

©Hero Images/Getty Images RF

Statistical Inference Concerning Variance

ANALYSIS OF VARIANCE (ANOVA)

Learning Objectives

In this chapter, you learn:

- The basic concepts of experimental design
- How to use the one-way analysis of variance to test for differences among the means of several groups
- How to use the two-way analysis of variance and interpret the interaction

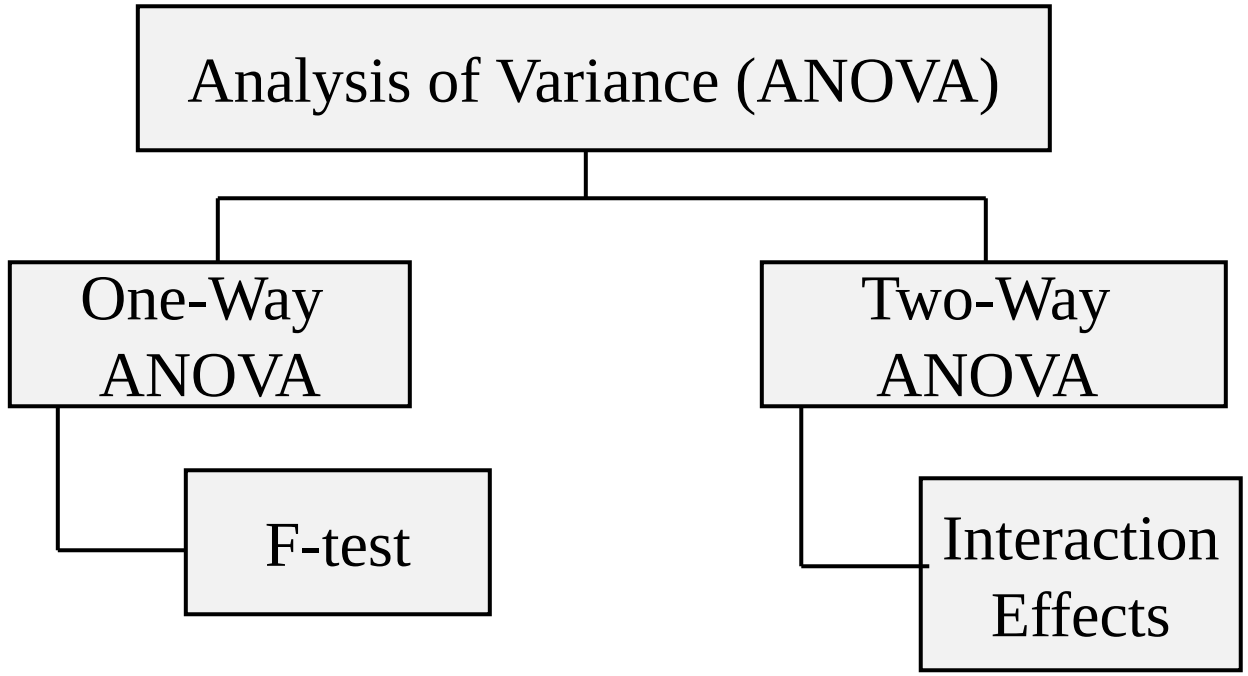
Chapter Overview

Example (one factor is tested)

1.To verify is there any significant difference between four fertilizers with respect to average yield

2.Is there any significant difference between three zonal managers with respect to average sales.

3.Is there any significant difference between four detergents with respect to average whiteness of clothing.



If there are two factors ,we would use a two way ANOVA

EXAMPLE:

If the four fertilizers are used in different plots of land then we can test

is there any significant difference between fertilizers &

is there any significant difference between plots of land with regard to average yield.

What is an Interaction Effect : The combined effect observed in response (output) variable on changing the specific levels of two or more.
1. Crispness of the bread depends on the interaction between temp and time 2. No of classes attended & No of hrs studied 3. If the no of people increase by 1% when the outside temp is hot, airconditioner performance drops by 4 %

The F Distribution

ANOVA

The probability distribution used in this chapter is F distribution. It was named to honor Sir Ronald Fisher, one of the founders of modern-day statistics. It is used to test whether two samples are from population having equal variances, and it is also applied when we want to compare several population means simultaneously.

The simultaneous comparison of several population means is called **analysis of variance (ANOVA).** In both of these situations, the populations must follow a normal distribution, and the data must be at least interval-scale.

1) There are a family of F distribution. A particular member of the family is determined by two parameters: the degrees of freedom in the numerator and the degrees of freedom in the denominator.

The F distribution cannot be negative. The smallest value F can assume is 0.

It is positively skewed. The long tail of the distribution is to the right-hand side. As the number of degrees of freedom increases in both the numerator and denominator the distribution approaches a normal distribution.

General ANOVA Setting

- Investigator controls one or more factors of interest
  - Each factor contains two or more levels
  - Levels can be numerical or categorical
  - Different levels produce different groups
  - Think of each group as a sample from a different population
- Observe effects on the dependent variable
  - Are the groups the same?
- Experimental design: the plan used to collect the data

- **One dependent (metric) variable**
- **There is only one categorical independent variable. Variable is called a Factor.**
- **Each category of an independent variable is called a level**
- The independent variable may be different levels of prices or different pack sizes or different product colours and the effect (dependent variable) could be sales preferences or attitudes towards the brand

Sales Manager (Factor)

Level

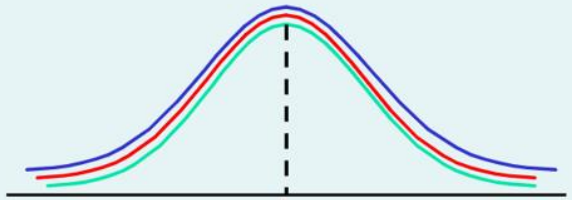
A	B	C
300	600	700
400	300	300
300	300	400
500	400	600
0	-	500

Sales revenue – Dependent variable

$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_c$

$H_1 :$ Not all μ_j are the same

The Null Hypothesis is True
All Means are the same:
(No Factor Effect)

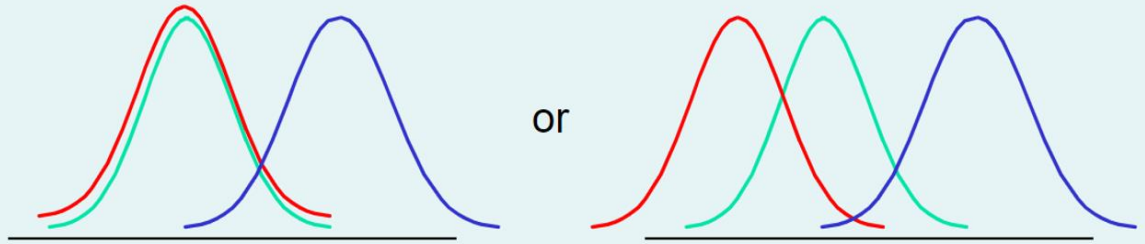


$\mu_1 = \mu_2 = \mu_3$

$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_c$

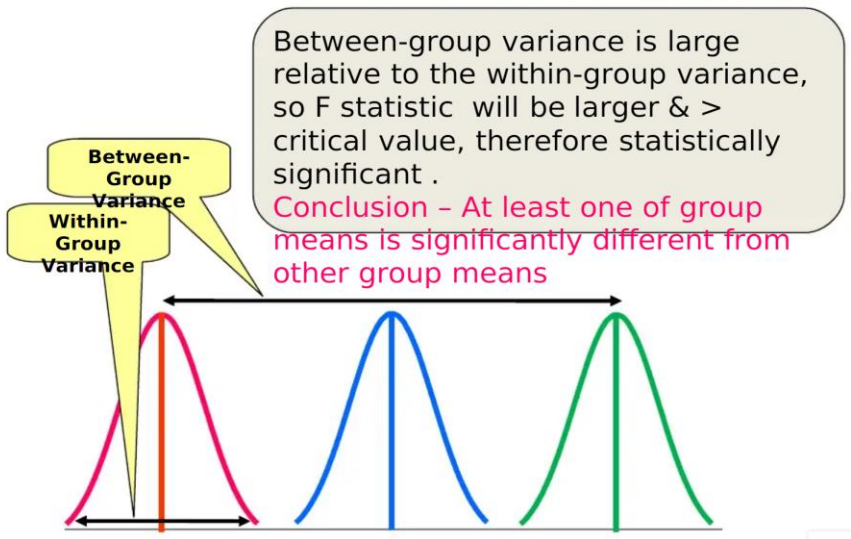
$H_1 :$ Not all μ_j are the same

The Null Hypothesis is NOT true
At least one of the means is different
(Factor Effect is present)

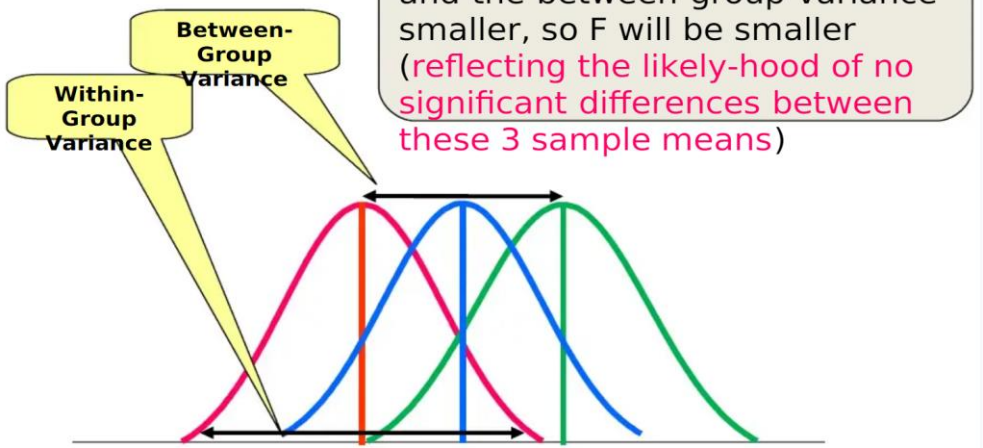


$\mu_1 = \mu_2 \neq \mu_3$

$\mu_1 \neq \mu_2 \neq \mu_3$



Between-group variance is large relative to the within-group variance, so F statistic will be larger & > critical value, therefore statistically significant .
Conclusion – At least one of group means is significantly different from other group means

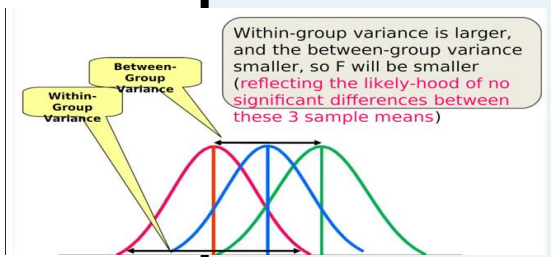


Within-group variance is larger, and the between-group variance smaller, so F will be smaller (reflecting the likely-hood of no significant differences between these 3 sample means)

Partitioning the Variation

- Total variation can be split into two parts:

$$SST = SSA + SSW$$



SST = Total Sum of Squares
(Total variation)

SSA = Sum of Squares Among Groups
(Among-group variation)

SSW = Sum of Squares Within Groups
(Within-group variation)

Partition of Total Variation

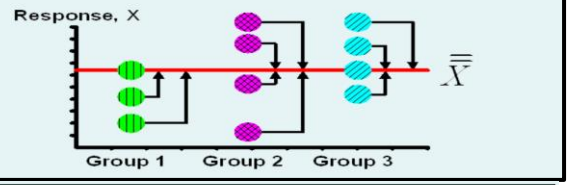
Total Variation (*SST*)

= **Variation Due to Factor (*SSA*)** + **Variation Due to Random Error (*SSW*)**

Total Variation

DCOVA (continued)

$$SST = (X_{11} - \overline{\overline{X}})^2 + (X_{12} - \overline{\overline{X}})^2 + \cdots + (X_{cn_c} - \overline{\overline{X}})^2$$

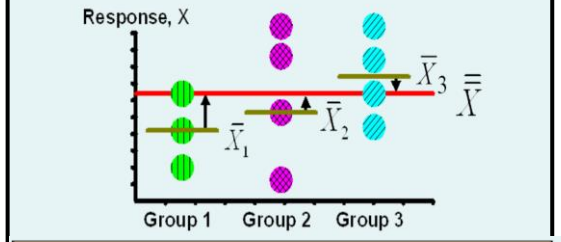


Total Variation (*SST*)

$$MST = \frac{SST}{n-1}$$

=

$$SSA = n_1(\overline{X}_1 - \overline{\overline{X}})^2 + n_2(\overline{X}_2 - \overline{\overline{X}})^2 + \cdots + n_c(\overline{X}_c - \overline{\overline{X}})^2$$



Variation Due to Factor (*SSA*)

$$MSA = \frac{SSA}{c-1}$$

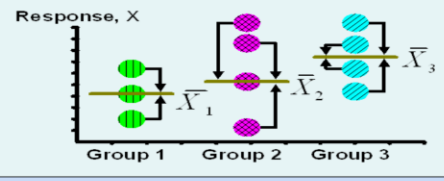
Mean Square Among = SSA/degrees of freedom

+

Within-Group Variation

DCOVA (continued)

$$SSW = (X_{11} - \overline{X}_1)^2 + (X_{12} - \overline{X}_2)^2 + \cdots + (X_{cn_c} - \overline{X}_c)^2$$



Variation Due to Random Error (*SSW*)

$$MSW = \frac{SSW}{n-c}$$

Mean Square Within = SSW/degrees of freedom

$$F_{STAT} = \frac{MSA}{MSW}$$

Obtaining the Mean Squares

$$MSA = \frac{SSA}{c-1}$$

Mean Squares Among

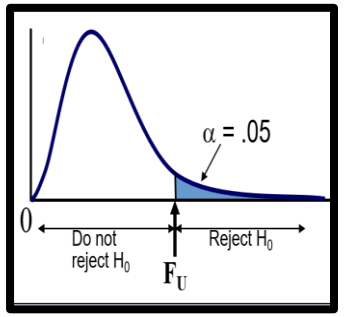
$$MSW = \frac{SSW}{n-c}$$

Mean Squares Within

$$MST = \frac{SST}{n-1}$$

Mean Squares Within

One-Way ANOVA Table



Decision Rule:

Reject H_0 if $F > F_U$,
otherwise do not reject H_0

Source of Variation	df	SS	MS (Variance)	F-Ratio
Among Groups	c-1	SSA	MSA	$F = \frac{MSA}{MSW}$
Within Groups	n-c	SSW	MSW	
Total	n-1	SST = SSA+SSW		

c = number of groups

n = sum of the sample sizes from all groups

df = degrees of freedom

- *MSA* is mean squares among variances
- *MSW* is mean squares within variances

- The F statistic is the ratio of the among variance to the within variance
  - The ratio must always be positive
  - $df_1 = c - 1$ will typically be small
  - $df_2 = n - c$ will typically be large

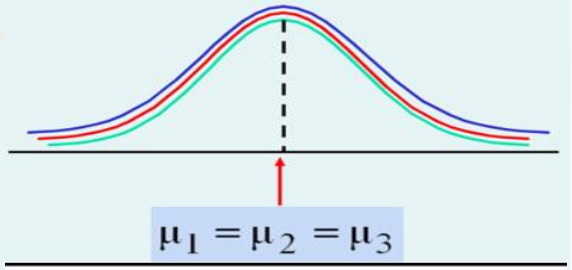
One-Way Analysis of Variance

- Evaluate the difference among the means of three or more groups

> Examples: Accident rates for 1st, 2nd, and 3rd shift
>
> Expected mileage for five brands of tires

- Assumptions
  - Populations are normally distributed
  - Populations have equal variances
  - Samples are randomly and independently drawn

Hypotheses of One-Way ANOVA



- $H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_c$
  - All population means are equal
  - i.e., no factor effect (no variation in means among groups)
- H_1 : Not all of the population means are the same
  - At least one population mean is different
  - i.e., there is a factor effect
  - Does not mean that all population means are different (some pairs may be the same)

One-way ANOVA in Excel

You may conduct a one-way ANOVA using Excel,

(Preliminary step) First, make sure that the “Analysis ToolPak” is installed.

Under "Tools" is the option "Data Analysis" present?

If yes - ToolPak is installed.

If no- select "Add-ins."

Check the boxes entitled “Analysis ToolPak” and “Analysis ToolPak - VBA" and click "OK". This will install the "Data Analysis ToolPak."

1. Under “Tools” select “Data Analysis”

   In the window that appears select “ANOVA: One factor" and click “OK.”

2. Using your mouse highlight the cells containing the data.

3. Select “Columns” if each treatment is its own column or "Row" if each treatment is its own row.

4. Set your level of significance. (The default is 5% or 0.05.)

5. Click “OK” and the ANOVA output will appear on a new worksheet.

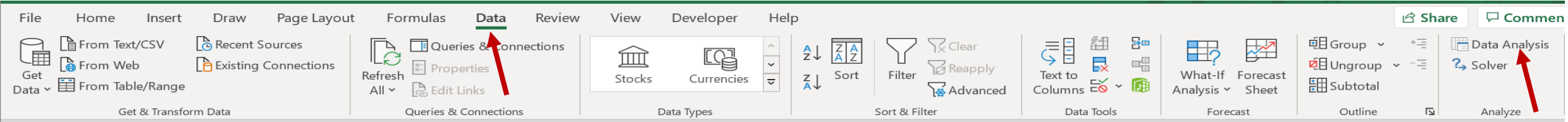
ANOVA – Problem 1

1. The Amit Merchandising Company wishes to test whether its three salesmen, A, B and C tend to make sales of the same size or whether they differ in their selling ability as measured by the average size of their sales. During the last week, out of 14 sales, A made 5, B made 4 and C made 5 calls. The following are the weekly sales (in Rs. Thousand) record of three salesmen :

A	B	C
300	600	700
400	300	300
300	300	400
500	400	600
0	-	500

Test whether the three salesmen's average sales differ in size.

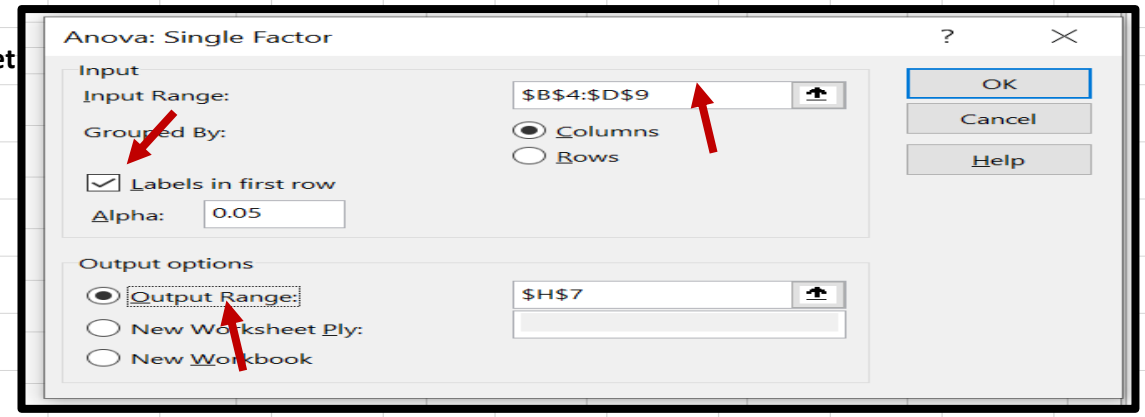
Objective : To find out whether there is no significant difference in the average sales value of the three salesmen.

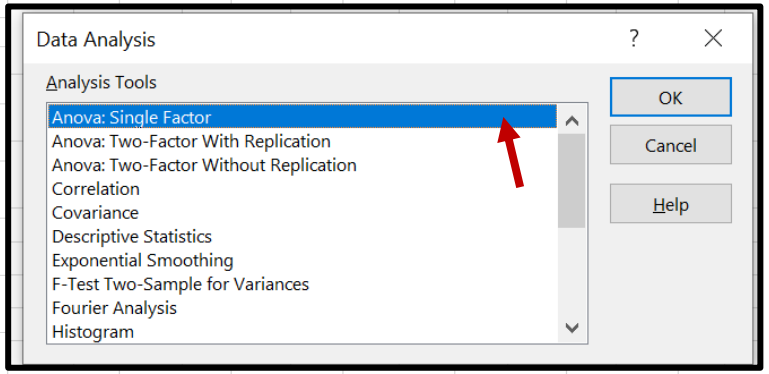


B4 | A

Step 1 :Enter the values in the work sheet

A	B	C
300	600	700
400	300	300
300	300	400
500	400	600
0		500





Decision :

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

H_1 : Not all Means are equal

Approach 1 : Since calculated value of F (1.833) is < $F_{critical}$(3.9823) Accept H_0

Approach 2 : Since p value (0.2055) is > alpha (0.05) Accept H_0

SUMMARY — Anova: Single Factor

Groups	Count	Sum	Average	Variance
A	5	1500	300	35000
B	4	1600	400	20000
C	5	2500	500	25000

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	100000	2	50000	1.83333	0.20551	3.9823
Within Groups	300000	11	27272.7			
Total	400000	13				

Conclusion :

There is no significant difference in the average sales of the 3 Salesman

ANOVA – Problem 2

1. Four machines A, B, C and D are used to produce a certain kind of cotton fabrics. Samples of size 4 with each unit as 100 square metres are selected from the outputs of the machines at random, and the number of flaws in each 100 square metres are counted with the following result.

A	B	C	D
8	6	14	20
9	8	12	22
11	10	18	25
12	4	9	23

Do you think that there is a significant difference in the performance of the four machines?

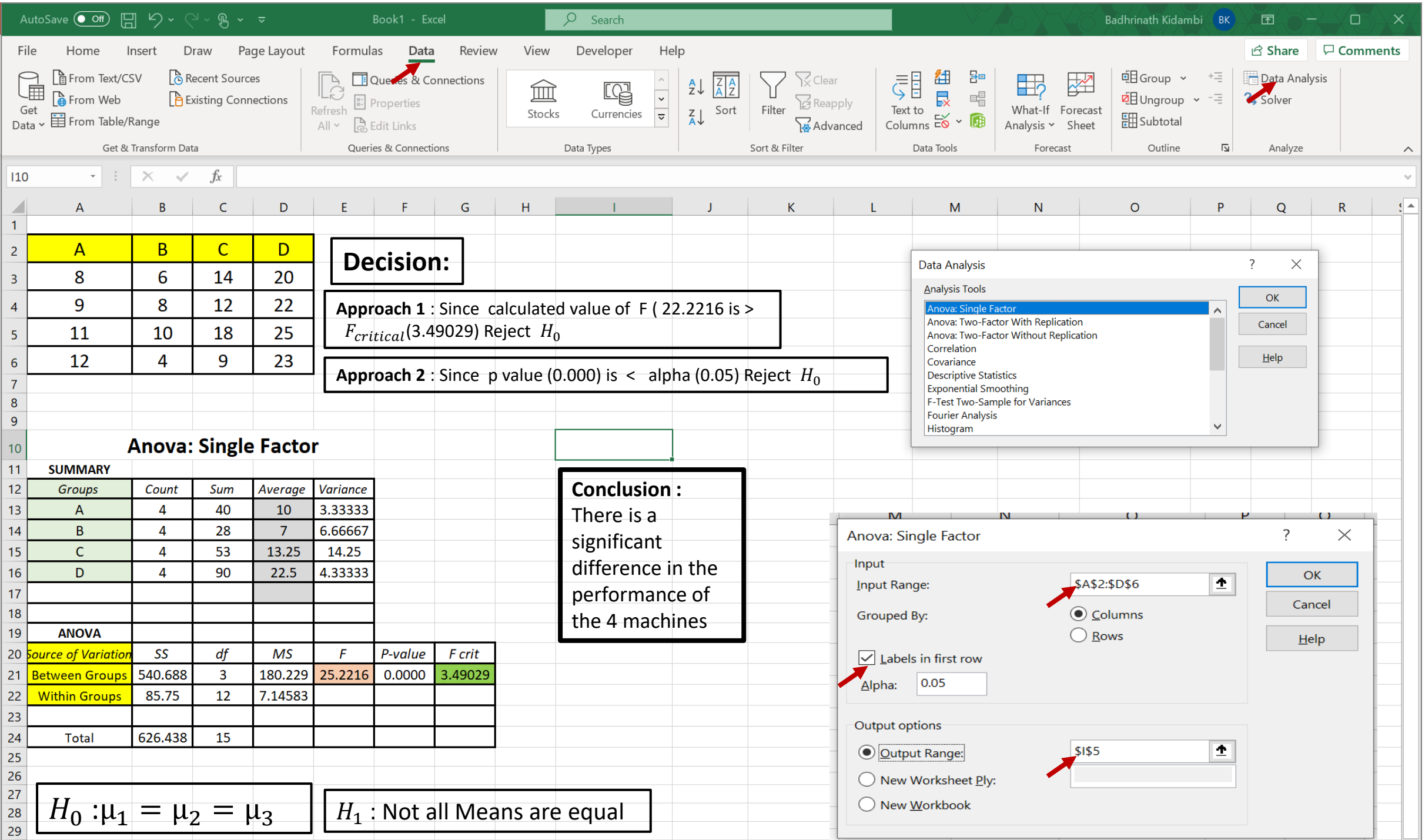
- **Objective : To find out whether the machines do not differ significantly in performance**

ANS

- ANOVA table

Source of variation	Sum of squares	df	Mean sum of squares	F Ratio
Between columns	540.69	3	540.69 / 3 = 180.23 = MSC	F = MSC/MSE
Within column	85.75	12	85.75 / 12 = 7.15 = MSE	180.23 / 7.15 = 25.207
Total	626.44	15		

- Table value of F — Nr. df = 3 | Dr. df = 12 | LOS = 1% | 3.49
- Decision — Since calculated value of F (25.207) is more than table value (3.49), reject H0.
- Conclusion — There is a significant difference in the performance of the four machines.



A	B	C	D
8	6	14	20
9	8	12	22
11	10	18	25
12	4	9	23

Decision:

Approach 1 : Since calculated value of F (22.2216 is > $F_{critical}$(3.49029) Reject H_0

Approach 2 : Since p value (0.000) is < alpha (0.05) Reject H_0

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
A	4	40	10	3.33333
B	4	28	7	6.66667
C	4	53	13.25	14.25
D	4	90	22.5	4.33333

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	540.688	3	180.229	25.2216	0.0000	3.49029
Within Groups	85.75	12	7.14583			
Total	626.438	15				

Conclusion : There is a significant difference in the performance of the 4 machines

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

H_1 : Not all Means are equal

Two way ANOVA

- A two-way ANOVA is used to estimate how the mean of a quantitative variable changes according to the levels of two categorical variables. Use a two-way ANOVA when you want to know how two independent variables, in combination, affect a dependent variable.
- You can use a two-way ANOVA when you have collected data on a quantitative dependent variable at multiple levels of two categorical independent variables.
- A two-way ANOVA with interaction tests three null hypotheses at the same time:
  - There is no difference in group means at any level of the first independent variable.
  - There is no difference in group means at any level of the second independent variable.
  - The effect of one independent variable does not depend on the effect of the other independent variable (a.k.a. no interaction effect).
  - A two-way ANOVA without interaction (a.k.a. an additive two-way ANOVA) only tests the first two of these hypotheses.

Assumptions of the two-way ANOVA

- To use a two-way ANOVA your data should meet certain assumptions.Two-way ANOVA makes all of the normal assumptions of a parametric test of difference:

1. **Homogeneity of variance**
    - The variation around the mean for each group being compared should be similar among all groups. If your data don't meet this assumption, you may be able to use a non-parametric alternative, like the Kruskal-Wallis test.

2. **Independence of observations**
    - Your independent variables should not be dependent on one another (i.e. one should not cause the other). This is impossible to test with categorical variables – it can only be ensured by good experimental design.
    - In addition, your dependent variable should represent unique observations – that is, your observations should not be grouped within locations or individuals.

- If your data don't meet this assumption (i.e. if you set up experimental treatments within blocks), you can include a blocking variable and/or use a repeated-measures ANOVA.

3. **Normally-distributed dependent variable**
    - The values of the dependent variable should follow a bell curve. If your data don't meet this assumption, you can try a data transformation.

Two-Way ANOVA
Sources of Variation

$$SST = SSA + SSB + SSAB + SSE$$

SST
Total Variation

n - 1

SSA
Factor A Variation

SSB
Factor B Variation

SSAB
Variation due to interaction between A and B

SSE
Random variation (Error)

Two Factors of interest: A and B

- r = number of levels of factor A
- c = number of levels of factor B
- n' = number of replications for each cell
- n = total number of observations in all cells ($n = rcn'$)
- X_{ijk} = value of the kth observation of level i of factor A and level j of factor B

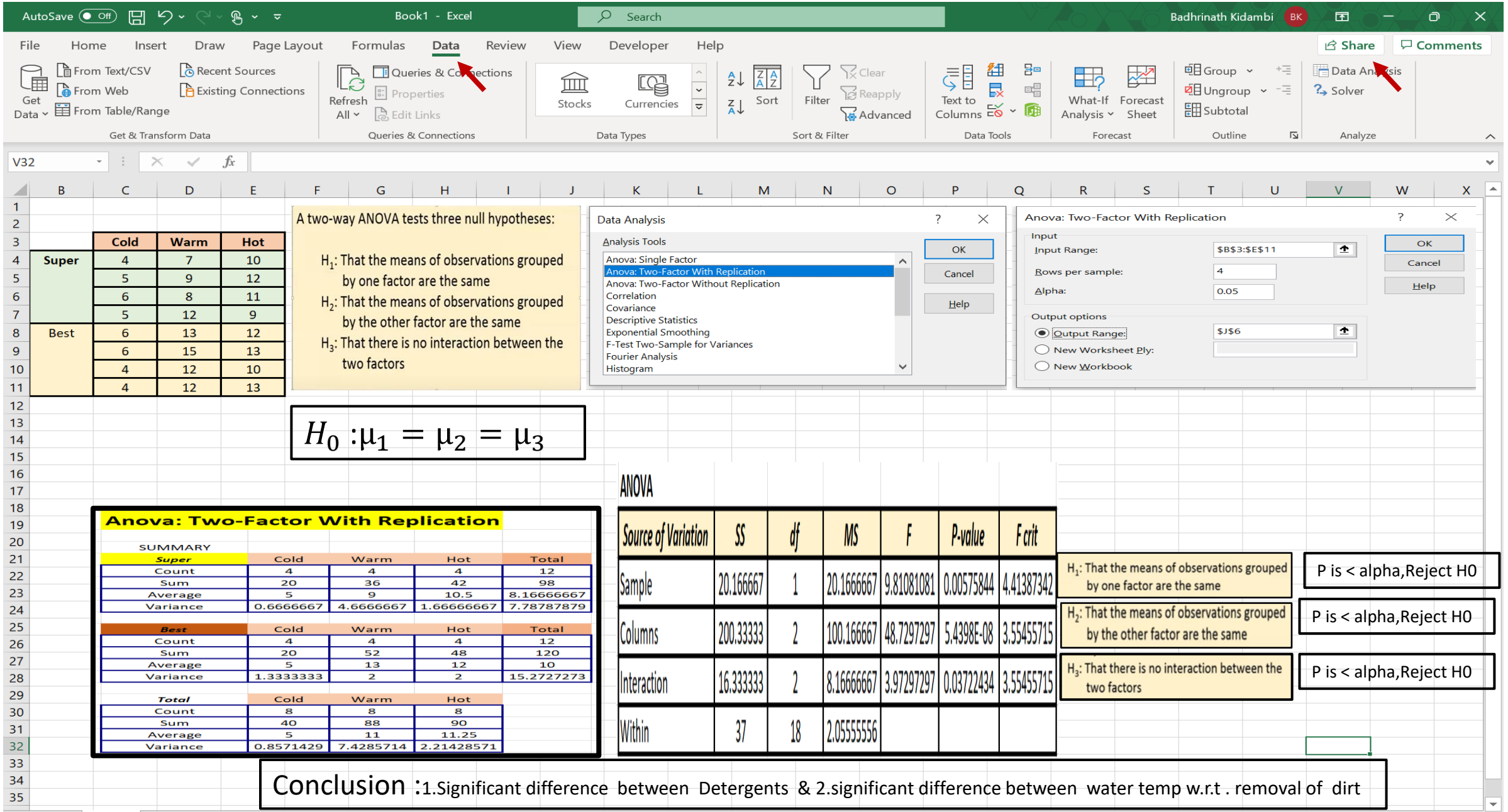
Degrees of Freedom:
r – 1
c – 1
(r – 1)(c – 1)
rc(n' – 1)

Two-Way ANOVA: Summary Table

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F Statistic
Factor A	**r – 1**	**SSA**	**MSA** = SSA /(r – 1)	**MSA/ MSE**
Factor B	**c – 1**	**SSB**	**MSB** = SSB /(c – 1)	**MSB/ MSE**
AB (Interaction)	**(r – 1)(c – 1)**	**SSAB**	**MSAB** = SSAB / (r – 1)(c – 1)	**MSAB/ MSE**
Error	**rc(n′ – 1)**	**SSE**	**MSE** = SSE/rc(n’ – 1)	
Total	**n – 1**	**SST**		

Problem

- Suppose you want to determine whether the brand of laundry detergent used and the temperature affects the amount of dirt removed from your laundry.
- To this end, you buy two different brand of detergent (“ Super” and “Best”) and choose three different temperature levels (“cold”, “warm”, and “hot”).
- Then you divide your laundry randomly into 6×r piles of equal size and assign each r piles into the combination of (“Super” and “Best”) and (”cold”,”warm”, and “hot”).
- In this example, we are interested in testing Null Hypotheses H0D : The amount of dirt removed does not depend on the type of detergent
- H0 : The amount of dirt removed does not depend on the temperature One says the experiment hastwo factors (Factor Detergent, Factor Temperature) at a = 2(Super and Best) and b = 3(cold,warm and hot) levels.
- Thus there are ab = 3 × 2 = 6 different combinations of detergent and temperature. With each combination you wash r = 4 loads.
- r is called the number of replicates. This sums up to n = abr = 24 loads in total.


AutoSave
Book1 - Excel
Search
Badhrinath Kidambi
File Home Insert Draw Page Layout Formulas Data Review View Developer Help
Share Comments
Get Data
From Text/CSV
From Web
From Table/Range
Recent Sources
Existing Connections
Refresh All
Queries & Connections
Properties
Edit Links
Stocks
Currencies
Sort
Filter
Clear
Reapply
Advanced
Text to Columns
What-If Analysis
Forecast Sheet
Group
Ungroup
Subtotal
Solver
Get & Transform Data
Queries & Connections
Data Types
Sort & Filter
Data Tools
Forecast
Outline
Analyze
V32
A two-way ANOVA tests three null hypotheses:
H1: That the means of observations grouped by one factor are the same
H2: That the means of observations grouped by the other factor are the same
H3: That there is no interaction between the two factors
Data Analysis
Analysis Tools
Anova: Single Factor
Anova: Two-Factor With Replication
Anova: Two-Factor Without Replication
Correlation
Covariance
Descriptive Statistics
Exponential Smoothing
F-Test Two-Sample for Variances
Fourier Analysis
Histogram
OK
Cancel
Help
Anova: Two-Factor With Replication
Input
Input Range: $B$3:$E$11
Rows per sample: 4
Alpha: 0.05
Output options
Output Range: $J$6
New Worksheet Ply:
New Workbook
OK
Cancel
Help
$H_0 : \mu_1 = \mu_2 = \mu_3$
Anova: Two-Factor With Replication
SUMMARY
Super Cold Warm Hot Total
Count 4 4 4 12
Sum 20 36 42 98
Average 5 9 10.5 8.16666667
Variance 0.6666667 4.6666667 1.66666667 7.78787879
Best Cold Warm Hot Total
Count 4 4 4 12
Sum 20 52 48 120
Average 5 13 12 10
Variance 1.3333333 2 2 15.2727273
Total Cold Warm Hot
Count 8 8 8
Sum 40 88 90
Average 5 11 11.25
Variance 0.8571429 7.4285714 2.21428571
ANOVA
Source of Variation SS df MS F P-value F crit
Sample 20.166667 1 20.1666667 9.81081081 0.00575844 4.41387342
Columns 200.33333 2 100.166667 48.7297297 5.4398E-08 3.55455715
Interaction 16.333333 2 8.16666667 3.97297297 0.03722434 3.55455715
Within 37 18 2.05555556
H1: That the means of observations grouped by one factor are the same
P is < alpha,Reject H0
H2: That the means of observations grouped by the other factor are the same
P is < alpha,Reject H0
H3: That there is no interaction between the two factors
P is < alpha,Reject H0
Conclusion :1.Significant difference between Detergents & 2.significant difference between water temp w.r.t . removal of dirt
Sheet1
Ready
100%

Chapter Summary

We have

- Described one-way analysis of variance
  - The logic of ANOVA
  - ANOVA assumptions
- Described two-way analysis of variance
  - Examined effects of multiple factors
  - Examined interaction between factors