

# DATA ANALYTICS

***Class # 12***

**ANOVA - I**

*Slides adapted from*

**Dr. Sreeja S R**

*Assistant Professor*

**Indian Institute of Information Technology  
IIIT Sri City**

## QUOTE OF THE DAY..

Try not to become a person of success, but rather try to become a person of value.

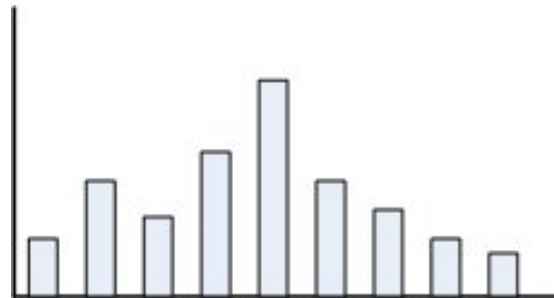
ALBERT EINSTEIN, Theoretical physicist

# THIS PRESENTATION INCLUDES...

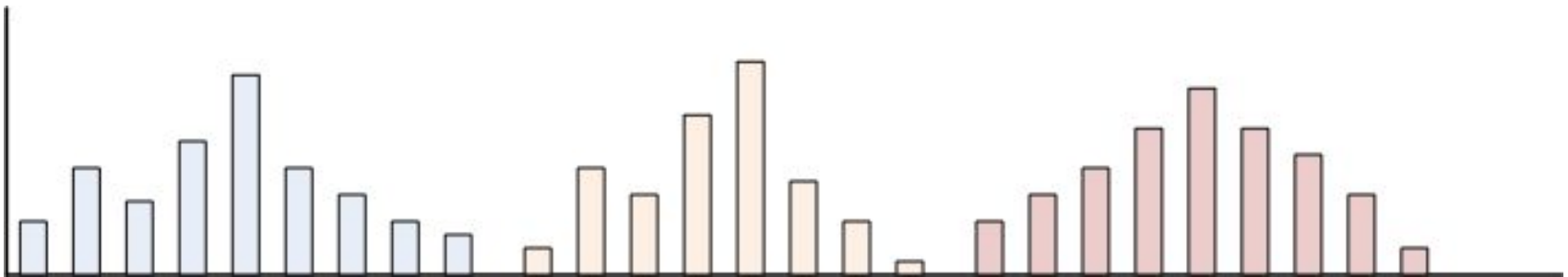
- What is “Analysis of variance”?
- Why ANOVA?
- How ANOVA?
  - *One – way ANOVA*
  - *Two–way ANOVA*

# What is Analysis of Variance?

# WHAT IS ANALYSIS OF VARIATION?



Single population



Multiple population

## EXAMPLE : SINGLE VS. MULTIPLE POPULATION



# WHAT IS THE ISSUE?

- Are the statistical inference valid?

$\mu$

$\sigma$

# EXAMPLE 1: THE ISSUE IN STATISTICAL TESTING

A recent study claims that using music in a class enhances the concentration and consequently helps students absorb more information.

- What if it affected the results of the students in a negative way?

or

- What kind of music would be a good choice for this?

We should have some proof that it actually works or not.



# DESIGN OF EXPERIMENT

- The teacher decided to implement it on a smaller group of randomly selected students from **three different** classes.

Three different groups of **ten randomly selected students** from three different classrooms were taken.

Each classroom was provided with **three different environments** for students to study.

- **Classroom A had constant music** being played in the background
  - **Classroom B had variable music** being played in the background
  - **Classroom C was a regular class with no music playing**
- A test was conducted after one month for all the three groups and their test scores were collected.

# TEST RESULT

	Test scores of students (out of 10)										Mean
Class A (constant music)	7	9	5	8	6	8	6	10	7	4	7
Class B (variable music)	4	3	6	2	7	5	5	4	1	3	4
Class C (no music)	6	1	3	5	3	4	6	5	7	3	4.3
								Grand Mean ->			5.1

# OBSERVATIONS FROM THE RESULTS

- It is noticed that the mean score of students from **Group A** is definitely greater than the other two groups, so the treatment must be helpful.
- Maybe it's true, but there is also a slight chance that we happened to select the best students from class A, which resulted in better test scores (remember, the selection was done at random).
- This leads to a few questions:
  1. How do we decide that these three groups performed differently because of the different situations and not merely by chance?
  2. In a statistical sense, how different are these three samples from each other?

# ANALYSIS OF VARIANCE (ANOVA)

## Definition 16.1

- Analysis of Variance (ANOVA) is derived from a partitioning of total variability into its component parts.
  - ANOVA is a statistical technique that is used to check if the means of two or more groups are significantly different from each other.
  - ANOVA checks the impact of one or more factors by comparing the means of different samples.
- 
- This technique was invented by Sir Ronald Aylmer Fisher (1921), and is often referred to as Fisher's ANOVA.

# Why ANOVA?

# STATISTICAL INFERENCES

- ANOVA is a statistical technique
  - It is similar in application to techniques such as t-test, z-test and  $\chi^2$ -test in that it is used to compare means and the relative variance between them.
- Why not use t-test, z-test and  $\chi^2$ -test ?
- Why analysis of variance for comparing means?



# USING T-TEST

t-test is used to:

- To infer **mean of a single population**
- T-test can be used to compare two populations

*However, t-test is not useful to compare mean of more than two populations*

# EXTENDING THE TWO POPULATION PROCEDURE

- Construct pairwise comparison on all means.

For 5 populations  $\rightarrow$  10 possible pairs. When all pairwise comparisons are made for  $n$  groups, the total number of possible combinations is  $n*(n-1)/2$ .

- Considering  $\alpha = 0.05$ , probability of correctly failing to reject the null hypothesis for all 10 tests is  $(0.95)^{10}$ , assuming that the tests are independent
- Thus, the true value of  $\alpha$  for this set of comparison is 0.4, instead of .05
- It inflates the Type 1 error.
- The probability that a Type I error occurs if  $k$  comparisons are made is  $1-(1-\alpha)^k$ ; if 10 comparisons are made, the Type I error rate increases to 40%.



# EXTENDING THE TWO POPULATION PROCEDURE

- Statistical Inference I
  - A car magazine wishes to compare the average petrol consumption of THREE models for car and has available SIX vehicles of each model.

Model 1	Model 2	Model 3

- There are THREE populations
- There are samples each of size six from each population

# EXTENDING THE TWO POPULATION PROCEDURE

- Statistical Inference II
  - A teacher is interested in a comparison of the average percentage marks obtained in the examinations of five different subjects and has available the marks of eight students who all completed each examination.

Subject 1	Subject 2	Subject 3	Subject 4	Subject 5

- What is the number of populations?
- How many samples? What are their sizes?? Are each sample independent to each other?

# EXAMPLE 2 : WHY ANOVA?

Consider the two sets of contrived data as shown below:

Set 1 (Benz)			Set 2 (Toyota)		
Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3
5.7	9.4	14.2	3.0	5.0	11.0
5.9	9.8	14.4	4.0	7.0	13.0
6.0	10.0	15.0	6.0	10.0	16.0
6.1	10.2	15.6	8.0	13.0	17.0
6.3	10.6	15.8	9.0	15.0	18.0
$\bar{y} = 6.0$	$\bar{y} = 10.0$	$\bar{y} = 15.0$	$\bar{y} = 6.0$	$\bar{y} = 10.0$	$\bar{y} = 15.0$

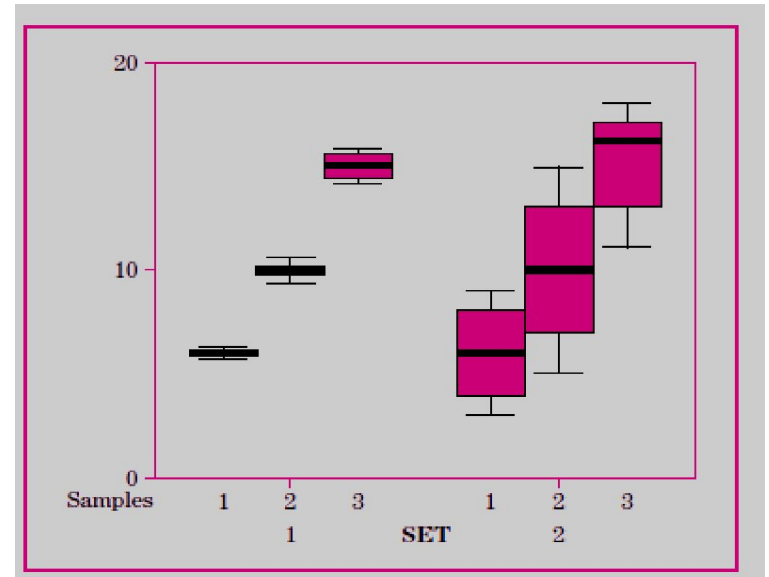
## Observations:

- Looking only at the means, we can see that they are identical for the three populations in both the sets.
- Using the means alone, we would state that there is no difference between the two sets.

# BOX PLOTS OF THE TWO EXPERIMENTS

## Observation from Box plots

- It appears that there is stronger evidence of differences among means in Set 1 than among means in Set 2.
- The observations *within* the samples are more closely bunched in Set 1 than they are in Set 2.
- We know that **sample means from populations with smaller variances** will also be less variable.  
(Central Limit Theorem)
- Thus, although the variances among the means for the two sets are identical, the variance among the observations within the individual samples is smaller for Set 1 and is the reason for the apparently stronger evidence of different means.
- This observation is the basis for using the analysis of variance for making inferences about differences among means.
- The analysis of variance is based on the **comparison of the variance among the means of the populations to the variance among sample observations within the individual populations.**



# REFERENCE

- The detail material related to this lecture can be found in

Design and Analysis of Experiments (8<sup>th</sup> Edition), Douglas C. Montgomery, John Wiley & Sons, 2013.

Any question?