ANOVA

Presented By:
Dr. Vinay Kulkarni

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

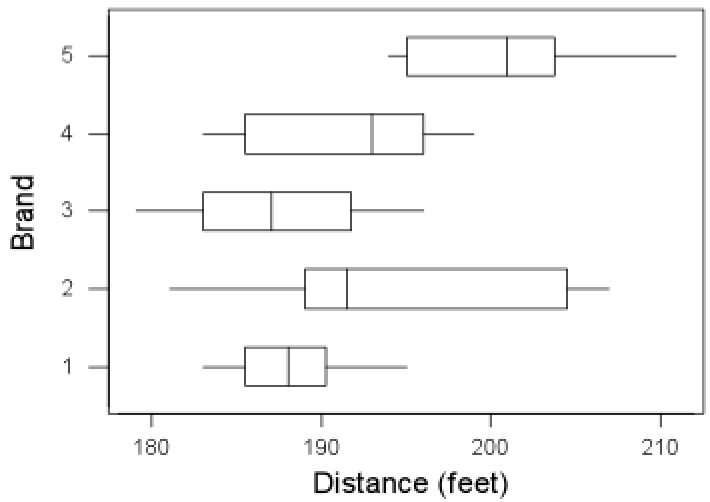
ANOVA: Analysis of Variance

- It is a procedure to test three or more population means
- It works by comparing:
  - Variability **within** the samples, and
  - Variability **between** the samples
- It uses
  - F-distribution (named after Ronal A. Fisher)
    - Similar to Chi-square distribution (it is a ratio of two chi-sq)
    - It is right skewed
    - Non-negative
    - Infinite number of curves
    - Characterized by d.o.f. of numerator & denominator

A Problem: Comparison of braking distances

- Breaking distance related to 5 tyres

Brand1	Brand2	Brand3	Brand4	Brand5
194	189	185	183	195
184	204	183	193	197
189	190	186	184	194
189	190	183	186	202
188	189	179	194	200
186	207	191	199	211
195	203	188	196	203
186	193	196	188	206
183	181	189	193	202
188	206	194	196	195



Brand	N	MEAN	SD
1	10	188.20	3.88
2	10	195.20	9.02
3	10	187.40	5.27
4	10	191.20	5.55
5	10	200.50	5.44

Are the means same? Or
Do they differ?

One Way ANOVA

- One way ANOVA used to determine
  - Significant differences between three or more independent populations
- ANOVA compares the means among the groups
- Determines if the means are significant from each other
- It tests the null hypothesis:

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

- The alternate hypothesis
  - At least two population “means” are significantly different from each other

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

ANOVA: The basis of the test

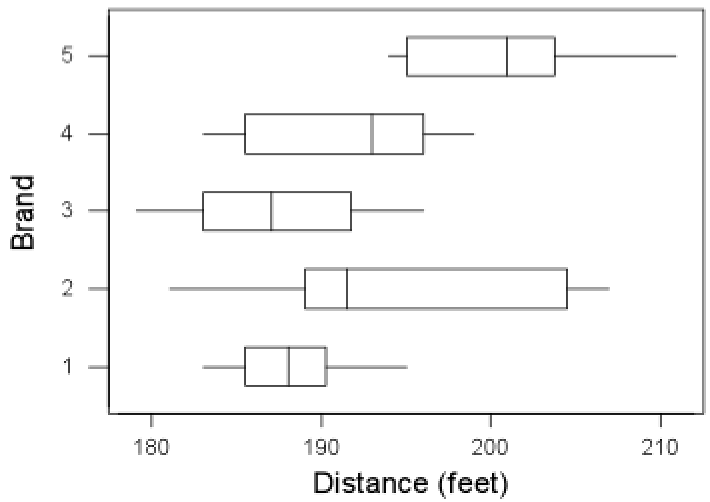
- ANOVA produces an F-statistic
  - Ratio of variance among the means to variance within the samples
  - Logic:
    - If groups are drawn from populations with the same mean values …
    - **Variance between the group mean should be lower than variance within the samples**
    - Higher ratio of variance between the means → variance within the sample ➔ drawn from different populations

SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

ANOVA: Example

- Breaking distance related to 5 tyres

Brand1	Brand2	Brand3	Brand4	Brand5
194	189	185	183	195
184	204	183	193	197
189	190	186	184	194
189	190	183	186	202
188	189	179	194	200
186	207	191	199	211
195	203	188	196	203
186	193	196	188	206
183	181	189	193	202
188	206	194	196	195



Brand	N	MEAN	SD
1	10	188.20	3.88
2	10	195.20	9.02
3	10	187.40	5.27
4	10	191.20	5.55
5	10	200.50	5.44

Analysis of Variance
for comparing all 5 brands

Source	DF	SS	MS	F	P
Brand	4	1174.8	293.7	7.95	0.000
Error	45	1661.7	36.9		
Total	49	2836.5			

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

The Anova Table

One-way Analysis of Variance

Source	DF	SS	MS	F	P
Factor	m-1	SS(Between)	MSB	MSB/MSE	
Error	n-m	SS(Error)	MSE		
Total	n-1	SS(Total)			

From F-distribution with m-1 numerator and n-m denominator d.f.

MSB = SS(Between)/(m-1)
MSE = SS(Error)/(n-m)

n-1 = (m-1) + (n-m)

SS(Total) = SS(Between) + SS(Error)

(1) **Source** means "the source of the variation in the data."

(2) **DF** means "the degrees of freedom in the source."

(3) **SS** means "the sum of squares due to the source."

(4) **MS** means "the mean sum of squares due to the source."

(5) **F** means "the *F*-statistic."

(6) **P** means "the *P*-value."

ANOVA: Method

Source of Variation	df	SS	MS	E(MS)	F-ratio
$Treatments(between Groups)$	k-1	$\sum_j^k \frac{y_{.j}^2}{n_j} - \frac{y_{..}^2}{N}$	$\frac{SS_{treat}}{k-1}$	$\sigma^2 + \frac{\sum_{j=1}^k n_j t_j^2}{k-1}$	$\frac{MS_{treat}}{MS_{error}}$
Error	N – k	Difference	$\frac{SSerror}{N-k}$	σ^2	
Total	N -1	$\sum\sum y^2{}_{ij} - \frac{y_{..}^2}{N}$			

The hypotheses to be tested are

$$H_0 : t_1 = t_2 = ... = t_k = 0 \quad agianst$$
$$H_1 : some\, t_j \neq 0, for\ some\ j.$$

The test statistic is $F = \frac{MS_{treat}}{MS_{error}}$, which has an f-distribution with (k-1, N-k) degrees of freedom, when H_0 is true.

Critical region of size: Reject H_0 if $F > f_{1-\alpha}$, where $f_{1-\alpha}$ is the (1- α) percentile point of the F-distribution with the above degrees of freedom.

ANOVA: Assumptions

- Response variables are normally distributed
- Samples are independent
- Variances of the population are equal
- Responses for a given group are independent

Anova: Example 2

- Teaching methods v/s Productivity data

Method 1	15	18	19	22	11	
Method 2	22	27	18	21	17	
Method 3	18	24	19	16	22	15

- Observations:
  1. Number of observations = 16
  2. Number of groups / features = 3

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Anova: Example 2

- Teaching methods v/s Productivity data

Method 1	15	18	19	22	11	
Method 2	22	27	18	21	17	
Method 3	18	24	19	16	22	15

- Steps:
  1. Means: Calculate the mean of every group, and the grand mean
  2. Variance of means: Calculate the weighted sum of squares of differences between the column means and the grand mean & then calculate the variance. Let this be **MSB**.
  3. Variance within the columns: First calculate variances within each group, and then weighted overall variance based on these numbers. **MSE**.
  4. F statistic = MSB / MSE
     - If the groups are from same / similar population, this ratio will be closer to 1.

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Anova: Step 1

METHOD 1	METHOD 2	METHOD 3	
		18	
15	22	24	
18	27	19	
19	18	16	
22	21	22	
11	17	15	
85	105	114	
÷5	÷5	÷6	
$17 = \bar{x}_1$	$21 = \bar{x}_2$	$19 = \bar{x}_3 \leftarrow$	sample means
$n_1 = 5$	$n_2 = 5$	$n_3 = 6 \leftarrow$	sample sizes

$$\bar{\bar{x}} = \frac{15+18+19+22+11+22+27+18+21+17+18+24+19+16+22+15}{16}$$

$$= \frac{304}{16}$$

$= 19 \leftarrow$ grand mean using all the data

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Anova: Step 2

Calculation of the between-column variance

n	$\bar{x}$	$\bar{\bar{x}}$	$\bar{x} - \bar{\bar{x}}$	$(\bar{x} - \bar{\bar{x}})^2$	$n(\bar{x} - \bar{\bar{x}})^2$
5	17	19	$17 - 19 = -2$	$(-2)^2 = 4$	$5 \times 4 = 20$
5	21	19	$21 - 19 = 2$	$(2)^2 = 4$	$5 \times 4 = 20$
6	19	19	$19 - 19 = 0$	$(0)^2 = 0$	$6 \times 0 = \underline{0}$

$$\Sigma n_j(\bar{x}_j - \bar{\bar{x}})^2 \rightarrow 40$$

$$\hat{\sigma}^2 = \frac{\Sigma n_j(\bar{x}_j - \bar{\bar{x}})^2}{k - 1} = \frac{40}{3 - 1} \qquad [10\text{-}6]$$

$$= \frac{40}{2}$$

$= 20 \leftarrow$ the between-column variance

ANOVA: Step 3

Calculation of variances within the samples and the within-column variance

Training method 1 Sample mean: $\bar{x} = 17$		Training method 2 Sample mean: $\bar{x} = 21$		Training method 3 Sample mean: $\bar{x} = 19$	
$x - \bar{x}$	$(x - \bar{x})^2$	$x - \bar{x}$	$(x - \bar{x})^2$	$x - \bar{x}$	$(x - \bar{x})^2$
$15 - 17 = -2$	$(-2)^2 = 4$	$22 - 21 = 1$	$(1)^2 = 1$	$18 - 19 = -1$	$(-1)^2 = 1$
$18 - 17 = 1$	$(1)^2 = 1$	$27 - 21 = 6$	$(6)^2 = 36$	$24 - 19 = 5$	$(5)^2 = 25$
$19 - 17 = 2$	$(2)^2 = 4$	$18 - 21 = -3$	$(-3)^2 = 9$	$19 - 19 = 0$	$(0)^2 = 0$
$22 - 17 = 5$	$(5)^2 = 25$	$21 - 21 = 0$	$(0)^2 = 0$	$16 - 19 = -3$	$(3)^2 = 9$
$11 - 17 = -6$	$(-6)^2 = 36$	$17 - 21 = -4$	$(-4)^2 = 16$	$22 - 19 = 3$	$(3)^2 = 9$
	$\Sigma(x - \bar{x})^2 = \mathbf{70}$		$\Sigma(x - \bar{x})^2 = \mathbf{62}$	$15 - 19 = -4$	$(-4)^2 = 16$
	$\frac{\Sigma(x - \bar{x})^2}{n-1} = \frac{70}{5-1} = \frac{70}{4}$		$\frac{\Sigma(x - \bar{x})^2}{n-1} = \frac{62}{5-1} = \frac{62}{4}$		$\Sigma(x - \bar{x})^2 = \mathbf{60}$
					$\frac{\Sigma(x - \bar{x})^2}{n-1} = \frac{60}{6-1} = \frac{60}{5}$
sample variance $\rightarrow s_1^2 = 17.5$		sample variance $\rightarrow s_2^2 = 15.5$		sample variance $\rightarrow s_3^2 = 12.0$	

And:

$$\hat{\sigma}^2 = \sum\left(\frac{n_j - 1}{n_T - k}\right)s_j^2 = (4/13)(17.5) + (4/13)(15.5) + (5/13)(12.0) \qquad [10\text{-}7]$$

$$= \frac{192}{13}$$

$= 14.769 \leftarrow$ Second estimate of the population variance based on the variances within the samples (the within-column variance)

ANOVA: Step 4

- F = between column variance/ Within column variance

  = 20 / 14.769

  = 1.354
- The ANOVA Table will be as follows:

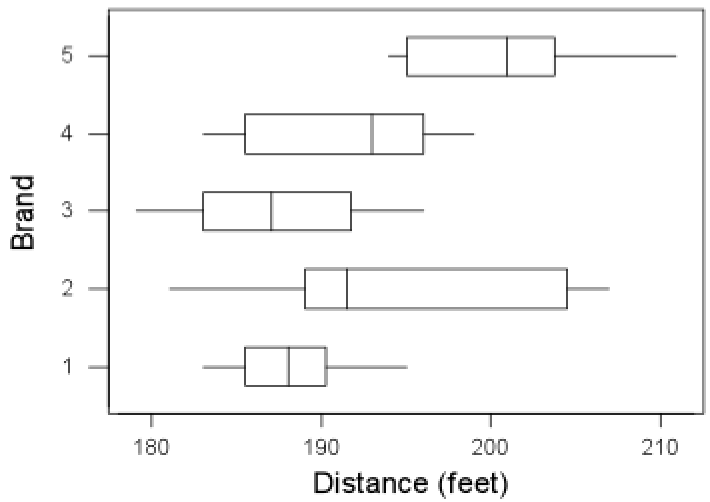
Source	DOF	SS	MS	F-Ratio	F-Critical
Factor	3-1 = 2	40	20	1.354	3.805
Error	16-3 = 13	192	14.769		
Total	15	132			

- Since F-Ratio is much less than F-Critical, we are not in the critical region. Hence NULL hypothesis (that the means of the groups are equal) cannot be rejected.

ANOVA: Using R

- Breaking distance related to 5 tyres

Brand1	Brand2	Brand3	Brand4	Brand5
194	189	185	183	195
184	204	183	193	197
189	190	186	184	194
189	190	183	186	202
188	189	179	194	200
186	207	191	199	211
195	203	188	196	203
186	193	196	188	206
183	181	189	193	202
188	206	194	196	195



Brand	N	MEAN	SD
1	10	188.20	3.88
2	10	195.20	9.02
3	10	187.40	5.27
4	10	191.20	5.55
5	10	200.50	5.44

Analysis of Variance for comparing all 5 brands

Source	DF	SS	MS	F	P
Brand	4	1174.8	293.7	7.95	0.000
Error	45	1661.7	36.9		
Total	49	2836.5			

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

ANOVA: Example-2

Four sections of the same elementary course in statistics were taught by the same instructor. The final grades out of 20 were recorded as follows:

Section	Grades						Totals
1	12	10	7	8	9	14	60
2	12	16	15	9			52
3	9	7	6	11	7		40
4	12	8	8	10			38
Total							190

Is there a significant difference in the average grades of the four sections? Use a 0.05 level of significance.

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Anova: Example-2 … contd

$S_{Stotal} = 148$, $SS_{treat} = 57$, thus $SS_{error} = 91$, which will make the following table:

ANOVA Table

Source	df	SS	MS	F-ratio
Treatment	3	57	19	3.13
Error	15	91	6.07	
Total	18	148		

With the critical value of $f_{.95}(3, 15) = 3.29$, H_0 is not rejected, since $3.13 < 3.29$. The conclusion is that there is no difference in the means of the four sections on that final exam.

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Problems

6.6 Astudy of the amount of violence viewed on telvesion as it relates to the age of the viewer yielded the following resuls as tabulated below

	Age		
Viewing	16-34	35-54	55 and over
Low Violence	8	12	21
High violence	18	15	7

Do the data indicate that viewing of violence is not independent of age of viewer, at the 5% significance level?

6.16 The following data show the effects of four operators, chosen at random from all operators at a certain factory, on the output of a particular machine:

I.	175.4	171.7	173.0	170.5
II.	168.5	162.7	165.0	164.1
III.	170.1	173.4	175.7	170.7
IV.	175.2	175.7	180.1	183.7

a) Perform the analysis of variance(random effects),