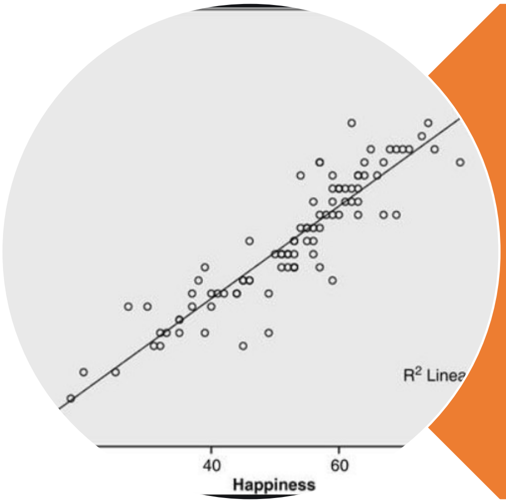
The Complete AI Course

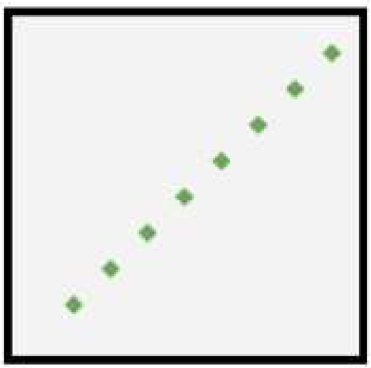
Bivariate Data Analysis



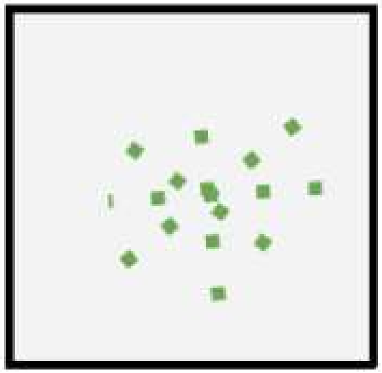
Bivariate Data Analysis

- Covariance
- Correlation
- Collinearity
- Multicollinearity
- Variance Inflation Factor
- Homoscedasticity
- Heteroscedasticity

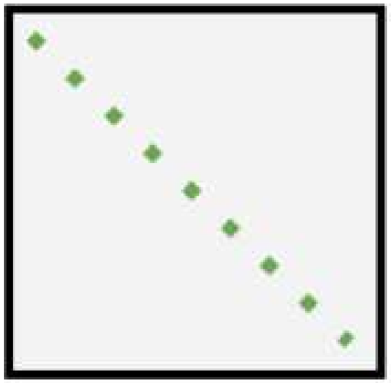
COVARIANCE



Large Positive Covariance



Nearly Zero Covariance



Large Negative Covariance

COVARIANCE

Population Covariance Formula

$$Cov(x,y) = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{N}$$

Sample Covariance

$$Cov(x,y) = \frac{\Sigma(x_i-\bar{x})(y_i-y)}{N-1}$$

Notations in Covariance Formulas

- x_i = data value of x
- y_i = data value of y
- $\bar{x}$ = mean of x
- $\bar{y}$ = mean of y
- N = number of data values.

CORRELATION

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where:

- r_{xy} – the correlation coefficient of the linear relationship between the variables x and y
- x_i – the values of the x-variable in a sample
- $\bar{x}$ – the mean of the values of the x-variable
- y_i – the values of the y-variable in a sample
- $\bar{y}$ – the mean of the values of the y-variable

#learnaiwithramisha

TYPES OF CORRELATION:: LINEAR TYPE WITH TWO VARIABLES

ALL ABOUT CORRELATION

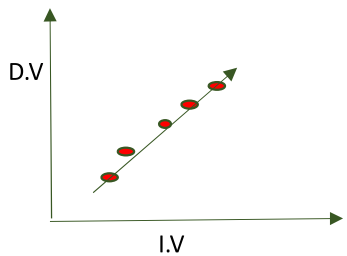
SEE THE PICTURE AND TELL THE STORY

Correlation explained in terms of degree(amount of change from one point to another point))

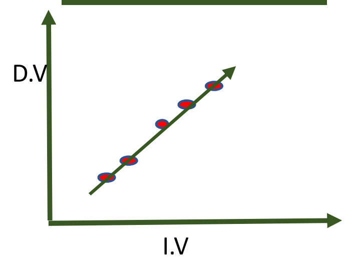
POSITIVE CORRELATION

Independent Variable is directly proportional to Dependant Variable

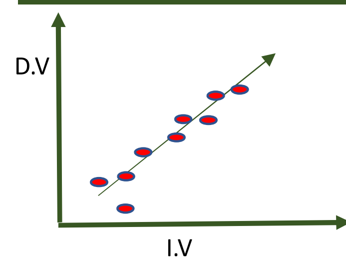
I.V ↑ D.V ↑ I.V α D.V



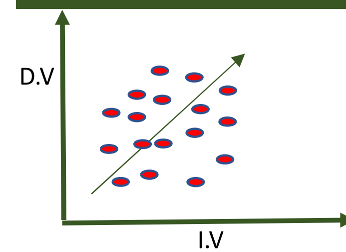
1. Perfect positive Correlation(r= 1)



2. High Degree of +Ve Correlation (r= + High or 0.95):



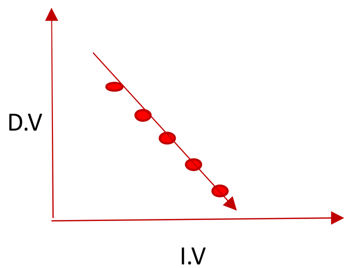
3. Low degree of +Ve Correlation (r= + Low or 0.54):



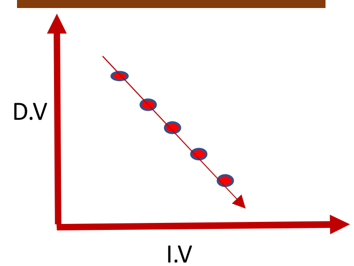
NEGATIVE CORRELATION

Independent Variable is indirectly proportional to Dependant Variable

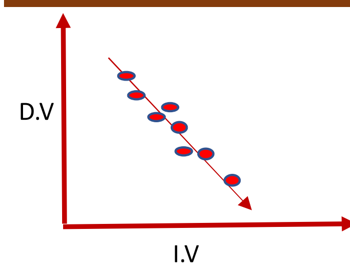
I.V ↑ D.V ↓ I.V α 1/D.V



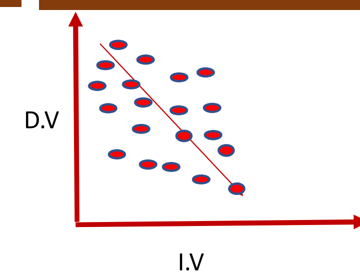
4. Perfect Negative Correlation (r=-1)



5.High Degree of –Ve Correlation (r= – low or -0.54):

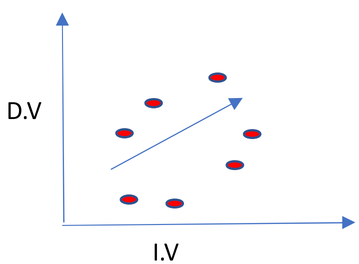


6. Low Degree of –Ve Correlation (r= - high or -0.94)

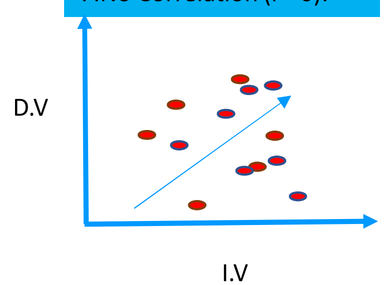


ZERO CORRELATION

No pattern between Independent Variable And Dependant Variable



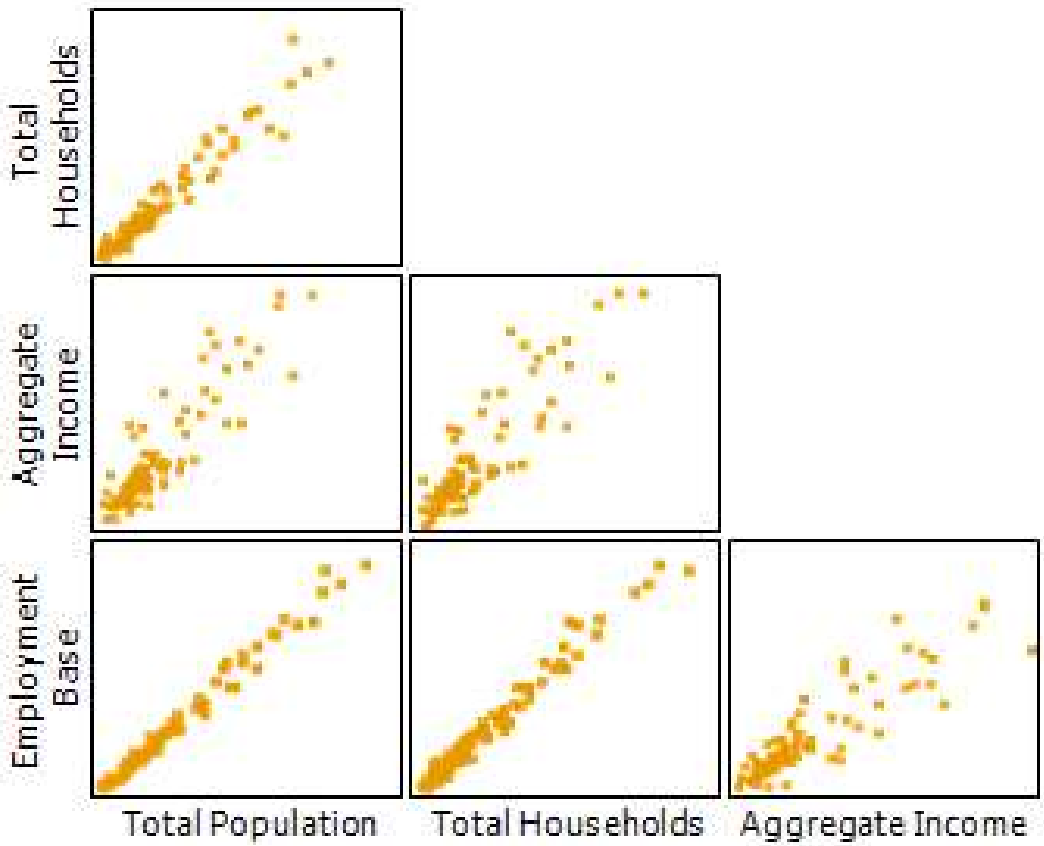
7.No Correlation (r= 0):



INFERENCE

According to problem statement, from scattered diagram can understand the relation between two variables.

MULTI-COLINEAR | TESTING



MULTI-COLINEAR | TESTING

- An easy way to detect multicollinearity is to calculate correlation coefficients for all pairs of predictor variables.
- If the correlation coefficient, r, is exactly +1 or -1, this is called perfect multicollinearity.
- If r is close to or exactly -1 or +1, one of the variables should be removed from the model if at all possible
- Multicollinearity generally occurs when there are high correlations between two or more predictor variables.
- In other words, one predictor variable can be used to predict the other.
- This creates redundant information, skewing the results in a regression model.
- Examples of correlated predictor variables (also called multicollinear predictors) are: a person's height and weight, age and sales price of a car, or years of education and annual income.

KINDS OF MULTICOLLINEARIT

Structural multicollinearity:

- This type occurs when we create a model term using other terms.
- In other words, it's a byproduct of the model that we specify rather than being present in the data itself.
- For example, if you square term X to model curvature, clearly there is a correlation between X and X^2.

Data multicollinearity:

- This type of multicollinearity is present in the data itself rather than being an artifact of our model. Observational experiments are more likely to exhibit this kind of multicollinearity.

Variance Inflation Factor(VIF)

- A variance inflation factor(VIF)detects multicollinearity in regression analysis.
- Multicollinearity is when there's correlation between predictors (i.e. independent variables) in a model;
- it's presence can adversely affect your regression results.
- The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

$$\mathrm{VIF} = \frac{1}{1 - R_i^2}$$

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

Example: Multicollinearity

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0705118	56.23%	54.22%	50.48%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.155	0.132	1.18	0.243	
%Fat	0.00557	0.00409	1.36	0.176	14.93
Weight kg	0.01447	0.00285	5.07	0.000	33.95
Activity	0.000022	0.000007	3.08	0.003	1.05
%Fat*Weight kg	-0.000214	0.000074	-2.90	0.005	75.06

Example: Multicollinearity

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0705118	56.23%	54.22%	50.48%

Coefficients

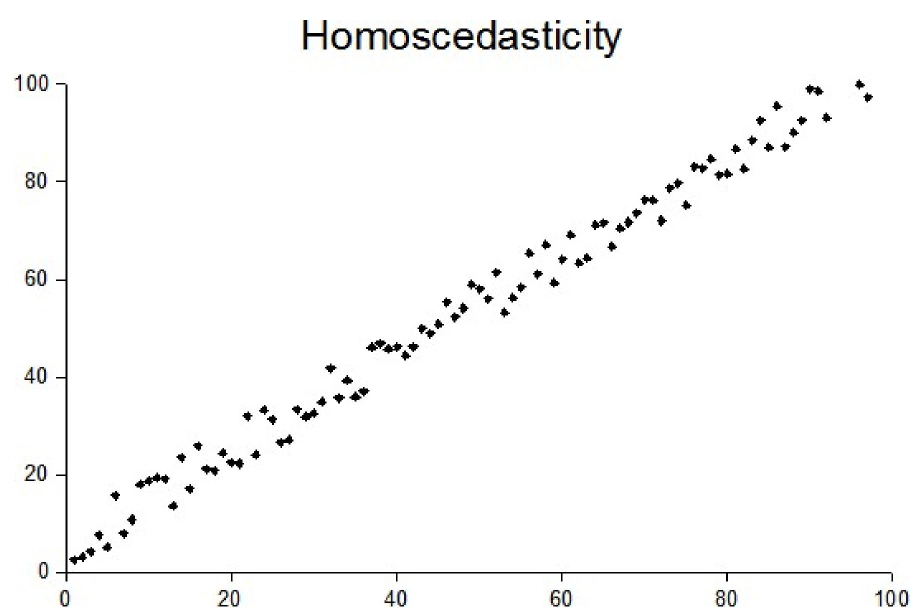
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.82161	0.00973	84.40	0.000	
%Fat S	-0.00598	0.00193	-3.10	0.003	3.32
Weight S	0.00835	0.00107	7.83	0.000	4.75
Activity S	0.000022	0.000007	3.08	0.003	1.05
%Fat S*Weight S	-0.000214	0.000074	-2.90	0.005	1.99

Homoscedasticity

- The assumption of homoscedasticity (meaning "same variance") is central to linear regression models.

- Homoscedasticity describes a situation in which the error term (that is, the "noise" or random disturbance in the relationship between the independent variables and the dependent variable) is the same across all values of the independent variables.

Homoscedasticity



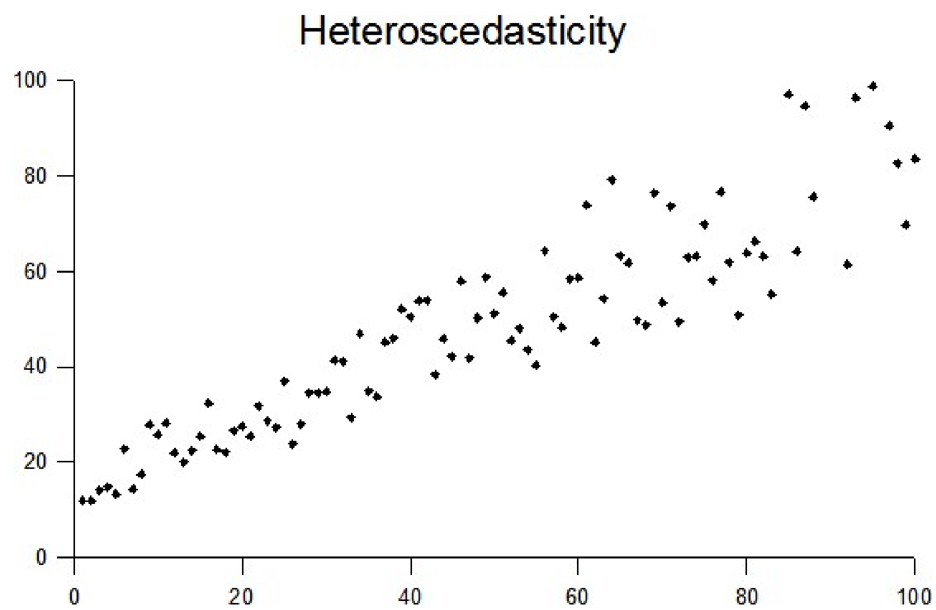
Homoscedasticity

Heteroscedasticity (the violation of homoscedasticity) is present when the size of the error term differs across values of an independent variable.

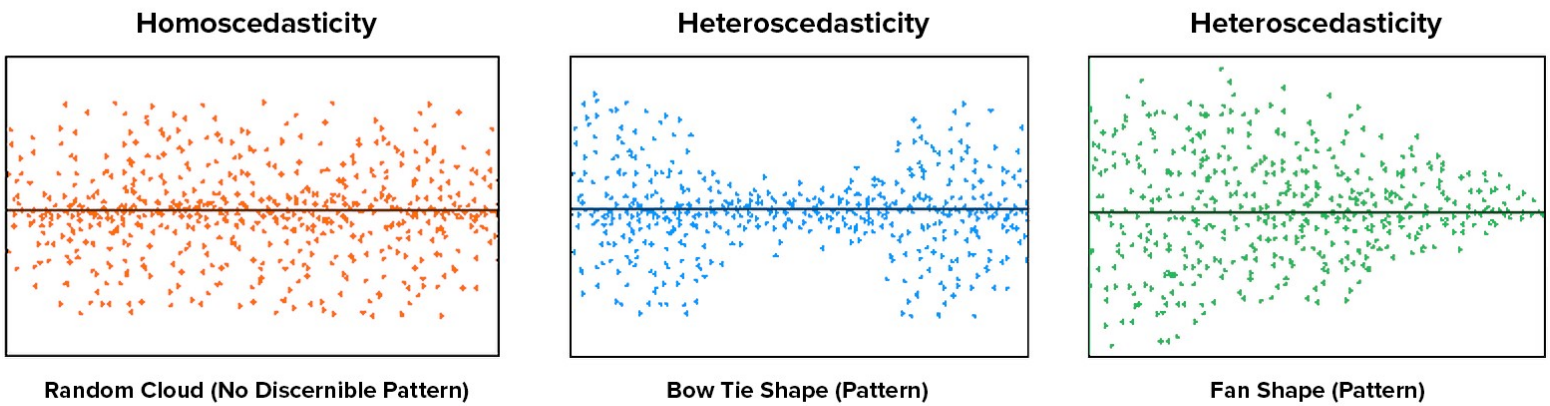
The impact of violating the assumption of homoscedasticity is a matter of degree, increasing as heteroscedasticity increases.

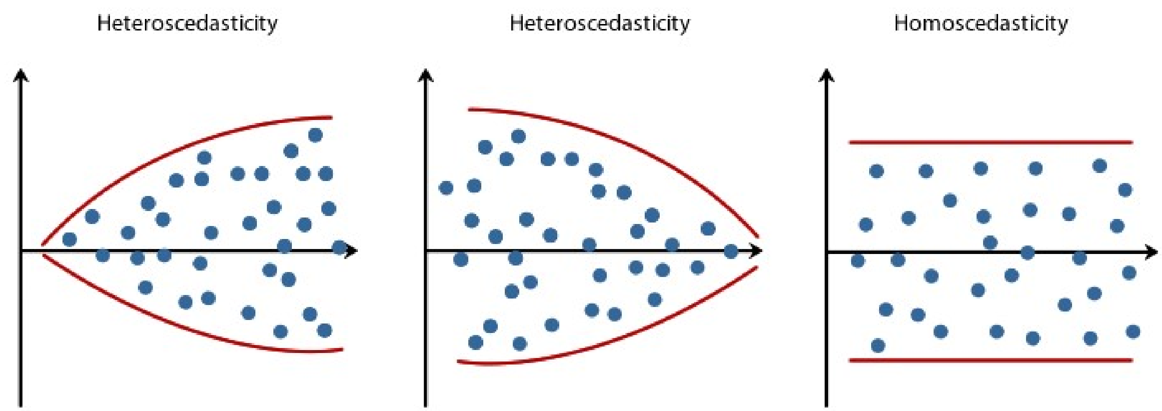
$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

Homoscedasticity



Homoscedasticity





Measure of spread |Z-Score

Comparing Values from Different Data Sets

Two students, John and Ali, from different high schools, wanted to find out who had the highest GPA when compared to his school. Which student had the highest GPA when compared to his school?

Student	GPA	School mean GPA	School standard deviation
John	2.85	3.0	0.7
Ali	77	80	10

For John, $z = \#ofSTDEVs = \frac{2.85 - 3.0}{0.7} = -0.21$

For Ali, $z = \#ofSTDEVs = \frac{77 - 80}{10} = -0.3$

Measure of spread |Z-Score

Comparing Values from Different Data Sets

For John, $z = \#ofSTDEVs = \frac{2.85-3.0}{0.7} = -0.21$

For Ali, $z = \#ofSTDEVs = \frac{77-80}{10} = -0.3$

John has the better GPA when compared to his school because his GPA is 0.21 standard deviations **below** his school's mean while Ali's GPA is 0.3 standard deviations **below** his school's mean.

John's z-score of −0.21 is higher than Ali's z-score of −0.3. For GPA, higher values are better, so we conclude that John has the better GPA when compared to his school.

T - TEST

T-Test: how significant the similarity between groups

T-Test : To find similarity between two groups based on Mean.

Paired T-Test: Independent Sample

Unpaired T-Test: Dependent Sample

T - TEST

T-Test : To find similarity between two groups based on Mean.

What is T-Test?

The t test tells you how significant the differences between groups are;
In other words it lets you know if those differences (measured in means/averages) could have happened by chance.

Example:

- ✓ Let's say you have a cold and you try a naturopathic remedy. Your cold lasts a couple of days.
- ✓ The next time you have a cold, you buy an over-the-counter pharmaceutical and the cold lasts a week.
- ✓ You survey your friends and they all tell you that their colds were of a shorter duration (an average of 3 days) when they took the homeopathic remedy.
- ✓ What you *really* want to know is, are these results repeatable?

A t test can tell you by comparing the means of the two groups and letting you know the probability of those results happening by chance.

T - TEST

T-Test : To find similarity between two groups based on Mean.

Interpret the T-Test Value

- A large t-score tells you that the groups are different.
- A small t-score tells you that the groups are similar.

T - TEST

T-Test : To find similarity between two groups based on Mean.

For example,

- p value of 5% is 0.05.
- **Low p-values are good**; They indicate your data did not occur by chance.
- For example, a p-value of .01 means there is only a 1% probability that the results from an experiment happened by chance.
- In most cases, a p-value of 0.05 (5%) is accepted to mean the data is valid.

Hypothesis Testing

Null Hypothesis: The mean is not same for both sample and population

Alternate Hypothesis: The Mean is same for both sample and population

Significance Level: 5%

Hypothesis Testing

Note : General procedure for Hypothesis tests

1. From the problem context, identify the parameter of interest.
2. State the null hypothesis, H_o.
3. Specify an appropriate alternative hypothesis, H_1
4. Choose a significance level α.
5. Determine an appropriate test statistic.
6. State the rejection region for the statistic.
7. Compute any necessary sample quantities, substitute these into the equation for the test statistic, and compute the value.
8. **Conclusion :** Decide whether or not H_o should be rejected and report that in the problem context.

Hypothesis Testing

■ (p) Level of significance ■ [A.U. N/D 2013]

The probability that the value of the statistic lies in the critical region is called the level of significance.

In general, these levels are chosen as 0.01 or 0.05, called 1% level and 5% level of significance respectively.

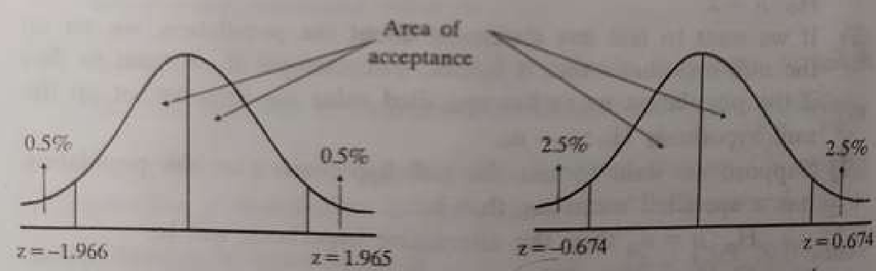


Fig. 1% level of significance

Fig. 5% level of significance

Hypothesis Testing

Note 2 : (i) For two-tailed test :

If $|z| < 1.96$ accept H_0 at 5% level of significance.

If $|z| > 1.96$ reject H_0 at 5% level of significance.

If $|z| < 2.58$ accept H_0 at 1% level of significance.

If $|z| > 2.58$ reject H_0 at 1% level of significance.

(ii) For single-tailed test : (Right or left)

If $|z| < 1.645$ accept H_0 at 5% level of significance.

If $|z| > 1.645$ reject H_0 at 5% level of significance.

If $|z| < 2.33$ accept H_0 at 1% level of significance.

If $|z| > 2.33$ reject H_0 at 1% level of significance.

Hypothesis Testing

Example

In one sample of 8 observations, the sum of the squares of deviati[ons] of the sample values from the sample mean was 84.4 and in the oth[er] sample of 10 observations it was 102.6. Test whether this differen[ce] is significant at 5% level, given that the 5 percent point of F f[or] $n_1 = 7$ and $n_2 = 9$ degrees of freedom is 3.29.

Solution : Given : $n_1 = 8,\ n_2 = 10$

$$\Sigma (X_1 - \overline{X}_1)^2 = 84.4$$

$$\Sigma (X_2 - \overline{X}_2)^2 = 102.6$$

$$S_1^2 = \frac{\Sigma (X_1 - \overline{X}_1)^2}{n_1 - 1} = \frac{84.4}{7} = 12.057$$

$$S_2^2 = \frac{\Sigma (X_2 - \overline{X}_2)^2}{n_2 - 1} = \frac{102.6}{9} = 11.42$$

1. The parameter of interest is σ_1^2 and σ_2^2
2. $H_o : \sigma_1^2 = \sigma_2^2$
3. $H_1 : \sigma_1^2 \neq \sigma_2^2$
4. $\alpha = 0.05$, d.f $(\nu_1) = n_1 - 1 = 7$, d.f $(\nu_2) = n_2 - 1 = 9$
5. **The test statistic is** $F = \dfrac{S_1^2}{S_2^2}$
6. **Reject** H_o if $F > 3.29$, [from table F]
7. **Computations :** $F = \dfrac{12.057}{11.42} = 1.057$
8. **Conclusion :** Since $F = 1.057 < 3.29$, we accept H_o at 5% level of significance.

#learnaiwithramisha

Hypothesis Testing

Example 1.2.b(2)

The sales manager of a large company conducted a sample survey in states A and B taking 400 samples in each case. The results were.

	State A	State B
Average sales	Rs. 2,500	Rs. 2,200
S.d	Rs. 400	Rs. 550

Hypothesis Testing

Test whether the average sales is the same in the 2 states at 1% level.

Solution : [A.U. M/J 2013]

Given : $\bar{x}_1 = 2500, \quad s_1 = 400, \quad n_1 = 400$

$\bar{x}_2 = 2200, \quad s_2 = 550, \quad n_2 = 400$

1. The parameter of interest is μ_1 and μ_2, difference of mean
2. $H_0 : \mu_1 = \mu_2$ [No significant difference between state A and State B]
3. $H_1 : \mu_1 \neq \mu_2$
4. $\alpha = 0.01$
5. The test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

6. Reject H_0 if $|Z| > 2.58$ at 1% level.
7. Computation :

$$Z = \frac{2500 - 2200}{\sqrt{\frac{(400)^2}{400} + \frac{(550)^2}{400}}} = \frac{300}{\sqrt{400 + 756.25}}$$

$$= \frac{300}{34.003} = 8.82$$

Conclusion :

Here $|Z| = 8.82 > 2.58$ So we reject $H_0 : \mu_1 = \mu_2$ at 1% level of significance.

Hence the average sales within two states differ significantly.

ANAVO

Analysis of Variance

One-Way Classification
→One Independent Variable

Two-Way Classification
→Two Independent Variable

One Way Classification

One-Way Classification →One Independent Variable

Working Procedure [One-way classification CRD]

1. H_0 : There is no significant difference between the treatments.

2. H_1 : There is significant difference between the treatments.

Step 1 : Find N, the number of observations

Step 2 : Find T, the total value of all observations

Step 3 : Find $\frac{T^2}{N}$, the correction factor

Step 4 : Calculate the total sum of squares.

$$TSS = \sum X_1^2 + \sum X_2^2 + \ldots - \frac{T^2}{N}$$

Step 5 : Calculate the column sum of squares

$$SSC = \frac{(\Sigma X_1)^2}{N_1} + \frac{(\Sigma X_2)^2}{N_1} + \frac{(\Sigma X_3)^2}{N_1} + \ldots - \frac{T^2}{N}$$

Here N_1 is the number of elements in each column.

$$SSE = TSS - SSC$$

Step 6 : Prepare the ANOVA table to calculate F-ratio.

Step 7 : Find the table value.

Step 8 : Conclusion :

One Way Classification

One-Way Classification →One Independent Variable

Example :1

There are three main brands of a certain powder. A set of 120 values is examined and found to be allocated among four groups (A, B, C and D) and three brands (I, II, III) as shown here under : [A.U. A/M.

Brands	Groups			
	A	B	C	D
I	0	4	8	15
II	5	8	13	6
III	8	19	11	13

Is there any significant difference in brands preference ? Answer at 5% level.

Solution : H_0 : There is no significant difference in brands.

H_1 : There is significant difference in brands.

Brands	Groups				Total	X_1^2	X_2^2	X_3^2
	A (X_1)	B (X_2)	C (X_3)	D (X_4)				
I (Y_1)	0	4	8	15	~~27~~	~~0~~	16	64
II (Y_2)	5	8	13	6	32	25	64	169
III(Y_3)	8	19	11	13	51	64	361	121
Total	13	31	32	34	110	89	441	354

Step 1 : $N = 12.$

Step 2 : $T = 110$

Step 3 : $\frac{T^2}{N} = \frac{(110)^2}{12} = 1008.3$

Step 4 : $TSS = \Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 + \Sigma X_4^4 - \frac{T^2}{N}$

$= 89 + 441 + 354 + 430 - 1008.3$

$= 305.7$

Step 5 : $SSR = \frac{(\Sigma Y_1)^2}{N_2} + \frac{(\Sigma Y_2)^2}{N_2} + \frac{(\Sigma Y_3)^2}{N_2} - \frac{T^2}{N}$

[N_2 → No. of elements in each row]

$= \frac{(27)^2}{4} + \frac{(32)^2}{4} + \frac{(51)^2}{4} - 1008.3$

$= 182.25 + 256 + 650.25 - 1008.3 = 80.2$

$SSE = TSS - SSR$

$= 305.7 - 80.2 = 225.50$

Step 6 : ANOVA

Source of variation	Sum of squares	d.f.	Mean square	Variance ratio	Table value at 5% level
between rows	SSR = 80.2	$r - 1 = 3 - 1 = 2$	$MSR = \frac{SSR}{r-1} = \frac{80.2}{2} = 40.1$	$F_R = \frac{MSR}{MSE} = \frac{40.1}{20.06} = 1.999$	F_R (2,9) = 4.26
Error	SSE = 225.5	$N - r = 12 - 3 = 9$	$MSE = \frac{SSE}{N-r} = \frac{225.5}{9} = 20.06$		
Total	305.7				

Conclusion : Cal F_R < Table F_R

So the accept H_0.

One Way Classification

One-Way Classification →One Independent Variable

Example :2

Example 2.2.4

The following table shows the lives in hours of four brands of elec lamps.

Brand A :	1610	1610	1650	1680	1700	1720	1800	
B :	1580	1640	1640	1700	1750			
C :	1460	1550	1600	1620	1640	1660	1740	18
D :	1510	1520	1530	1570	1600	1680		

Perform an analysis of variance test the homogneity of the mean li of the four brands of Lamps. [A.U. A/M. 2008] [A.U N/D 20 [A.U Tvli M/J 20

Solution :

H_0 : There is no significant difference between the four brands.

H_1 : There is a significant difference between the four brands.

Subtract 1600 and then divided by 10 we get

X_1 A	X_2 B	X_3 C	X_4 D	Total	X_1^2	X_2^2	X_3^3	X_4^2
1	−2	−14	−9	−24	1	4	196	81
1	4	−5	−8	−8	1	16	25	64
5	4	0	−7	2	25	16	0	49
8	10	2	−3	17	64	100	4	9
10	15	4	0	29	100	225	16	0
12	–	6	8	26	144	–	36	64
20	–	14	–	34	400	–	196	–
–	–	22	–	22	–	–	484	–
57	31	29	−19	98	735	361	957	267

Step 1 : N = 26

2.15

Step 2 : T = 98

Step 3 : C.F $= \frac{T^2}{N} = \frac{9604}{26} = 369.39$

Step 4 : TSS $= \Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 + \Sigma X_4^2 - \frac{T^2}{N}$

$= 735 + 361 + 957 + 267 - 369.39$

$= 1950.61$

Step 5 : SSC $= \frac{(\Sigma X_1)^2}{N_1} + \frac{(\Sigma X_2)^2}{N_1} + \frac{(\Sigma X_3)^2}{N_1} + \frac{(\Sigma X_4)^2}{N_1} - \frac{T^2}{N}$

$N_1 \rightarrow$ Number of elements in their respective columns.

$= \frac{(57)^2}{7} + \frac{(31)^2}{5} + \frac{(29)^2}{8} + \frac{(-19)^2}{6} - 369.39$

$= \frac{3249}{7} + \frac{961}{5} + \frac{841}{8} + \frac{361}{6} - 369.39$

$= 464.14 + 192.2 + 105.13 + 60.17 - 369.39 = 452.25$

SSE = TSS − SSC

$= 1950.61 - 452.25 = 1498.36$

Step 6 : ANOVA

Source of Variation	Sum of squares	d.f.	Mean squre	Variance Ratio	Table value 5% level
Between columns	SSC = 452.25	C − 1 = 4 − 1 = 3	$MSC = \frac{SSC}{C-1} = \frac{452.25}{3} = 150.75$	$F_C = \frac{MSC}{MSE} = \frac{150.75}{68.11} = 2.21$	F_C (3.22) = 3.05
Error	SSE = 1498.36	N − C = 26 − 4	$MSE = \frac{SSE}{N-C}$	Since $\frac{MSE}{MSC} < 1$	

Step 7 : Conclusion : Cal F_c < Table F_c

∴ So we accept H_0

Two-Way Classification

Example 2.3.3

Five doctors, each test five treatments for a certain disease and observe the number of days each patient takes to recover. The results are as follows :

Given Recovery time in days.

Doctors	Treatments 1	2	3	4	5
Jebasingh	10	14	23	19	20
Niranjan Kumar	11	15	24	17	21
Deivanai	9	11	20	16	19
Sathyapriya	8	13	17	17	20
Kanimozhi	12	15	19	15	22

Discuss the difference between (i) Doctors, (ii) Treatments.

Solution :

1. H_0 : There is no significant difference between doctors.
2. H_1 : There is a significant difference between doctors.

	Doctors	Treatments X_1	X_2	X_3	X_4	X_5	Total (Row wise)	X_1^2	X_2^2	X_3^2	X_4^2	X_5^2
Y_1	Jebasingh	−6	−2	7	3	4	6	36	4	49	9	16
Y_2	Niranjan Kumar	−5	−1	8	1	5	8	25	1	64	1	25
Y_3	Deivanai	−7	−4	4	0	3	−4	49	16	16	0	9
Y_4	Sathyapriya	−8	−3	1	1	4	−5	64	9	9	1	16
Y_5	Kanimozhi	−4	−1	3	−1	6	3	16	1	1	1	36
	Total (column wise)	−30	−11	23	4	22	8 (T)	190	31	139	12	102

Two-Way Classification
→Two Independent Variable

Example :1

Two-Way Classification

Two-Way Classification
→Two Independent Variable

Example :1

2.28

Step 1 : N = 25

Step 2 : T = 8

Step 3 : $\frac{T^2}{N} = \frac{64}{25} = 2.56$

Step 4 : $TSS = \Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 + \Sigma X_4^2 + \Sigma X_5^2 - \frac{T^2}{N}$

$= 190 + 31 + 139 + 12 + 102 - 2.56$

$= 474 - 2.56 = 471.44$

Step 5 : $SSC = \frac{(\Sigma X_1)^2}{N_1} + \frac{(\Sigma X_2)^2}{N_1} + \frac{(\Sigma X_3)^2}{N_1} + \frac{(\Sigma X_4)^2}{N_1} + \frac{(\Sigma X_5)^2}{N_1} - \frac{T^2}{N}$

[N_1 = Number of elements in each column]

$= \frac{(30)^2}{5} + \frac{(11)^2}{5} + \frac{(23)^2}{5} + \frac{(4)^2}{5} + \frac{(22)^2}{5} - \frac{64}{25}$

$= 410 - 2.56 = 407.44$

Step 6 : $SSR = \frac{(\Sigma Y_1)^2}{N_2} + \frac{(\Sigma Y_2)^2}{N_2} + \frac{(\Sigma Y_3)^2}{N_2} + \frac{(\Sigma Y_4)^2}{N_2} + \frac{(\Sigma Y_5)^2}{N_2} - \frac{T^2}{N}$

[N_2 = Number of elements in each row]

$= \frac{(6)^2}{5} + \frac{(8)^2}{5} + \frac{(-4)^2}{5} + \frac{(-5)^2}{5} + \frac{(3)^2}{5} - \frac{64}{25}$

$= 30 - 2.56 = 27.44$

$SSE = TSS - SSC - SSR$

$= 471.44 - 407.44 - 27.44$

Two-Way Classification

Step 7 : ANOVA

ANOVA Table

Source of Variation	SS	DF	MSS	VR	Table value at 5% level
Between rows	$SSR = 27.44$	$r-1 = 5-1 = 4$	$MSR = \frac{SSR}{r-1}$ $\frac{27.44}{4} = 6.86$	$F_R = \frac{MSR}{MSE} = \frac{6.86}{2.28} = 3.01$	$F_R(4,16) = 3.01$
Between columns	$SSC = 407.44$	$C-1 = 5-1 = 4$	$MSC = \frac{SSC}{C-1} = \frac{407.44}{2} = 101.86$	$F_c = \frac{MSC}{MSE} = \frac{101.86}{2.28} = 44.67$	$F_c(4, 16) = 3.01$
Error	$SSE = 36.56$	$N-c-r+1 = 16$	$MSE = \frac{SSE}{N-c-r+1} = \frac{36.56}{16} = 2.28$		
Total	$TSS = 471.44$	24			

Step 8 : Conclusion :

Cal F_c < tab F_c H_0 is accepted.

Cal F_R > tab F_R H_0 is rejected.

Two-Way Classification
→Two Independent Variable

Example :1

Two-Way Classification

Two-Way Classification
→Two Independent Variable

Example :2

The following table gives monthly sales (in thousand rupees) of a certain firm in the three states by its four salesmen.

States	Salesmen I	II	III	IV
A	6	5	3	8
B	8	9	6	5
C	10	7	8	7

Setup the analysis of variance table and test whether there is any significant difference (i) between sales by the firm salesmen and (ii) between sales in the three states.

Solution : 1. H_0 : (i) there is no significant difference between the sales by the firm's salesmen and (ii) there is no significant difference between sales in the three states.

H_1 : Significant difference

States		Salesmen I (X_1)	II (X_2)	III (X_3)	IV (X_4)	Total	X_1^2	X_2^2	X_3^2	X_4^2
Y_1	A	6	5	3	8	22	36	25	9	64
Y_2	B	8	9	6	5	28	64	81	36	25
Y_3	C	10	7	8	7	32	100	49	64	49
Total		24	21	17	20	82	200	155	109	138

Step 1 : N = 12

Step 2 : T = 82

Step 3 : $\frac{T^2}{N} = \frac{(82)^2}{12} = 560.333$

Step 4 : $TSS = \Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 + \Sigma X_4^2 - \frac{T^2}{N}$

$= 200 + 155 + 109 + 138 - 560.333 = 41.667$

Step 5 : $SSC = \frac{(\Sigma X_1)^2}{N_1} + \frac{(\Sigma X_2)^2}{N_1} + \frac{(\Sigma X_3)^2}{N_1} + \frac{(\Sigma X_4)^2}{N_1} - \frac{T^2}{N}$

[N_1 = Number of elements in each column]

$= \frac{(24)^2}{3} + \frac{(21)^2}{3} + \frac{(17)^2}{3} + \frac{(20)^2}{3} - 560.333$

$= \frac{576}{3} + \frac{441}{3} + \frac{289}{3} + \frac{400}{3} - 560.333$

Two-Way Classification

Two-Way Classification
→Two Independent Variable

Example :2

$$= \frac{1}{3}[576 + 441 + 289 + 400] - 560.333 = 8.334$$

Step 6 : $SSR = \frac{(\Sigma Y_1)^2}{N_2} + \frac{(\Sigma Y_2)^2}{N_2} + \frac{(\Sigma Y_3)^2}{N_2} - \frac{T^2}{N}$

[N_2 = Number of elements in each row]

$$= \frac{1}{4}[(22)^2 + (28)^2 + (32)^2] - \frac{T^2}{N}$$

$$= 573 - 56.333 = 12.667$$

$$SSE = TSS - SSC - SSR$$

$$= 41.667 - 8.334 - 12.667 = 20.666$$

Step 7 : ANOVA Table

Source of Variation	SS	DF	MSS	VR	Table value at 5% level
Between columns	SSC $= 8.334$	$C-1$ $= 4-1$ $= 3$	$MSC = \frac{SSC}{C-1}$ $= \frac{8.334}{3}$ $= 2.778$	$F_c = \frac{MSE}{MSC}$ $= \frac{3.444}{2.778}$ $= 1.23$	$F_c(6, 3)$ $= 8.94$
Between rows	SSR $= 12.667$	$r-1$ $= 3-1$ $= 2$	$MSR = \frac{SSR}{r-1}$ $\frac{12.667}{2} = 6.334$	$F_R = \frac{MSR}{MSE}$ $= \frac{6.334}{3.444}$ $= 1.84$	$F_R(2,6)$ $= 5.14$
Residual	SSE $= 20.666$	$N-c-r+1$ $= 6$	$MSE = \frac{SSE}{N-c-r+1}$ $= \frac{20.666}{6} = 3.444$		
Total	TSS $= 41.667$	11			

Step 8 : Conclusion : (i) Cal F_c < Table F_c accept H_0

(ii) Cal F_R < Table F_R accept H_0.