populations (for each condition) are Normally Distributed
3. Homogeneity of Variance
   - Populations (for each condition) have Equal Variances

About CRD

- Random samples of size $n_1, n_2, \ldots, n_k$ are drawn from *k* populations with means $\mu_1, \mu_2, \ldots, \mu_k$ and with common variance σ^2.
- Let x_{ij} be the *j*-th measurement (replication) and the *i*-th sample.
- The total variation in the experiment is measured by the **total sum of squares**:

$$\text{Total SS} = \sum (x_{ij} - \bar{x})^2$$

Example

Is the attention duration of children affected by whether or not they had a good breakfast? Twelve children were randomly divided into three groups and assigned to a different meal plan. The response was attention duration in minutes during the morning reading time.

No Breakfast		Light Breakfast		Full Breakfast	
8	(x_{11})	14	(x_{21})	10	(x_{31})
7	(x_{12})	16	(x_{22})	12	(x_{32})
9	(x_{13})	12	(x_{23})	16	(x_{33})
13	(x_{14})	17	(x_{24})	15	(x_{34})

k = 3 treatments. Are the average attention spans different?

The ANOVA Table of the CRD

Source	df	SS	MS	F
Treatments	k -1	SST	SST/(k-1)	MST/MSE
Error	n - k	SSE	SSE/(n-k)	
Total	n -1	Total SS		

SS(Sum of Squares)

$$\text{SST} = \frac{\sum_{i=1}^{k} T_i^2}{n_i} - \frac{\left(\sum x_{ij}\right)^2}{n}$$

$$\text{Total SS} = \sum x_{ij}^2 - \frac{\left(\sum x_{ij}\right)^2}{n}$$

$$\text{SSE} = \text{Total SS} - \text{SST}$$

MS(Mean Squares)

MST = SST/(*k*-1)

MSE = SSE/(*n*-*k*)

Computing

$$\text{Total SS} = \Sigma x_{ij}^2 - \text{CM}$$

$$= (\text{Sum of squares of all } x\text{-values}) - \text{CM}$$

with

$$\text{CM} = \frac{(\Sigma x_{ij})^2}{n} = \frac{G^2}{n}$$

$$\text{SST} = \Sigma \frac{T_i^2}{n_i} - \text{CM}$$

$$\text{SSE} = \text{Total SS} - \text{SST}$$

and

G = Grand total of all n observations

T_i = Total of all observations in sample i

n_i = Number of observations in sample i

$n = n_1 + n_2 + \cdots + n_k$

Note:
CM = Common Mean
CM = CF
CF =Correction Factor

The ANOVA Table of the CRD

The **Total SS** is divided into two parts:

- ✓ **SST** (sum of squares for treatments): measures the variation among the *k* sample means.
- ✓ **SSE** (sum of squares for error): measures the variation within the *k* samples.

$$\text{Total SS} = \text{SST} + \text{SSE}$$

- These **sums of squares** behave like the numerator of a sample variance. When divided by the appropriate **degrees of freedom**, each provides a **mean square**, an estimate of variation in the experiment.
- **Degrees of freedom** are additive, just like the sums of squares.

$$\text{Total } df = \text{Trt } df + \text{Error } df$$

The Breakfast Problem

No Breakfast	Light Breakfast	Full Breakfast
8	14	10
7	16	12
9	12	16
13	17	15
$T_1 = 37$	$T_2 = 59$	$T_3 = 53$

G = 149

$$CM = \frac{149^2}{12} = 1850.0833$$

$$\text{Total SS} = 8^2 + 7^2 + \ldots + 15^2 - CM = 1973 - 1850.0833 = 122.9167$$

$$SST = \frac{37^2}{4} + \frac{59^2}{4} + \frac{53^2}{4} - CM = 1914.75 - CM = 64.6667$$

$$SSE = \text{Total SS} - SST = 58.25$$

The Breakfast Problem

$$CM = \frac{149^2}{12} = 1850.0833$$

$$\text{Total SS} = 8^2 + 7^2 + \ldots + 15^2 - CM = 1973 - 1850.0833 = 122.9167$$

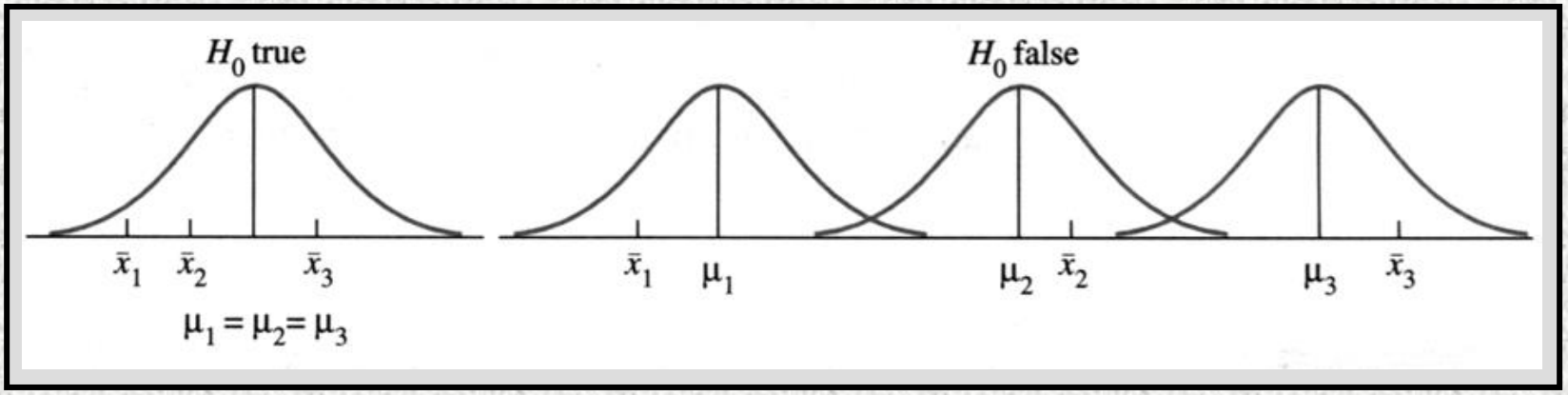
$$SST = \frac{37^2}{4} + \frac{53^2}{4} + \frac{59^2}{4} - CM = 1914.75 - CM = 64.6667$$

$$SSE = \text{Total SS} - SST = 58.25$$

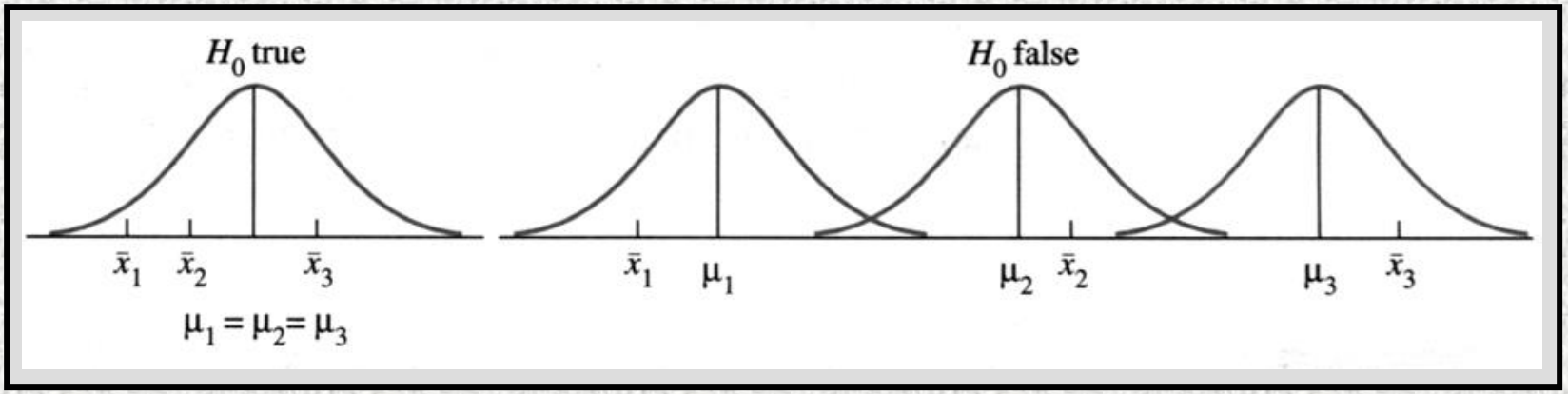
Source	df	SS	MS	F
Treatments	2	64.6667	32.3333	5.00
Error	9	58.25	6.4722	
Total	11	122.9167		

Testing the Treatment Means

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k \text{ versus}$$
$$H_1 : \text{at least one mean is different}$$

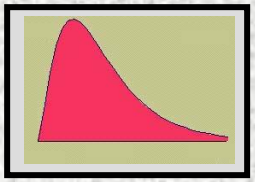


Remember that σ^2 is the common variance for all *k* populations. The quantity MSE = SSE/($n - k$) is a pooled estimate of σ^2, a weighted average of



- If H_0 is true, then the variation in the sample means, measured by MST = [SST/ (k – 1)], also provides an unbiased estimate of σ^2.
- However, if H_0 is false and the population means are different, then MST— which measures the variance in the sample means — is unusually **large.** The test statistic **F = MST/ MSE** tends to be larger that usual.

The F Test

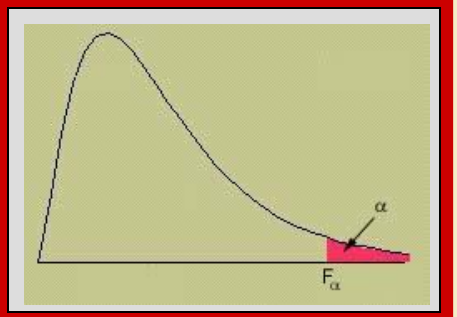


- Hence, you can reject H_0 for large values of F, using a **right-tailed** statistical test.
- When H_0 is true, this test statistic has an F distribution with $df_1 = (k - 1)$ and $df_2 = (n - k)$ degrees of freedom and **right-tailed** critical values of the F distribution can be used.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$$

$$\text{Test Statistic} : F = \frac{MST}{MSE}$$

Reject H_0 if $F > F_\alpha$ with $k - 1$ and n-k df.



The Breakfast Problem

Source	df	SS	MS	F
Treatments	2	64.6667	32.3333	5.00
Error	9	58.25	6.4722	
Total	11	122.9167		

$$H_0 : \mu_1 = \mu_2 = \mu_3 \text{ versus}$$

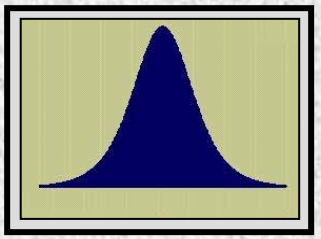
$$H_1 : \text{at least one mean is different}$$

$$F = \frac{MST}{MSE} = \frac{32.3333}{6.4722} = 5.00$$

$$\text{Rejection region} : F > F_{.05} = 4.26.$$

We reject H_0 and conclude that there is a difference in average of attention spans

Confidence Intervals



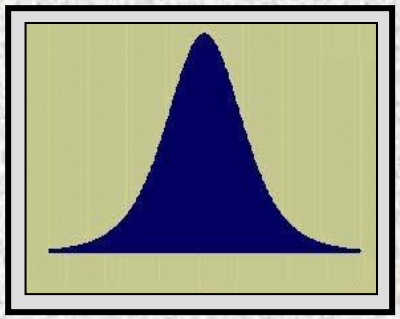
If a difference exists between the treatment means, we can explore it with confidence intervals.

A single mean, $\mu_i : \bar{x}_i \pm t_{\alpha/2} \dfrac{s}{\sqrt{n_i}}$

Difference $\mu_i - \mu_j : (\bar{x}_i - \bar{x}_j) \pm t_{\alpha/2} \sqrt{s^2 \left(\dfrac{1}{n_i} + \dfrac{1}{n_j} \right)}$

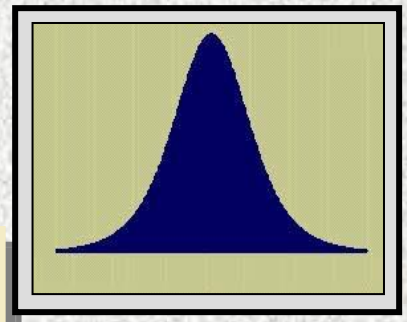
where $s = \sqrt{\text{MSE}}$ and t is based on error df.

Tukey's Method for Paired Comparisons



- Designed to test all pairs of population means simultaneously, with an **overall error rate of α**.
- Based on the **studentized range**, the difference between the largest and smallest of the k sample means.
- Assume that the **sample sizes are equal** and calculate a "ruler" that measures the distance required between any pair of means to declare a significant difference.

Tukey's Method



Calculate : $\omega = q_{\alpha}(k, df)\dfrac{s}{\sqrt{n_i}}$

where k = number of treatment means

$s = \sqrt{\text{MSE}}$ df = error df

n_i = common sample size

$q_{\alpha}(k, n-k)$ = value from Table 11.

If any pair of means differ by more than ω, they are declared different.

The Breakfast Problem

Use Tukey's method to determine which of the three population means differ from the others.

	No Breakfast	Light Breakfast	Full Breakfast
	$T_1 = 37$	$T_2 = 59$	$T_3 = 53$
Means	37/4 = 9.25	59/4 = 14.75	53/4 = 13.25

Percentage Points of the Studentized Range, $q(k, df)$; Upper 5% Points

df	k = 2	3	4	5	6
1	17.97	26.98	32.82	37.08	40.41
2	6.08	8.33	9.80	10.88	11.74
3	4.50	5.91	6.82	7.50	8.04
4	3.93	5.04	5.76	6.29	6.71
5	3.64	4.60	5.22	5.67	6.03
6	3.46	4.34	4.90	5.30	5.63
7	3.34	4.16	4.68	5.06	5.36
8	3.26	4.04	4.53	4.89	5.17
9	3.20	3.95	4.41	4.76	5.02

$$\omega = q_{.05}(3,9)\frac{s}{\sqrt{4}} = 3.95\frac{\sqrt{6.4722}}{\sqrt{4}} = 5.02$$

The Breakfast Problem

List the sample means from smallest to largest.

$\bar{x}_1$	$\bar{x}_3$	$\bar{x}_2$
9.25	13.25	14.75

$\omega = 5.02$

Since the difference between 9.25 and 13.25 is less than $\omega = 5.02$, there is no significant difference. There is a difference between population means 1 and 2 **however…**

There is no difference between 13.25 and 14.75.

We can declare a significant difference in average attention spans between “no breakfast” and “light breakfast”, but not between the other pairs.

Exercise

Twelve students were randomly assigned to three groups with three different methods of teaching statistics. At the end of the semester, the same test was given to all 12 students. The table gives the scores of students in the three groups. Construct the ANOVA table for this problem. (Hint: $\sum x_{ij}^{2} = 60551$)

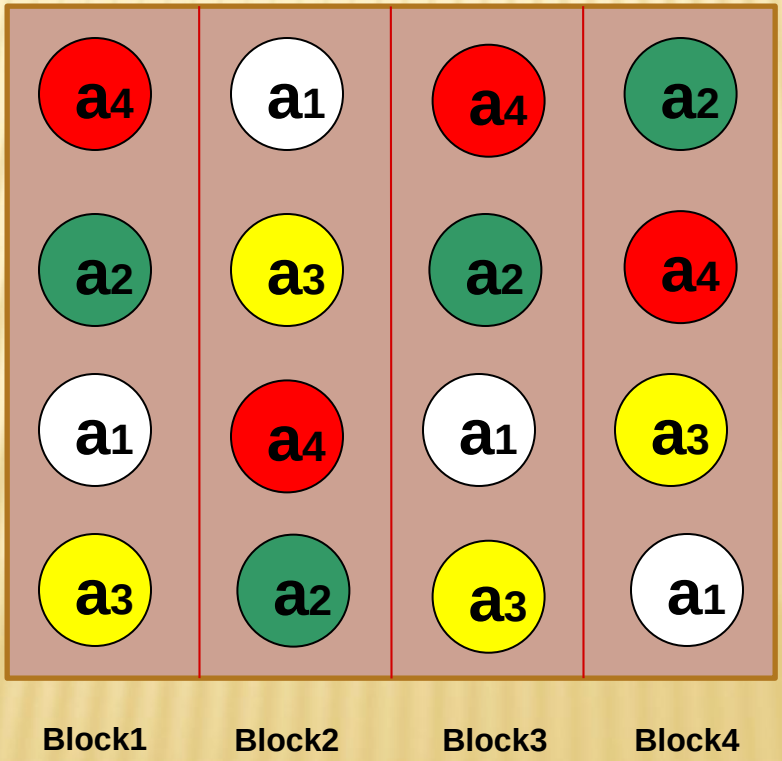
	Method I	Method II	Method III	
	48	55	84	
	73	85	68	
	51	70	95	
	65	69	74	
T_i	237	279	321	837

2. THE COMPLETELY RANDOMIZED BLOCK DESIGN (CRBD)

- A direct extension of the paired difference or matched pairs design.
- A **two-way classification** in which *k* treatment means are compared.
- The design uses **blocks** of ***k* experimental units** that are relatively similar or homogeneous, with one unit within each block randomly assigned to each treatment.

RANDOMIZATION IN CRBD

- Factor : A
- A has 4 levels (a_1, a_2,a_3, and a_4)
- Block : 4 blocks *(in CRD, this is like replication)*
- So, there fill be 4 x 4 = 16 EU



ABOUT THE CRBD

- Let x_{ij} be the response for the i-th treatment applied to the j-th block.

  $i = 1, 2, \ldots k$ $\quad$ $j = 1, 2, \ldots, b$

- The total variation in the experiment is measured by the **total sum of squares**:

$$\text{Total SS} = \sum (x_{ij} - \bar{x})^2$$

THE ANOVA TABLE OF THE CRBD

Source	df	SS	MS	F
Treatments	k -1	SST	SST/(k-1)	MST/MSE
Blocks	b -1	SSB	SSB/(b-1)	MSB/MSE
Error	(b-1)(k-1)	SSE	SSE/(b-1)(k-1)	
Total	n -1	Total SS		

SS(Sum of Squares)

$$\text{SST} = \frac{\sum_{i=1}^{k} T_i^2}{b} - \frac{\left(\sum x_{ij}\right)^2}{n}$$

$$\text{SSB} = \frac{\sum B_j^{\,2}}{k} - \frac{\left(\sum x_{ij}\right)^2}{n}$$

$$\text{Total SS} = \sum x_{ij}^2 - \frac{\left(\sum x_{ij}\right)^2}{n}$$

$$\text{SSE} = \text{Total SS} - \text{SST} - \text{SSB}$$

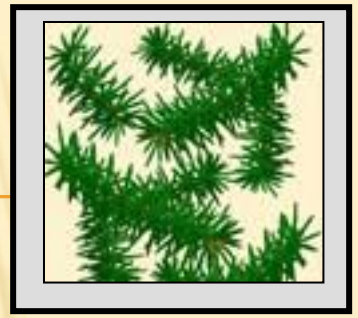
MS(Mean Squares)

MST = SST/(k-1)

MSB = SSB/(b-1)

MSE = SSE/(n-k)

THE ANOVA TABLE OF THE CRBD



The **Total SS** is divided into 3 parts:

- **SST** (sum of squares for treatments): measures the variation among the *k* treatment means
- **SSB** (sum of squares for blocks): measures the variation among the *b* block means
- **SSE** (sum of squares for error): measures the random variation or experimental error

in such a way that:

$$\text{Total SS} = \text{SST} + \text{SSB} + \text{SSE}$$

$$df\ \text{Total} = df\ \text{T} + df\ \text{B} + df\ \text{E}$$

COMPUTING FORMULAS

$$\text{CM} = \frac{\text{G}^2}{n} \quad \text{where} \quad \text{G} = \sum x_{ij}$$

$$\text{Total SS} = \sum x_{ij}^2 - \text{CM}$$

$$\text{SST} = \frac{\sum T_i^2}{b} - \text{CM} \quad \text{where } T_i = \text{total for treatment } i$$

$$\text{SSB} = \frac{\sum B_j^2}{k} - \text{CM} \quad \text{where } B_j = \text{total for block } j$$

$$\text{SSE} = \text{Total SS} - \text{SST} - \text{SSB}$$

THE SEEDLING PROBLEM

	Locations				
Soil Prep	**1**	**2**	**3**	**4**	T_i
A	11	13	16	10	50
B	15	17	20	12	64
C	10	15	13	10	48
B_j	36	45	49	32	162

$$CM = \frac{162^2}{12} = 2187$$

$$\text{Total SS} = \left(11^2 + 15^2 + \ldots + 10^2\right) - 2187 = 111$$

$$SST = \frac{50^2 + 64^2 + 48^2}{4} - 2187 = 38$$

$$SSB = \frac{36^2 + 45^2 + 49^2 + 32^2}{3} - 2187 = 61.6667$$

$$SSE = 111 - 38 - 61.6667 = 11.3333$$

THE SEEDLING PROBLEM-ANOVA

Total *df* = $bk - 1 = n - 1$

Treatment *df* = $k - 1$

Block *df* = $b - 1$

Error *df* = $bk - (k - 1) - (b-1) = (k-1)(b-1)$

Mean Squares

$MST = SST/(k-1)$

$MSB = SSB/(b-1)$

$MSE = SSE/(k-1)(b-1)$

Source	**df**	**SS**	**MS**	**F**
Treatments	k -1	SST	SST/(k-1)	MST/MSE
Blocks	b -1	SSB	SSB/(b-1)	MSB/MSE
Error	(b-1)(k-1)	SSE	SSE/(b-1)(k-1)	
Total	n -1	Total SS		

THE SEEDLING PROBLEM-ANOVA

$$CM = \frac{162^2}{12} = 2187$$

$$\text{Total SS} = 11^2 + 15^2 + \ldots + 10^2 - 2187 = 111$$

$$SST = \frac{50^2 + 64^2 + 48^2}{4} - 2187 = 38$$

$$SSB = \frac{36^2 + 45^2 + 49^2 + 32^2}{3} - 2187 = 61.6667$$

$$SSE = 111 - 38 - 61.6667 = 11.3333$$

Source	df	SS	MS	F
Treatments	2	38	19	10.06
Blocks	3	61.6667	20.5556	10.88
Error	6	11.3333	1.8889	
Total	11	122.9167		

TESTING THE TREATMENT AND BLOCK MEANS

For either treatment or block means, we can test:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots \quad \text{versus}$$

$$H_1 : \text{at least one mean is different}$$

To test the H_0 that treatment (or block) means are equal

$$\text{Test Statistic} : F = \frac{MST}{MSE} \left(\text{or } F = \frac{MSB}{MSE}\right)$$

Reject H_0 if $F > F_\alpha$ with k-1 (or $b-1$) and $(b-1)(k-1)$ df.

TESTING THE TREATMENT AND BLOCK MEANS

- Remember that σ^2 is the common variance for all ***bk*** treatment/block combinations. MSE is the best estimate of σ^2, whether or not H_0 is true.

- If H_0 is false and the population means are different, then MST or MSB— whichever you are testing— will unusually **large.**

  The test statistic **F = MST/ MSE**(or **F = MSB/ MSE)** tends to be larger that usual.

- We use a right-tailed F test with the appropriate degrees of freedom.

THE SEEDLING PROBLEM-ANOVA

Source	df	SS	MS	F
Soil Prep (Trts)	2	38	19	10.06
Location (Blocks)	3	61.6667	20.5556	10.88
Error	6	11.3333	1.8889	
Total	11	122.9167		

To test for a difference due to soil preparation:

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

H_a: at least one mean is different

$$F = \frac{MST}{MSE} = 10.06$$

Rejection region: $F > F_{.05} = 5.14$.

We reject H_0 and conclude that there is a difference of seedling growth due to soil preparation.

Although not of primary importance, notice that the blocks (locations) were also significantly different (F = 10.88)

df_2	a	1	2
1	.100	39.86	49.50
	.050	161.4	199.5
	.025	647.8	799.5
	.010	4052	4999.5
	.005	16211	20000
2	.100	8.53	9.00
	.050	18.51	19.00
	.025	38.51	39.00
	.010	98.50	99.00
	.005	198.5	199.0
3	.100	5.54	5.46
	.050	10.13	9.55
	.010	21.20	18.00
	.005	31.33	26.28
5	.100	4.06	3.78
	.050	6.61	5.79
	.025	10.01	8.43
	.010	16.26	13.27
	.005	22.78	18.31
6	.100	3.78	3.46
	.050	5.99	5.14
	.025	8.81	7.26
	.010	13.75	10.92
	.005	18.63	14.54

CONFIDENCE INTERVALS

If a difference exists between the treatment means or block means, we can explore it with confidence intervals or using Tukey's method.

Difference in treatment means :

$$(\bar{T}_i - \bar{T}_j) \pm t_{\alpha/2}\sqrt{s^2\left(\frac{2}{b}\right)}$$

Difference in block means :

$$(\bar{B}_i - \bar{B}_j) \pm t_{\alpha/2}\sqrt{s^2\left(\frac{2}{k}\right)}$$

where $\bar{T}_i = T_i / b$ and $\bar{B}_i = B_i / k$ are the necessary treatment or block means.

$s = \sqrt{\text{MSE}}$ and t is based on error df.

TUKEY'S METHOD

For comparing treatment means : $\omega = q_\alpha(k, df)\dfrac{s}{\sqrt{b}}$

For comparing block means : $\omega = q_\alpha(b, df)\dfrac{s}{\sqrt{k}}$

where :

$s = \sqrt{\text{MSE}}$ df = error df

$q_\alpha(k, df)$ = value from Table 11.

If any pair of means differ by more than ω, they are declared different.

THE SEEDLING PROBLEM – TUKEY'S METHOD

Use Tukey's method to determine which of the three soil preparations differ from the others.

	A (no prep)	B (fertilization)	C (burning)
	$T_1 = 50$	$T_2 = 64$	$T_3 = 48$
Means	50/4 = 12.5	64/4 = 16	48/4 = 12

$$\omega = q_{.05}(3,6)\frac{s}{\sqrt{4}}$$

$$= 4.34\frac{\sqrt{1.8889}}{\sqrt{4}} = 2.98$$

TABLE 11(a)
Percentage Points of the Studentized Range, $q(k, df)$; Upper 5% Points

	k				
df	2	3	4	5	6
1	17.97	26.98	32.82	37.08	40.41
2	6.08	8.33	9.80	10.88	11.74
3	4.50	5.91	6.82	7.50	8.04
4	3.93	5.04	5.76	6.29	6.71
5	3.64	4.60	5.22	5.67	6.03
6	3.46	4.34	4.90	5.30	5.63
7	3.34	4.16	4.68	5.06	5.36
8	3.26	4.04	4.53	4.89	5.17
9	3.20	3.95	4.41	4.76	5.02

THE SEEDLING PROBLEM – TUKEY'S METHOD

List the sample means from smallest to largest.

$\bar{T}_C$	$\bar{T}_A$	$\bar{T}_B$
12	12.5	16.0

$$\omega = 2.98$$

Since the difference between 12 and 12.5 is less than $\omega = 2.98$, there is no significant difference. There is a difference between population means C and B however.

There is a significant difference between A and B.

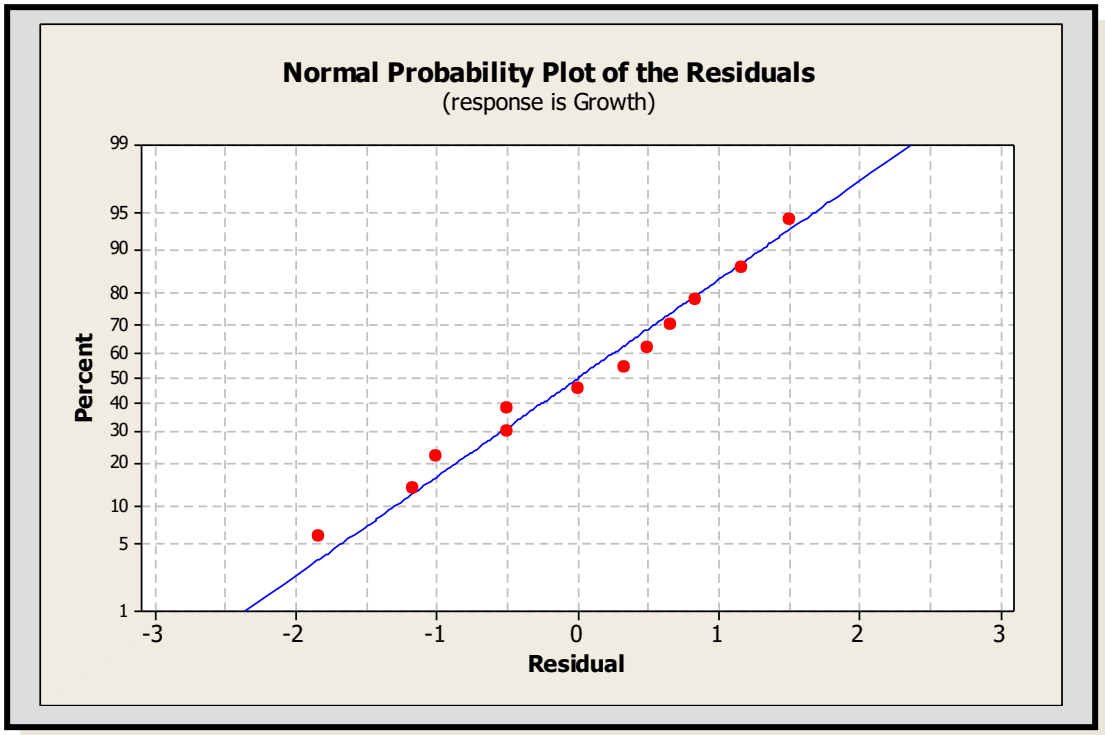
A significant difference in average growth only occurs when the soil has been fertilized.

CAUTIONS ABOUT BLOCKING

- A randomized block design should not be used when **both treatments and blocks** correspond to experimental factors of interest to the researcher
- Remember that **blocking may not always be beneficial**.
- Remember that you **cannot construct confidence intervals for individual treatment means** unless it is reasonable to assume that the *b* blocks have been randomly selected from a population of blocks.

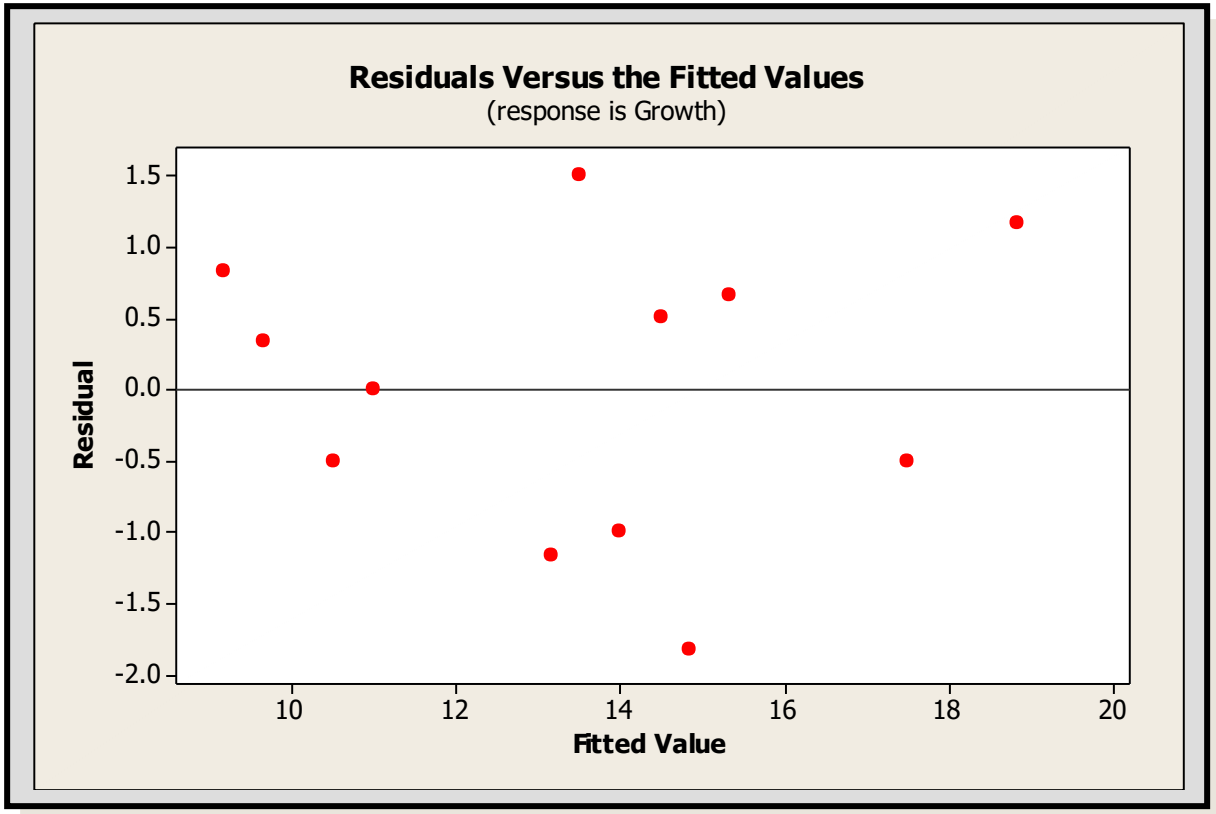
Normal Probability Plot

- ✓ If the normality assumption is valid, the plot should resemble a straight line, sloping upward to the right.
- ✓ If not, you will often see the pattern fail in the tails of the graph.



Residuals versus Fits

- ✓ If the equal variance assumption is valid, the plot should appear as a random scatter around the zero center line.
- ✓ If not, you will see a pattern in the residuals.



Exercise

Suppose you wish to compare the means of four populations based on independent random samples, each of which contain 6, 8, 10 and 9 observations respectively. What is the degrees of freedom for the F statistic?

Exercise

Complete the ANOVA table.

Source	SS	df	MS	F
Treatments	a	2	5.7	f
Blocks	17.1	c	e	2.41
Error	b	d	1.42	
Total	42.7	17		