

# Correlation

## Introduction:

- Two variables are said to be correlated if the change in one variable results in a corresponding change in the other variable.
- The correlation is a statistical tool which studies the relationship between two variables.
- Correlation analysis involves various methods and techniques used for studying and measuring the extent of the relationship between the two variables.
- Correlation is concerned with the measurement of “*strength of association between variables*”.
- The degree of association between two or more variables is termed as correlation.

## Contd...

- Correlation analysis helps us to decide the strength of the linear relationship between two variables.
- The word correlation is used to decide the degree of association between variables.
- If two variables 'x' and 'y' are so related, the variables in the magnitude of one variable tend to be accompanied by variations in the magnitude of the other variable, they are said to be correlated.
- Thus, correlation is a statistical tool, with the help of which, we can determine whether or not two or more variables are correlate and if they are correlated, what is the degree and direction of correlation.

## Definition

➤ *The correlation is the measure of the extent and the direction of the relationship between two variables in a bivariate distribution.*

Example:

- (i) Height and weight of children.
- (ii) An increase in the price of the commodity by a decrease in the quantity demanded.

Types of Correlation: The following are the types of correlation

- (i) Positive and Negative Correlation
- (ii) Simple, Partial and Multiple Correlation
- (iii) Linear and Non-linear Correlation

## Contd...

- Correlation first developed by Sir Francis Galton (1822 – 1911) and then reformulated by Karl Pearson (1857 – 1936)
- Note: The degree of relationship or association is known as the degree of relationship.



# Types of Correlation

- i. Positive and Negative correlation: If both the variables are varying in the same direction i.e. if one variable is increasing and the other on an average is also increasing or if as one variable is decreasing, the other on an average, is also decreasing, correlation is said to be positive. If on the other hand, the variable is increasing, the other is decreasing or vice versa, correlation is said to be negative.

Example 1: a) heights and weights (b) amount of rainfall and yields of crops (c) price and supply of a commodity (d) income and expenditure on luxury goods (e) blood pressure and age

Example 2: a) price and demand of commodity (b) sales of woolen garments and the days temperature.

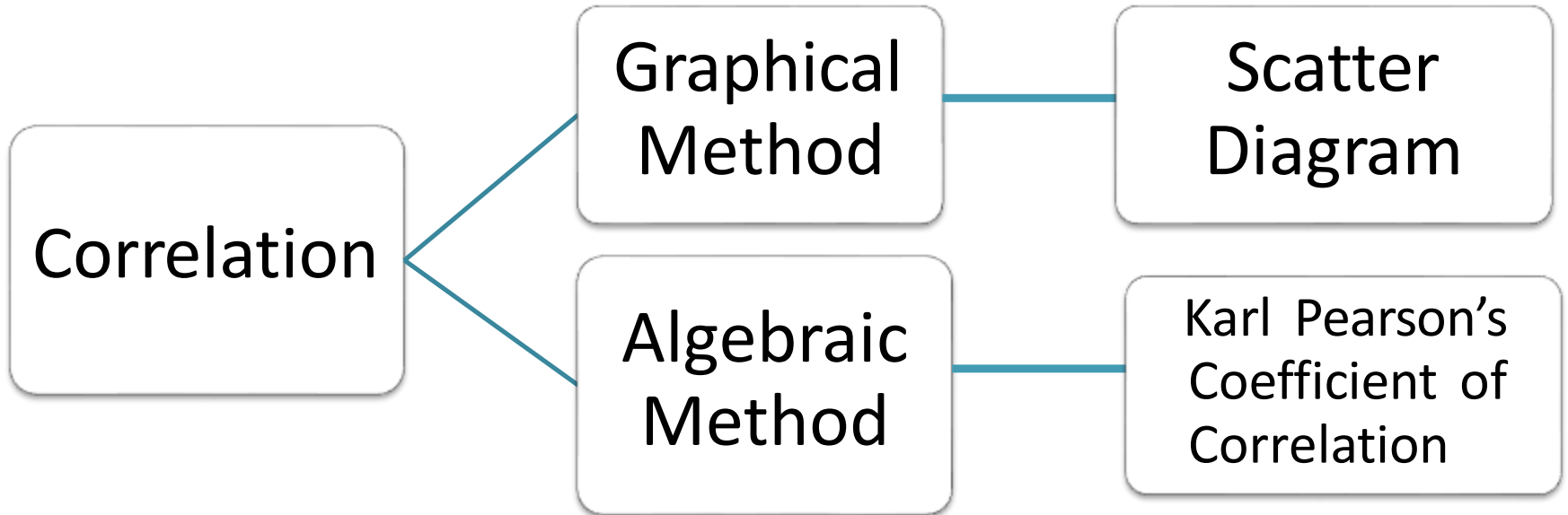
Contd...

- ii. **Simple, Partial and Multiple Correlation:**  
When only two variables are studied, it is a case of simple correlation. In partial and multiple correlation, three or more variables are studied. In multiple correlation three or more variables are studied simultaneously. In partial correlation, we have more than two variables, but consider only two variables to be influencing each other, the effect of the other variables being kept constant.

Contd...

iii. Linear and Non-linear Correlation: If the change in one variable tends to bear a constant ratio to the change in the other variable, the correlation is said to be linear. Correlation is said to be non-linear if the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable.

# Methods of Studying Correlation

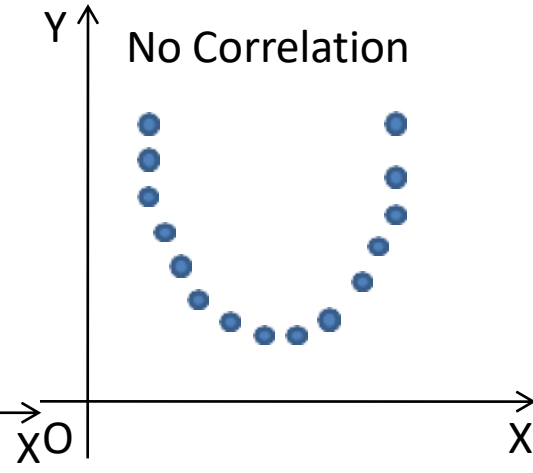
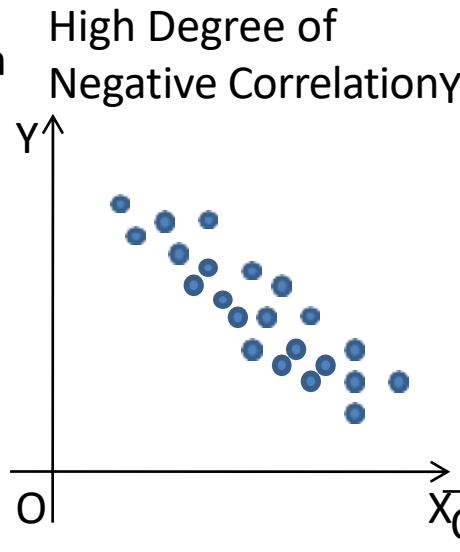
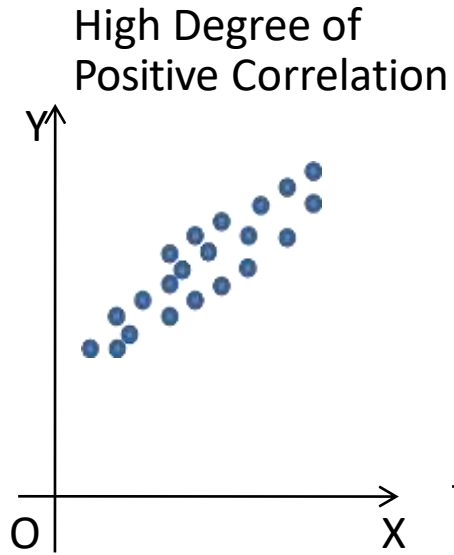
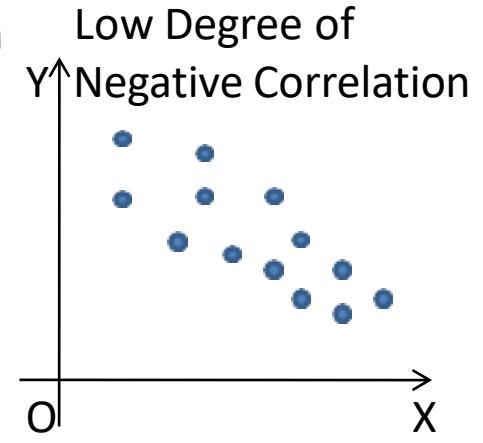
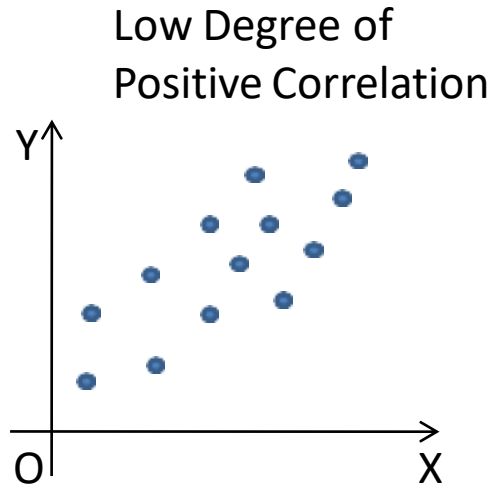
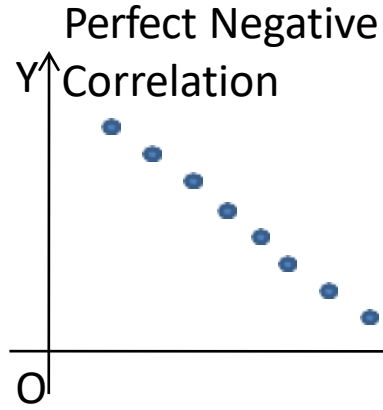
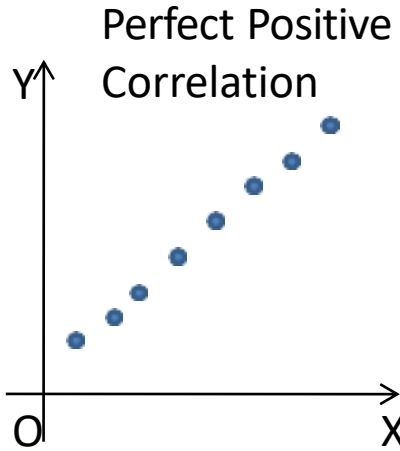




# Methods of Studying Correlation

- The following are the methods of determining correlation
  1. Scatter diagram method
  2. Karl Pearson's Coefficient of Correlation
    1. Scatter Diagram:
      - This is a graphic method of finding out relationship between the variables.
      - Given data are plotted on a graph paper in the form of dots i.e. for each pair of  $x$  and  $y$  values we put a dot and thus obtain as many points as the number of observations.
      - The greater the scatter of points over the graph, the lesser the relationship between the variables.

# Scatter Diagram



# Interpretation

- If all the points lie in a straight line, there is either perfect positive or perfect negative correlation.
- If all the points lie on a straight line falling from the lower left hand corner to the upper right hand corner then the correlation is perfect positive.
- Perfect positive if  $r = +1$ .
- If all the points lie on a straight line falling from the upper left hand corner to the lower right hand corner then the correlation is perfect negative.
- Perfect negative if  $r = -1$ .
- The nearer the points are to be straight line, the higher degree of correlation.
- The farthest the points from the straight line, the lower degree of correlation.
- If the points are widely scattered and no trend is revealed, the variables may be un-correlated i.e.  $r = 0$ .

# The Coefficient of Correlation:

- A scatter diagram give an idea about the type of relationship or association between the variables under study. It does not tell us about the quantification of the association between the two.
- In order to quantify the relation ship between the variables a measure called correlation coefficient developed by Karl Pearson.
- It is defined as the measure of the degree to which there is linear association between two intervably scaled variables.
- Thus, the coefficient of correlation is a number which indicates to what extent two variables are related , to what extent variations in one go with the variations in the other

## Contd...

- The symbol 'r' or ' $r_{xy}$ ' or ' $r_{yx}$ ' is denoted in this method and is calculated by:
- $r = \{\text{Cov}(X, Y) \div S_x S_y\} \dots\dots\dots(i)$
- Where  $\text{Cov}(X, Y)$  is the Sample Covariance between X and Y. Mathematically it is defined by
- $\text{Cov}(X, Y) = \{\sum(X - \bar{X})(Y - \bar{Y})\} \div (n - 1)$
- $S_x$  = Sample standard deviation of X, is given by
- $S_x = \{\sum(X - \bar{X})^2 \div (n - 1)\}^{1/2}$
- $S_y$  = Sample standard deviation of Y, is given by
- $S_y = \{\sum(Y - \bar{Y})^2 \div (n - 1)\}^{1/2}$  and  $\bar{X} = \sum X \div n$  and  $\bar{Y} = \sum Y \div n$

# Interpretation

- i. If the covariance is positive, the relationship is positive.
  - ii. If the covariance is negative, the relationship is negative.
  - iii. If the covariance is zero, the variables are said to be not correlated.
- Hence the covariance measures the strength of linear association between considered numerical variables.
  - Thus, covariance is an absolute measure of linear association .
  - In order to have relative measure of relationship it is necessary to compute correlation coefficient .
  - Computation of correlation coefficient a relation developed by Karl Pearson are as follows:

Contd...

➤ The formula for sample correlation coefficient (r) is calculated by the following relation:

$$r = \frac{\sum(X - \bar{X}) \cdot (Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2} \cdot \sqrt{\sum(Y - \bar{Y})^2}}$$

$$r = \frac{N \sum XY - \sum X \cdot \sum Y}{\sqrt{[N \sum X^2 - (\sum X)^2] \cdot [N \sum Y^2 - (\sum Y)^2]}}$$

$$r = \frac{\sum XY - n \cdot \bar{X}\bar{Y}}{\sqrt{[\sum X^2 - n(\bar{X})^2] \cdot [\sum Y^2 - n(\bar{Y})^2]}}$$

➤ If  $(X - \bar{X}) = x$  and  $(Y - \bar{Y}) = y$  then above formula reduces to:

$$r = \frac{\sum xy}{\sqrt{\sum x^2} \cdot \sqrt{\sum y^2}}$$

➤ Example:

## Properties of Karl Pearson's Correlation Coefficient

1. The coefficient of correlation 'r' is always a number between -1 and +1 inclusive.
2. If  $r = +1$  or  $-1$ , the sample points lie on a straight line.
3. If 'r' is near to +1 or -1, there is a strong linear association between the variables.
4. If 'r' is small(close to zero), there is low degree of correlation between the variables.
5. The coefficient of correlation is the geometric mean of the two regression coefficients.

Symbolically:  $r = \sqrt{b_{xy} \cdot b_{yx}}$

*Note:* It is clear that correlation coefficient is a measure of the degree to which the association between the two variables approaches a linear functional relationship.



# Interpretation of Correlation Coefficient

- i. The coefficient of correlation, as obtained by the above formula shall always lie between +1 to -1.
- ii. When  $r = +1$ , there is perfect positive correlation between the variables.
- iii. When  $r = -1$ , there is perfect negative correlation between the variables.
- iv. When  $r = 0$ , there is no correlation.
- v. When  $r = 0.7$  to  $0.999$ , there is high degree of correlation.
- vi. When  $r = 0.5$  to  $0.699$ , there is a moderate degree of correlation.
- vii. When  $r$  is less than  $0.5$ , there is a low degree of correlation.
- viii. The value of correlation lies in between  $-1$  to  $+1$  i.e.  
 $-1 \leq r \leq +1$ .
- ix. The correlation coefficient is independent of the choice of both origin and scale of observation.
- x. The correlation coefficient is a pure number. It is independent of the units of measurement.

# Simple Linear Regression

- Regression is concerned with the “Prediction” of the most likely value of one variable when the value of the other variable is known.
- The term regression literally means “stepping back towards the average”.
- It was first used by British biometrician *Sir Francis Galton (1822 – 1911)*.

*Definition: Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data.*

- Thus term regression is used to denote *estimation* or *prediction* of the average value of one variable for a specified value of the other variable.
- The estimation is done by means of *suitable equation*, derived on the basis of available bivariate data. Such an equation and its geometrical representation is called regression curve.

## Contd...

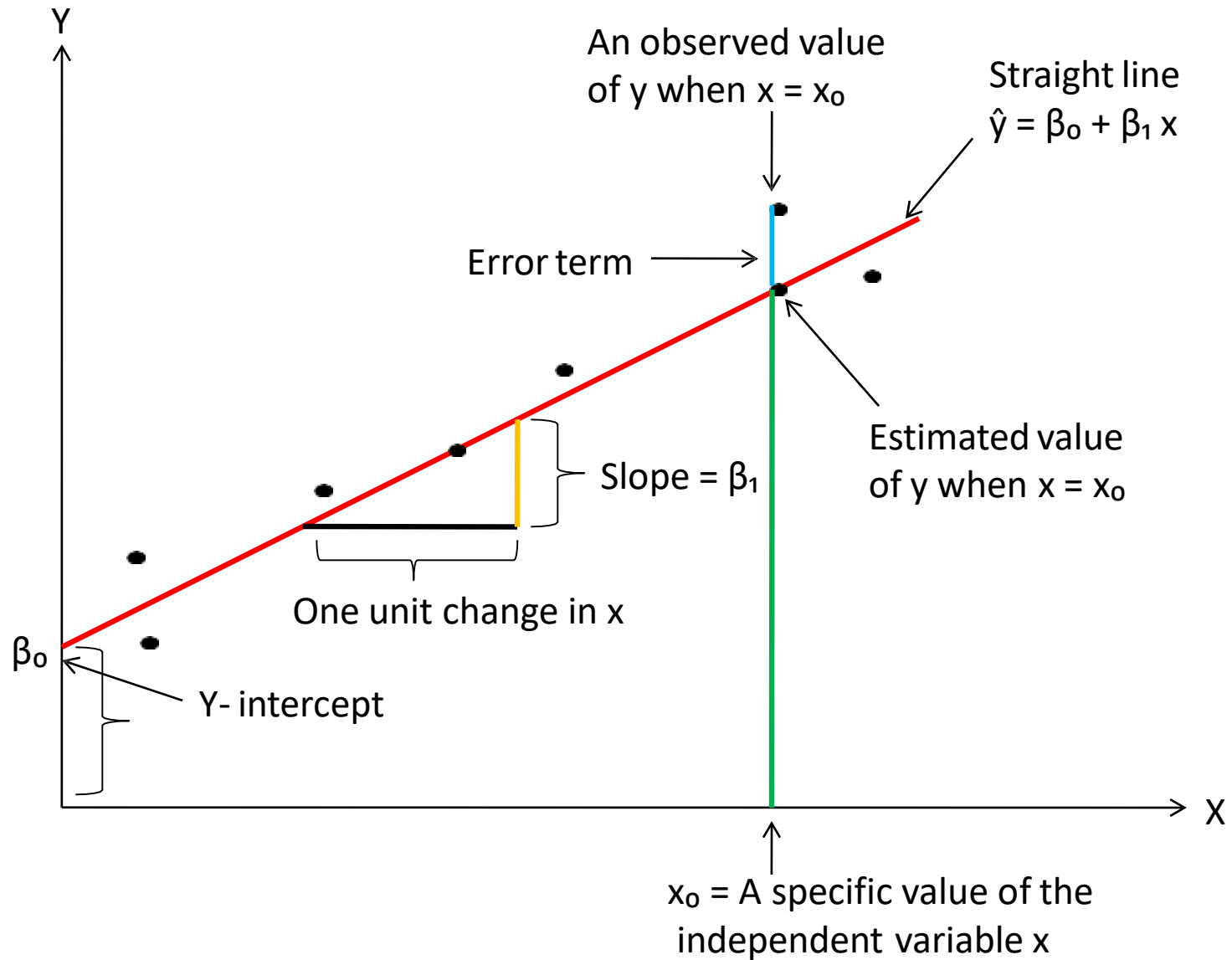
- In regression analysis there are two types of variables and they are:
- i. Independent and ii. Dependent.
- Dependent variable(Y): The variable whose value is influenced or is to be predicted is called dependent variable.
- Independent variable(X): The variable which influences the values or is used for prediction is called independent variable.
- In regression analysis independent variable is known as regressor or predictor or explanatory variable.
- In regression analysis dependent variable is known as regressed or explained variable.
- Thus the term regression is used to denote estimation or prediction of the average value of one variable for a specified value of the other variable.

# The lines of Regression

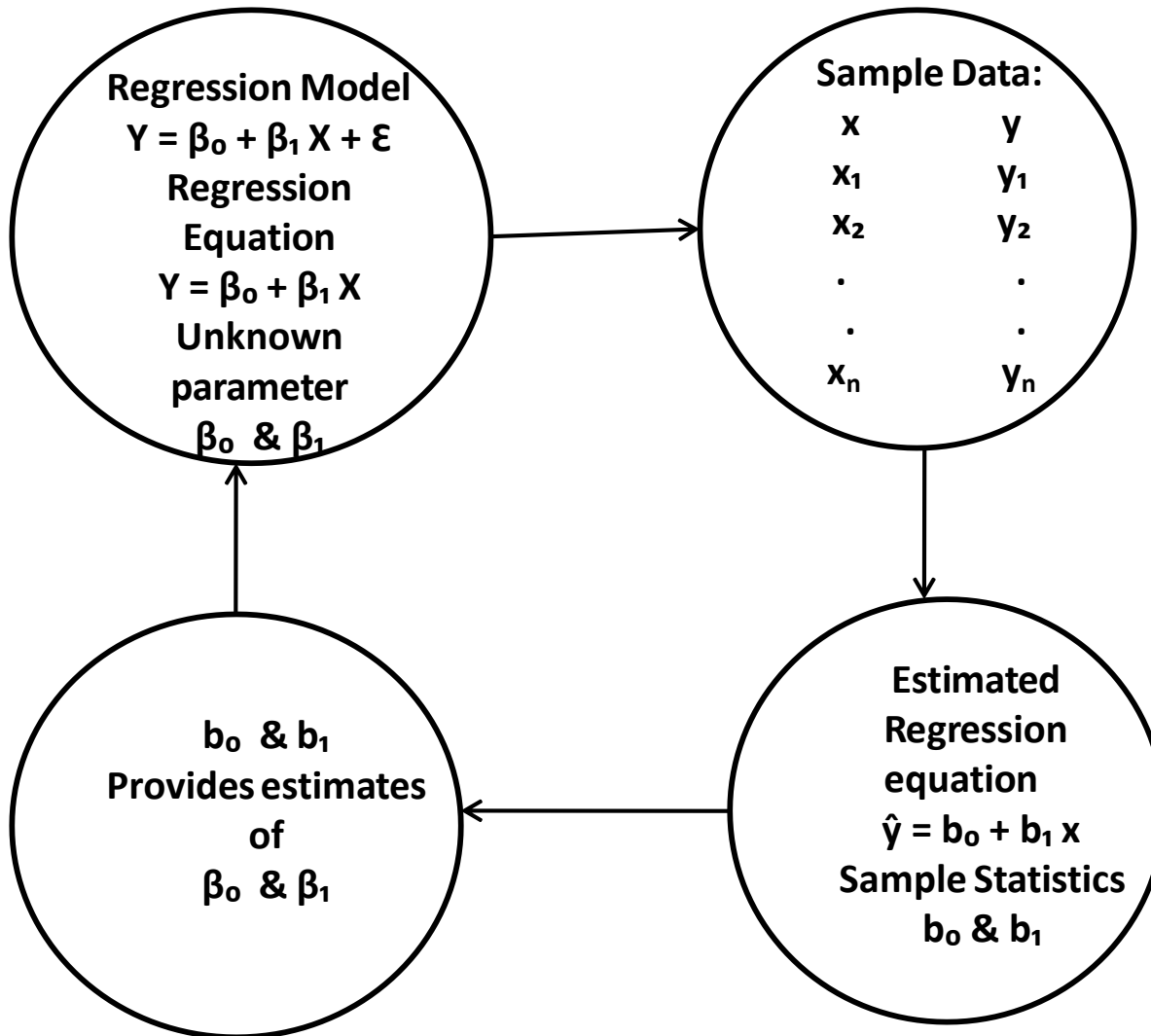
- A line fitted to a set of data points to estimate the relationship between two variables is called a regression line.
- The regression equation of  $Y$  on  $X$  describes the changes in the value of  $Y$  for given changes in the value of  $X$ .
- The regression equation of  $X$  on  $Y$  describes the changes in the value of  $X$  for given changes in the value of  $Y$ .
- Hence, an equation for estimating a dependent variable  $Y$  for  $X$  from the independent variable  $X$  or  $Y$ , is called a regression equation of  $Y$  on  $X$  or  $X$  on  $Y$  respectively.
- The regression equations of the regression lines, also called least squares lines, are determined by the least square method.

# Simple Regression Model

- Simple regression line is a straight line that describe about the dependence of the average value of one variable on the other.
- $Y = \beta_0 + \beta_1 X + \varepsilon \dots\dots\dots(*)$
- Where  $Y$  = Dependent or response or outcome variable (Population)
- $X$  = Independent or explanatory or predictor variable (Population)
- $\beta_0$  =  $Y$ - intercept of the model for the population
- $\beta_1$  = population slope coefficient or population regression coefficient. It measures the average rate of change in dependent variable per unit change in independent variable.
- $\varepsilon$  = Population error in  $Y$  for observation.



# Estimation of Regression Equation



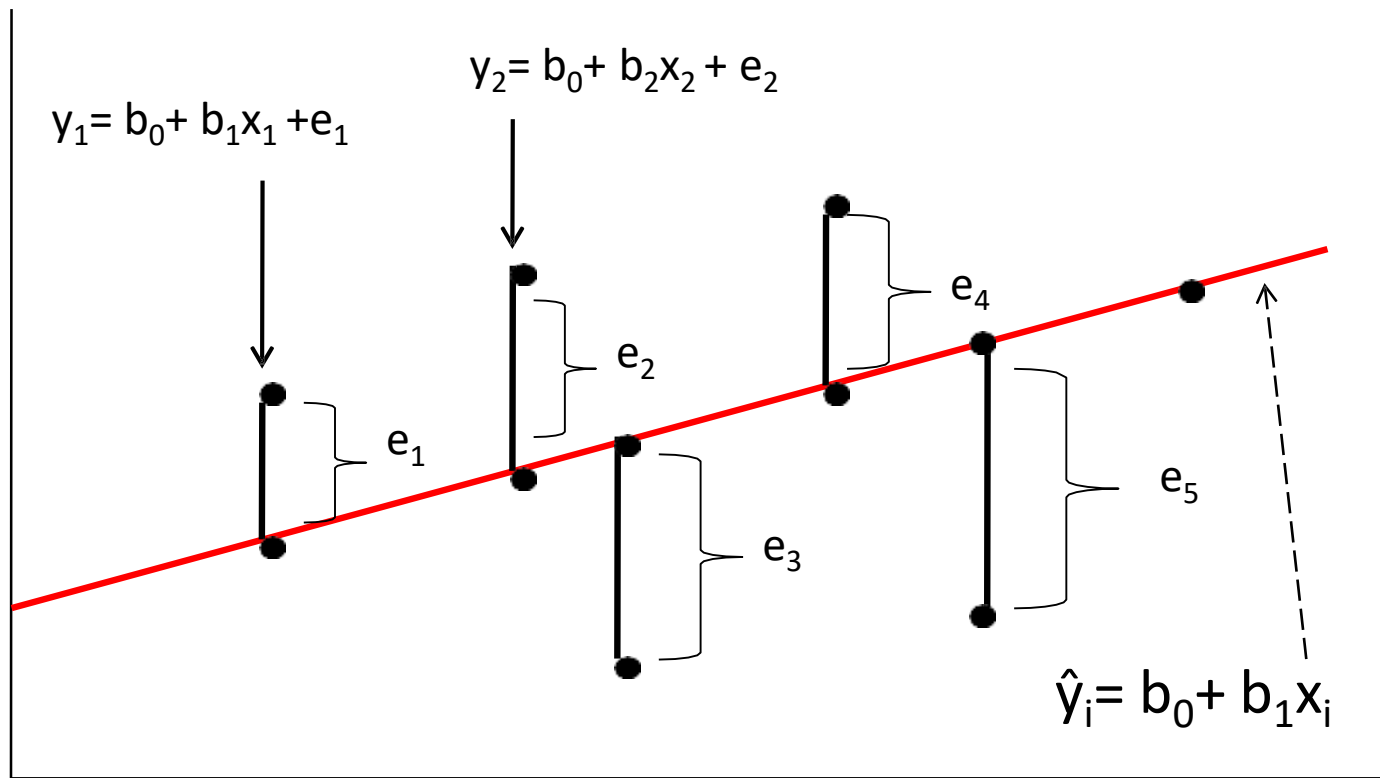
# Model

- Linear regression model is
- $Y = \beta_0 + \beta_1 X + \varepsilon$
- Linear regression equation is:
- $Y = \beta_0 + \beta_1 X$
- Sample regression model is
- $\hat{y} = b_0 + b_1x + e$
- Sample regression equation is
- $\hat{y} = b_0 + b_1x$
- Where  $b_0$  = sample  $y$  intercept,
- $b_1$  = sample slope coefficient
- $x$  = independent variable
- $y$  = dependent variable
- $\hat{y}$  = estimated value of dependent variable for a given value of independent variable.
- $e$  = error term =  $y - \hat{y}$



# Least square graphically

least square minimizes  $\sum_{i=1}^n e_i^2 = e_1^2 + e_2^2 + e_3^2 + e_4^2 + e_5^2$



## Least squares methods

- Let  $\hat{y} = b_0 + b_1x \dots\dots(1)$  be estimated linear regression equation of  $y$  on  $x$  of the regression equation  $Y = \beta_0 + \beta_1 X$ .
- By using the principles of least square, we can get two normal equations of regression equation (1) are as:
  - $\sum y = nb_0 + b_1 \sum x \dots\dots\dots(2)$
  - $\sum xy = b_0 \sum x_1 + b_1 \sum x_2 \dots\dots\dots(3)$
- By solving equations (2) & (3) we get the value of  $b_0$  &  $b_1$  as:

## Contd...

$$b_1 = \frac{SS_{xy}}{SS_x} = \frac{\sum(x - \bar{x}) \cdot (y - \bar{y})}{\sum(x - \bar{x})^2} = \frac{\sum xy - \frac{\sum x \cdot \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{n \cdot \sum xy - \sum x \cdot \sum y}{n \cdot \sum x^2 - (\sum x)^2} = \frac{\sum xy - n \bar{x} \cdot \bar{y}}{\sum x^2 - n \cdot (\bar{x})^2}$$

- The computational formula for y intercept  $b_0$  as follows:

$$b_0 = \bar{y} - b_1 \bar{x} = \frac{\sum y}{n} - b_1 \frac{\sum x}{n}$$

- After finding the value of  $b_0$  &  $b_1$ , we get the required fitted regression model of  $y$  on  $x$  as  $\hat{y} = b_0 + b_1 x$ .

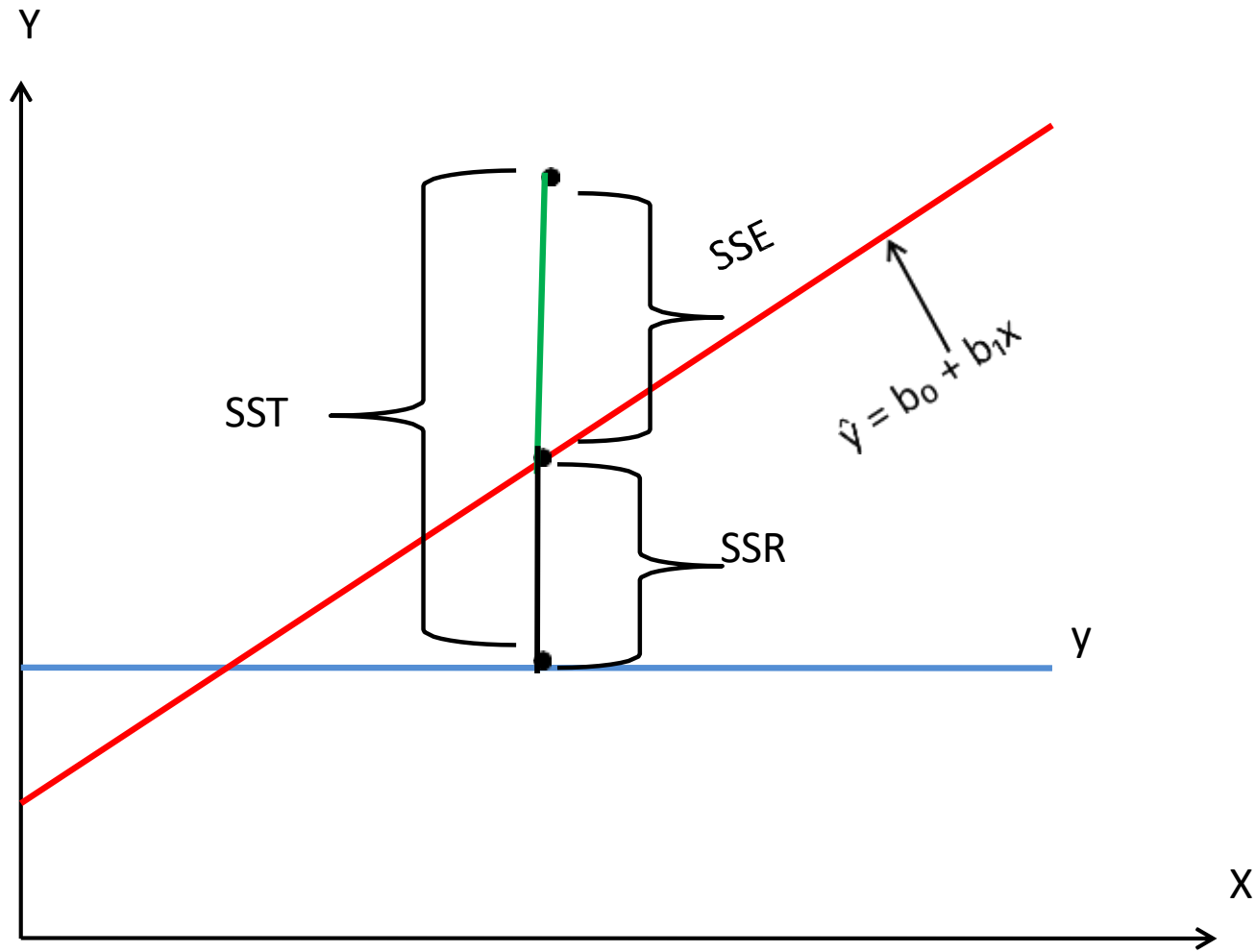
# Measures of variation

- There are three measures of variations.
- They are as follows:
  - i. Total Sum of Squares (SST): It is a measure of variation in the values of dependent variable ( $y$ ) around their mean value ( $\bar{y}$ ). That is
    - $SST = \sum(y - \bar{y})^2 = \sum y^2 - (\sum y)^2/n = \sum y^2 - n \cdot \bar{y}^2$ .
    - Note: The total sum of squares or the total variation is divided into the sum of two components. One is explained variation due to the relationship between the considered dependent variable ( $y$ ) and the independent variable ( $x$ ) and the other is unexplained variation which might be developed due to some other factors other than the relationship between variable  $x$  and  $y$ .

Contd...

- ii. Regression Sum of Squares( SSR): The regression sum of squares is the sum of the squared differences between the predicted value of  $y$  and the mean value of  $y$ .
  - $SSR = \sum(\hat{y} - \bar{y})^2 = b_0 \cdot \sum y + b_1 \sum xy - (\sum y)^2/n = b_0 \cdot \sum y + b_1 \sum xy - n \cdot \bar{y}^2$
- ii. Error Sum of Squares (SSE): The error sum of square is computed as the sum of the squared differences between the observed value of  $y$  and the predicted value of  $y$  i.e.
  - $SSE = \sum(y - \hat{y})^2 = \sum y^2 - b_0 \sum y - b_1 \sum xy.$

# Contd...



## Contd...

Relationship: From the above figures the relationship of SST, SSR and SSE are as follows

$$SST = SSR + SSE \dots \dots \dots (i)$$

Where: SST = Total sum of square

SSR = Regression sum of squares

SSE = Error sum of squares

- The fit of the estimated regression line would be best if every value of the dependent variable  $y$  falls on the regression line.

## Contd....

- If  $SSE = 0$  i. e.  $e = (y - \hat{y}) = 0$  then  $SST = SSR$ .
- For the perfect fit of the regression model, the ratio of SSR to SST must be equal to unity i. e. If  $SSE = 0$  then the model would be perfect.
- If SSE would be larger, the fit of the regression line would be poor.
- Note: Largest value of SSE the regression line would be poor and if  $SSE = 0$  the regression line would be perfect.



## Coefficient of Determination ( $r^2$ )

- The coefficient of determination measures the strength or extent of the association that exists between dependent variable ( $y$ ) and independent variable ( $x$ ).
- It measures the proportion of variation in the dependent variable ( $y$ ) that is explained by independent variable of the regression line.
- Coefficient of variation measures the total variation in the dependent variable due to the variation in the independent variable and it is denoted by  $r^2$ .
- $r^2 = SSR/SST$  but  $SST = SSE + SSR$
- then  $SSR = SST - SSE$
- $r^2 = 1 - (SSE/SST) = (b_0 \cdot \sum y + b_1 \sum xy - n \cdot \bar{y}^2) / (\sum y^2 - n \bar{y}^2)$ .

## Contd...

- Note:
  - i. Coefficient of determination is the square of coefficient of correlation.  
then  $r = \pm\sqrt{r^2}$
  - ii. If the regression coefficient ( $b_1$ ) is negative then take the negative sign
  - iii. If the regression coefficient ( $b_1$ ) is positive then take the positive sign
- Adjusted coefficient of determination: The adjusted coefficient of determination is calculated by using the following relation:

$$r^2_{adj} = 1 - \frac{n-1}{n-2}(1-r^2)$$

## (i) t- test for Significance in Simple Linear Regression

- t-test is applied whether the regression coefficient  $\beta_1$  is statistically significant or not.
- The process of setting Hypothesis are as follows:
- Setting of Hypothesis:
- Null Hypothesis,  $H_0: \beta_1=0$  (The population slope ( $\beta_1$ ) is zero between two variables X and Y in the population.)
- Alternative Hypothesis,  $H_1: \beta_1 \neq 0$  (The population slope ( $\beta_1$ ) is not zero between two variables X and Y in the population.) or  $H_1: \beta_1 > 0$  or  $H_1: \beta_1 < 0$

# Confidence Interval Estimating for $\beta_1$

- Another way for the linear relationship between the variables X and Y, we can construct confidence Interval (C. I.) estimate of  $\beta_1$ .
- By the help of C. I. we conclude that whether the hypothesized value ( $\beta_1= 0$ ) is included or not.
- For this the following formula is used:
- C. I. for  $\beta_1 = b_1 \pm t_{(n-2)} \cdot S_{b_1}$
- Conclusion: If this confidence interval does not include 0(zero), then we can conclude that there is significant relationship between the variables X and Y.
- Example:

## Contd...

- Test statistic: F is defined as the ratio of regression mean square (MSR) to the error mean square (MSE).
- Where,  $MSR = SSR/k$  and  $MSE = SSE/(n-k-1)$
- $k$  = No. of independent variables in the regression model. The value of  $k = 1$  for simple linear regression model as it has only one predictor variable  $x$ .
- $SSR$  = regression sum of squares =  $\sum(\hat{y} - \bar{y})^2$
- $SSE$  = error sum of squares =  $\sum(y - \hat{y})^2$
- The test statistic F- follows F-distribution with  $(n - k - 1)$  i. E.  $(n - 2)$  d. f. With  $k = 1$ .

## Contd...

- iii. Using p- value we reject  $H_0$  if p- value is less than  $\alpha$ .
- Note:
  - i. F- test will provide same conclusion as provided by the t-test for only one independent variable.
  - ii. For simple linear regression; if t-test indicates  $\beta_1 \neq 0$  and hence the F-test will also show a significance relationship.
  - iii. However only the F-test can be used to test for an overall significant relationship for the regression with more than one independent variable.

## Contd...

- The formula to compute the confidence interval estimate for the mean value of y is:

$$\hat{y} \pm t_{(n-2)} \cdot S_{YX} \sqrt{h}$$

- The formula to compute the prediction interval estimate of an individual value of y is:  $\hat{y} \pm t_{(n-2)} \cdot S_{YX} \sqrt{1 + h^2}$
- Where;  $\hat{y}$  = estimated or predicted value of the dependent variable for a given value of independent variable.
- $S_{YX}$  = standard error of estimate
- $t_{(n-2)}$  = tabulated value of t for (n- 2) d. f. and  $\alpha$  level of significance.
- $h$  = hat matrix element.
- $n$  = number of pairs of observations or sample size.

$$S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum(Y - \hat{y})^2}{n-2}}$$

$$h = \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x - \bar{x})^2} = \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum x^2 - n \cdot (\bar{x})^2} = \frac{1}{n} + \frac{(x - \bar{x})^2}{SS_X}$$

# CHI-SQUARE TEST

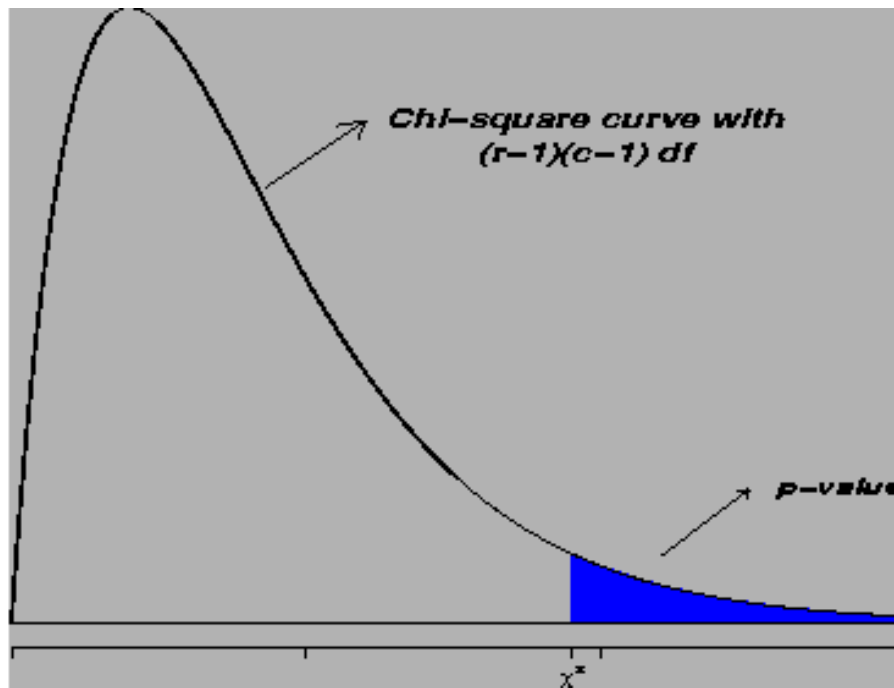


# Introduction

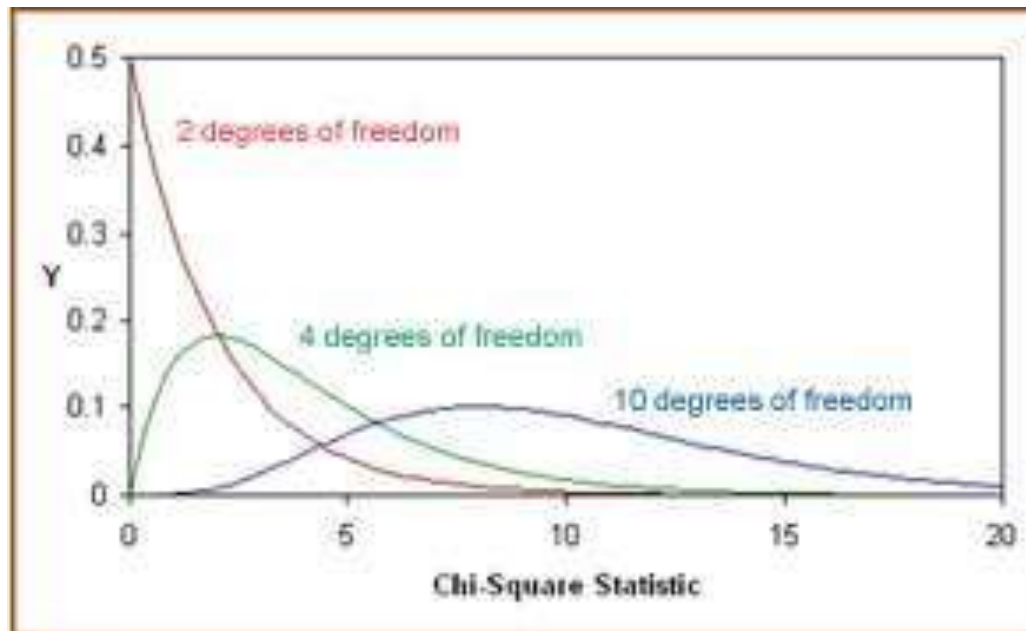
- The Chi-square test is one of the most commonly used non-parametric test, in which the sampling distribution of the test statistic is a **chi-square distribution**, when the null hypothesis is true.
- It was introduced by **Karl Pearson** as a test of association. The Greek Letter  $\chi^2$  is used to denote this test.
- It can be applied when there are few or no assumptions about the population parameter.
- It can be applied on categorical data or qualitative data using a contingency table.
- Used to evaluate ***unpaired/unrelated samples and proportions***.

# Chi-squared distribution

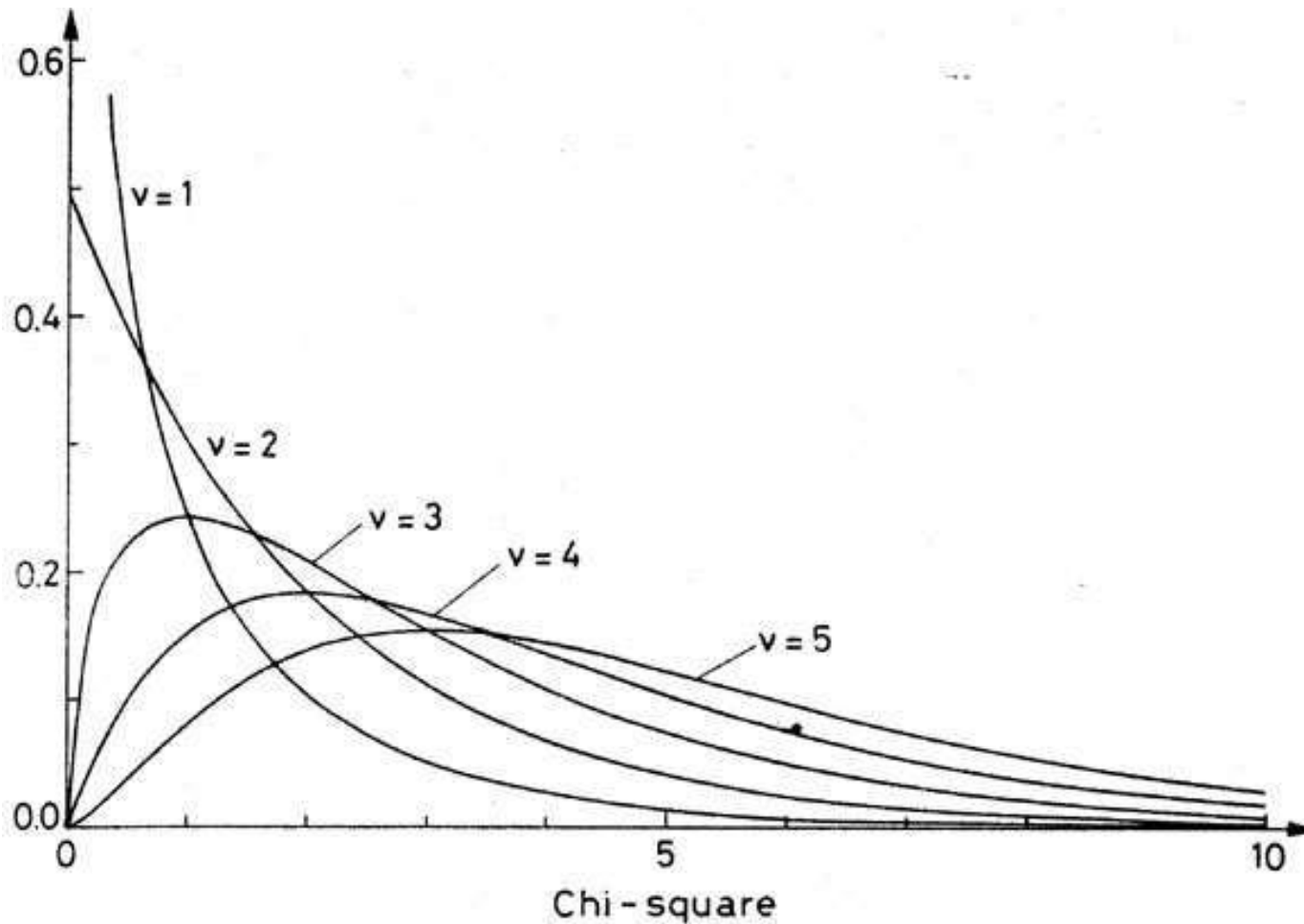
- The distribution of the chi-square statistic is called the chi-square distribution.
- The **chi-squared distribution** with  $k$  degrees of freedom is the distribution of a sum of the squares of  $k$  independent standard normal random variables. It is determined by the *degrees of freedom*.
- The simplest chi-squared distribution is the square of a standard normal distribution.
- The chi-squared distribution is used primarily in hypothesis testing.



- The chi-square distribution has the following properties:
  1. The mean of the distribution is equal to the number of degrees of freedom:  $\mu = \nu$ .
  2. The variance is equal to two times the number of degrees of freedom:  $\sigma^2 = 2 * \nu$



3. The  $\chi^2$  distribution is not symmetrical and all the values are positive. The distribution is described by degrees of freedom. For each degrees of freedom we have asymmetric curves.



4. As the degrees of freedom increase, the chi-square curve approaches a normal distribution.

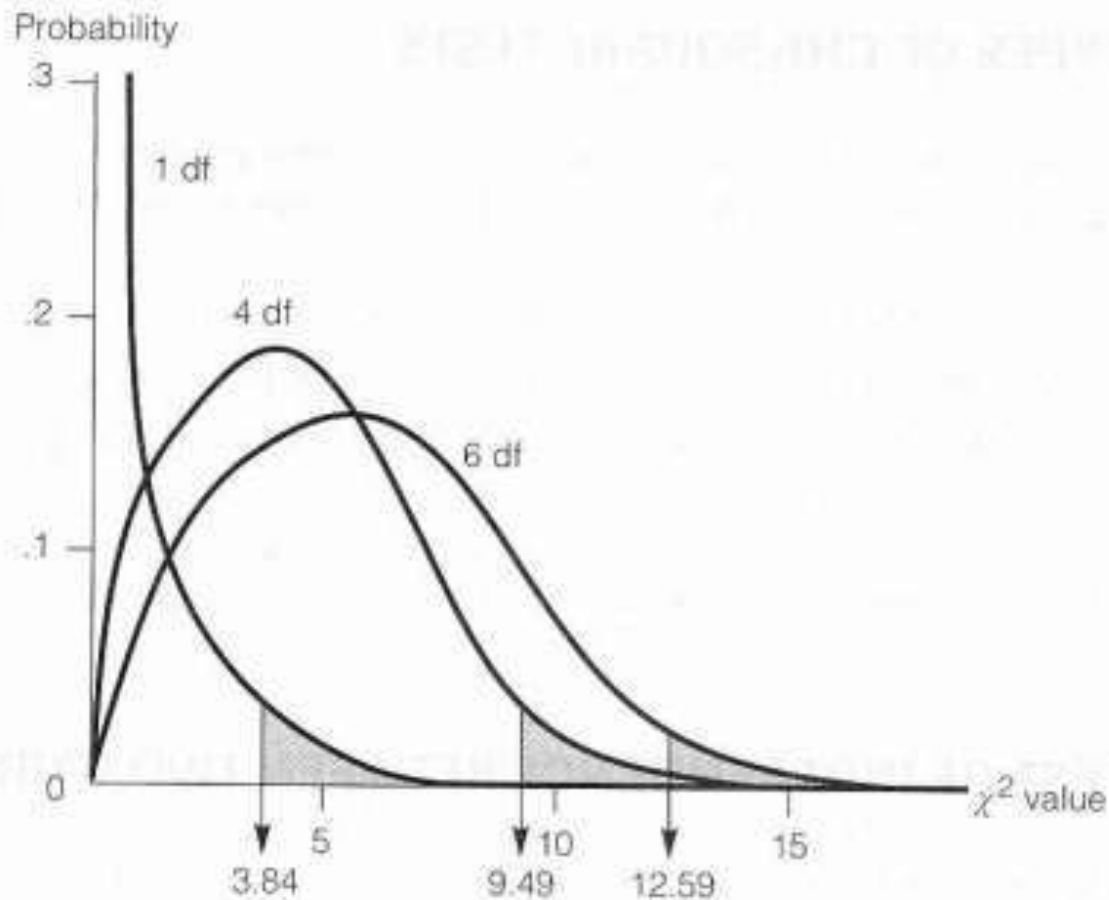
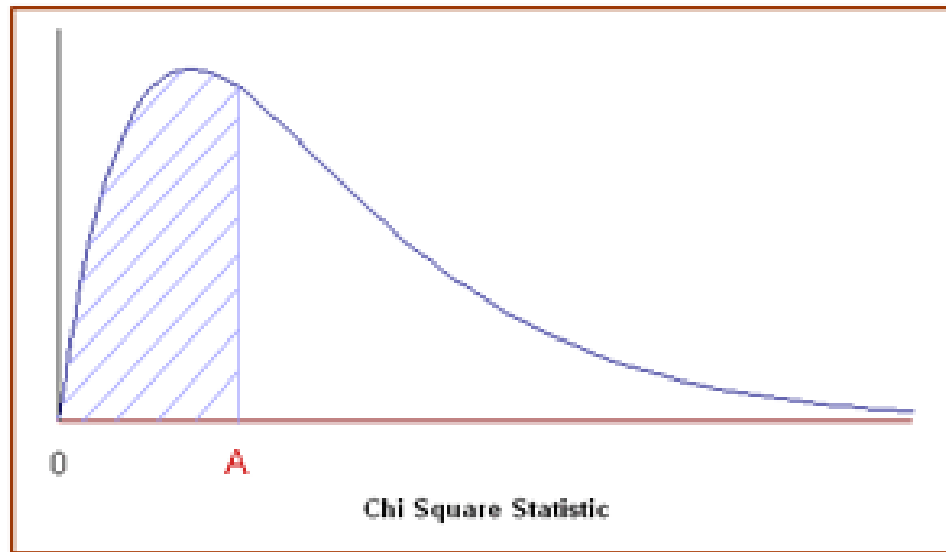


Figure 12.1 The Chi-Square Distribution for Varying Degrees of Freedom.

# Cumulative Probability and the Chi-Square Distribution

- The chi-square distribution is constructed so that the total area under the curve is equal to 1. The area under the curve between 0 and a particular chi-square value is a ***cumulative probability associated with that chi-square value.***
- Ex: The shaded area represents a cumulative probability associated with a chi-square statistic equal to  $A$ ; that is, it is the probability that the value of a chi-square statistic will fall between 0 and  $A$ .



# Contingency table

- A **contingency table** is a type of table in a matrix format that displays the frequency distribution of the variables.
- They provide a basic picture of the interrelation between two variables and can help find interactions between them.

	Column 1	Column 2	Totals
Row 1	A	B	R1
Row 2	C	D	R2
Totals	C1	C2	N

- The chi-square statistic compares the observed count in each table cell to the count which would be expected ***under the assumption of no association between the row and column classifications.***

# Degrees of freedom

- The number of independent pieces of information which are free to vary, that go into the estimate of a parameter is called the degrees of freedom.
- In general, the degrees of freedom of an estimate of a parameter is equal to ***the number of independent scores that go into the estimate minus the number of parameters used as intermediate steps in the estimation of the parameter itself*** (i.e. the sample variance has  $N-1$  degrees of freedom, since it is computed from  $N$  random scores minus the only 1 parameter estimated as intermediate step, which is the sample mean).
- The number of degrees of freedom for 'n' observations is 'n-k' and is usually denoted by 'v', where 'k' is the number of independent linear constraints imposed upon them. It is the only parameter of the chi-square distribution.
- The degrees of freedom for a chi squared contingency table can be calculated as:

$$v = (\text{Number of rows} - 1) * (\text{Number of columns} - 1)$$



# Chi Square formula

- The chi-squared test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories.
- The value of  $\chi^2$  is calculated as:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} + \dots + \frac{(O_n - E_n)^2}{E_n}$$

Where,  $O_1, O_2, O_3 \dots O_n$  are the observed frequencies and  $E_1, E_2, E_3 \dots E_n$  are the corresponding expected or theoretical frequencies.

The observed frequencies are the frequencies obtained from the observation, which are sample frequencies.

The expected frequencies are the calculated frequencies.

# Alternate $\chi^2$ Formula

Disease			
Exposure	Yes	No	Total
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	n

$$\chi_1^2 = \frac{n(ad - bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

The alternate  $\chi^2$  formula applies only to 2x2 tables

# Characteristics of Chi-Square test

1. It is often regarded as a ***non-parametric test*** where no parameters regarding the rigidity of populations are required, such as mean and SD.
2. It is based on ***frequencies***.
3. It encompasses the ***additive property*** of differences between observed and expected frequencies.
4. It tests the hypothesis about the ***independence of attributes***.
5. It is preferred in analyzing complex contingency tables.

# Steps in solving problems related to Chi-Square test

STEP 1

- Calculate the expected frequencies

$$E = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

STEP 2

- Take the difference between the observed and expected frequencies and obtain the squares of these differences  $(O-E)^2$

STEP 3

- Divide the values obtained in Step 2 by the respective expected frequency, E and add all the values to get the value according to the formula given by:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

# Conditions for applying Chi-Square test

1. The data used in Chi-Square test must be **quantitative** and in the form of **frequencies**, which must be **absolute** and not in relative terms.
2. The total number of observations collected for this test must be **large** ( at least 10) and should be done on a **random** basis.
3. Each of the observations which make up the sample of this test must be **independent** of each other.
4. The expected frequency of any item or cell must not be **less than 5**; the frequencies of adjacent items or cells should be polled together in order to make it more than 5.
5. This test is used only for **drawing inferences** through test of the hypothesis, so it **cannot be used for estimation** of parameter value.

# Practical applications of Chi-Square test

- The applications of Chi-Square test include testing:
  1. The significance of *sample & population variances* [ $\sigma^2 s$  &  $\sigma^2 p$ ]
  2. The *goodness of fit* of a theoretical distribution: Testing for goodness of fit determines if an observed frequency distribution fits/matches a theoretical frequency distribution (**Binomial distribution, Poisson distribution or Normal distribution**). These test results are helpful to know whether the samples are drawn from identical distributions or not. **When the calculated value of  $\chi^2$  is less than the table value at certain level of significance, the fit is considered to be good one and if the calculated value is greater than the table value, the fit is not considered to be good.**

# Table/Critical values of $\chi^2$

Degrees of Freedom	Probability										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
	Nonsignificant								Significant		

3. The ***independence*** in a contingency table:

- Testing independence determines whether two or more observations across two populations are dependent on each other.
- If the **calculated value is less than the table value** at certain level of significance for a given degree of freedom, then it is concluded that null hypothesis is true, which means that two attributes are independent and hence not associated.
- If **calculated value is greater than the table value**, then the null hypothesis is rejected, which means that two attributes are dependent.

4. The chi-square test can be used to test the strength of the association between exposure and disease in a ***cohort study, an unmatched case-control study, or a cross-sectional study***.



# Chi-Square Test

```
graph TD; A[Chi-Square Test] --> B[Parametric]; A --> C[Non-Parametric]; B --> D[Test for comparing variance]; C --> E[Testing Independence<br/>Test for Goodness of Fit];
```

The diagram is a flowchart with a top-level box labeled 'Chi-Square Test'. Two arrows point downwards from this box to two separate boxes: 'Parametric' on the left and 'Non-Parametric' on the right. From the 'Parametric' box, an arrow points down to a larger box containing the text 'Test for comparing variance'. From the 'Non-Parametric' box, an arrow points down to a larger box containing the text 'Testing Independence' and 'Test for Goodness of Fit' on two separate lines.

**Parametric**

**Test for  
comparing  
variance**

**Non-Parametric**

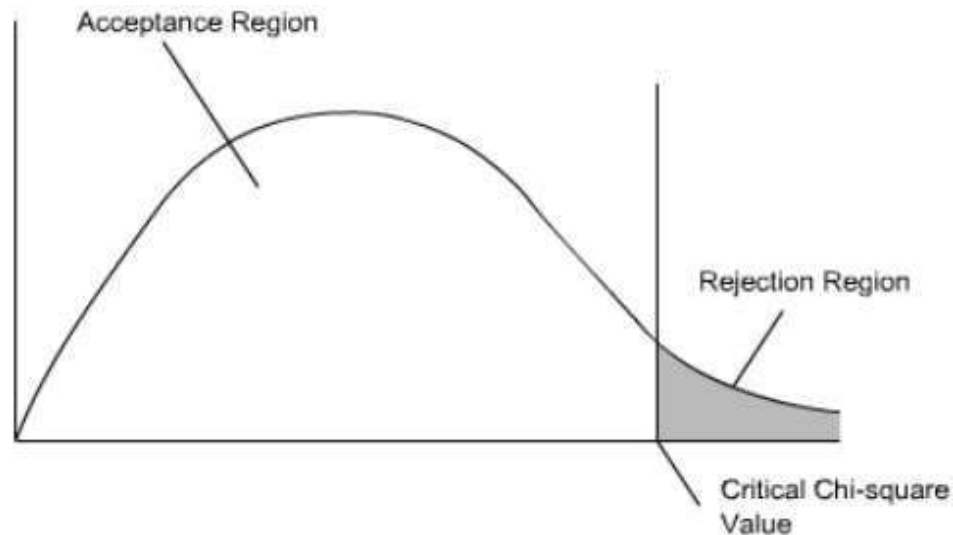
**Testing  
Independence  
Test for Goodness of  
Fit**

# Interpretation of Chi-Square values

- The  $\chi^2$  statistic is calculated under the *assumption of no association* .„
- **Large value of  $\chi^2$  statistic**  $\Rightarrow$  ***Small probability*** of occurring by chance alone ( $p < 0.05$ )  $\Rightarrow$  Conclude that ***association exists*** between disease and exposure. „(Null hypothesis rejected)
- **Small value of  $\chi^2$  statistic**  $\Rightarrow$  ***Large probability*** of occurring by chance alone ( $p > 0.05$ )  $\Rightarrow$  Conclude that ***no association exists*** between disease and exposure. (Null hypothesis accepted)

# Interpretation of Chi-Square values

- The left hand side indicates the degrees of freedom. If the calculated value of  $\chi^2$  falls in the acceptance region, the null hypothesis 'Ho' is accepted and vice-versa.



# Limitations of the Chi-Square Test

1. The chi-square test does ***not give us much information about the strength of the relationship***. It only conveys the existence or nonexistence of the relationships between the variables investigated.
2. The chi-square test is ***sensitive to sample size***. This may make a weak relationship statistically significant if the sample is large enough. Therefore, chi-square should be used together with measures of association like ***lambda, Cramer's V or gamma*** to guide in deciding whether a relationship is important and worth pursuing.
3. The chi-square test is also ***sensitive to small expected frequencies***. It can be used only when not more than **20%** of the cells have an ***expected frequency of less than 5***.
4. Cannot be used when samples are ***related or matched***.

# Modifications/alternatives to chi square test

1. Yates continuity correction
2. Fisher's exact test
3. McNemar's test

# Yates continuity correction

- The Yates correction is a correction made to account for the fact that chi-square test is **biased upwards** for a 2 x 2 contingency table. An upwards bias tends to make results larger than they should be.
- Yates correction should be used:
  - If the expected cell frequencies are below 5
  - If a 2 x 2 contingency table is being used
- With large sample sizes, Yates' correction makes little difference, and the chi-square test works well. With small sample sizes, chi-square is not accurate, with or without Yates' correction.
- The chi-square test is only an **approximation**. Though the **Yates continuity correction** makes the chi-square approximation better, but in this process it over corrects so as to give a P value that is too large. When conditions for approximation of the chi-square tests is not held, **Fisher's exact test** is applied.

# Fisher's exact test

- **Fisher's exact test** is an alternative statistical significance test to chi square test used in the analysis of 2 x 2 contingency tables.
- It is one of a class of *exact tests*, so called because the *significance of the deviation from a null hypothesis ( P-value) can be calculated exactly*, rather than relying on an approximation that becomes exact as the sample size grows to infinity, as seen with chi-square test.
- It is used to examine the significance of the association between the two kinds of classification.
- It is valid for all sample sizes, although in practice it is employed when *sample sizes are small ( $n < 20$ ) and expected frequencies are small ( $n < 5$ )*.

# McNemar's test

- **McNemar's test** is a statistical test used on *paired nominal data*.
- It is applied to  $2 \times 2$  contingency tables with a dichotomous trait, with *matched pairs of subjects*, to determine whether the row and column marginal frequencies are equal (that is, whether there is "marginal homogeneity").

	Test 2 positive	Test 2 negative	Row total
Test 1 positive	$a$	$b$	$a + b$
Test 1 negative	$c$	$d$	$c + d$
Column total	$a + c$	$b + d$	$n$



	Test 2 positive	Test 2 negative	Row total
Test 1 positive	$a$	$b$	$a + b$
Test 1 negative	$c$	$d$	$c + d$
Column total	$a + c$	$b + d$	$n$

- The null hypothesis of marginal homogeneity states that the two marginal probabilities for each outcome are the same, i.e.  $p_a + p_b = p_a + p_c$  and  $p_c + p_d = p_b + p_d$ .

- Thus the null and alternative hypotheses are:
 
$$H_0 : p_b = p_c$$

$$H_1 : p_b \neq p_c$$

The McNemar test statistic is:

$$\chi^2 = \frac{(b - c)^2}{b + c}.$$

Comparing two proportions  
(2 by 2 table)

unpaired

paired

Fisher's exact  
Test

McNemar's test<sup>a</sup>

Chi-square test

## EXAMPLES:

Estrogen supplementation to delay or prevent the onset of Alzheimer's disease in postmenopausal women.

		Alzheimer's onset during 5-year period		
		No	Yes	
received estrogen	Yes	147	9	156
	No	810	158	968
		957	167	1,124

**The null hypothesis ( $H_0$ ):** Estrogen supplementation in postmenopausal women is unrelated to Alzheimer's onset.

**The alternate hypothesis ( $H_A$ ):** Estrogen supplementation in postmenopausal women delays/prevents Alzheimer's onset.

		Alzheimer's onset during 5-year period		
		No	Yes	
received estrogen	Yes	147	9	156
	No	810	158	968
		957	167	1,124

Of the women who did not receive estrogen supplementation, 16.3% (158/968) showed signs of Alzheimer's disease onset during the five-year period; whereas, of the women who did receive estrogen supplementation, only 5.8% (9/156) showed signs of disease onset.

- Next step: To calculate expected cell frequencies

		Alzheimer's onset during 5-year period		
		No	Yes	
received estrogen	Yes	147	9	156
	No	810	158	968
		957	167	1,124

		Alzheimer's onset during 5-year period		
		No	Yes	
received estrogen	Yes	$E_a = \frac{156 \times 957}{1124}$ $= 132.82$	$E_b = \frac{156 \times 167}{1124}$ $= 23.18$	156
	No	$E_c = \frac{968 \times 957}{1124}$ $= 824.18$	$E_d = \frac{968 \times 167}{1124}$ $= 143.82$	968
		957	167	1,124

		Alzheimer's onset during 5-year period		
		No	Yes	
received estrogen	Yes	147 132.82	9 23.18	156
	No	810 824.18	158 143.82	968
		957	167	1,124

$$\chi^2 = \sum \frac{(|O - E| - .5)^2}{E}$$

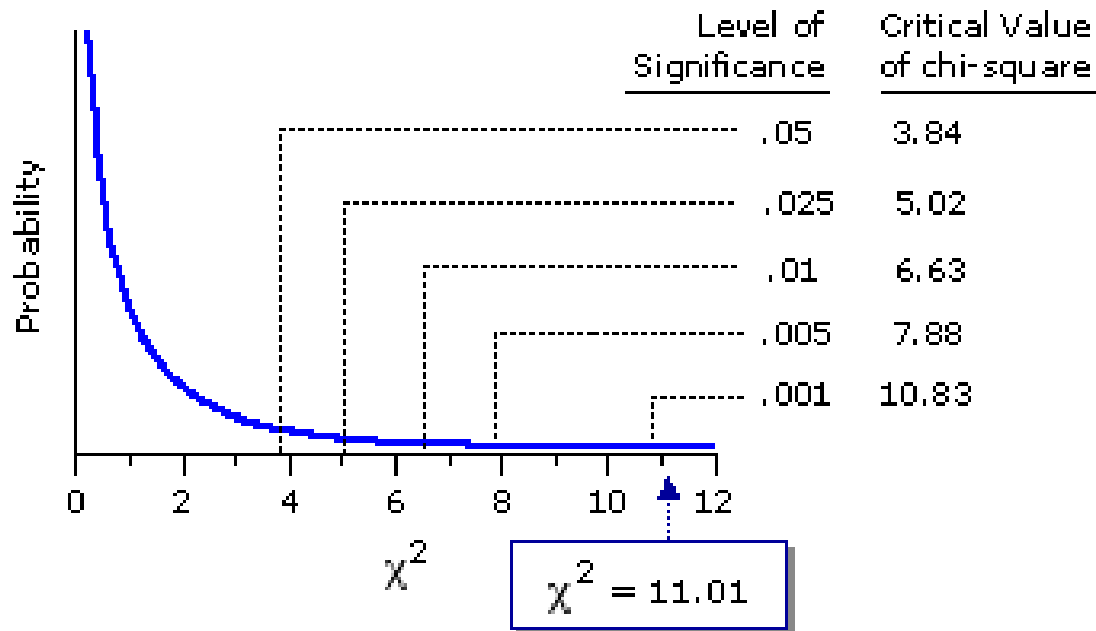
		Alzheimer's onset during 5-year period	
		No	Yes
received estrogen	Yes	$\frac{( 147 - 132.82  - .5)^2}{132.82} = 1.41$	$\frac{( 9 - 23.18  - .5)^2}{23.18} = 8.07$
	No	$\frac{( 810 - 824.18  - .5)^2}{824.18} = 0.23$	$\frac{( 158 - 143.82  - .5)^2}{143.82} = 1.3$
			sum: $\chi^2 = 11.01$

The next step is to refer calculated value of chi-square to the appropriate sampling distribution, which is defined by the applicable number of degrees of freedom.

- For this example, there are 2 rows and 2 columns. Hence,

$$df = (2-1)(2-1) = 1$$

## Sampling Distribution of Chi-Square for $df=1$

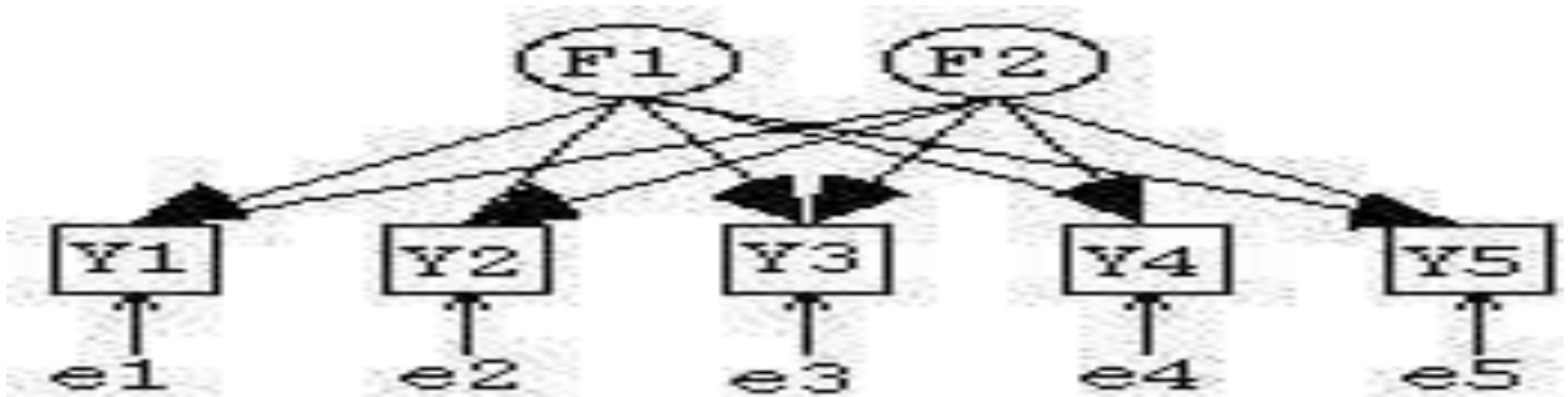




# Factor Analysis

- Variable reduction technique
  - Reduces a set of variable in terms of a small number of latent factors(unobservable).
- 
- Factor analysis is a correlational method used to find and describe the underlying factors driving data values for a large set of variables.

## SIMPLE PATH DIAGRAM FOR A FACTOR ANALYSIS MODEL



- F1 and F2 are two common factors. Y1, Y2, Y3, Y4, and Y5 are observed variables, possibly 5 subtests or measures of other observations such as responses to items on a survey.
- e1, e2, e3, e4, and e5 represent residuals or unique factors, which are assumed to be uncorrelated with each other.

# Uses of Factor Analysis

- ❖ Questionnaire construction
- ❖ Test Battery construction



# Conducting Factor Analysis

Testing the Assumptions



Construction of correlation Matrix



Method of Factor Analysis



Determination of Number of Factors

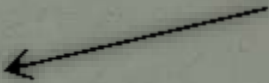


Rotation of Factors



Interpretation of Factors

# Assumptions to be fulfilled for running Factor analysis

1. No outliers in the data set.
2. Normality of the data set. KMO test is used
3. Adequate sample size. 
4. Multi collinearity and singularity among the variables does not exist.
5. 

---

Homoscedasticity does not exist between the variables because factor analysis is a linear function of measured variables.
6. Variables should be linear in nature.
7. Data should be metric in nature i.e. on interval and ratio scale.

## Bartlett test of sphericity

It tests the null hypothesis that all the correlations between the variables are zero.

It also tests whether the correlation matrix is an identity matrix or not.

If it is an identity matrix then factor analysis becomes inappropriate.

---

## Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy

This test checks the adequacy of data for running the factor analysis. The value of KMO ranges from 0 to 1. The larger the value of KMO, the more adequate is the sample for running the factor analysis. Kaiser recommends accepting values greater than 0.5 as acceptable.

**Problem formulation**



**Testing the Assumptions**



**Construction of correlation Matrix**



**Method of Factor Analysis**



**Determination of Number of Factors**



**Rotation of Factors**



**Interpretation of Factors**

# Construction of the Correlation Matrix

- Analyses the pattern of correlations between variables in the correlation matrix
- Which variables tend to correlate highly together?
- If variables are highly correlated, likely that they represent the same underlying dimension

Factor analysis pinpoints the clusters of high correlations between variables and for each cluster, it will assign a factor



# Correlation Matrix

	Q1	Q2	Q3	Q4	Q5	Q6
Q1	1					
Q2	<b>.987</b>	1				
Q3	<b>.801</b>	<b>.765</b>	1			
Q4	-.003	-.088	0	1		
Q5	-.051	.044	.213	<b>.968</b>	1	
Q6	-.190	-.111	0.102	<b>.789</b>	<b>.864</b>	1

- **Q1-3** correlate strongly with each other and hardly at all with 4-6
- **Q4-6** correlate strongly with each other and hardly at all with 1-3
- Two factors!

**Problem formulation**



**Testing the Assumptions**



**Construction of correlation Matrix**



**Method of Factor Analysis**



**Determination of Number of Factors**



**Rotation of Factors**



**Interpretation of Factors**

# Method of Factor Analysis

## (A) Principal component analysis

- Provides a unique solution, so that the original data can be reconstructed from the results
- It looks at the total variance among the variables that is the unique as well as the common variance.
- In this method, the factor explaining the maximum variance is extracted first.

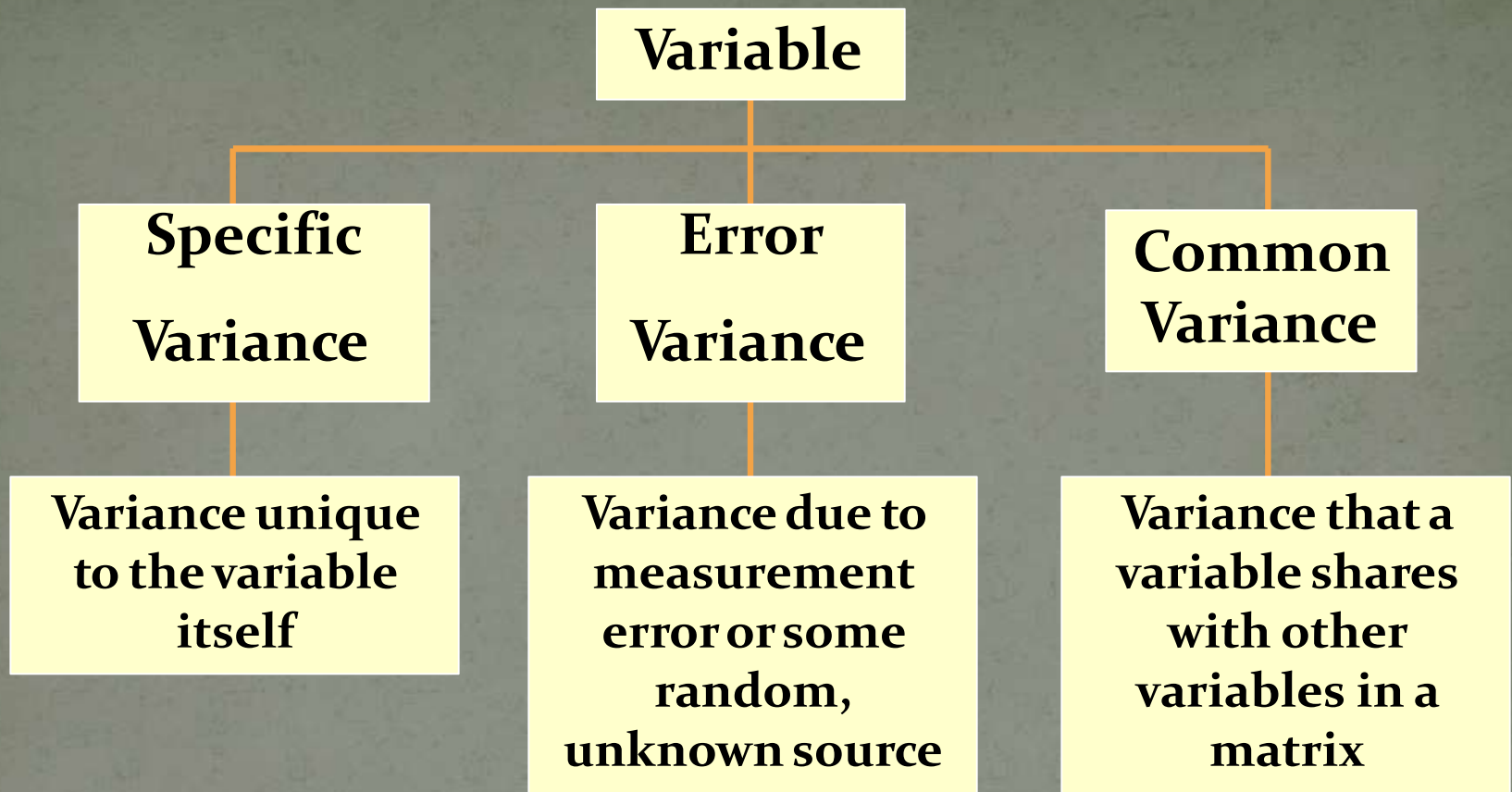
## *(B) Common factor analysis*

Uses an estimate of common variance among the original variables to generate factor solution.

Because of this, the number of factors will always be less than the number of original variables

### Other Methods Includes:-

Un weighted least squares, Generalized least squares, Maximum likelihood, Principal axis factoring, Alpha factoring, and Image factoring.



**Total Variance = common variance + specific variance + error variance**

*When searching for the factors underlying the relationships between a set of variables, we are interested in detecting and explaining the common variance*

# Determination of Number of Factors

## *EIGEN VALUE*

- The Eigen value for a given factor measures the variance in all the variables which is accounted for by that factor.
- It is the amount of variance explained by a factor. It is also called as characteristic root.

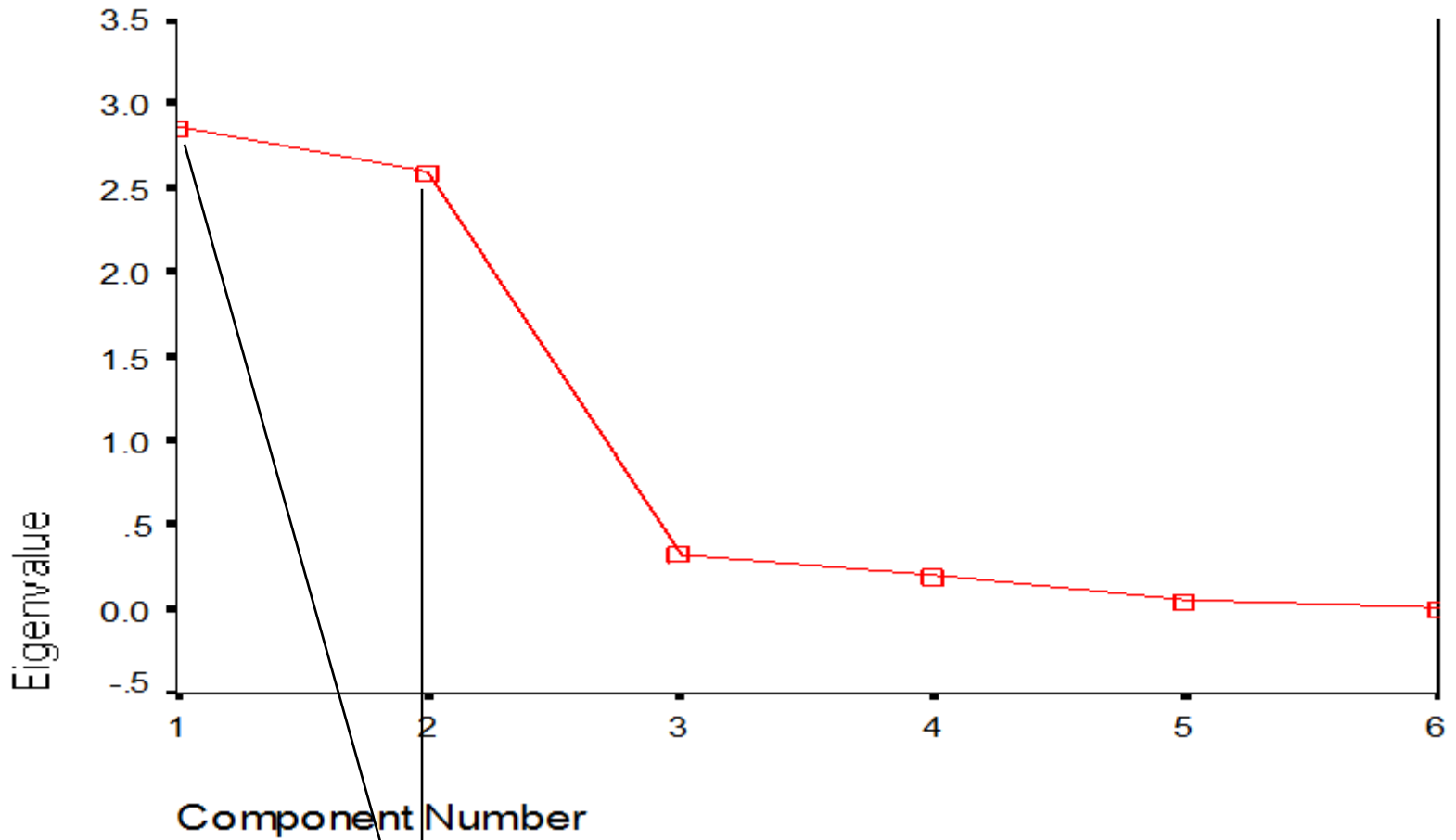
## *Kaiser Guttman Criterion*

This method states that the number of factors to be extracted should be equal to the number of factors having an Eigen value of 1 or greater than 1.

## *The Scree Plot*

- The examination of the **Scree plot** provides a visual of the total variance associated with each factor.
- The steep slope shows the large factors.
- The gradual trailing off (scree) shows the rest of the factors usually lower than an Eigen value of 1.

## Scree Plot

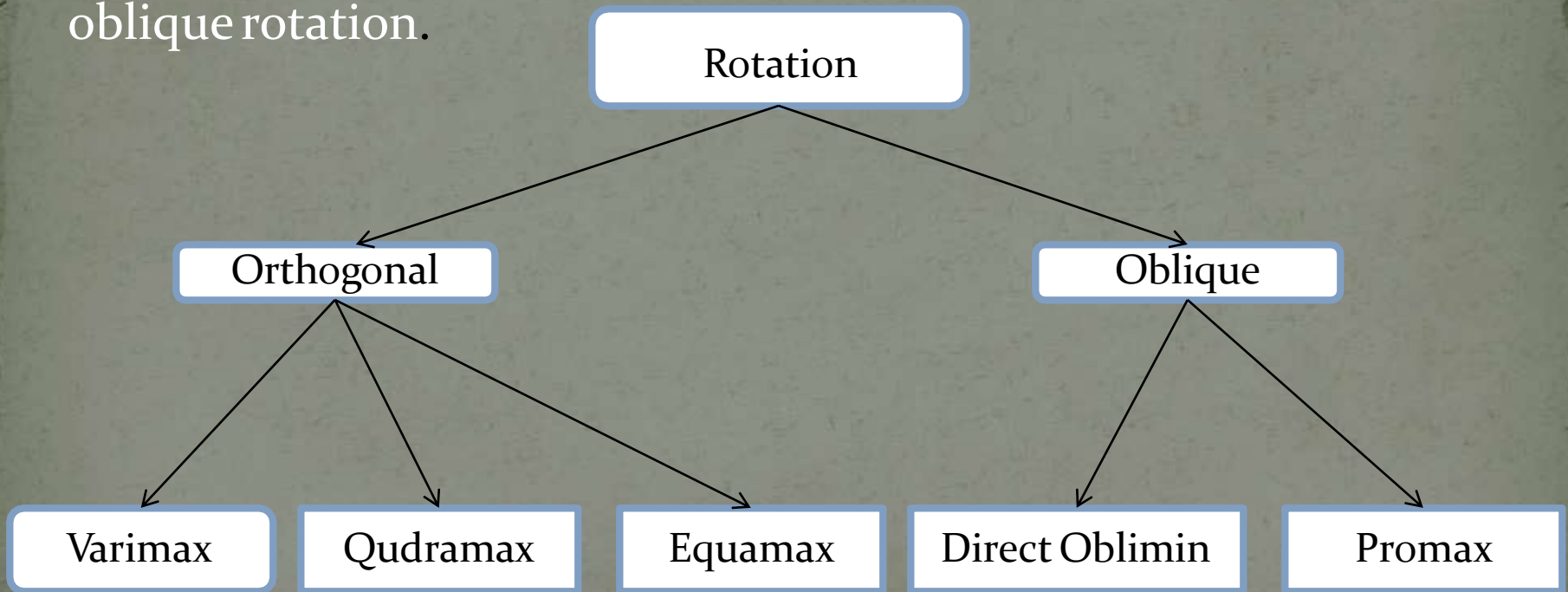


Take the components *above* the elbow

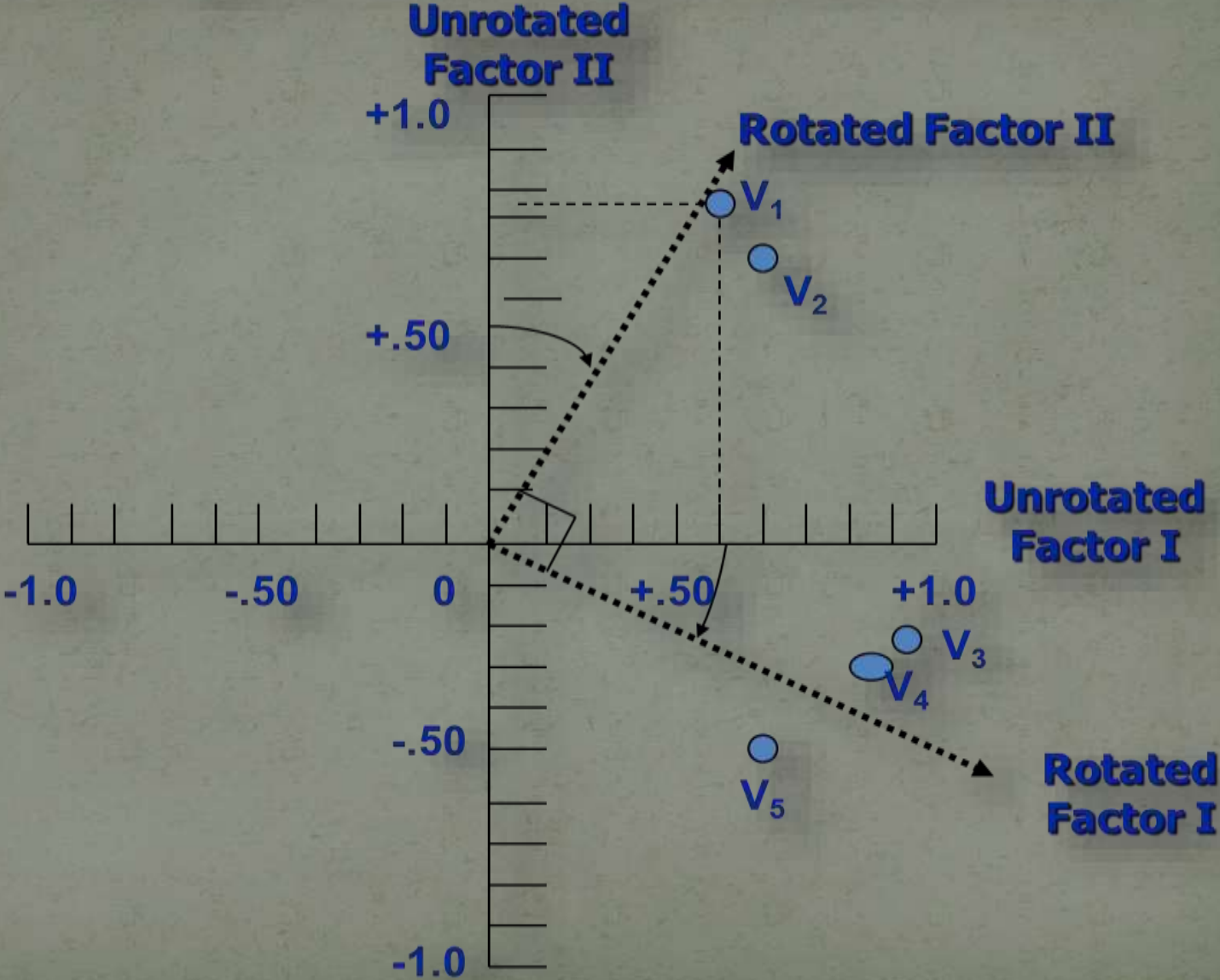


# Rotation of Factors

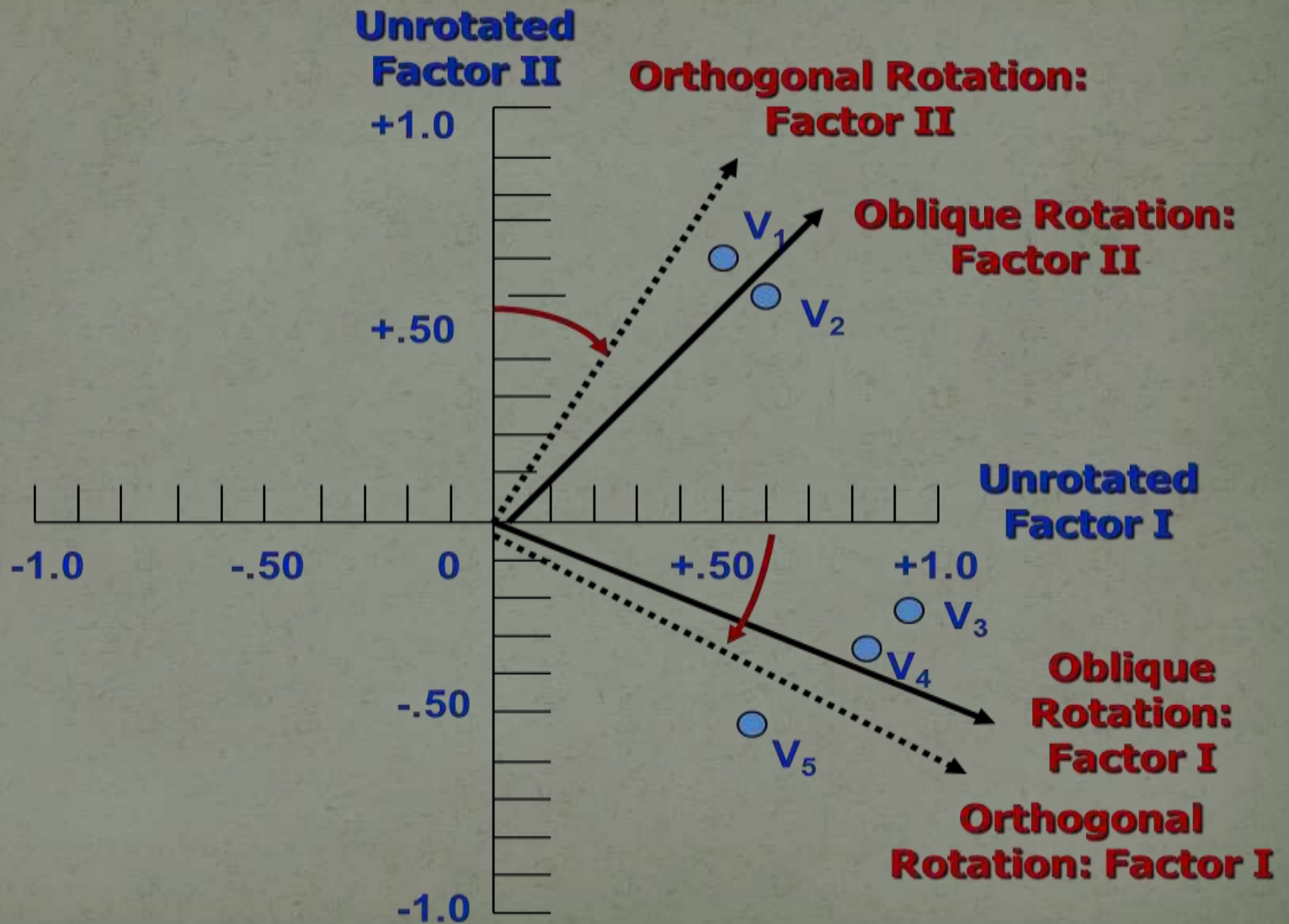
- Maximizes high item loadings and minimizes low item loadings, thereby producing a more interpretable and simplified solution.
- Two common rotation techniques orthogonal rotation and oblique rotation.



# Orthogonal Factor Rotation



# Oblique Factor Rotation



# KEY TERMINOLOGIES TO KNOW

## Factor Loading

- It can be defined as the correlation coefficient between the variable and the factor.
- The squared factor loading of a variable indicates the percentage variability explained by the factor in that variable. A factor loading of 0.7 is considered to be sufficient.

## COMMUNALITY

- The communality is the amount of variance each variable in the analysis shares with other variables.
- Squared multiple correlation for the variable as dependent using the factors as predictors and is denoted by  $h^2$ .
- The value of communality may be considered as the indicator of reliability of a variable.

# Principal Component Analysis

## Overview:

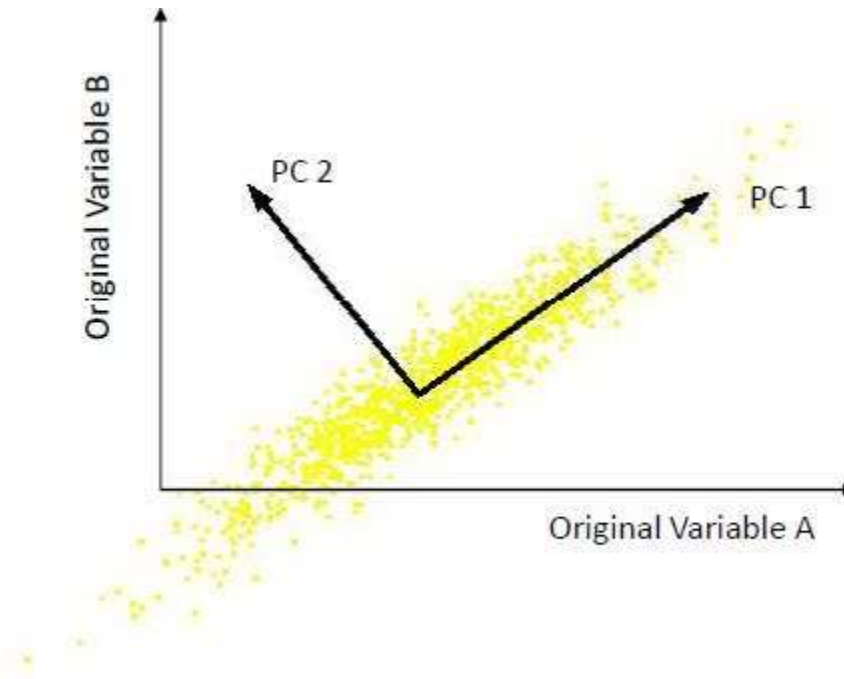
- What is Principal Component Analysis
- Computing the components in PCA
- Dimensionality Reduction using PCA
- A 2D example in PCA
- Applications of PCA in computer vision
- Importance of PCA in analysing data in higher dimensions
- Questions

## Principal Component Analysis

- Most common form of factor analysis
- The new variables/dimensions
  - Are linear combinations of the original ones
  - Are uncorrelated with one another
- Orthogonal in original dimension space
  - Capture as much of the original variance in the data as possible
  - Are called Principal Components



# What are the new axes?



- Orthogonal directions of greatest variance in data
- Projections along PC1 discriminate the data most along any one axis

# Principal Components

- First principal component is the direction of greatest variability (covariance) in the data
- Second is the next orthogonal (uncorrelated) direction of greatest variability
  - So first remove all the variability along the first component, and then find the next direction of greatest variability
- And so on ...

# Principal Components Analysis (PCA)

- Principle

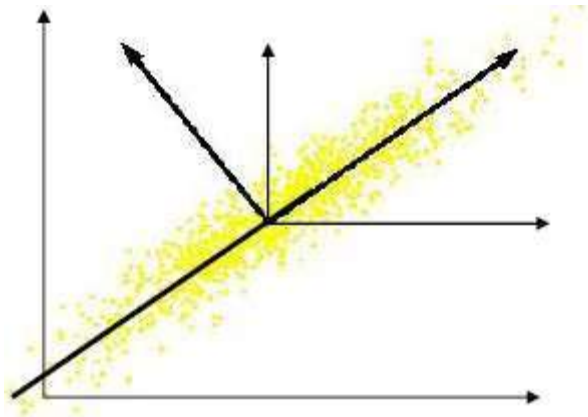
- Linear projection method to reduce the number of parameters
- Transfer a set of correlated variables into a new set of uncorrelated variables
- Map the data into a space of lower dimensionality
- Form of unsupervised learning

- Properties

- It can be viewed as a rotation of the existing axes to new positions in the space defined by original variables
- New axes are orthogonal and represent the directions with maximum variability

# Computing the Components

- Data points are vectors in a multidimensional space
- Projection of vector  $\mathbf{x}$  onto an axis (dimension)  $\mathbf{u}$  is  $\mathbf{u} \cdot \mathbf{x}$
- Direction of greatest variability is that in which the average square of the projection is greatest
  - I.e.  $\mathbf{u}$  such that  $E((\mathbf{u} \cdot \mathbf{x})^2)$  over all  $\mathbf{x}$  is maximized
  - (we subtract the mean along each dimension, and center the original axis system at the centroid of all data points, for simplicity)
  - This direction of  $\mathbf{u}$  is the direction of the first Principal Component



# Computing the Components

- $E((\mathbf{u} \cdot \mathbf{x})^2) = E((\mathbf{u} \cdot \mathbf{x})(\mathbf{u} \cdot \mathbf{x})^T) = E(\mathbf{u} \cdot \mathbf{x} \cdot \mathbf{x}^T \cdot \mathbf{u}^T)$
- The matrix  $\mathbf{S} = \mathbf{x} \cdot \mathbf{x}^T$  contains the correlations (similarities) of the original axes based on how the data values project onto them
- So we are looking for  $w$  that maximizes  $\mathbf{u}^T \mathbf{S} \mathbf{u}$ , subject to  $\mathbf{u}$  being unit-length
- It is maximized when  $w$  is the principal eigenvector of the matrix  $\mathbf{S}$ , in which case
  - $\mathbf{u}^T \mathbf{C} \mathbf{u} = \mathbf{u}^T \lambda \mathbf{u} = \lambda$  if  $\mathbf{u}$  is unit-length, where  $\lambda$  is the principal eigenvalue of the correlation matrix  $\mathbf{C}$
  - The eigenvalue denotes the amount of variability captured along that dimension

# Why the Eigenvectors?

Maximise  $\mathbf{u}^T \mathbf{x} \mathbf{x}^T \mathbf{u}$  s.t  $\mathbf{u}^T \mathbf{u} = 1$

Construct Lagrangian  $\mathbf{u}^T \mathbf{x} \mathbf{x}^T \mathbf{u} - \lambda \mathbf{u}^T \mathbf{u}$

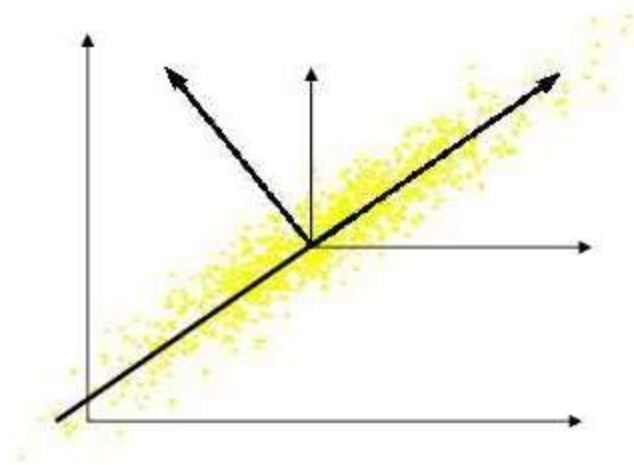
Vector of partial derivatives set to zero

$$\mathbf{x} \mathbf{x}^T \mathbf{u} - \lambda \mathbf{u} = (\mathbf{x} \mathbf{x}^T - \lambda \mathbf{I}) \mathbf{u} = 0$$

As  $\mathbf{u} \neq \mathbf{0}$  then  $\mathbf{u}$  must be an eigenvector of  $\mathbf{x} \mathbf{x}^T$  with eigenvalue  $\lambda$

## Computing the Components

- Similarly for the next axis, etc.
- So, the new axes are the eigenvectors of the matrix of correlations of the original variables, which captures the similarities of the original variables based on how data samples project to them



- Geometrically: centering followed by rotation
- – Linear transformation

## PCs, Variance and Least-Squares

- The first PC retains the greatest amount of variation in the sample
- The  $k$ th PC retains the  $k$ th greatest fraction of the variation in the sample
- The  $k$ th largest eigenvalue of the correlation matrix  $C$  is the variance in the sample along the  $k$ th PC
- The least-squares view: PCs are a series of linear least squares fits to a sample, each orthogonal to all previous ones

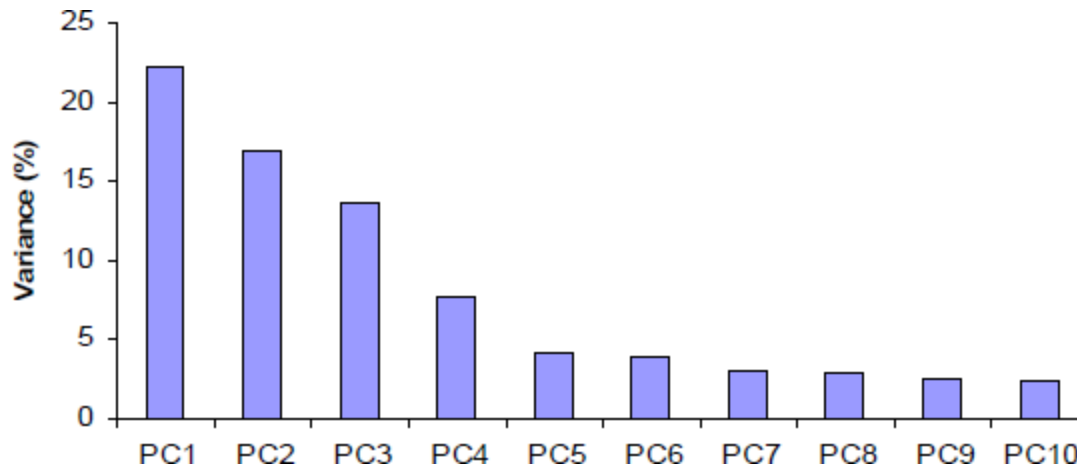


## How Many PCs?

- For  $n$  original dimensions, correlation matrix is  $n \times n$ , and has up to  $n$  eigenvectors. So  $n$  PCs.
- Where does dimensionality reduction come from?

# Dimensionality Reduction

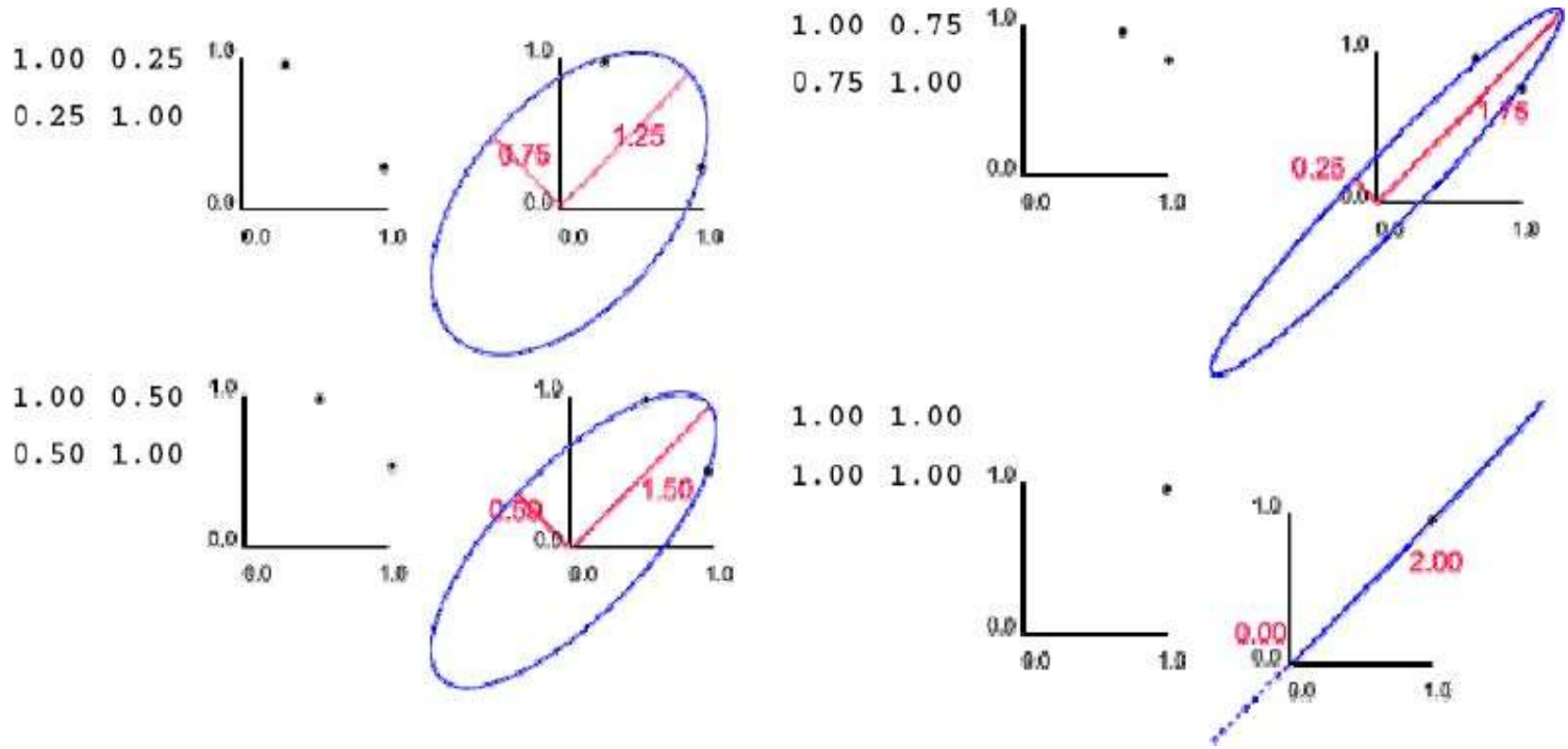
Can *ignore* the components of lesser significance.



You do lose some information, but if the eigenvalues are small, you don't lose much

- n dimensions in original data
- calculate n eigenvectors and eigenvalues
- choose only the first p eigenvectors, based on their eigenvalues
- final data set has only p dimensions

# Eigenvectors of a Correlation Matrix



# PCA Example –STEP 1

---

- Subtract the mean

from each of the data dimensions. All the x values have  $\bar{x}$  subtracted and y values have  $\bar{y}$  subtracted from them. This produces a data set whose mean is zero.

Subtracting the mean makes variance and covariance calculation easier by simplifying their equations. The variance and co-variance values are not affected by the mean value.

# PCA Example –STEP 2

---

- Calculate the covariance matrix

$$\text{cov} = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix}$$

- since the non-diagonal elements in this covariance matrix are positive, we should expect that both the x and y variable increase together.

# PCA Example –STEP 3

---

- Calculate the eigenvectors and eigenvalues of the covariance matrix

$$\text{eigenvalues} = \begin{pmatrix} .0490833989 \\ 1.28402771 \end{pmatrix}$$

$$\text{eigenvectors} = \begin{pmatrix} -.735178656 & -.677873399 \\ .677873399 & -.735178656 \end{pmatrix}$$

# PCA Example –STEP 4

---

- Reduce dimensionality and form *feature vector*

the eigenvector with the *highest* eigenvalue is the *principle component* of the data set.

In our example, the eigenvector with the largest eigenvalue was the one that pointed down the middle of the data.

Once eigenvectors are found from the covariance matrix, the next step is to **order them by eigenvalue**, highest to lowest. This gives you the components in order of significance.

# PCA Example –STEP 4

---

Now, if you like, you can decide to *ignore* the components of lesser significance.

You do *lose some information*, but if the eigenvalues are small, you don't lose much

- $n$  dimensions in your data
- calculate  $n$  eigenvectors and eigenvalues
- choose only the first  $p$  eigenvectors
- final data set has only  $p$  dimensions.



# PCA Example –STEP 4

---

- Feature Vector

$$\text{FeatureVector} = (\text{eig}_1 \text{ eig}_2 \text{ eig}_3 \dots \text{eig}_n)$$

We can either form a feature vector with both of the eigenvectors:

$$\begin{pmatrix} -.677873399 & -.735178656 \\ -.735178656 & .677873399 \end{pmatrix}$$

or, we can choose to leave out the smaller, less significant component and only have a single column:

$$\begin{pmatrix} -.677873399 \\ -.735178656 \end{pmatrix}$$

# PCA Example –STEP 5

---

- Deriving the new data

**FinalData = RowFeatureVector x RowZeroMeanData**

**RowFeatureVector** is the matrix with the eigenvectors in the columns *transposed* so that the eigenvectors are now in the rows, with the most significant eigenvector at the top

**RowZeroMeanData** is the mean-adjusted data *transposed*, ie. the data items are in each column, with each row holding a separate dimension.

# PCA Example –STEP 5

---

FinalData transpose: dimensions  
along columns

x	y
-.827970186	-.175115307
1.77758033	.142857227
-.992197494	.384374989
-.274210416	.130417207
-1.67580142	-.209498461
-.912949103	.175282444
.0991094375	-.349824698
1.14457216	.0464172582
.438046137	.0177646297
1.22382056	-.162675287

# Reconstruction of original Data

---

- If we reduced the dimensionality, obviously, when reconstructing the data we would lose those dimensions we chose to discard. In our example let us assume that we considered only the x dimension...

## PCA for image compression:

- Compile a dataset of 20 images
- Build the covariance matrix of 20 dimensions
- Compute the eigenvectors and eigenvalues
- Based on the eigenvalues, 5 dimensions can be left out, those with the least eigenvalues.
- $1/4^{\text{th}}$  of the space is saved.

# Importance of PCA

- In data of high dimensions, where graphical representation is difficult, PCA is a powerful tool for analysing data and finding patterns in it.
- Data compression is possible using PCA
- The most efficient expression of data is by the use of perpendicular components, as done in PCA.

# Introduction: What is SPSS?

“Statistical package for social sciences”

One of the most popular statistical packages which can perform highly complex **data management** and **analysis** with simple instructions.

# Statistics

Statistics is the science of collecting, organizing, presenting, analyzing and interpreting numerical data to assist in making more effective decisions/conclusions.



# Uses of SPSS

- Data management
- Data analysis

# What is a variable

- Any characteristic which is subject to change and can have more than one value.
- Anything that has a quantity or quality that varies.

# Variables affect each other



# Dependent variable

- Variable affected by the independent variable
- It responds to the independent variable



# Independent variable

- Variable that is presumed to influence other variable
- It is the presumed cause whereas the dependent variable is the presumed effect

Independent  
Variable (IV)



Dependent  
Variable (DV)

# Example

- You are interested in how stress affects mental state of human beings
- Independent variable : **stress**
- Dependent variable : **mental state of H.B**

You can manipulate stress level and measure how those stress levels change mental state

# Example 2

Prizes affects student's motivation

Prize :

**Independent variable**

Student's motivation:

**dependent variable**



# Extraneous variable

- Any variable other than dependent or independent variable

# Extraneous variable

- Extraneous variables are undesirable variables that influence the relationship between the variables an experimenter is examining

# example

- Two groups of students are made.
- All of the students study text material on a biology topic for thirty minutes. One group uses a new strategy and other uses a strategy of its own choice
- Then all students complete a test over the material.
- Extraneous knowledge... pre knowledge of the topic.

# Intervening/confounding variable

- It is a variable whose existence is inferred but it can not be measured

# Example

- Determining the effects of video clips on learning ability of students
- The association between the video clips and learning ability is to be explained

**other variables intervene**

- Such as anxiety , fatigue etc.

# Nominal or categorical variable

- They can be measured only in terms of whether the individual items belong to certain distinct categories
- Different numbers for different objects
- We can not quantify or rank/order the categories
- Nominal data has no order

# Example

- Gender
  1. male
  2. female
- Martial status
  1. married
  2. unmarried

# Ordinal variable

- Ordinal data has an order but the intervals between the scale points may be uneven.
- Numbers have meaningful order
- A typical example of ordinal data is the socioeconomic status of the families
  - Lower
  - Middle
  - Upper



# Interval

- Numbers have orders but there are also equal intervals between adjacent categories
- Example : temperature in degrees
- 43
- 44
- 45

# Ratio

- Differences are meaningful ( like interval) plus now ratios are also meaningful and there is a true zero point
- Example: weight in pounds
- 10 lbs is twice as much as 5 lbs (ratios are meaningful)  $10/5=2$  and
- Zero pounds means no weight or an absence of weight