Analysis of Variance and Covariance

Dr. Shambhavi Mishra

Department of Statistics

University of Lucknow

Outline

- Introduction
- What is ANOVA
- Assumptions in ANOVA
- Classification of ANOVA
  - One-way (or single factor) ANOVA
  - Two-way ANOVA
- Analysis of Co-Variance(ANCOVA)
  - Why ANCOVA
- ANCOVA techniques
- Assumptions in ANCOVA

INTRODUCTION

- This statistical technique first developed by R.A. Fisher was extensively used for agricultural experiments.

- Analysis of variance (ANOVA) is a method for testing the hypothesis that there is no difference between two or more population means.

- The significance of the difference of means of the two samples can be judged through either z-test or t- test.

- When there are more than two means, it is possible to compare means for each pair using multiple t- tests.

- Conducting multiple t-tests can lead to severe inflation of the Type I error rate.

Contd…

- In such circumstances, we do not want to consider all possible combinations of two populations at a time for that we would require a great number of tests before we would be able to arrive at a decision.
- ANOVA can be used to test differences among several means for significance without increasing the Type I error rate.
- An extremely useful technique concerning researches in the many fields of economics, biology, education, psychology, sociology, business/industry and in researches of several other disciplines.

EXAMPLES

- A group of psychiatric patients are trying three different therapies: **counseling, medication** and **biofeedback.** We want to see if one therapy is better than the others.

- A manufacturer has two different processes to make light bulbs. They want to know if **one process is better than the other.**

- Students from different colleges take the same exam. You want to see if one college **outperforms the other**.

WHAT IS ANOVA?

- ANOVA is a procedure for testing the difference among different groups of data for homogeneity.

- The essence of ANOVA is that the total amount of variation in a set of data is broken down into two types, that amount which can be attributed to chance and that amount which can be attributed to specified causes.

- Thus, the basic principle of ANOVA is to test for differences among the means of the populations by examining the amount of variation within each of these samples/groups, relative to the amount of variation between the samples/groups.

ASSUMPTIONS IN ANOVA

- The parent population from which samples/groups are taken is normal.

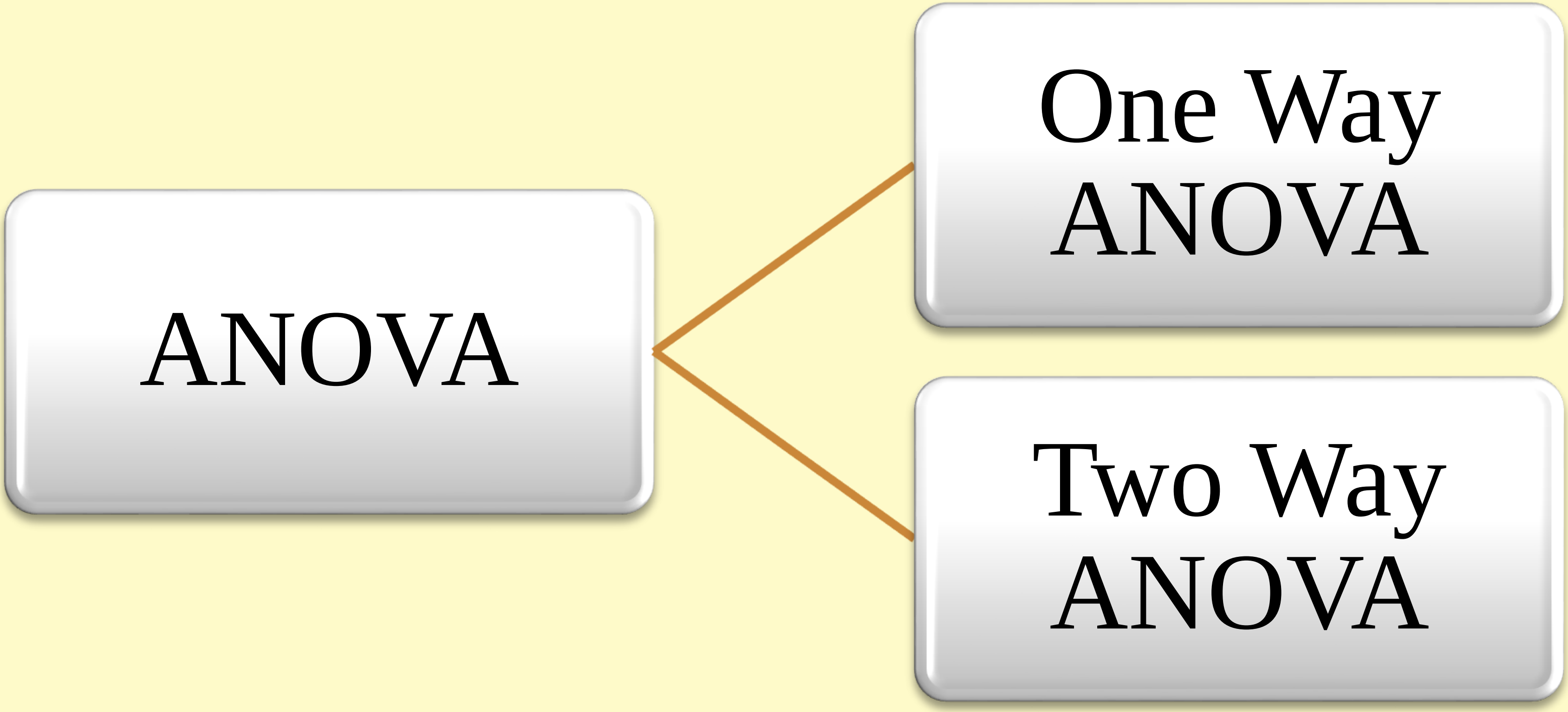
- Independence of samples i.e. each sample is randomly selected and independent.

- Homogeneity of variances or Homoscedasticity of the groups i.e.. equal variances among groups.

- The experimental errors of data are normally distributed with mean zero and variance σ^2.

- The effects are additive in nature.

CLASSIFICATION OF ANOVA



ANOVA

One Way ANOVA

Two Way ANOVA

ONE-WAY ANOVA

- When only one factor is considered, then it is called one way classification.

- The total variance is equal to the variance between groups and variance within groups or residual variance.

- **Example:** In an experiment, three groups are selected for an experimental treatment on one factor i.e., evaluation of performance of the three groups on the basis of attitude scales (factor).

LAYOUT

- Let us suppose that we have p groups/levels of a factor A and i^{th} group has n_i $(i = 1,2, \dots, p)$ observations.
- The layout is given as:

Factor A	Replication				Total $(y_{i.})$
1	y_{11}	y_{12}	$\cdots$	y_{1n_1}	$y_{1.}$
2	y_{21}	y_{22}	$\cdots$	$y_{2\mathrm{n}_2}$	$y_{2.}$
$\dots$	$\dots$	$\dots$	$\cdots$	$\dots$	$\dots$
p	y_{p1}	y_{p2}	$\cdots$	$y_{p\mathrm{n}_p}$	$y_{p.}$

THE MODEL

- Let us suppose that we have p groups/levels of a factor A and each group has n_i observations .
- The model is given

$$\boldsymbol{y_{ij}} = \boldsymbol{\mu} + \boldsymbol{\alpha_i} + \boldsymbol{\epsilon_{ij}} \; ; \quad i = 1,2, \ldots p \; ; j = 1,2, \ldots , n_i$$

where

- y_{ij} is the j^{th} observation of i^{th} group.
- μ is general mean effect.
- α_i is the effect of i^{th} level of factor A.
- ϵ_{ij} is the error and it is independently normally distributed with mean zero and variance σ^2.
- Also $n = \sum n_i$ is the total number of observations.

HYPOTHESES

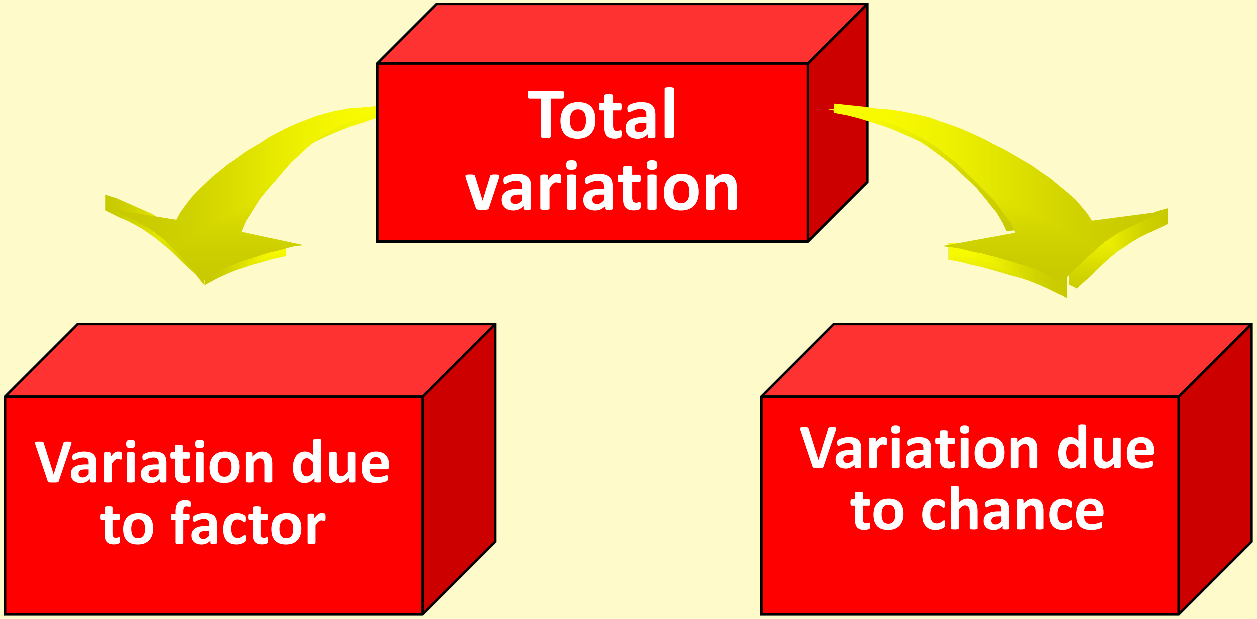
- **Null Hypothesis**

  $H_0: \mu_1 = \mu_2 = \cdots = \mu_p$ i.e. the mean of all the groups is same.

- **Alternative Hypothesis**

  H_1: mean’s of at least two are not same.

TOTAL VARIATION PARTITIONING



PROCEDURE

- Obtain the mean of each sample.
- Work out the mean of the sample means.
- Calculate sum of squares for variance between the samples (or SS between).
- Obtain variance or mean sum of square (MS) between samples.
- Calculate sum of squares for variance within samples (or SS within).
- Obtain the variance or mean sum of square (MS) within samples.
- Find sum of squares of deviations for total variance.
- Finally, find F-ratio.

METHODOLOGY

- Total Sum of Squares = Sum of squares due to factor A + Sum of squares due to error

  i.e., $TSS = SSA + SSE$

  d.f. $n-1 = (p-1) + (n-p)$

where

- $TSS = \sum_{i=1}^{p} \sum_{j=1}^{n_i} y_{ij}^2 - \frac{G^2}{n}$
- $SSA = \sum_{i=1}^{p} \frac{y_{i.}{}^2}{n_i} - \frac{G^2}{n}$
- here $y_{i.}$ = total for the i^{th} group of the factor A.
- $G = \sum_{i=1}^{p} \sum_{j=1}^{n_i} y_{ij}$ = Grand total

ANOVA Table for ONE-WAY CLASSIFIED DATA

Source of variation	d.f.	Sum of squares	Mean sum of squares	F- ratio
Between groups	$p-1$	SSA	$MSA = \frac{SSA}{p-1}$	$F_A = \frac{MSA}{MSE}$
Within groups	$n-\mathrm{p}$	SSE	$MSE = \frac{SSE}{n-p}$	
Total	$n-1$	TSS		

$$F_A = \frac{MSA}{MSE} \sim F_{p-1,n-p}$$

- If $F_A > F_{(p-1,n-p)}(\alpha)$, then H_0 is rejected at 100α% level of significance and we conclude that groups differ significantly, otherwise H_0 accepted.

NUMERICAL EXAMPLE

- In a comparison of the cleaning action of four detergents, 20 pieces of white cloth were first soiled with India ink. The cloths were then washed under controlled conditions with 5 pieces washed by each of the detergents. Unfortunately three pieces of cloth were lost in the course of the experiment. Whiteness readings, made on the 17 remaining pieces of cloth, are shown below:

Detergent			
A	**B**	**C**	**D**
77	74	73	76
81	66	78	85
61	58	57	77
76		69	64
69		63	

Assuming all whiteness readings to be normally distributed with common variance, test the hypothesis of no difference between the four brands on the basis of mean whiteness readings after washing.

SOLUTION

- H_0: No difference in mean readings i.e. $\mu_i = \mu \, \forall i$
- H_1: a difference in mean readings i.e., $\mu_i \neq \mu$ for some i
- We have, $p = 4, n_1 = 5, n_2 = 3, n_3 = 5, n_4 = 4$ and $n = 17$ therefore,
- $TSS = \sum_{i=1}^{p} \sum_{j=1}^{n_i} y_{ij}^2 - \frac{G^2}{n}$

$$= (77^2 + 81^2 + \cdots + 77^2 + 64^2) - \frac{1204^2}{17} = 1090.47$$

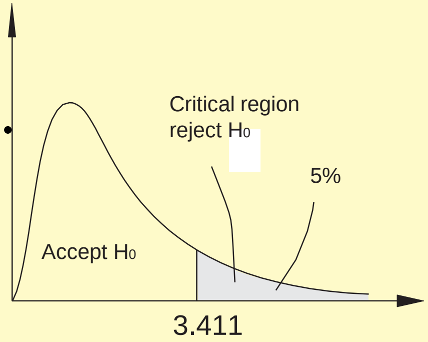
- $SSA = \sum_{i=1}^{p} \frac{{y_{i.}}^2}{n_i} - \frac{G^2}{n}$

$$= \left(\frac{364^2}{5} + \frac{198^2}{3} + \frac{340^2}{5} + \frac{302^2}{4}\right) - \frac{1204^2}{17} = 216.67$$

- $SSE = TSS - SSA = 1090.47 - 216.67 = 873.80$

ANOVA TABLE

Source of variation	d.f.	Sum of squares	Mean sum of squares	F- ratio
Between detergents	3	216.67	72.22	1.07
Within detergents	13	873.80	76.22	
Total	16	1090.47		



- Critical region is > 3.411 (since $F_{(3,13)}(0.05) = 3.411$).
- The *F* ratio 1.07 does not lie in the critical region.
- So, we do not reject null hypothesis.
- Thus, there is no evidence at the 5% significance level to suggest a difference between the four brands on the basis of mean whiteness after washing.

TWO-WAY ANOVA

- Two way ANOVA technique is used when the data are classified on the basis of two factors.

- For example, the agricultural output may be classified on the basis of different varieties of seeds and also on the basis of different varieties of fertilizers used.

- A statistical test used to determine the effect of two nominal predictor variables on a continuous outcome variable.

- Two way ANOVA test analyzes the effect of the independent variables on the expected outcome along with their relationship to the outcome itself.

- Two-way ANOVA may have repeated measurements (more than one observation per cell) of each factor or may not have repeated values (one observation per cell).

LAYOUT

- Let us suppose that we have p groups/levels of a factor A and q groups/levels of factor B.
- The layout is given as

Factor B Factor A	1	2	…	q	Total ($y_{i.}$)
1	y_{11}	y_{12}	…	y_{1q}	$y_{1.}$
2	y_{21}	y_{22}	…	y_{2q}	$y_{2.}$
…	…	…	…	…	…
p	y_{p1}	y_{p2}	…	y_{pq}	$y_{p.}$
Total ($y_{.j}$)	$y_{.1}$	$y_{.2}$	…	$y_{.q}$	$y_{..}$

THE MODEL

- The model is

$$\boldsymbol{y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}}$$

where

- y_{ij} is the observation of (i, $j)^{th}$ cell i^{th} group ($i = 1,2, \dots p\,; j = 1,2, \dots, q$).
- μ is general mean effect
- α_i is the effect of i^{th} group of factor A.
- β_j is the effect of j^{th} group of factor B.
- ϵ_{ij} is the error and it is independently normally distributed with mean zero and variance σ^2.
- Also, $n = pq$ is the total number of observations

Hypotheses

- **Null Hypothesis**

  $H_{0A}: \mu_{1A} = \mu_{2A} = \cdots = \mu_{pA}$ i.e. the mean of all the groups of factor A is same.

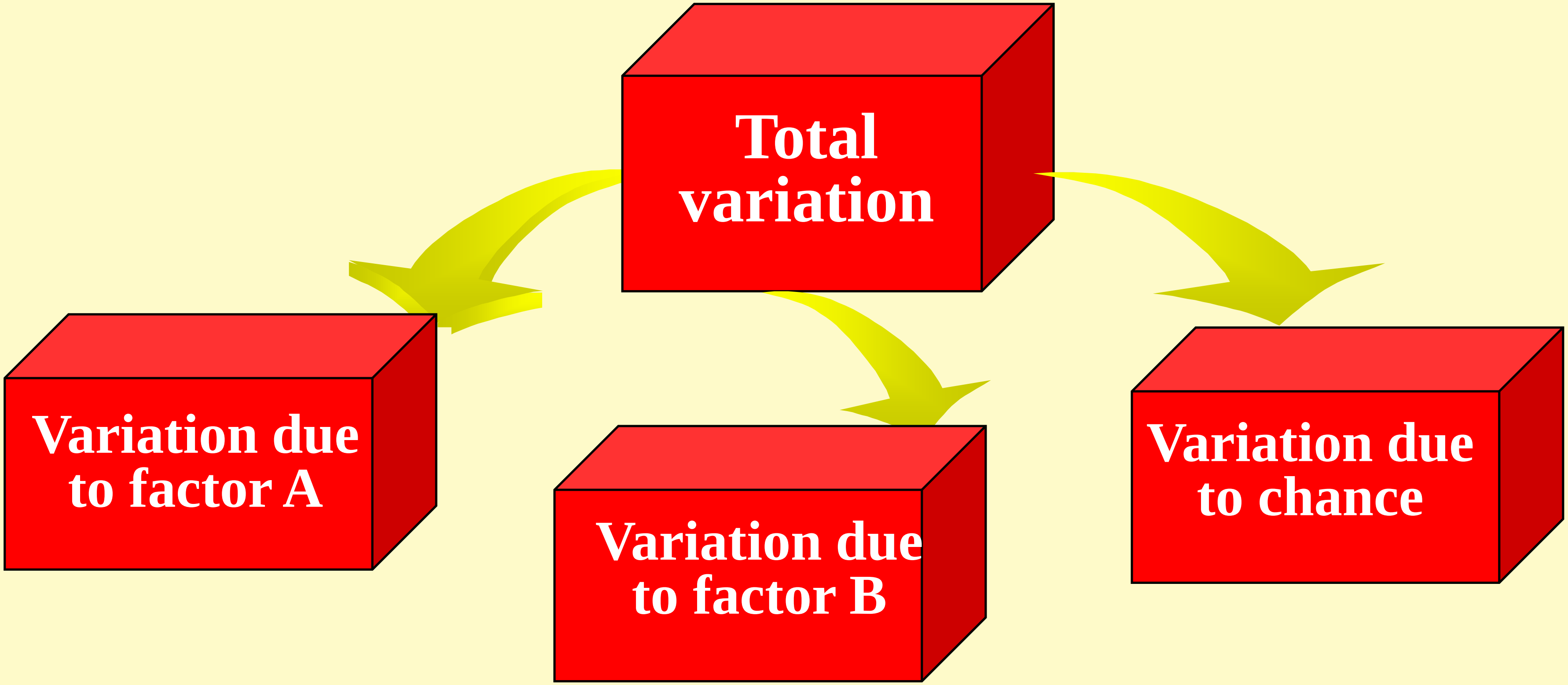
  $H_{0B}: \mu_{1B} = \mu_{2B} = \cdots = \mu_{pB}$ i.e., the mean of all the groups of factor B is same.

- **Alternative Hypothesis**

  H_{1A}: Mean of at least two groups of factor A are not same.

  H_{1B}: Mean of at least two groups of factor B are not same.

TOTAL VARIATION PARTITIONING



Total variation

Variation due to factor A

Variation due to factor B

Variation due to chance

METHODOLOGY

- Total Sum of Squares = Sum of squares due to factor A
  + Sum of squares due to factor B
  + Sum of squares due to error

i.e., $TSS = SSA + SSB + SSE$

d.f. $n-1 = (p-1) + (q-1) + (p-1)(q-1)$

where

- $TSS = \sum_{i=1}^{p}\sum_{j=1}^{q} y_{ij}^2 - \frac{G^2}{n}$; $SSA = \frac{1}{q}\sum_{i=1}^{p} y_{i.}^2 - \frac{G^2}{n}$
- $SSB = \frac{1}{p}\sum_{i=1}^{q} y_{.j}^2 - \frac{G^2}{n}$; $SSE = TSS - SSA - SSB$
- Here $y_{i.}$ = total for the i^{th} level of the factor A.
- $y_{.j}$ = total for the j^{th} level of the factor B.
- $G = \sum_{i=1}^{p}\sum_{j=1}^{n_i} y_{ij} = y_{..}$ =Grand total

ANOVA Table for TWO-WAY CLASSIFIED DATA

Source of variation	d.f.	Sum of squares	Mean sum of squares	F- ratio
Factor A	$p-1$	SSA	$MSA=\frac{SSA}{p-1}$	$F_A=\frac{MSA}{MSE}$
Factor B	$q-1$	SSB	$MSB=\frac{SSB}{q-1}$	$F_B=\frac{MSB}{MSE}$
Error	$(p-1)(q-1)$	SSE	$MSE=\frac{SSE}{n-p}$	
Total	$n-1$	TSS		

We know that

$$F_A=\frac{MSA}{MSE}\sim F_{p-1,(p-1)(q-1)}\ ;\quad F_B=\frac{MSB}{MSE}\sim F_{q-1,(p-1)(q-1)}$$

Test Criteria

- The F-ratio is used to judge whether the difference among several sample means is significant or is just due to random causes/chance.

- If $F_A > F_{\{p-1,(p-1)(q-1)\}}(\alpha)$, then H_{0A} is rejected at 100α% level of significance and we conclude that groups differ significantly, otherwise H_{0A} accepted.

- If $F_B > F_{\{q-1,(p-1)(q-1)\}}(\alpha)$, then H_{0B} is rejected at 100α% level of significance and we conclude that groups differ significantly, otherwise H_{0B} accepted.

NUMERICAL EXAMPLE

- Three experimenters determine the moisture content of samples of body lotion. For this purpose each experimenter has taken 4 consignments. The results are given below:

Experimenter	**Consignment**			
	I	II	III	IV
A	9	10	9	10
B	12	11	9	11
C	11	12	10	12

- Test whether there is any significant difference among consignments and among experimenters at 5% significance of level.

SOLUTION

- $\boldsymbol{H_{0A}}$: There is no difference among the means of experimenters

  i.e., $\mu_{1A} = \mu_{2A} = \mu_{3A}$

v/s $\boldsymbol{H_{1A}}$: At least two μ_{iA} are not same $(i = 1,2,3)$

- $\boldsymbol{H_{0B}}$: There is no difference among the means of consignments

  i.e., $\mu_{1B} = \mu_{2B} = \mu_{3B} = \mu_{4B}$

v/s $\boldsymbol{H_{1B}}$: At least two μ_{jB} are not same $(j = 1,2,3,4)$

- We have, $p = 3, q = 4$, and $n = pq = 12$, therefore,
- $TSS = \sum_{i=1}^{p} \sum_{j=1}^{q} y_{ij}^2 - \frac{G^2}{n}$

$$= (9^2 + 12^2 + \cdots + 11^2 + 12^2) - \frac{126^2}{12} = 15.0$$

Contd…

- $SSA = \frac{1}{q}\sum_{i=1}^{p} y_{i}.^{2} - \frac{G^{2}}{n}$

  $= \frac{1}{4}(38^{2} + 43^{2} + 45^{2}) - \frac{126^{2}}{12} = 6.5$

- $SSB = \frac{1}{p}\sum_{i=1}^{q} y_{.j}^{2} - \frac{G^{2}}{n}$

  $= \frac{1}{3}(32^{2} + 33^{2} + 28^{2} + 33^{2}) - \frac{126^{2}}{12} = 5.67$

- $SSE = TSS - SSA - SSB$

  $= 15.0 - 6.5 - 5.67 = 2.83$

ANOVA Table

Source of variation	d.f.	Sum of squares	Mean sum of squares	F- ratio
Between experimenters	2	6.5	3.25	6.89
Between Consignments	3	5.67	1.89	4.0
Error	6	2.83	0.47	
Total	11	15.0		

- The tabulated values are $F_{(2,6)}(0.05) = 5.14$ and $F_{(3,6)}(0.05) = 4.76$.
- The *F* ratio 6.89 (>5.14) does lie in the critical region at 5% level of significance, therefore, we conclude that the mean moisture content as determined by 3 experimenters are not equal.
- The F-ratio 4.0 (<4.76) does not lie in the critical region at 5% level of significance, therefore, we conclude that the moisture content of the 4 consignments may not different from one another.

USES OF ANOVA

- Interpretation of the significance of means and their interactions.
- To test the significance between the variance of two samples.
- To test fitting of regression model.
- To study the homogeneity in case of two-way classification.
- To test the linearity of regression.

ADVANTAGES AND DISADVANTAGES OF ANOVA

- **ADVANTAGES**
  - It is an improved technique over **t-test** or **z-test**.
  - Suitable for multidimensional variables.
  - Analyze various factors at a time.
  - Economical method of parametric testing.
  - Can be used in 3 or more than 3 groups at a time.

- **DISADVANTAGES**

  - It is difficult to analyze ANOVA under strict assumptions regarding the nature of data.

  - It is not so helpful in comparison with t-test. There is no special interpretation of ANOVA for testing the significance of two means.

  - There is a requirement of post-hoc **t-test** for further testing.

APPLICATIONS OF ANOVA

- Comparing the gas mileage of different vehicles, or the same vehicle under different fuel types, or road types.

- Understanding the impact of temperature, pressure or chemical concentration on some chemical reaction (power reactors, chemical plants, etc).

- Studying whether advertisements of different kinds solicit different numbers of customer responses.

- Recommendation of a fertilizer against others for the improvement of crop yield.

Contd…

- ANOVA has immensely useful practical applications in business, particularly Lean-Six Sigma/operational efficiency.

- Understanding the impact of different catalysts on chemical reaction rates.

- Understanding the performance, quality or speed of manufacturing processes based on the number of cells or steps they're divided into.

ANALYSIS OF COVARIANCE (ANCOVA)

- Analysis of covariance (ANCOVA) is a technique in which it is possible to control some sources of variation by taking the additional observations on each of the experimental units.

- ANCOVA allows to compare one variable in 2 or more groups taking into account (or to correct for) variability of other variables, called covariates.

- It tests whether there is a significant difference between groups after controlling for variance explained by a covariate.

WHY ANCOVA?

- The object of experimental design in general happens to be to ensure that the results observed may be attributed to the treatment variable and to no other causal circumstances.

- For instance, the researcher studying one independent variable, *X*, *may wish to control the influence of some uncontrolled* variable (sometimes called the covariate or the concomitant variables), *Z*, *which is known to be* correlated with the dependent variable, *Y*, *then he should use the technique of analysis of covariance* for a valid evaluation of the outcome of the experiment.

- In psychology and education primary interest in the analysis of covariance rests in its use as a procedure for the statistical control of an uncontrolled variable.

ANCOVA Techniques

- The influence of uncontrolled variable is usually removed by simple linear regression method and the residual sums of squares are used to provide variance estimates which in turn are used to make tests of significance.

- covariance analysis consists in subtracting from each individual score (Y_i) that portion of it Y_i that is predictable from uncontrolled variable (Y_i) and then computing the usual analysis of variance on the resulting $(Y - Y')'$.

- The adjustment to the degree of freedom is made because of the fact that estimation using regression method required loss of degree of freedom.

ASSUMPTIONS IN ANCOVA

- There is some sort of relationship between dependent variable and the uncontrolled variables.

- We also assume that this form of relationship is the same in the various treatment groups.

- Other assumptions are:
  - Various treatment groups are selected at random from the population.
  - The groups are homogeneous in variability.
  - The regression is linear and is same from group to group.

THANK you!