RESEARCH METHODOLOGY

UE20CS506A

Unit-03:

Testing of hypotheses and Data Analysis

Raghu B. A
Priya B.
Santhosh Kumar V
Department of Computer Science & Engineering

RESEARCH METHODOLOGY

Topic: Basic concepts - Procedure for hypothesis testing, flow diagram for hypothesis testing

Santhosh Kumar V
Department of Computer Science & Engineering

ANOVA: Analysis of Variance

- Used of hypothesis testing when >2 population/samples cases are involved

1. *Population normal, population infinite, sample size may be large or small but variance of the population is known, H_a may be one-sided or two-sided:*

In such a situation z-test is used for testing hypothesis of mean and the test statistic z is worked our as under:

$$z = \frac{\bar{X} - \mu_{H_0}}{\sigma_p / \sqrt{n}}$$

3. *Population normal, population infinite, sample size small and variance of the population unknown, H_a may be one-sided or two-sided:*

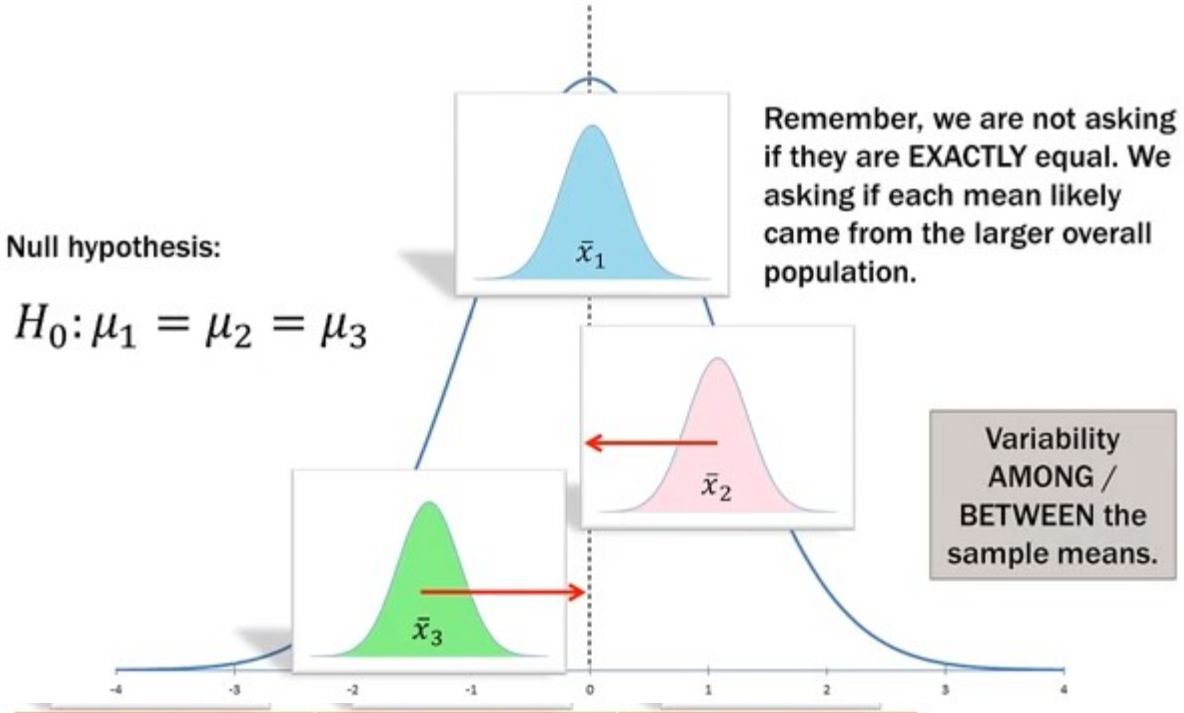
In such a situation t-test is used and the test statistic t is worked out as under:

$$t = \frac{\bar{X} - \mu_{H_0}}{\sigma_s / \sqrt{n}} \text{ with d.f.} = (n-1)$$

$$\sigma_s = \sqrt{\frac{\Sigma\left(X_i - \bar{X}\right)^2}{(n-1)}}$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\Sigma\left(X_{1i} - \bar{X}_1\right)^2 + \Sigma\left(X_{2i} - \bar{X}_2\right)^2}{n_1 + n_2 - 2}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

ANOVA

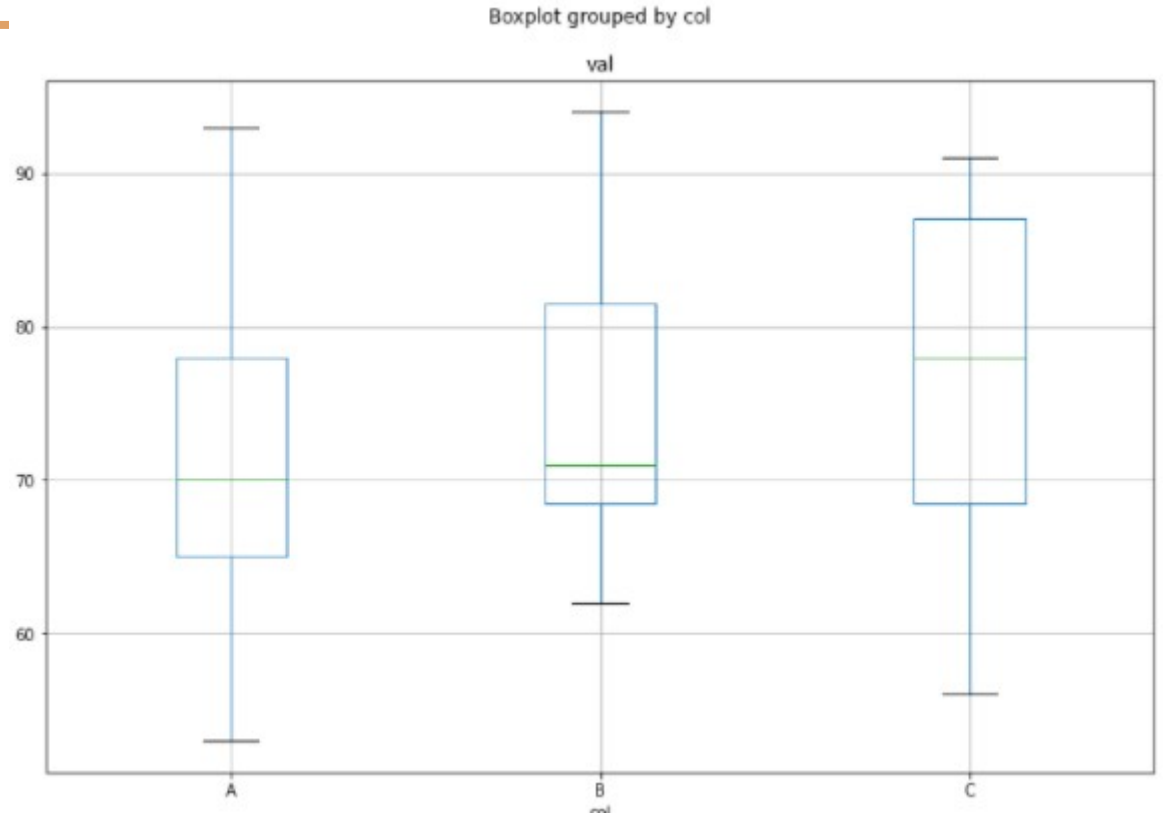


Year 1 Scores	Year 2 Scores	Year 3 Scores
82	71	64
93	62	73
61	85	87
74	94	91
69	78	56
70	66	78
53	71	87
$\bar{x}_1 = 71.71$	$\bar{x}_2 = 75.29$	$\bar{x}_3 = 76.57$

Overall Mean:

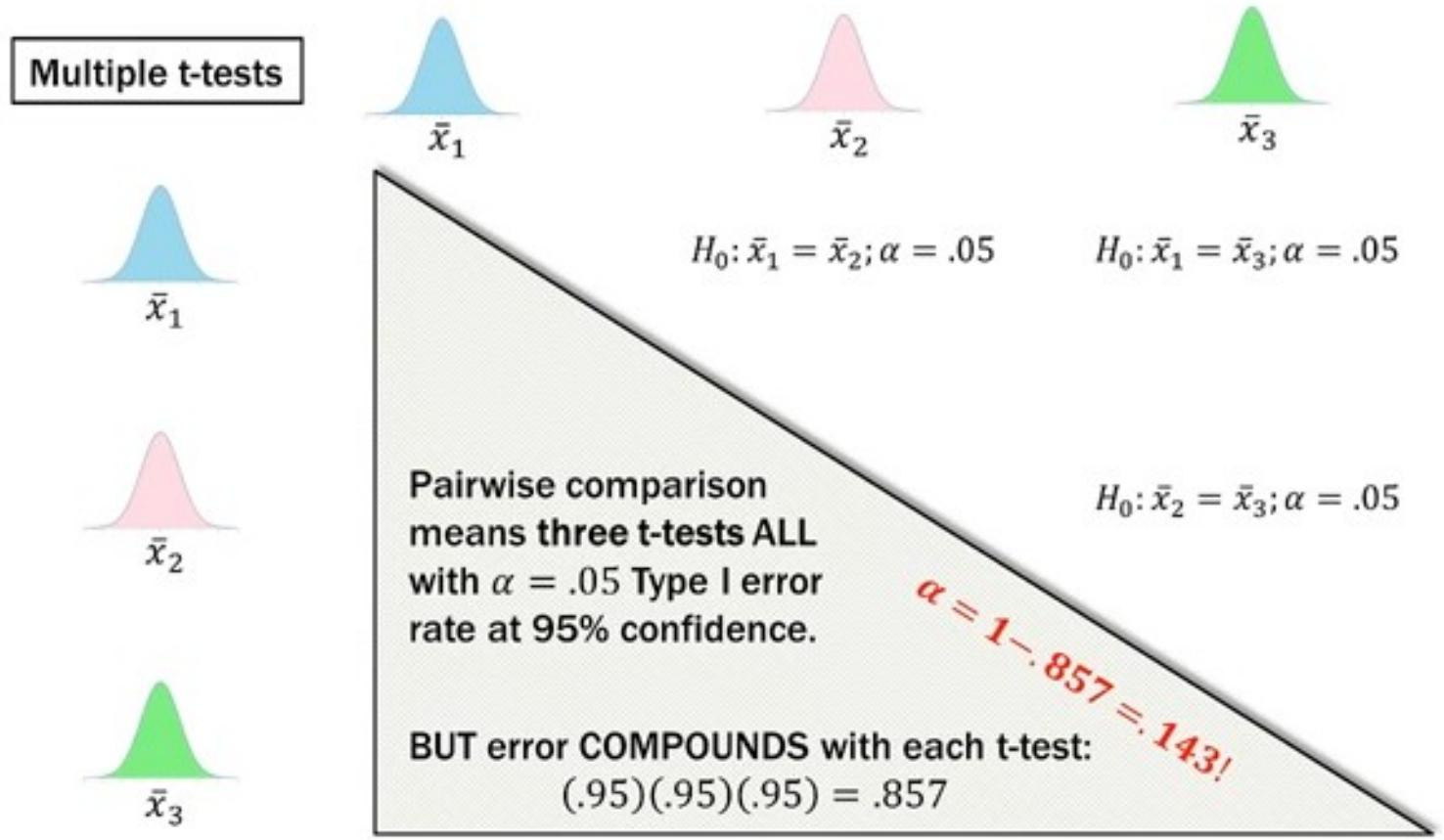
The mean of all 21 scores taken together.

$$\bar{\bar{x}} = 74.52$$

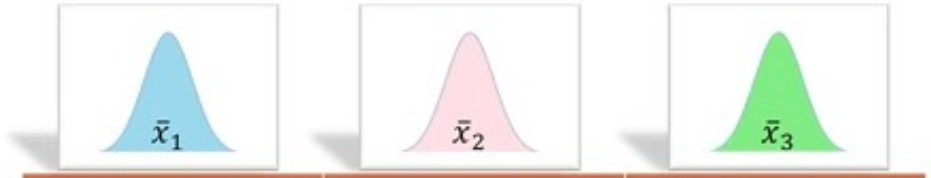


Ref: statistics 101: Anova visual

Multiple t-test (not a solution)



ANOVA



Year 1 Scores	Year 2 Scores	Year 3 Scores
82	71	64
93	62	73
61	85	87
74	94	91
69	78	56
70	66	78
53	71	87
$\bar{x}_1 = 71.71$	$\bar{x}_2 = 75.29$	$\bar{x}_3 = 76.57$

Overall Mean:

The mean of all 21 scores taken together.

$$\bar{\bar{x}} = 74.52$$

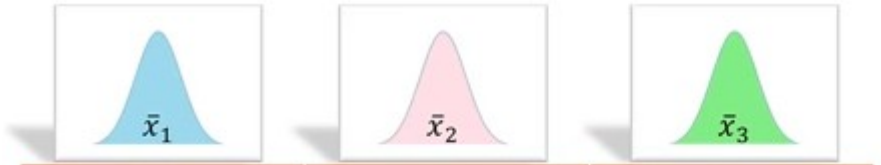
Year 1 Scores	Year 2 Scores	Year 3 Scores
82	71	64
93	62	73
61	85	87
74	94	91
69	78	56
70	66	78
53	71	87
$\bar{x}_1 = 71.71$	$\bar{x}_2 = 75.29$	$\bar{x}_3 = 76.57$

SST
(total / overall) sum of squares

1. Find difference between each data point and the overall mean.
2. Square the difference.
3. Add them up

$$\bar{\bar{x}} = 74.52$$

ANOVA

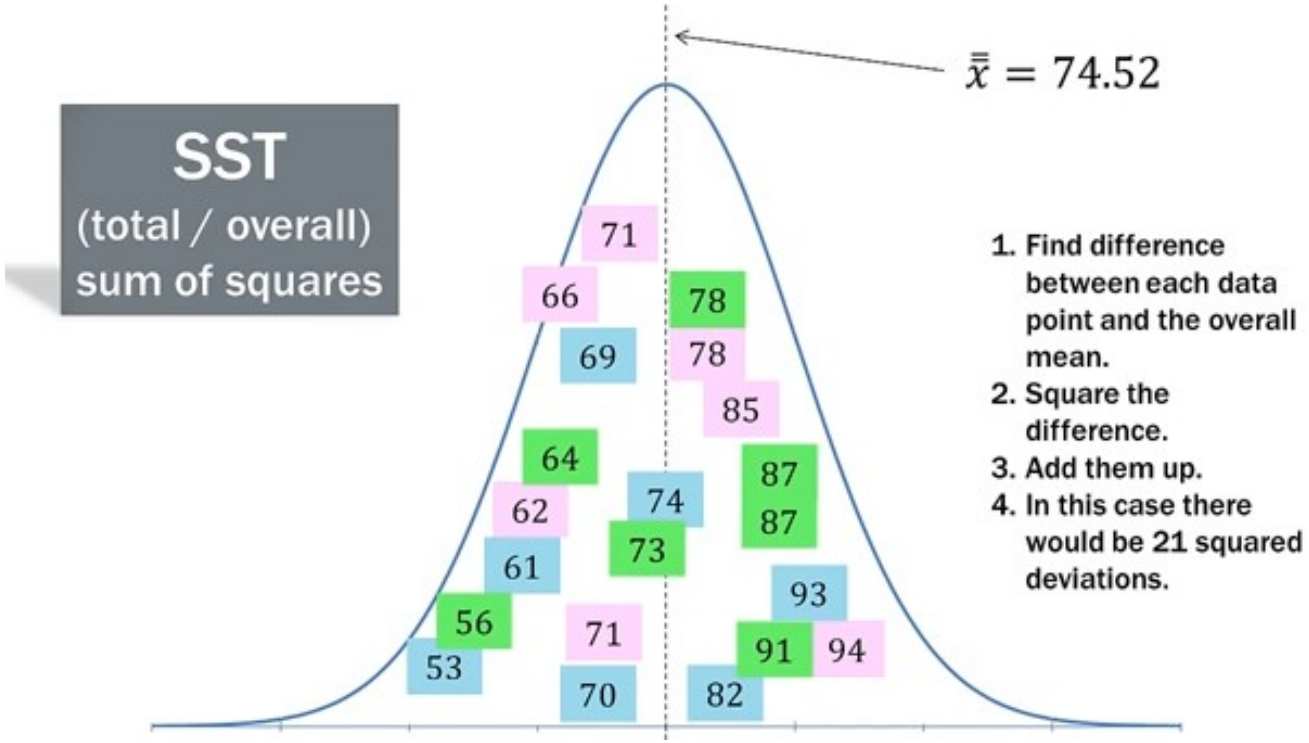


Year 1 Scores	Year 2 Scores	Year 3 Scores
82	71	64
93	62	73
61	85	87
74	94	91
69	78	56
70	66	78
53	71	87
$\bar{x}_1 = 71.71$	$\bar{x}_2 = 75.29$	$\bar{x}_3 = 76.57$

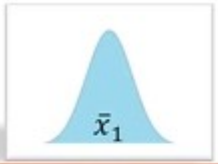
SST
(total / overall)
sum of squares

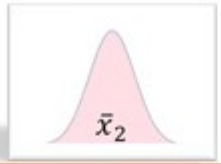
1. Find difference between each data point and the overall mean.
2. Square the difference.
3. Add them up

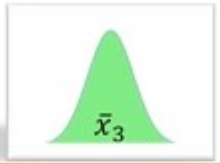
$\bar{\bar{x}} = 74.52$



ANOVA





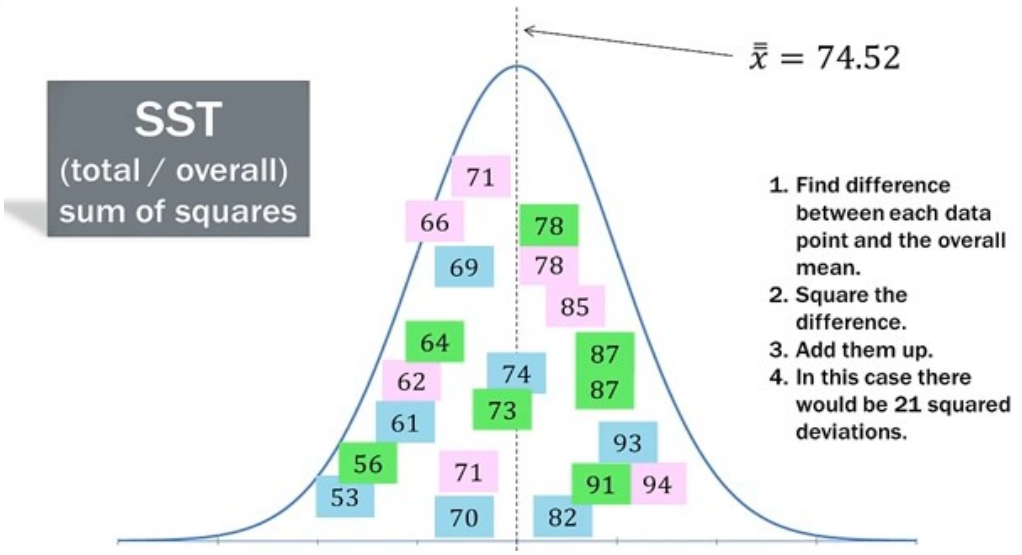


Year 1 Scores	Year 2 Scores	Year 3 Scores
82	71	64
93	62	73
61	85	87
74	94	91
69	78	56
70	66	78
53	71	87
$\bar{x}_1 = 71.71$	$\bar{x}_2 = 75.29$	$\bar{x}_3 = 76.57$

SST
(total / overall)
sum of squares

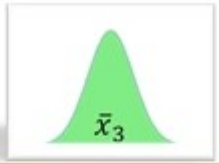
1. Find difference between each data point and the overall mean.
2. Square the difference.
3. Add them up

$\bar{\bar{x}} = 74.52$



SST
(total / overall)
sum of squares

1. Find difference between each data point and the overall mean.
2. Square the difference.
3. Add them up.
4. In this case there would be 21 squared deviations.

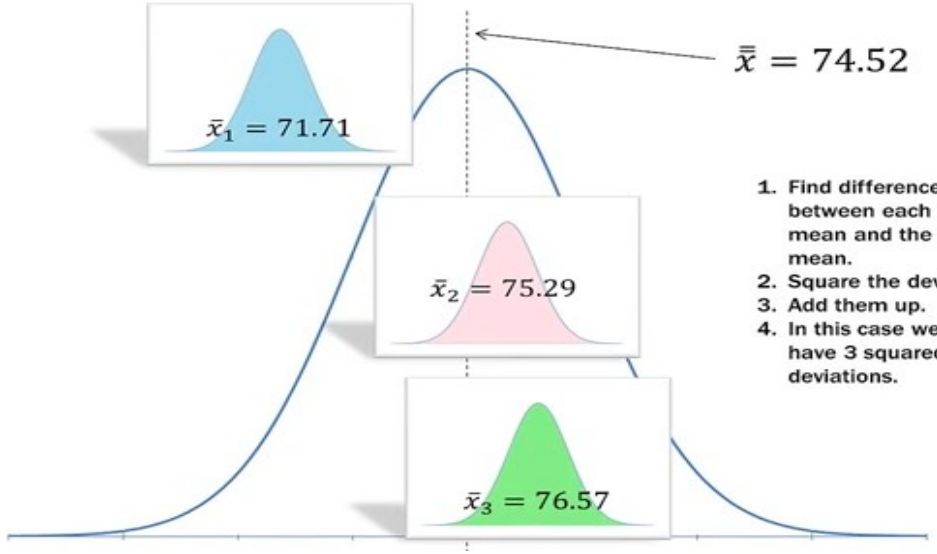


Year 1 Scores	Year 2 Scores	Year 3 Scores
82	71	64
93	62	73
61	85	87
74	94	91
69	78	56
70	66	78
53	71	87
$\bar{x}_1 = 71.71$	$\bar{x}_2 = 75.29$	$\bar{x}_3 = 76.57$

SSC
(column/ between)
sum of squares

$\bar{\bar{x}} = 74.52$

1. Find difference between each group mean and the overall mean.
2. Square the deviations.
3. Add them up.
4. In this case we would have 3 squared deviations.



1. Find difference between each sample mean and the overall mean.
2. Square the deviations.
3. Add them up.
4. In this case we would have 3 squared deviations.

ANOVA

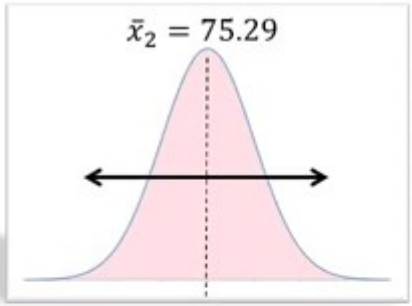
SSE
(within / error)
sum of squares

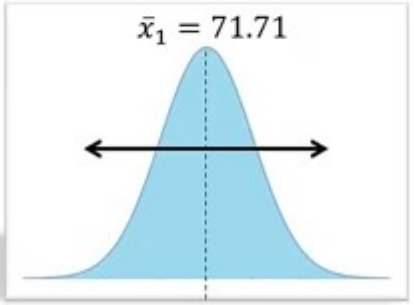
$\bar{\bar{x}} = 74.52$

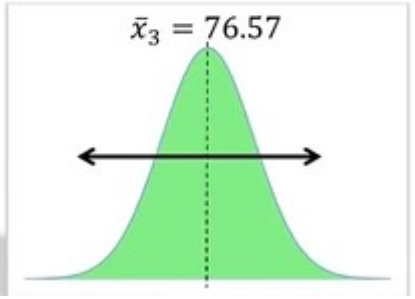
1. Find difference between each data point and its **column** mean.
2. Square each deviation.
3. Add them up the squared deviations.
4. In this case we would have 21 squared deviations.

SSE
(within / error)
sum of squares

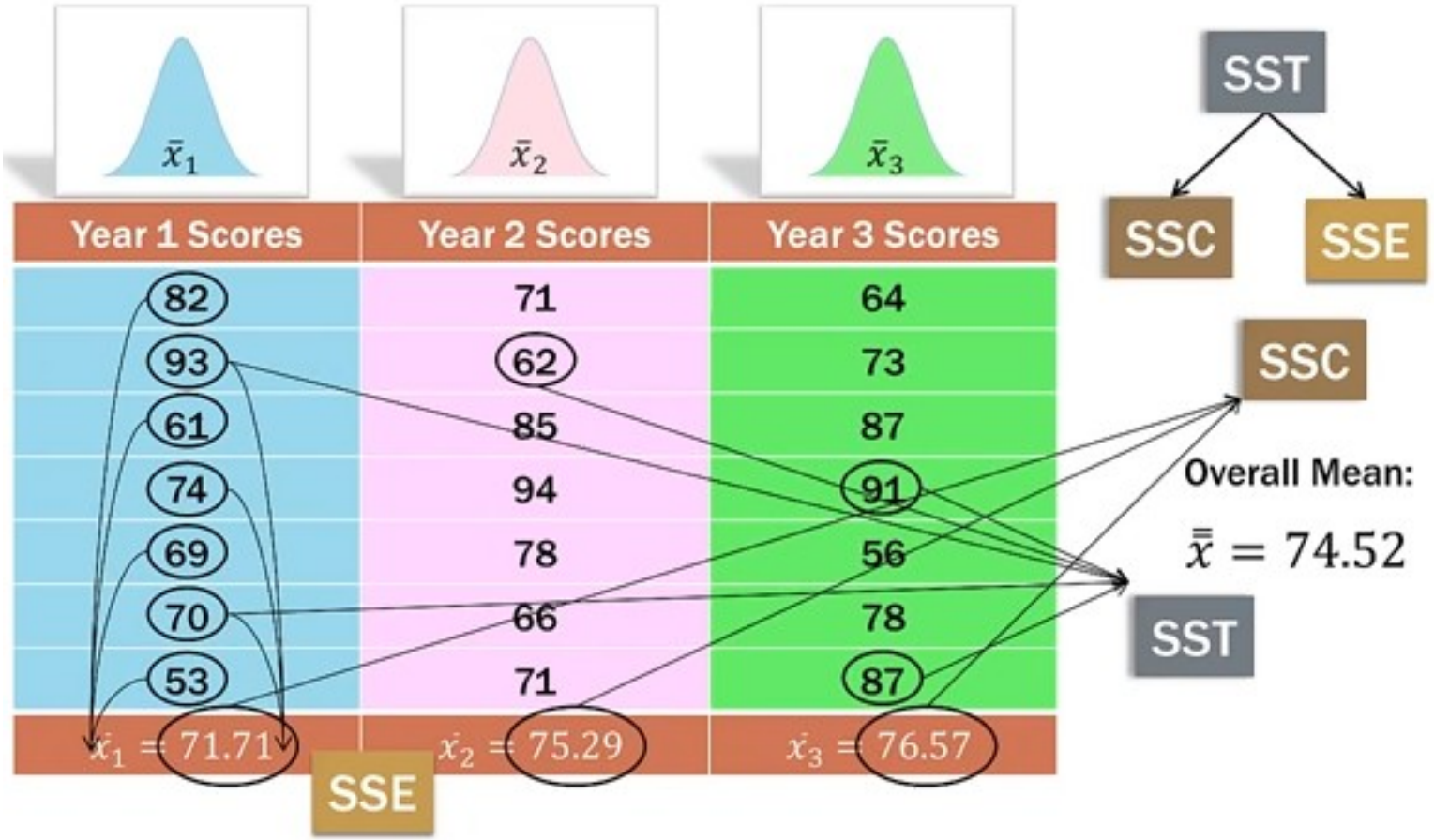
1. Find difference between each data point and its **column** mean.
2. Square each deviation.
3. Add them up the squared deviations.
4. In this case we would have 21 squared deviations.







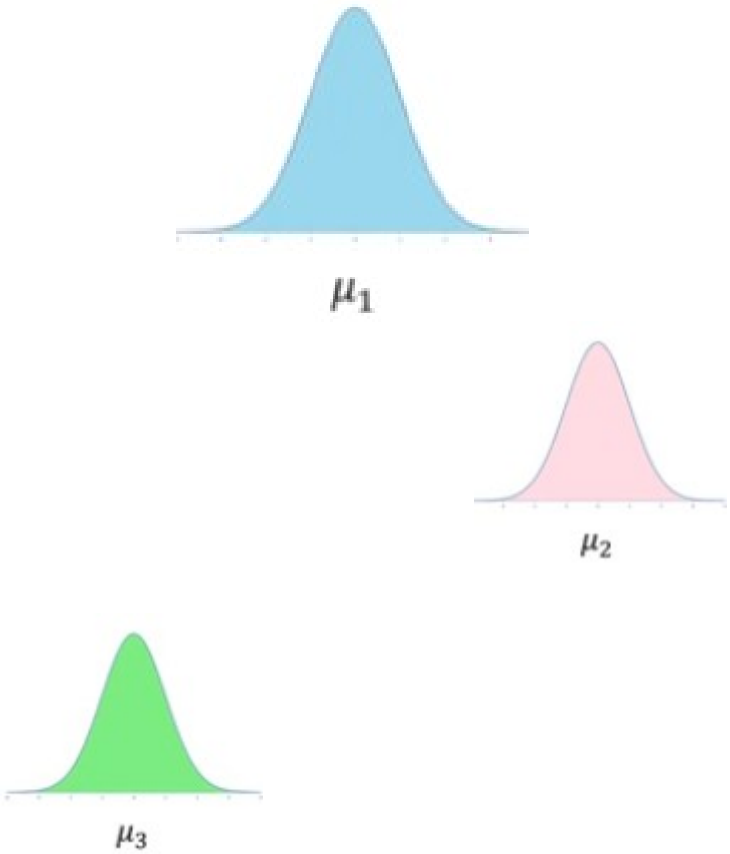
ANOVA


$\bar{x}_1$
$\bar{x}_2$
$\bar{x}_3$
Year 1 Scores
Year 2 Scores
Year 3 Scores
82
93
61
74
69
70
53
71
62
85
94
78
66
71
64
73
87
91
56
78
87
$\bar{x}_1 = 71.71$
$\bar{x}_2 = 75.29$
$\bar{x}_3 = 76.57$
SSE
SST
SSC
SSE
SSC
Overall Mean:
$\bar{\bar{x}} = 74.52$
SST

ANOVA

Compare 3 population means to see if they are different

Do all the 3 means come from the same population



Is one mean so far away , it is from a different population

Do all of these come from different population

Per Acre yeild			
Plot of land	Variety of Wheat		
	A	B	C
1	6	5	5
2	7	5	4
3	3	3	3
4	8	7	4

ANOVA

Per Acre yield			
Plot of land	Variety of Wheat		
	A	B	C
1	6	5	5
2	7	5	4
3	3	3	3
4	8	7	4
	6	5	4

$\bar{X}_1, \bar{X}_2, \bar{X}_3, \ldots, \bar{X}_k$ $\qquad \bar{\bar{X}} = \dfrac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \ldots + \bar{X}_k}{\text{No. of samples } (k)}$

$$SS \text{ between} = n_1\left(\bar{X}_1 - \bar{\bar{X}}\right)^2 + n_2\left(\bar{X}_2 - \bar{\bar{X}}\right)^2 + \ldots + n_k\left(\bar{X}_k - \bar{\bar{X}}\right)^2$$

$$MS \text{ between} = \frac{SS \text{ between}}{(k-1)}$$

$$SS \text{ within} = \sum\left(X_{1i} - \bar{X}_1\right)^2 + \sum\left(X_{2i} - \bar{X}_2\right)^2 + \ldots + \sum\left(X_{ki} - \bar{X}_k\right)^2$$

$$MS \text{ within} = \frac{SS \text{ within}}{(n-k)}$$

$$F\text{-ratio} = \frac{MS \text{ between}}{MS \text{ within}}$$

$$SS \text{ for total variance} = \sum\left(X_{ij} - \bar{\bar{X}}\right)^2$$

SS for total variance = SS between + SS within.

$(n-1) = (k-1) + (n-k)$

n = total number of items in all the samples
i.e., $n_1 + n_2 + \ldots + n_k$

k = number of samples

ANOVA

Source of Variation	Sum of Squares (SS)	Deg of Freedom	Mean Sqaure(MS)	F- Ratio
Between	SS Between	(k-1)	MS Between = SS Between/(k-1)	MS between / MS Within
Within	SS Within	(n-k)	MS Within = SS within/(n-k)	
Total	SS Total	(n -1)		

$$SS \text{ between} = n_1\left(\bar{X}_1 - \bar{\bar{X}}\right)^2 + n_2\left(\bar{X}_2 - \bar{\bar{X}}\right)^2 + \ldots + n_k\left(\bar{X}_k - \bar{\bar{X}}\right)^2$$

$$SS \text{ within} = \sum\left(X_{1i} - \bar{X}_1\right)^2 + \sum\left(X_{2i} - \bar{X}_2\right)^2 + \ldots + \sum\left(X_{ki} - \bar{X}_k\right)^2$$

$$SS \text{ for total variance} = \sum\left(X_{ij} - \bar{\bar{X}}\right)^2$$

ANOVA

Per Acre yield			
Plot of land	Variety of Wheat		
	A	B	C
1	6	5	5
2	7	5	4
3	3	3	3
4	8	7	4
	6	5	4

$$\bar{X}_1 = \frac{6+7+3+8}{4} = 6$$

$$\bar{X}_2 = \frac{5+5+3+7}{4} = 5$$

$$\bar{X}_3 = \frac{5+4+3+4}{4} = 4$$

$$\bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3}{k}$$

$$= \frac{6+5+4}{3} = 5$$

$$SS \text{ for total variance} = \sum \left(X_{ij} - \bar{\bar{X}} \right)^2 \qquad i = 1, 2, 3\ldots \quad j = 1, 2, 3\ldots$$

$$= (6-5)^2 + (7-5)^2 + (3-5)^2 + (8-5)^2 + (5-5)^2 + (5-5)^2 + (3-5)^2 + (7-5)^2 + (5-5)^2 + (4-5)^2 + (3-5)^2 + (4-5)^2$$

$$= 1+4+4+9+0+0+4+4+0+1+4+1$$

$$= 32$$

$$SS \text{ between} = n_1 \left(\bar{X}_1 - \bar{\bar{X}} \right)^2 + n_2 \left(\bar{X}_2 - \bar{\bar{X}} \right)^2 + n_3 \left(\bar{X}_3 - \bar{\bar{X}} \right)^2$$

$$= 4(6-5)^2 + 4(5-5)^2 + 4(4-5)^2$$

$$= 4 + 0 + 4$$

$$= 8$$

$$SS \text{ within} = \sum \left(X_{1i} - \bar{X}_1 \right)^2 + \sum \left(X_{2i} - \bar{X}_2 \right)^2 + \sum \left(X_{3i} - \bar{X}_3 \right)^2.$$

$$= \{(6-6)^2 + (7-6)^2 + (3-6)^2 + (8-6)^2\} + \{(5-5)^2 + (5-5)^2 + (3-5)^2 + (7-5)^2\} + \{(5-4)^2 + (4-4)^2 + (3-4)^2 + (4-4)^2\}$$

$$= \{0+1+9+4\} + \{0+0+4+4\} + \{1+0+1+0\}$$

$$= 14 + 8 + 2$$

$$= 24$$

n = total number of items in all the sample
i.e., $n_1 + n_2 + \ldots + n_k$

k = number of samples

ANOVA

Per Acre yield			
Plot of land	ariety of Wheat		
	A	B	C
1	6	5	5
2	7	5	4
3	3	3	3
4	8	7	4
	6	5	4

Source of variation	*SS*	*d.f.*	*MS*	*F-ratio*	*5% F-limit (from the F-table)*
Between sample	8	$(3-1)=2$	$8/2=4.00$	$4.00/2.67=1.5$	$F(2,9)=4.26$
Within sample	24	$(12-3)=9$	$24/9=2.67$		
Total	32	$(12-1)=11$			

n = total number of items in all the samples
i.e., $n_1 + n_2 + \ldots + n_k$

k = number of samples

ANOVA

Per Acre yield			
Plot of land	ariety of Wheat		
	A	B	C
1	6	5	5
2	7	5	4
3	3	3	3
4	8	7	4
	6	5	4

n = total number of items in all the samples
i.e., $n_1 + n_2 + \ldots + n_k$

k = number of samples

Source of variation	*SS*	*d.f.*	*MS*	*F-ratio*	*5% F-limit (from the F-table)*
Between sample	8	$(3-1)=2$	$8/2=4.00$	$4.00/2.67=1.5$	$F(2,9)=4.26$
Within sample	24	$(12-3)=9$	$24/9=2.67$		
Total	32	$(12-1)=11$			

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
A	4	24	6	4.666667
B	4	20	5	2.666667
C	4	16	4	0.666667

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	8	2	4	1.5	0.274016	4.256495
Within Groups	24	9	2.666667			
Total	32	11				

THANK YOU

Raghu B.A
Priya B.
Santhosh Kumar V.
Department of Computer Science & Engineering