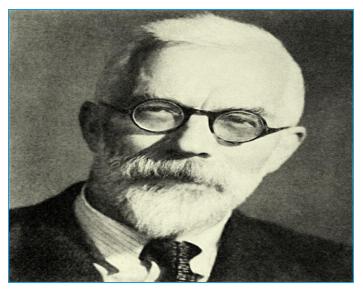
# Statistical Inference Analysis of Variance



## Analysis of Variance (ANOVA)

• Analysis of variance (ANOVA) is a collection of statistical models used to analyze the differences among more than two group means developed by statistician and evolutionary biologist **Ronald Fisher**.



• Example: There are 20 plots of wheat and 5 fertilizers are applied to four different plots. The yield of wheat is recorded for each of 20 plots.

ANOVA can be used to find out whether effect of these fertilisers on yields is equal or significantly

different.

## ANOVA

- Note that although the name is 'Analysis of Variance', the method is used to analyze the differences among group means.
- Variation in the variable is inherent in nature. In general, the observed variance in a particular variable is partitioned into components attributable to different sources of variation.
- The total variance in any variable is due to a number of causes which may be classified "assignable causes (which can be detected and measured)" and "chance causes (which is beyond control of human and cannot be traced separately)".
- Hence, ANOVA is the separation of variance ascribable to one group of causes from the variance ascribable to other group.



## Assumptions of ANOVA

- The assumptions of ANOVA are listed below:
  - The samples drawn are random samples.
- -The populations from which samples are drawn have equal & unknown variances.
  - The populations follow normal distribution.



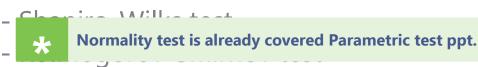
## **Testing Normality assumption**

- An assessment of the normality of data is a prerequisite for many statistical tests because normal data is an underlying assumption in parametric testing.
- Normality can be assessed using two approaches: graphical and numerical.

## - Graphical approach

- Box-Whisker plot (It is used to asses symmetry rather than normality.)
  - Quantile-Quantile plot (Q-Q plot).

## - Statistical approach





# One Way ANOVA

- One Way Anova can be considered as an extension of the t test for independent samples.
- One Way Anova is used to test the equality of K population means.
   (when K=2, t-test can be used.)
- For two levels (K=2), the t test and One Way Anova provide identical results.

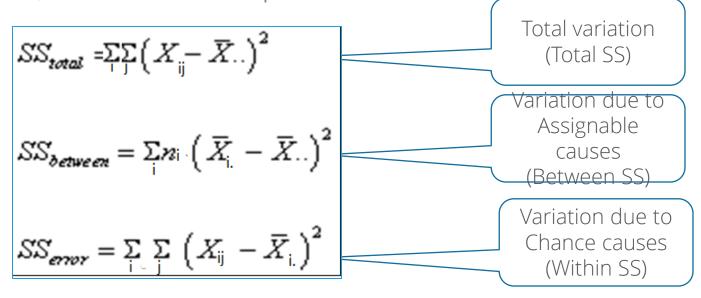
The null hypothesis is

$$H_0: \mu_1 = \mu_2 = \dots = \mu_K = \mu$$



# Partitioning Total Variance

Total variation is partitioned into two parts:
 Total SS= Between Groups SS + Within Groups SS where, SS stands for sum of squares



- Total SS is calculated using squared deviations of each value from overall mean.
- Between SS is calculated using squared deviation of each group mean from overall mean.
- Within Group SS can be obtained by subtracting Between SS from Total SS



# Case Study

To execute analysis of Variance in R, we shall consider the below case as an example.

## **Background**

A large company is assessing the difference in 'Satisfaction Index' of employees in Finance, Marketing and Client-Servicing departments.

## **Objective**

To test whether **mean satisfaction index** for employees in three departments (CS, Marketing, Finance) are equal.

## **Sample Size**

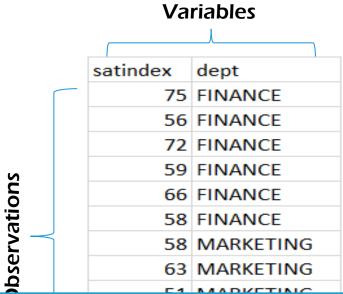
Sample size: 37

Variables: satindex, dept



# Data Snapshot

## One way anova



Columns	Description	Type	Measurement	Possible values
satindex	Satisfaction Index	Numeric		Positive Values
dept	Department	Character	MARKETING, CS, FINANCE	3



# One Way ANOVA

Testing equality of means in one factor with more than two levels.

Objective	To test whether <b>mean satisfaction index</b> for employees in three departments (CS, Marketing, Finance) are equal.
-----------	---

Null Hypothesis ( $H_0$ ): Mean satisfaction index for 3 departments are equal i.e.  $\mu 1 = \mu 2 = \mu 3$ Alternate Hypothesis ( $H_1$ ): Mean satisfaction index for 3 departments are not equal

Test Statistic	The test statistic is denoted as F and is based on F distribution.
Decision Criteria	Reject the null hypothesis <b>if p-value &lt; 0.05</b>



## Calculation

**Total SS** =  $(75-65.59)^2+(56-65.59)^2+....+(65-65.59)^2+(76-65.59)^2$ = 1840.92

**Between Groups SS** =  $12*(64.42-65.59)^2+12*(63.25-65.59)^2+13*(68.85-65.59)^2$ = 220.0599

Within Groups SS = Total SS - Between SS

Overall Mean	65.59 n=37
Mean for Finance	64.42 n1=12
Mean for Marketing	63.25 n2=12
Mean for CS	68.85 n3=13



# One Way ANOVA table

Sources of variation	Degrees of freedom (df)	Sum of Squares (SS)	Mean Sum of Squares (MS=SS/df)	F-Value
Between groups	K-1=3-1 =2	SSA=220.0599	MSA=110.03	F=2.3080
Within groups (error)	n-k=37-3 =34	SSE=1620.86	MSE=47.6724	
TOTAL	n-1=37-1 =36	TSS=1840.92		



## One Way ANOVA in R

# Import data

```
data<-read.csv("One way anova.csv",header=TRUE)</pre>
```

# ANOVA table

```
anovatable<-aov(formula=satindex~dept, data=data) ←
summary(anovatable)</pre>
```

- $\Box$  'aov' is the R function for ANOVA.
- ☐ formula specifies 'satindex' as analysis (dependent) variable and 'dept' as factor (independent) variable.
- □ *anovatable* is user defined object name created to store output.
- □ *summary* function displays the ANOVA table output.

#### # Output:

```
Df Sum Sq Mean Sq F value Pr(>F)
dept 2 220.1 110.03 2.308 0.115
Residuals 34 1620.9 47.67
```

#### Interpretation:

Since p-value is >0.05, do not reject H0. There is no significant difference in satisfaction index among 3 different departments.



# Two Way ANOVA

- Two Way Anova is used when there are 2 factors under study.
- Each factor can have 2 or more levels . Example: Gender and Age can be 2 factors. Gender with 2 levels as Male and Female Age with 3 levels as 18-30,31-50 and >50
- Three hypothesis are tested.

Factor A H0: All group means are equal

H1: At least one mean is different from other means

Factor B H0: All group means are equal

H1: At least one mean is different from other means

Interaction H0: The interaction is not significant

H1: The interaction is significant



# Two Way ANOVA

Total variation is partitioned as below:

**Total SS**= Between Groups SS due to factor A (SSA)

- + Between Groups SS due to factor B (SSB)
- + Interaction SS due to factor A and B (SSAB)
- + Error SS (SSE)

where, SS stands for sum of squares



# Case Study

We will illustrate Two Way Anova in R using following case study

## **Background**

A large company is assessing the difference in 'Satisfaction Index' of employees in Finance, Marketing and Client-Servicing departments. Experience level is also considered in the study.( <=5 years and >5 years)

## **Objective**

To test the equality of the satisfaction index among employees of three departments (CS, Marketing, Finance) and among different experience bands.

### **Sample Size**

Sample size: 36

Variables: satindex, dept, exp



# Data Snapshot

Two Way Anova

Observations

satindex	dept	exp
75	FINANCE	lt5
56	FINANCE	lt5
62	FINANCE	gt5
66	FINANCE	gt5
58	FINANCE	gt5
58	MARKETIN	lt5
63	MARKETIN	lt5
53	MARKETIN	lt5
74	MARKETIN	lt5
77	MARKETIN	lt5
69	MARKETIN	lt5
57	MARKETIN	gt5
70	MARKETIN	gt5
68	MARKETIN	gt5
77	CS	lt5

Columns	Description	Type	Measurement	Possible values
Satindex	Satisfaction Index	Numeric	-	Positive Values
Dept	Department	Character	MARKETING, CS, FINANCE	3
Exp	Years of Experience (grouped)	Character	lt5 = less than 5, gt5 = greater than 5	2



## Two Way ANOVA

Testing equality of means in two factors.

Objective

To compare employee satisfaction index in three departments (CS, Marketing, Finance) and two experience level based groups.

## Null Hypothesis

 $(H_{01})$ : Average satisfaction index is equal for 3 departments.

 $(H_{02})$ : Average satisfaction index is equal for 2 experience levels.

 $(H_{03})$  Interaction effect(dept\*exp) is not significant on satisfaction index.

The test statistic is computed for each of these null hypothesis.

Reject the null hypothesis if p-value < 0.05



## Two Way ANOVA in R

```
# Import data
data<-read.csv("Two Way Anova.csv", header=TRUE)</pre>
# ANOVA Table
 anovatable<-aov(formula=satindex~dept+exp+dept*exp,data=data)</pre>
 summary(anovatable)
\Box 'aov' is the R function for ANOVA.
□ formula specifies 'satindex' as analysis (dependent) variable and
    'dept' and 'exp' as factor (independent) variables.
 □ dept*exp specifies the interaction effect.
   anovatable is user defined object name created to store output.
 □ summary function displays the ANOVA table output.
```



## Two Way ANOVA in R

#### # Output:

	Df	Sum Sq	Mean Sq	F	value	Pr(>F)	
dept	2	164.2	82.11		1.679	0.204	
exp	1	78.0	78.03		1.595	0.216	
dept:exp	2	20.2	10.11		0.207	0.814	
Residuals	30	1467.2	48.91				

## Interpretation:

- Since p-value is >0.05 for all three (dept, exp and dept\*exp), do not reject H0 for all three tests. There is no significant difference in satisfaction index among 3 different departments and 2 experience levels.
- > Also interaction effect is not significant.



## Knowledge check question

- A large retailer is testing a marketing campaign on 24 stores. 8 stores are selected randomly from each of 3 zones.
- The variable of interest is 'sales increment(%) during campaign month'. Objective is to test whether the campaign is equally effective in 3 regions. Data is given below.

NORTH	WEST	SOUTH
8	10.2	5.3
12.5	9.3	5.8
9.2	9.9	6
6.7	8.7	7.1
9.4	9.1	7
5.9	10.2	6.1
7.7	9.5	6.3
6.9	10	7.3

Is this One-way ANOVA problem or Two-way ANOVA problem?

**ANSWER:** One-way ANOVA

**EXPLANATION:** There is only one factor (zone) with 3 levels (North, West, South).



# **THANK YOU!**

