

**LECTURE 24**  
**STATISTICAL REPRESENTATION**  
**MEASURES OF CENTRAL TENDENCY**  
**PART 1**

**OBJECTIVES**

The objectives of the lecture are to learn about:

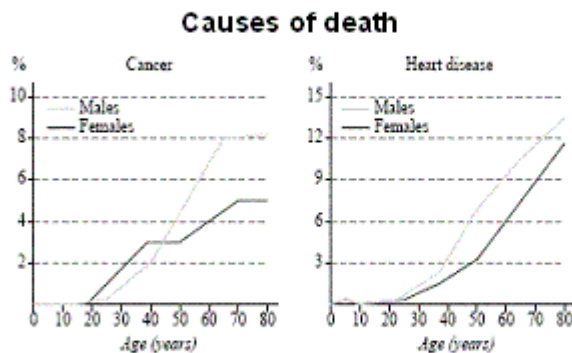
- Review Lecture 18
- Statistical Representation
- Measures of Central Tendency

**LINE GRAPHS**

Line graphs are the most commonly used graphs. In the following graph, you can see the occurrence of causes of death due to cancer in males and females. You can see that after the age of 40, the occurrence of cancer is much greater in the case of males. The line graph of heart diseases also shows that the disease is more prominent in the case of males.

As you see line graphs help us to understand the trends in data very clearly.

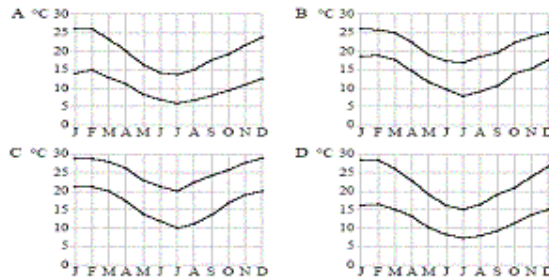
## LINE GRAPHS



Another line graph of temperature in 4 cities A, B, C and D shows that although the general pattern is similar, the temperature in city A is lowest followed by D, B and C. In city C the highest temperature is close to 30 whereas in city A and B it is about 25. The highest temperature in city D is about 28 degrees.

## LINE GRAPHS

Temperature of cities A, B, C and D



### MEAN

The most common average is the mean. The mean is used for things like marks and scores (e.g. sport), and is found by adding all the scores and dividing by the number of scores.

### Marks

58 69 73 67 76 88 91 and 74 (8 marks).

**Sum** = 596

**Mean** =  $596/8 = 74.5$

Please note that the mean is affected by extreme values.

### MEDIAN

Another typical value is the median. The median is the middle value when the data are arranged in order.

The median is easier to find than the mean, and unlike the mean it is not affected by values that are unusually high or low

### Data

3 6 11 14 19 19 21 24 31 (9 values)

The median is the middle score, or the mean of the two middle scores, when the scores are placed in order. In the above data there are 9 values. The middle value is 19.

When there is no middle value, the median is obtained by taking the average of the two middle values.

### MODE

The most common score in a set of scores is called the mode.

There may be more than one mode, or no mode at all

2 2 1 2 0 3 2 1 1 4 1 1 1 2 2 0 3 2 1

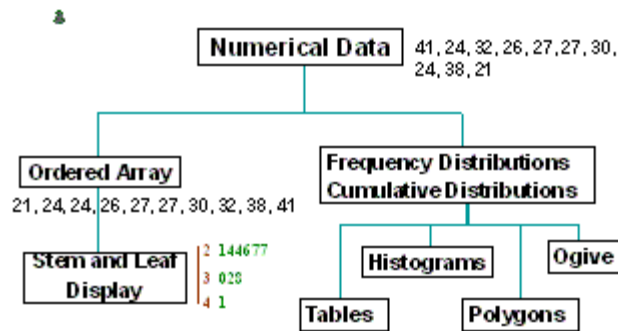
The mode, or most common value, is 1.

### ORGANISING DATA

There are many different ways of organizing data.

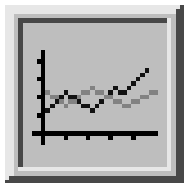
#### Organising Numerical Data

## Organizing Numerical Data



Numerical data can be organized in any of the following forms:

- The Ordered Array and Stem-leaf Display
- Tabulating and Graphing Numerical Data
- Frequency Distributions: Tables, Histograms, Polygons
- Cumulative Distributions: Tables, the Ogive



## ORGANISING NUMERICAL DATA

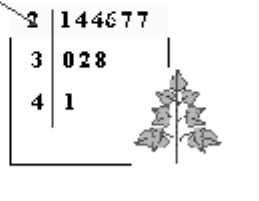
Data in Raw form (as collected)

24, 26, 24, 21, 27, 27, 30, 41, 32, 38

Data Ordered from Smallest to Largest:

21, 24, 24, 26, 27, 27, 30, 32, 38, 41

Stem and Leaf display:



**Tabulating and Graphing Univariate Categorical Data**

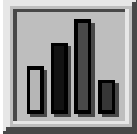
There are different ways of organizing univariate categorical data:

- The Summary Table
- Bar and Pie Charts, the Pareto Diagram

**Tabulating and Graphing Bivariate Categorical Data**

Bivariate categorical data can be organized as :

- Contingency Tables
- Side by Side Bar charts

**GRAPHICAL EXCELLENCE AND COMMON ERRORS IN PRESENTING DATA**

It is important that data is organised in a professional manner and graphical excellence is achieved in its presentation. High quality and attractive graphs can be used to explain and highlight facts which otherwise may go unnoticed in descriptive presentations. That is why all companies in their annual reports use different types of graphs to present data.

**Tabulating Numerical Data: Frequency Distributions**

The process of developing frequency distributions is described below.

**Step 1: Sort Raw Data in Ascending Order**

Data: 12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

**Step 2: Find Range**

Range:  $58 - 12 = 46$

**Step 3: Select Number of Classes**

Select the number of classes. (The classes are usually selected between 5 and 15)  
Say 5.

**Step 4: Compute Class Interval (width)**

=  $10$  ( $46/5$  then round up)

**Step 5: Determine Class Boundaries (limits)**

Start with 10 as the first limit.

Then add 10 to each limit: 10,  $20(=10+10)$ ,  $30(=20+10)$ ,  $40(=30+10)$ ,  $50(=40+10)$

**Step 6: Compute Class Midpoints**

First midpoint is  $10+20/2=15$ .

Midpoints:  $15((10+20)/2)$ ,  $25((20+30)/2)$ ,  $35((30+40)/2)$ ,  $45((40+50)/2)$ ,  $55((50+60)/2)$

**Step 7: Count Observations & Assign to Classes**

First class: Lower limit is 10. Higher limit is 20. We read it as "10 but under 20". In reality a value greater than 19.5 will be treated as above 20.

Frequency: Looking through the data shows that there are three values between 10 and 20. Hence frequency is 3. Similarly, frequency in other intervals can be found as follows:

20 - 30 : 6

30 - 40 : 5

40 - 50 : 4

50 - 60 : 2

Total : 20

Relative frequency: There are 3 observations in class interval 10 – 20. The relative frequency is  $3/20 = 0.15$ . Similarly frequency for other class intervals was calculated.

Percentage Frequency: If we multiply 0.15 by 100, then the % Relative Frequency 15% is obtained.

**TABULATING NUMERICAL DATA  
FREQUENCY DISTRIBUTIONS**

Data in ordered array

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Class	Frequency	Relative Frequency	Percentage
10 but under 20	3	.15	15
20 but under 30	6	.30	30
30 but under 40	5	.25	25
40 but under 50	4	.20	20
50 but under 60	2	.10	10
<b>Total</b>	<b>20</b>	<b>1</b>	<b>100</b>

Cumulative Frequency: If we add frequency of the second interval to the frequency of the second interval, then the cumulative frequency for the second interval is obtained. The cumulative frequency of the last interval is 100% as all observations have been added.

10 – 20 : 15  
 20 – 30 : 45  
 30 – 40 : 70  
 40 – 50 : 90  
 50 – 60 : 100

**LECTURE 25**  
**STATISTICAL REPRESENTATION**  
**MEASURES OF CENTRAL TENDENCY**  
**PART 2**

**OBJECTIVES**

The objectives of the lecture are to learn about:

- Review Lecture 24
- Statistical Representation
- Measures of Central Tendency

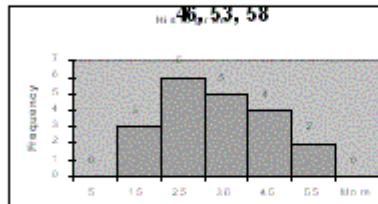
**Part 2****GRAPHING NUMERICAL DATA: THE HISTOGRAM**

When frequency is plotted in the form of bars or columns for each class interval a Histogram is obtained as shown below. The data is ordered in array form and frequency is counted for each class interval as explained under lecture 24.

**GRAPHING NUMERICAL  
DATA: THE HISTOGRAM**

Data in ordered array

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44,



**No Gaps  
Between  
Bars**

**Class Midpoints**

**MEASURES OF CENTRAL TENDENCY**

Measures of central tendency can be summarized as under:

- Arithmetic Mean
- Arithmetic Mean for Grouped Data
- Weighted Mean
- Median
- Median for Grouped Data
- Median for Discrete Data
- Graphic Location of Median
- Quintiles (Quartiles, Deciles, Percentiles)
- Quintiles from Grouped Data
- Quintiles from Discrete Data
- Graphic Location of Quintiles
- Mode
- Mode from Grouped Data
- Mode from Discrete Data
- Empirical Relation Between mean, Median and Mode

As you see it is a long list. However, if you look closely you will find that the main measures are Arithmetic Mean, Median, Mode and Quintiles.

All the above measures are used in different situations to understand the behaviour of data for decision making. It may be interesting to know the average, median or mode salary in an organization before you the company decides to increase the salary level. Comparisons with other companies are also important. The above measures provide a useful summary measure to consolidate large volumes of data. Without such summaries it is not possible to compare large selections of data.

EXCEL has a number of useful functions for calculating different measures of central tendency. Some of these are explained below. You are encouraged to go through EXCEL Help file for detailed descriptions of different functions. For selected functions, the help file has been included in the handouts. The examples are also from the help files.

### **AVERAGE**

Returns the average (arithmetic mean) of the arguments.

#### **Syntax**

**AVERAGE(number1,number2,...)**

Number1, number2, ... are 1 to 30 numeric arguments for which you want the average.

#### **Remarks**

- The arguments must either be numbers or be names, arrays, or references that contain numbers.
- If an array or reference argument contains text, logical values, or empty cells, those values are ignored; however, cells with the value zero are included

#### **Example**

An example of AVERAGE is shown below. Data was entered in cells A4 to A8. The formula was =AVERAGE(A4:A8). The 11 is shown in cell A10.

	A	B	C	D	E	F	G
1							
2							
3	<b>AVERAGE(number1,number2, ...)</b>						
4	<b>10</b>						
5	<b>7</b>						
6	<b>9</b>						
7	<b>27</b>						
8	<b>2</b>						
9	<b>=AVERAGE(A4:A8)</b>						
10	<b>11</b>						

### **AVERAGEA**

Calculates the average (arithmetic mean) of the values in the list of arguments. In addition to numbers, text and logical values such as TRUE and FALSE are included in the calculation.

**Syntax**

**AVERAGEA**(value1,value2,...)

Value1, value2, ... are 1 to 30 cells, ranges of cells, or values for which you want the average.

**Remarks**

- The arguments must be numbers, names, arrays, or references.
- Array or reference arguments that contain text evaluate as 0 (zero). Empty text ("") evaluates as 0 (zero). If the calculation must not include text values in the average, use the AVERAGE function.
- Arguments that contain TRUE evaluate as 1; arguments that contain FALSE evaluate as 0 (zero).

**Example**

	<b>A</b>
1	<b>Data</b>
2	10
3	7
4	9
5	2
6	Not available
7	

<b>Formula</b>	<b>Description (Result)</b>
=AVERAGEA(A2:A6)	Average of the numbers above, and the text "Not Available". The cell with the text "Not available" is used in the calculation. (5.6)
=AVERAGEA(A2:A5,A7)	Average of the numbers above, and the empty cell. (7)

**MEDIAN**

Returns the median of the given numbers. The median is the number in the middle of a set of numbers; that is, half the numbers have values that are greater than the median, and half have values that are less.

**Syntax**

**MEDIAN**(number1,number2,...)

Number1, number2, ... are 1 to 30 numbers for which you want the median.

**Remarks**

- The arguments should be either numbers or names, arrays, or references that contain numbers. Microsoft Excel examines all the numbers in each reference or array argument.



- If an array or reference argument contains text, logical values, or empty cells, those values are ignored; however, cells with the value zero are included.
- If there is an even number of numbers in the set, then MEDIAN calculates the average of the two numbers in the middle. See the second formula in the example.

### **Example**

The numbers are entered in cells A14 to A19.

In the first formula =MEDIAN(1,2,3,4,5) the actual values are specified. The median as you see is 3, in the middle.

In the next formula =MEDIAN(A14:A19), the entire series was specified. There is no middle value in the middle. Therefore the average of the two values 3 and 4 in the middle was used as the median 3.5.

	A	B	C	D	E	F	G
12	<b>MEDIAN(number1,number2, ...)</b>						
13							
14	<b>1</b>						
15	<b>2</b>						
16	<b>3</b>						
17	<b>4</b>						
18	<b>5</b>						
19	<b>6</b>						
20	<b>3 =MEDIAN(1, 2, 3, 4, 5)</b>						
21	<b>3,5 =MEDIAN(A14:A19)</b>						
22							

### **MODE**

Returns the most frequently occurring, or repetitive, value in an array or range of data. Like MEDIAN, MODE is a location measure.

#### **Syntax**

**MODE(number1,number2,...)**

Number1, number2, ... are 1 to 30 arguments for which you want to calculate the mode. You can also use a single array or a reference to an array instead of arguments separated by commas.

#### **Remarks**

- The arguments should be numbers, names, arrays, or references that contain numbers.
- If an array or reference argument contains text, logical values, or empty cells, those values are ignored; however, cells with the value zero are included.
- If the data set contains no duplicate data points, MODE returns the #N/A error value. In a set of values, the mode is the most frequently occurring value; the median

- is the middle value; and the mean is the average value. No single measure of central tendency provides a complete picture of the data. Suppose data is clustered in three areas, half around a single low value, and half around two large values. Both AVERAGE and MEDIAN may return a value in the relatively empty middle, and MODE may return the dominant low value.

**Example**

The data was entered in cells A27 to A32. The formula was =MODE(A27:A32). The answer 4 is the most frequently occurring value.

	A	B	C	D	E	F	G
24							
25	<b>MODE(number1,number2, ...)</b>						
26							
27	<b>5.6</b>						
28	<b>4</b>						
29	<b>4</b>						
30	<b>3</b>						
31	<b>2</b>						
32	<b>4</b>						
33	<b>4</b>	<b>=MODE(A27:A32)</b>					
34							

**COUNT FUNCTION**

Counts the number of cells that contain numbers and also numbers within the list of arguments. Use COUNT to get the number of entries in a number field that's in a range or array of numbers.

**Syntax**

**COUNT(value1,value2,...)**

Value1, value2, ... are 1 to 30 arguments that can contain or refer to a variety of different types of data, but only numbers are counted.

**Remarks**

- Arguments that are numbers, dates, or text representations of numbers are counted; arguments that are error values or text that cannot be translated into numbers are ignored.
- If an argument is an array or reference, only numbers in that array or reference are counted. Empty cells, logical values, text, or error values in the array or reference are ignored. If you need to count logical values, text, or error values, use the COUNTA function.

**Example**

1	A
2	Data

3	Sales
4	12/8/2008
5	
6	19
7	22.24
8	TRUE

#DIV/0!

**Formula****Description (Result)**

=COUNT(A2:A8)

Counts the number of cells that contain numbers in the list above (3)

=COUNT(A5:A8)

Counts the number of cells that contain numbers in the last 4 rows of the list (2)

=COUNT(A2:A8,2)

Counts the number of cells that contain numbers in the list, and the value 2 (4)

**FREQUENCY**

**Calculates how often values occur within a range of values, and then returns a vertical array of numbers. For example,** use FREQUENCY to count the number of test scores that fall within ranges of scores. Because FREQUENCY returns an array, it must be entered as an array formula.

**Syntax****FREQUENCY(data\_array,bins\_array)**

**Data\_array** is an array of or reference to a set of values for which you want to count frequencies. If data\_array contains no values, FREQUENCY returns an array of zeros.

**Bins\_array** is an array of or reference to intervals into which you want to group the values in data\_array. If bins\_array contains no values, FREQUENCY returns the number of elements in data\_array.

**Remarks**

- FREQUENCY is entered as an array formula after you select a range of adjacent cells into which you want the returned distribution to appear.
- The number of elements in the returned array is one more than the number of elements in bins\_array. The extra element in the returned array returns the count of any values above the highest interval. For example, when counting three ranges of values (intervals) that are entered into three cells, be sure to enter FREQUENCY into four cells for the results. The extra cell returns the number of values in data\_array that are greater than the third interval value.
- FREQUENCY ignores blank cells and text.
- Formulas that return arrays must be entered as array formulas.

**Example**

	<b>A</b>	<b>B</b>
	<b>Scores</b>	<b>Bins</b>
<b>1</b>	79	70
<b>2</b>	85	79
<b>3</b>	78	89
<b>4</b>	85	
<b>5</b>	50	
<b>6</b>	81	
<b>7</b>	95	
<b>8</b>	88	
<b>9</b>	97	
<b>10</b>		
<b>Formula</b>	<b>Description (Result)</b>	
=FREQUENCY(A2:A10,B2:B5)	Number of scores less than or equal to 70 (1)	
	Number of scores in the bin 71-79 (2)	
	Number of scores in the bin 80-89 (4)	
	Number of scores greater than or equal to 90 (2)	

**Note** The formula in the example must be entered as an array formula. After copying the example to a blank worksheet, select the range A13:A16 starting with the formula cell. Press F2, and then press CTRL+SHIFT+ENTER. If the formula is not entered as an array formula, the single result is 1.

**ARITHMETIC MEAN GROUPED DATA**

Below is an example of calculating arithmetic mean of grouped data. Here the marks and frequency are given. The class marks are the mid points calculated as average of lower and higher limits. For example, the average of 20 and 24 is 22. The frequency  $f$  is multiplied by the class mark to obtain the total number. In first row the value of  $fx$  is  $1 \times 22 = 22$ . The sum of all  $fx$  is 1950. The total number of observations is 50. Hence the arithmetic mean is  $1950/50 = 39$ .

Marks	Frequency	Class Marks	$fX$
20-24	1	22	22
24-29	4	27	108
30-34	8	32	256
35-39	11	37	407
40-44	15	42	630
45-49	9	47	423
50-54	2	52	104
TOTAL	50		1950

$n = 50$ ;  $\text{Sum}(fX) = 1950$ ;  $\text{Mean} = 1950/50 = 39$  Marks

**EXCEL Calculation**

The above calculation would be common in business life. Let us see how we can do it using EXCEL.

The basic data of lower limits is entered in cell range A54:A60. The data of higher limit is entered in cells B54:B60. Frequency is given in cell range D54:D60. Class mids were calculated in cells F54:F60. In cell F54 the formula  $=A54+B54/2$  was used to calculate the class mark. This formula was copied in other cells (F55 to F60). The value of  $fx$  was calculated in cell H54 using the formula  $=D54*F54$ . This formula was copied to other cells H55 to H60. Total frequency was calculated in cell D61 using the formula  $=\text{SUM}(D54:D60)$ . Sum of  $fx$  was calculated in cell H61 using the formula  $=\text{SUM}(H54:H60)$ . Mean was calculated in cell H62 using the formula  $=\text{ROUND}(H61/D61;0)$ . Watch for the “;” sign. It may be “,” on your computer.

	A	B	C	D	E	F	G	H	I	J	K
53	<b>Marks</b>	<b>Frequency(f)</b>		<b>Class Mark(X)</b>		<b>fX</b>					
54	20	24		1		22		22	D54*F54		
55	25	29		4		27		108			
56	30	34		8		32		256			
57	35	39		11		37		407			
58	40	44		15		42		630			
59	45	49		9		47		423			
60	50	54		2		52		104			
61	<b>TOTAL</b>			50				1950	=SUM(H54:H60)		
62						<b>Mean=</b>		39	=H61/D61		
63											

**LECTURE 26**  
**STATISTICAL REPRESENTATION**  
**MEASURES OF DISPERSION AND SKEWNESS**  
**PART 1**

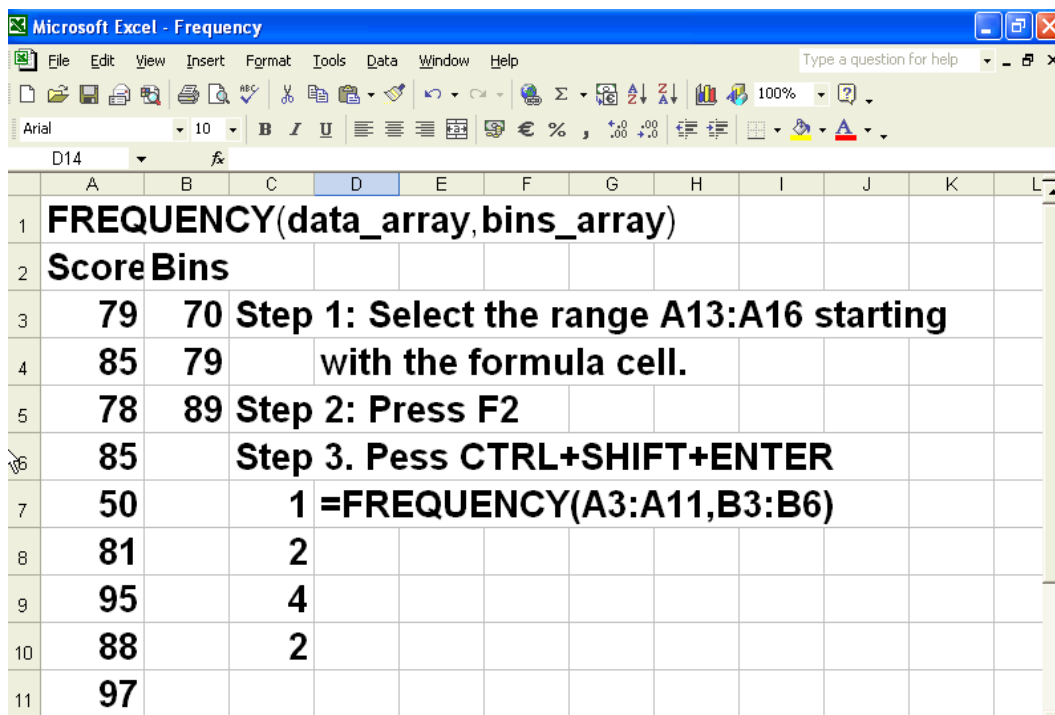
**OBJECTIVES**

The objectives of the lecture are to learn about:

- Review Lecture 25
- Statistical Representation
- Measures of Dispersion and Skewness

**FREQUENCY-EXAMPLE**

FREQUENCY Function calculates how often values occur within a range of values, and then returns a vertical array of numbers. For details see handout for lecture 25. The syntax is **FREQUENCY(data\_array,bins\_array)**.



	A	B	C	D	E	F	G	H	I	J	K	L
1	<b>FREQUENCY(data_array,bins_array)</b>											
2	<b>Score Bins</b>											
3	79	70	<b>Step 1: Select the range A13:A16 starting</b>									
4	85	79	<b>with the formula cell.</b>									
5	78	89	<b>Step 2: Press F2</b>									
6	85		<b>Step 3. Press CTRL+SHIFT+ENTER</b>									
7	50		1	<b>=FREQUENCY(A3:A11,B3:B6)</b>								
8	81		2									
9	95		4									
10	88		2									
11	97											

The data was entered in cells A3 to A11. The Bins array which gives the limits 70, 79 and 89 were entered in cells B3 to B5. The Bin array always requires one additional blank cell, B6 in our case.

Cells B7 to B10 (one more than the limits) were used for the results. Cell B7 was used for the formula. First the formula =FREQUENCY(A3:A11;B3:B5) was entered. Then, F2 followed by CTRL+Shift+Enter were pressed to indicate that we are entering an array formula.

The result is given in cells B7 to B10. It means that the frequency is as under:

Less than or equal to 70 : 1

71 to 79: 2

80 to under 89: 4

90 and above: 2

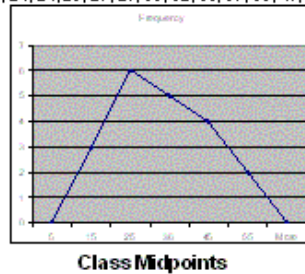
**FREQUENCY POLYGONS**

Numerical data can be represented in the form of Frequency Polygons after calculation of frequency for each interval. A typical frequency polygon is shown in the slide below.

**GRAPHING NUMERICAL DATA:  
THE FREQUENCY POLYGON**

Data in ordered array

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

**CUMULATIVE FREQUENCY**

Relative frequency can be converted into cumulative frequency by adding the current frequency to the previous total. In the slide below, the first interval has the relative as well as cumulative frequency as 3. In the next interval the relative frequency was 6. It was added to the previous value to arrive at 9 as cumulative frequency for interval 20 to 30. What it really means is that 9 values are equal to or less than 30. Similarly, the other cumulative frequencies were calculated. The total cumulative frequency 20 is the total number of observations.

Percent Cumulative frequency is calculated by dividing the cumulative frequency by the total number of observations and multiplying by 100. For the first interval the % cumulative frequency is  $3/20 \times 100 = 15\%$ . Similarly other values were calculated.

**TABULATING NUMERICAL  
DATA:**

**CUMULATIVE FREQUENCY**

Data in ordered array

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Class	Cumulative Frequency	Cumulative % Frequency
10 but under 20	3	15
20 but under 30	9	45
30 but under 40	14	70
40 but under 50	18	90
50 but under 60	20	100

**CUMULATIVE % POLYGON-OGIVE**

From the % cumulative frequency polygon that starts from the first limit (not mid point as in the case of relative frequency polygons) can be drawn. Such a polygon is called

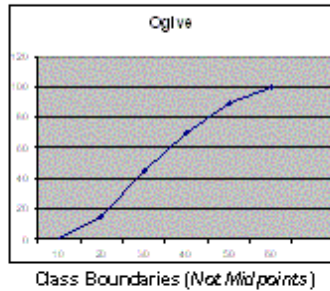


Ogive. The maximum value in an Ogive is always 100%. Ogives are determining cumulative frequencies at different values (not limits).

## GRAPHING NUMERICAL DATA:

### THE OGIVE (CUMULATIVE % POLYGON) Data in ordered array

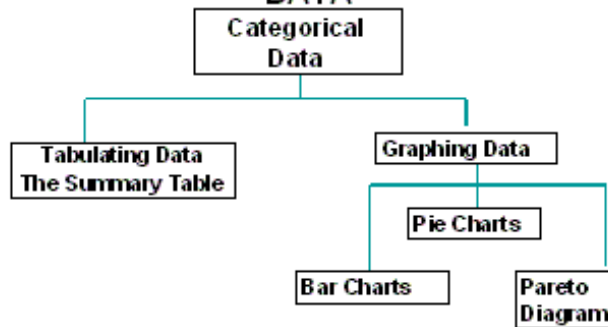
12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58



### TABULATING AND GRAPHING UNIVARIATE DATA

Univariate data (one variable) can be tabulated in Summary form or in graphical form. Three types of charts, namely, Bar Charts, Pie Charts or Pareto Diagrams can be prepared.

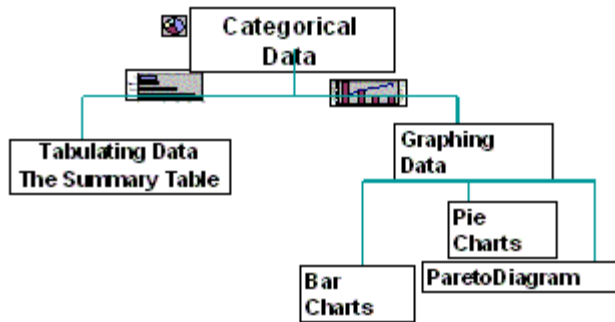
### TABULATING AND GRAPHING CATEGORICAL DATA: UNIVARIATE DATA



### SUMMARY TABLE

A typical Summary Table for an investor's portfolio is given in the slide. The variables such as stocks etc. are the categories. The table shows to amount and percentage.

## GRAPHING CATEGORICAL DATA: UNIVARIATE DATA

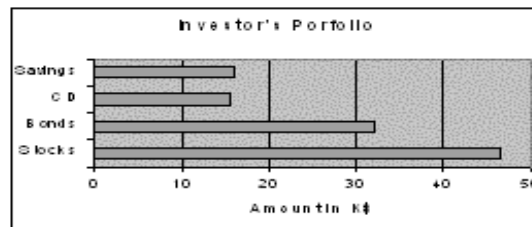


A typical Summary Table for an investor's portfolio is given in the slide. The variables such as stocks etc. are the categories. The table shows to amount and percentage.

### BAR CHART

The data of Investor's portfolio can be shown in the form of Bar Chart as shown below. This chart was prepared using EXCEL Chart Wizard. The Wizard makes it very simple to prepare such graphs. You must practice with the Chart Wizard to prepare different types of graphs.

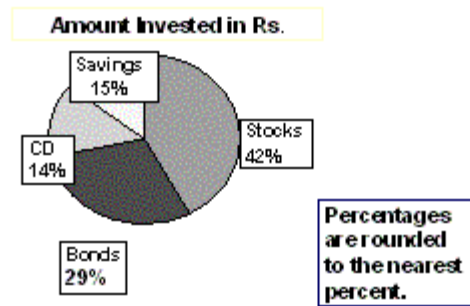
### BAR CHART (FOR AN INVESTOR'S PORTFOLIO)



### PIE CHARTS

Pie Charts are very useful charts to show percentage distribution. These charts are made with the help of Chart Wizard. You may notice how Stocks and bonds stand out.

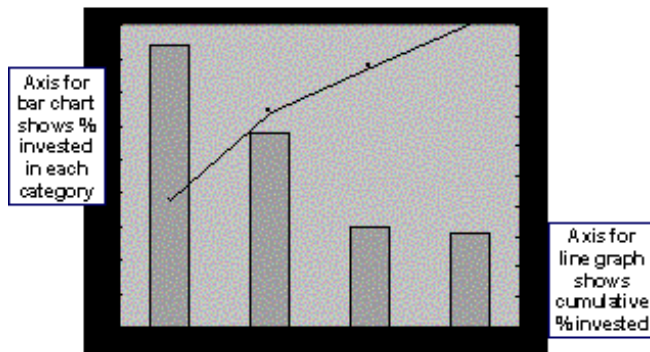
## PIE CHART (FOR AN INVESTOR'S PORTFOLIO)



### PARETO DIAGRAMS

A Pareto diagram is a cumulative distribution with the first value as first relative frequency, in this case 42%. The point is drawn in the middle of bar for the first category stocks. Next the category Bonds was added. The total is 71%. Next the savings 15% were added to 71% to obtain cumulative frequency 86%. Adding the 14% for CD gives 100%. Thus, the Pareto diagram gives both relative and cumulative frequency.

## PARETO DIAGRAM



### CONTINGENCY TABLES

Another form of presentation of data is the contingency table. An example is shown in the slide below. The table shows a comparison of three investors along with their combined total investment.

## TABULATING CATEGORICAL DATA: BIVARIATE DATA

### Contingency Table

Investment in Thousands of Rupees

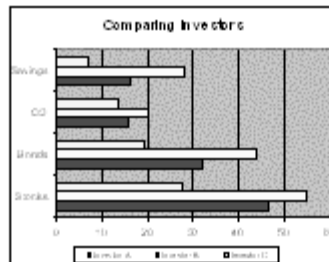
Investment Category	Investor A	Investor B	Investor C	Total
Stocks	46.5	55	27.5	129
Bonds	32	44	19	95
CD	15.5	20	13.5	49
Savings	16	28	7	51
Total	110	147	67	324

### SIDE BY SIDE CHARTS

The same investor data can be shown in the form of side by side charts where different colours were used to differentiate the investors. This graph is a complete representation of the contingency table.

## GRAPHING CATEGORICAL DATA: BIVARIATE DATA

Side by  
Side  
Chart



### GEOMETRIC MEAN

Geometric mean is defined as the root of product of individual values. Typical syntax is as under:

$$G = (x_1 \cdot x_2 \cdot x_3 \cdot \dots \cdot x_n)^{1/n}$$

#### Example

Find GM of 130, 140, 160

$$\begin{aligned} GM &= (130 \cdot 140 \cdot 160)^{1/3} \\ &= 142.8 \end{aligned}$$

### HARMONIC MEAN

Harmonic mean is defined as under:

$$\begin{aligned} HM &= n / (1/x_1 + 1/x_2 + \dots + 1/x_n) \\ &= n / \text{Sum}(1/x_i) \end{aligned}$$

**Example**

Find HM of 10, 8, 6

$$HM = 3 / (1/10 + 1/8 + 1/6)$$

$$= 7.66$$

**QUARTILES**

Quartiles divide data into 4 equal parts

**Syntax**

1st Quartile  $Q1 = (n+1)/4$

2nd Quartile  $Q2 = 2(n+1)/4$

3rd Quartile  $Q3 = 3(n+1)/4$

**Grouped data**

$$Q_i = \text{ith Quartile} = l + h/f[\text{Sum } f/4 * i - cf]$$

l = lower boundary

h = width of CI

cf = cumulative frequency

**DECILES**

Deciles divide data into 10 equal parts

**Syntax**

1st Decile  $D1 = (n+1)/10$

2nd Decile  $D2 = 2(n+1)/10$

9th Decile  $D9 = 9(n+1)/10$

**Grouped data**

$$Q_i = \text{ith Decile } (i=1,2,..,9) = l + h/f[\text{Sum } f/10 * i - cf]$$

l = lower boundary

h = width of CI

cf = cumulative frequency

**PERCENTILES**

Percentiles divide data into 100 equal parts

**Syntax**

1st Percentile  $P1 = (n+1)/100$

2nd Decile  $D2 = 2(n+1)/100$

99th Decile  $D9 = 99(n+1)/100$

**Grouped data**

$$Q_i = \text{ith Decile } (i=1,2,..,9) = l + h/f[\text{Sum } f/100 * i - cf]$$

l = lower boundary

h = width of CI

cf = cumulative frequency

**EMPIRICAL RELATIONSHIPS**

Symmetrical Distribution

**mean = median = mode**

Positively Skewed Distribution

(Tilted to left)

mean &gt; median &gt; mode

Negatively Skewed Distribution

mode &gt; median &gt; mean

(Tilted to right)

Moderately Skewed and Unimodal Distribution

Mean – Mode = 3(Mean – Median)Example

mode = 15, mean = 18, median = ?  
 Median =  $\frac{1}{3}[\text{mode} + 2 \text{ mean}]$   
 =  $\frac{1}{3}[15 + 2(18)]$   
 =  $\frac{[15+36]}{3} = \frac{51}{3} = 17$

### **MODIFIED MEANS TRIMMED MEAN**

Remove all observations below 1st quartile and above 3<sup>rd</sup> Quartile

### **Winsorized MEAN**

Replace each observation below first quartile with value of first quartile

Replace each observation above the third quartile with value of 3<sup>rd</sup> quartile

### **TRIMMED AND WINSORIZED MEAN**

#### **Example**

Find trimmed and winsorized mean.

9.1, 9.1, 9.2, 9.3, 9.2, 9.2

Array the data

9.1, 9.2, 9.2, 9.2, 9.2, 9.3, 9.9

$Q1 = \frac{(6+1)}{4} = 1.75$  (2<sup>nd</sup> value) = 9.2

$Q3 = \frac{3(6+1)}{4} = 5.25$  (6<sup>th</sup> value) = 9.3

$TM = \frac{(9.2+9.2+9.2+9.2+9.3)}{5} = 9.22$

$WM = \frac{(9.2+ 9.2+9.2+9.2+9.2+9.3+9.3)}{7}$

### **DISPERSION OF DATA**

#### **Definition**

The degree to which numerical data tend to spread about an average is called the dispersion of data

### **TYPES OF MEASURES OF DISPERSION**

Absolute measures

Relative measures (coefficients)

### **DISPERSION OF DATA**

*Types Of Absolute Measures:*

- **Range**
- **Quartile Deviation**
- **Mean Deviation**
- **Standard Deviation or Variance**

*Types Of Relative Measures*

- **Coefficient of Range**
- **Coefficient of Quartile Deviation**
- **Coefficient of Mean Deviation**
- **Coefficient of Variation**

**LECTURE 27**  
**STATISTICAL REPRESENTATION**  
**MEASURES OF DISPERSION AND SKEWNESS**  
**PART 2**

**OBJECTIVES**

The objectives of the lecture are to learn about:

- Review Lecture 26
- Measures of Dispersion and Skewness

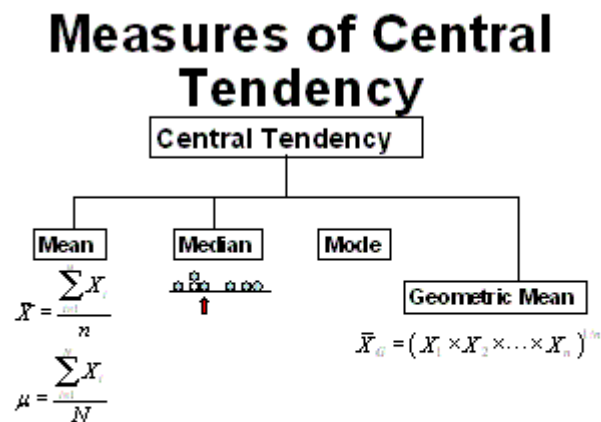
**MEASURES OF CENTRAL TENDENCY, VARIATION AND SHAPE FOR A SAMPLE**

There are many different measures of central tendency as discussed in the last lecture handout. These include:

- Mean, Median, Mode, Midrange, Quartiles, Midhinge
- Range, Interquartile Range
- Variance, Standard Deviation, Coefficient of Variation
- Right-skewed, Left-skewed, Symmetrical Distributions
- Measures of Central Tendency, Variation and Shape Exploratory Data Analysis
- Five-Number Summary
- Box-and-Whisker Plot
- Proper Descriptive Summarisation
- Exploring Ethical Issues
- Coefficient of Correlation

**MEANS**

The most common measure of central tendency is the mean. The slide below shows the Mean (Arithmetic), Median, Mode and Geometric mean. Another mean not shown is the Harmonic mean. Each of these has its own significance and application. The mean is the arithmetic mean and represents the overall average. The median divides data in two equal parts. Mode is the most common value. Geometric mean is used in compounding such as investments that are accumulated over a period of time. Harmonic mean is the mean of inverse values. Each has its own utility. The slide shows the formulas for mean and geometric mean.



**THE MEAN**

The formula for Arithmetic Mean is given in the slide. It is the sum of all values divided by the number. In the case of mean of a sample, the number  $n$  is the total sample size.

When the sample data is to be used for estimating the value of mean, then the number is reduced by 1 to improve the estimate. In reality this will be a slight overestimation of the population mean. This is done to avoid errors in estimation based on sample data that may not be truly represented of the population.

**The Mean (Arithmetic Average)**

The Arithmetic Average of data values:

$$\bar{X} = \sum_{i=1}^n X_i / n = \frac{X_1 + X_2 + \dots + X_n}{n}$$

Sample Mean Sample Size

$$\mu = \sum_{i=1}^N X_i / N = \frac{X_1 + X_2 + \dots + X_N}{N}$$

Population Mean Population Size

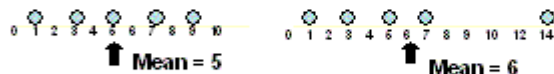
**EXTREME VALUES**

An important point to remember is that arithmetic mean is affected by extreme values. In the following slide mean of 5 values 1, 3, 5, 7 and 9 is 5. In the second case where the data values are 1, 3, 6, 7 and 14, the value 14 is an outlier as it is considerably different from the other values. In this case the mean is 6. In other words the mean increased by 1 or about 20% due to the outlier. While preparing data for mean, it is important to spot and eliminate outlier

**The Mean**

The Most Common Measure of Central Tendency

Affected by Extreme Values (Outliers)



values.



**THE MEDIAN**

The Median is derived after ordering the array in ascending order. If the number of

## The Median

Important Measure of Central  
Tendency

In an ordered array, the median is the  
"middle" number.

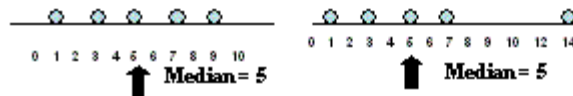
If  $n$  is *odd*, the median is the *middle*  
*number*

If  $n$  is *even*, the median is the *average of*  
*the 2 middle numbers*

*observations is odd, it is the middle value otherwise it is the the average of the the two middle values. It is not affected by extreme values.*

## The Median

*Not Affected by Extreme Values*

**THE MODE**

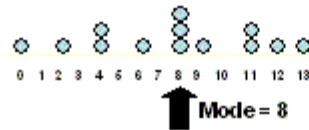
The mode is the value that occurs most frequently. In the example shown on the slide, 8 is the most frequently occurring value. Hence the mode is 8. Mode is also not affected by extreme values.

## The Mode

A Measure of Central Tendency

Value that Occurs Most Often

*Not Affected by Extreme Values*



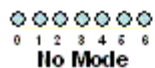
An important point about Mode is that there may not be a Mode at all (no value is occurring frequently). There may be more than one mode. The mode can be used for numerical or categorical data. The slide shows two examples where there is no mode or there are two modes.

## The Mode

There May Not be a Mode

There May be Several Modes

Used for Either Numerical or  
Categorical Data



### RANGE

Another measure of dispersion of data is the Range. It is the difference between the largest and smallest value. The slides shows an example where the value of range was calculated as 31.

## DISPERSION OF DATA

### Range

$R = \text{Largest} - \text{Smallest Value}$

### Example

Find range:

31, 26, 15, 43, 19, 27, 22, 12, 36, 33, 30, 24, 20

Largest value = 43

Smallest = 12

Range =  $43 - 12 = 31$

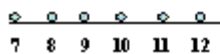
## The Range

Measure of Variation Difference  
Between Largest & Smallest  
Observations.

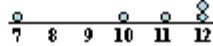
Range =

Ignores How Data Are Distributed:

$$\text{Range} = 12 - 7 = 5$$



$$\text{Range} = 12 - 7 = 5$$



### MIDRANGE

Midrange is the average of smallest and largest value. In other words it is half of a range. Midrange is affected by extreme values as it is based on smallest and largest values

## Midrange

A Measure of Central Tendency

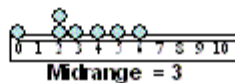
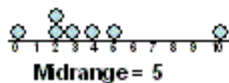
Average of Smallest and Largest

Observation:

$$\text{Midrange} = \frac{X_{\text{largest}} + X_{\text{smallest}}}{2}$$

## Midrange

Affected by Extreme Value



### QUARTILES

Quartiles are not exclusively measures of central tendency. However, they are useful for dividing the data in 4 equal parts. In working out quartiles divide the number of data items by 4 and use it as the position of the first quartile. Multiply by 2 for the second and 3 for the 3<sup>rd</sup> quartile. Say there are 12 items. Then the position of the first quartile is  $12/4 = 3$ . Supposing there were 14 values then the first quartile would be in  $14/4 = 3.5^{\text{th}}$  position. How do you calculate the value at  $3.5^{\text{th}}$  position? Obviously, you take the difference between the 4<sup>th</sup> and 3<sup>rd</sup> value and multiply by 0.5 and add it to the 3<sup>rd</sup> value. Let the 3<sup>rd</sup> and 4<sup>th</sup> values be 5 and 7. Then the difference is 2. The 1<sup>st</sup> quartile is then  $5 + 0.5 \times 2 = 6$ . In a similar fashion you can calculate any value.

## Quartiles

Not a measure of central tendency  
Split ordered data into 4 quarters

25%	25%	25%	25%
$Q_1$	$Q_2$	$Q_3$	

Position of  $i$ -th quartile:  $Q = \frac{i(n+1)}{4}$

Data in Ordered Array: 11 12 13 16 16 17 18 21 22

Position of  $Q_1 = \frac{1(9+1)}{4} = 2.50$        $Q_1 = 12.5$

### QUARTILE DEVIATION

Quartile Deviation is the average of 1<sup>st</sup> and 3<sup>rd</sup> Quartile.

$$Q.D = (Q_3 - Q_1)/2$$

#### Example

Find Q.D

14, 10, 17, 5, 9, 20, 8, 24, 22, 13

$$Q_1 = (n+1)/4\text{th value} = (10+1)/4 = 2.75\text{th}$$

$$= 8 + 0.75(9 - 8) = 8 + 0.75 \times 1 = 8.75$$

$$Q_3 = 3(2.75) = 8.25\text{th value}$$

$$= 8\text{th value} + 0.25(9\text{th value} - 8\text{th value})$$

$$= 20 + 0.25(22 - 20) = 20.50$$

$$Q.D = (20.50 - 8.75)/2 = 5.875$$

### BOX AND WHISKER PLOTS

Box and Whisker plots show the 5 number summary:

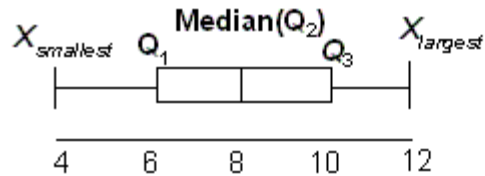
- Smallest value
- 1<sup>st</sup> Quartile ( $Q_1$ )
- Median( $Q_2$ )
- 3<sup>rd</sup> Quartile ( $Q_3$ )
- Largest value

The plots give a good idea about the shape of the distribution as detailed below. Box and whisker plots for symmetrical, left skewed and right skewed distributions are shown below.

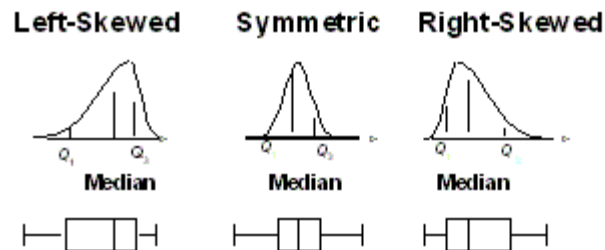
## Exploratory Data Analysis

### Box-and-whisker:

Graphical display of data using 5-number summary



### Distribution Shape & Box-and-whisker Plots



#### Data is perfectly symmetrical if:

**Distance from Q1 to Median = Distance from Median to Q3**

Distance from  $X_{\text{smallest}}$  to Q1 = Distance from Q3 to  $X_{\text{largest}}$

Median = Midhinge = Midrange

#### **Right-skewed distribution**

Median < Midhinge < Midrange  
Distance from  $X_{\text{largest}}$  to Q3 greatly exceeds distance from Q1 to  $X_{\text{smallest}}$

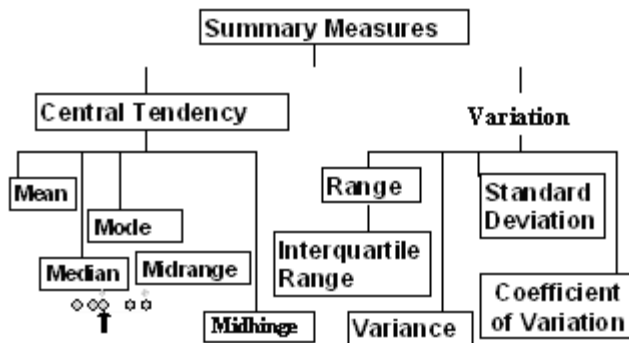
#### **Left-skewed distribution**

Median > Midhinge > Midrange  
Distance from Q1 to  $X_{\text{smallest}}$  greatly exceeds distance from  $X_{\text{largest}}$  to Q3

#### SUMMARY MEASURES

The slide shows summary of measures of central tendency and variation. In variation there are range, Interquartile range, standard deviation, variance, and coefficient of variation. The measures of central tendency have been discussed already

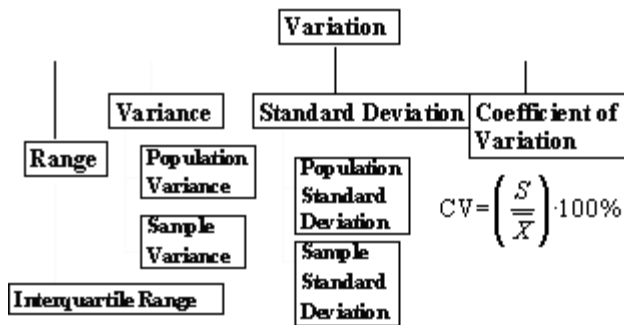
## Summary Measures



### MEASURES OF VARIATION

In measures of variation, there are the sample and population standards deviation and variance the most important measures. The coefficient of variation is the ratio of standard deviation to the mean in %.

## Measures of Variation



### INTERQUARTILE RANGE

Interquartile range is the difference between the 1st and 3<sup>rd</sup> quartile.

## Interquartile Range

Measure of Variation

Also Known as Midspread:

Spread in the Middle 50%

Difference Between Third & First  
Quartiles: Interquartile Range =

Data in Ordered Array: 11 12 13 16 16 17 17 18 21

$$Q_3 - Q_1 = 17.5 - 12.5 = 5$$

Not Affected by Extreme Values



**LECTURE 28**  
**MEASURES OF DISPERSION**  
**CORRELATION**  
**PART 1**

**OBJECTIVES**

The objectives of the lecture are to learn about:

- Review Lecture 27
- Measures of Dispersion
- Correlation

**MODULE 6**

Module 6 covers the following:

Correlation

(Lecture 28-29)

Line Fitting

(Lectures 30-31)

Time Series and Exponential Smoothing

(Lectures 32-33)

**VARIANCE**

Variance is the one of the most important measures of dispersion. Variance gives the average square of deviations from the mean. In the case of the population, the

## Variance

### Important Measure of Variation

Shows Variation About the Mean:

**For the Population:** 
$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{N}$$

**For the Sample:** 
$$s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

**For the Population: use N in the denominator.**

**For the Sample : use n - 1 in the denominator.**

sum of square of deviations is divided by N the number of values in the population. In the case of variance for the sample the number of observations less 1 is used.

**STANDARD DEVIATION**

Standard deviation is the most important and widely used measure of dispersion. The square root of square of deviations divided by the number of values for the population and number of observations less 1 gives the standard deviation.

## Standard Deviation

**Most Important Measure of Variation**

**Shows Variation About the Mean**

**Same unit of measurement as the observations**

**For the Population:** 
$$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}$$

$$s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n - 1}}$$

**For the Sample**

**For the Population: use N in the denominator.**

**For the Sample : use n - 1 in the denominator**

### COMPARING STANDARD DEVIATIONS

In many situations it becomes necessary to calculate population standard deviation (SD) on the basis of SD of the sample where n-1 is used for

#### Comparing Standard

#### Deviations

**Data :** 10 12 14 15 17 18 18  
24  
**N= 8      Mean=16**

$$s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n - 1}} = 4.2426$$

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}} = 3.9686$$

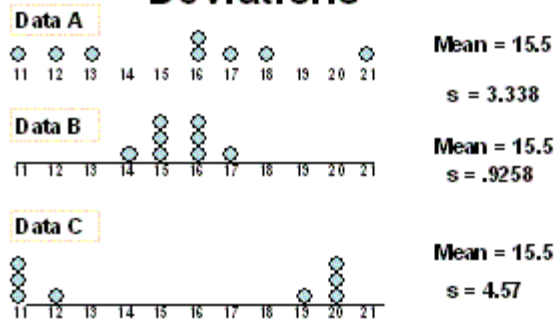
**Value for the Standard Deviation is larger for data considered as a Sample.**

division. In the slide the same data is first treated as the sample and the value of SD is 4.2426. When we treat it as the population the SD is 3.9686, which is slightly less than the SD for the sample. You can see how the sample SD will be overestimated if used for the population.

### COMPARING STANDARD DEVIATIONS

The slide shows three sets of data A, B and C. All the three datasets have the same mean 15.5 but different standard deviations (A: s=3.338; B: s=0.9258 and C: s=4.57). It is clear that SD is an important measure to understand how different sets of data differ from each other. Mean and SD together form a complete description of the central tendency of data.

## Comparing Standard Deviations



### COEFFICIENT OF VARIATION

## Comparing Coefficient of Variation

- **Stock A:** average price last year = Rs. 50
- **Standard deviation = Rs. 5**
- $CV = \left(\frac{s}{\bar{x}}\right) \cdot 100\%$  **average price last year = Rs. 100**
- **Standard deviation = Rs. 5**

**Coefficient of Variation:**

**Stock A: CV = 10%**

**Stock B: CV = 5%**

Coefficient of variation (CV) shows the dispersion of the standard deviation about the mean. In the slide you see two stocks A and B with CV=10% and 5% respectively. This comparison shows that in the case of stock A there was a much greater variation in price with reference to the mean.

### MEAN DEVIATION ABOUT MEAN

Other useful measures are Deviation about the Mean and median. The formulas for normal or grouped data are as follows:

#### **Mean Deviation About Mean – Normal data**

$$MD (\text{mean}) = \text{Sum } (x_i - \text{mean})/n$$

#### **For Grouped data – Grouped data**

$$MD (\text{mean}) = \text{Sum } f_i (x_i - \text{mean})/\text{Sum } f_i$$

#### **Mean Deviation About Median – Normal data**

$$MD (\text{median}) = \text{Sum } (x_i - \text{median})/n$$

#### **Mean Deviation About Median – Grouped data**

$$MD (\text{median}) = \text{Sum } f_i (x_i - \text{median})/\text{Sum } f_i$$

**CORRELATION**

In regression Analysis, we shall encounter different types of regression models. One of the main functions of regression analysis is determining the simple linear regression equation. What are the different Measures of variation in regression and correlation? What are the Assumptions of regression and correlation? What is Residual analysis? How do we make

Inferences about the slope? How can you estimate predicted values? What are the Pitfalls in regression? What are the ethical issues?

Correlation is measuring the strength of the association.

An important point in regression analysis is the purpose of the analysis.

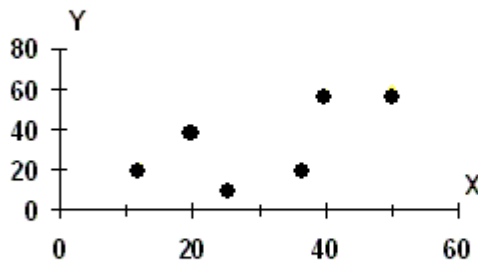
**SCATTER DIAGRAM**

The first step in regression analysis is to plot the values of the dependent and independent variable in the form of a scatter diagram as shown below. The form of the scatter of the points indicates whether there is any degree of association between them. In the scatter diagram below you can see that there seems to be a fairly distinct correlation between the two variables. It appears as if the points were located around a straight line.

Once the degree of association is established, it makes sense to proceed further and carry out regression analysis using a regression model.

## The Scatter Diagram

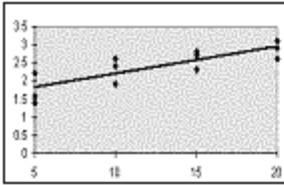
Plot of all  $(X_i, Y_i)$  pairs

**Types of Regression Models**

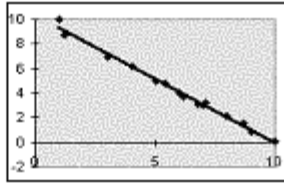
There are two types of linear regression models as shown in the slide below. These are positive and negative linear relationships. In the positive relationship, the value of the dependent variable increases as the value of the independent variable increases. In the case of negative linear relationship, the value of the dependent variable decreases with increase in the value of independent variable.

# Types of Regression Models

**Positive Linear Relationship**



**Negative Linear Relationship**



**LECTURE 29**  
**MEASURES OF DISPERSION**  
**CORRELATION**  
**PART 2**

**OBJECTIVES**

The objectives of the lecture are to learn about:

- Review Lecture 28
- Correlation

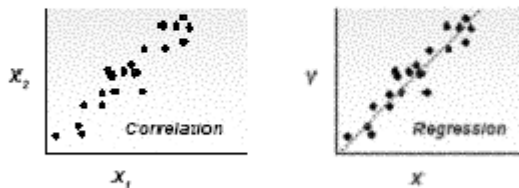
**CORRELATION****When do we use correlation?**

It will be used when we wish to establish whether there is a degree of association between two variables. If this association is established, then it makes sense to proceed further with regression analysis. Regression analysis determines the constants of the regression. You can not make any predictions with results of correlation analysis. Predictions are based on regression equations.

## CORRELATION

### When do we use correlation?

**Do use it to determine the strength of association between two variables**  
**Do not use it if you want to predict the value of X given Y, or vice versa**

**SIMPLE LINEAR CORRELATION VERSUS SIMPLE LINEAR REGRESSION**

The calculations for linear correlation analysis and regression analysis are the same.

In correlation analysis, one must sample randomly both X and Y.

Correlation deals with the association (importance) between variables whereas

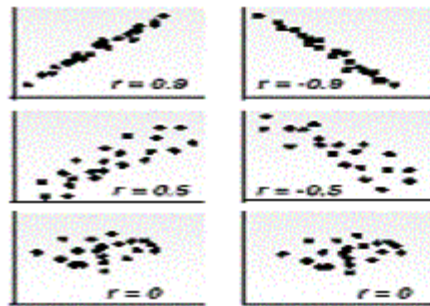
Regression deals with prediction (intensity).

The slide shows three types of correlation for both positive and negative linear relationships. In the first figure ( $r = .9$ ), the data points are practically in a straight line. This kind of association or correlation is near perfect. This applies to negative correlation also.

The graphs where  $r = 0.5$ , the points are more scattered, there is a clear association but this association is not very pronounced.

In graphs where  $r = 0$ , there is no association between variables.

## TYPICAL CORRELATION y vs x



5

### CORRELATION COEFFICIENT

#### For calculation of correlation coefficient:

1. A standardised transform of the covariance ( $s_{xy}$ ) is calculated by dividing it by the product of the standard deviations of X ( $s_x$ ) & Y ( $s_y$ ).
2. It is called the population correlation coefficient is defined as:

$$r = s_{xy}/s_x s_y$$

#### Properties

1. For the population:  
-1 = r = +1
2.  $r = 0$  means no linear relationship  
 $r = -1$  perfect negative relationship  
 $r = +1$  perfect positive relationship

#### Important points about Correlation

- $r$  always lies between  $-1$  and  $1$
- $r^2$  is the coefficient of determination, which measures the proportion of the variance in X1 (or X2) "explained" by variation in X2 or X1
- $r$  always lies between  $-1$  and  $1$ .

#### Strength of association

- It measures the strength of the association between X & Y on a scale from  $-1$ , through  $0$ , up to  $+1$ .
- This gives an intuitive feel for how strong the association is, regardless of the original units of X & Y.
- Near  $+1$  or  $-1$  means very strong.
- Near  $0$  means very weak.

#### Warning

- Existence of a high correlation does not mean there is causation, which means that there may be a correlation but it does not make things happen because of that.
- There can exist spurious correlations. And correlations can arise because of the action of a third unmeasured or unknown variable. In many situations correlation

can be high without any solid foundation.

### **MEASURING THE STRENGTH OF A CORRELATION**

**Test statistic** is the product-moment correlation coefficient  $r$

$$r = \frac{\text{covariance}(x,y)}{s(x) \cdot s(y)}$$

$$s(x) \cdot s(y)$$

$$\text{covariance}(x,y) = \frac{\sum[(x-x_m)(y-y_m)]}{n}$$

$$s(x) = \left[ \frac{\sum(x^2)}{n} - (x_m)^2 \right]^{1/2}$$

$$s(y) = \left[ \frac{\sum(y^2)}{n} - (y_m)^2 \right]^{1/2}$$

### **EXCEL Tools**

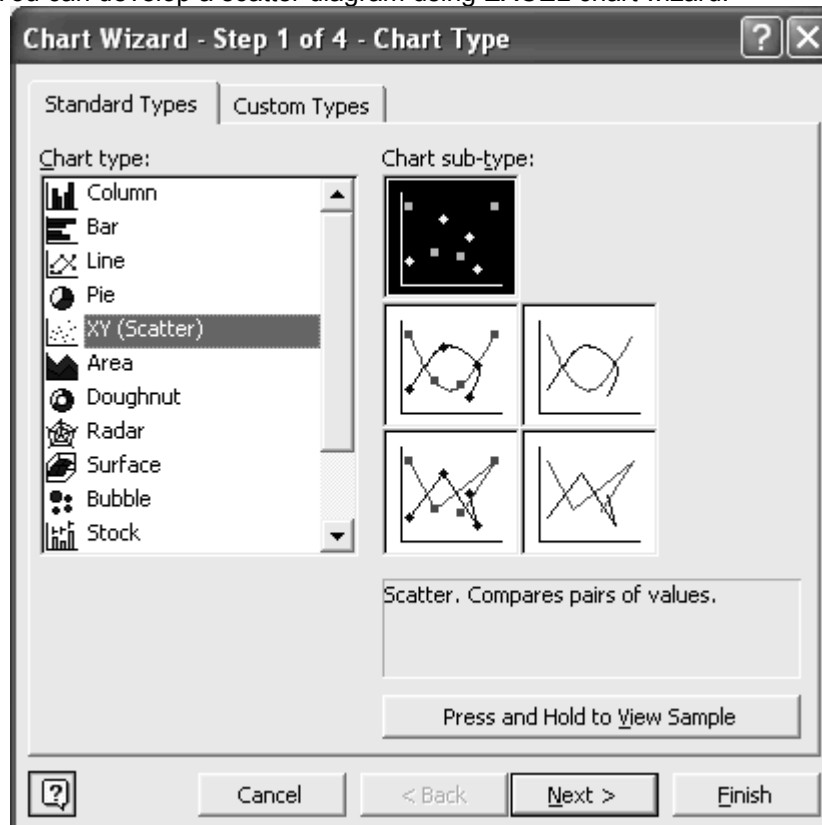
- For summary of sample statistics, use:  
**Tools / Data Analysis / Descriptive Statistics**
- For individual sample statistics, use:  
**Insert / Function / Statistical**  
and **select the function** you need

### **EXCEL Functions**

- In **EXCEL**, use the **CORREL** function to calculate correlations
- The correlation coefficient is also given on the output from **TOOLS, DATA ANALYSIS, CORRELATION or REGRESSION**

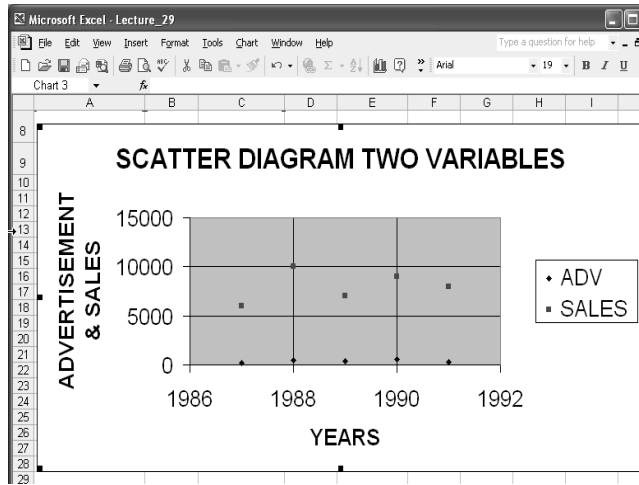
### **Scatter Diagram Two Variables**

You can develop a scatter diagram using EXCEL chart wizard.



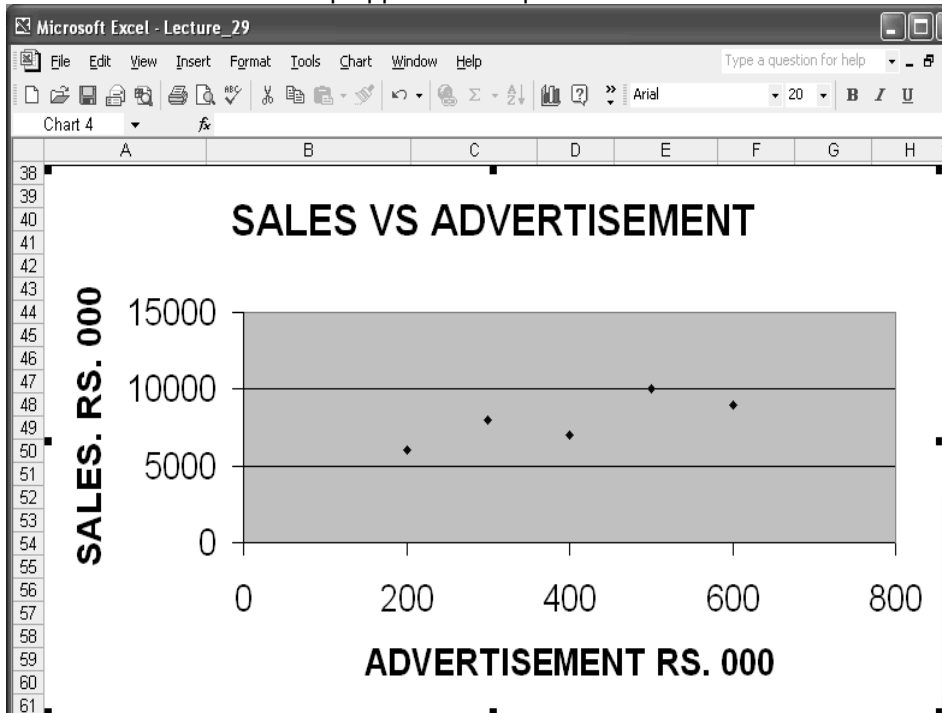


The slide shows a scatter diagram of Advertisement and Sales over the years. The graph was made using EXCEL chart Wizard. As you can see one cannot draw any conclusions about the degree of association between advertising from this graph.



**SALES VERSUS ADVERTISEMENT**

The scatter diagram for sale versus advertisement shows a fairly high degree of association. The relationship appears to be positive and linear.



**CORRELATION COEFFICIENT USING EXCEL**

Correlation Coefficient for correlation between two streams of data was calculated using the formula  $Cov(x,y)/S_x.S_y$  as given above.

The data for variable x was entered in cells A67 to A71. Data for variable y was entered in cells B67 to B71. Calculations for square of x, square of y, product of x and y,  $X_m$ ,  $Y_m$  and  $cov(x,y)$  were made in columns C, D, E, F and G respectively. Other calculations were made as follows:

Cell A72: Sum of x (=SUM(A67:A71))  
 Cell B72: Sum of y (=SUM(B67:B71))  
 Cell C72: Sum of square of x (=SUM(C67:C71))  
 Cell D72: Sum of square of y (=SUM(D67:D71))  
 Cell E72: Sum of product of x and y (=SUM(E67:E71))  
 Cell F72: Mean of x (=A72/5), where 5 is the number of observations  
 Cell G72: Mean of y (=B72/5), where 5 is the number of observations  
 Cell F73:  $S_x$  (=SQRT(C72/5-F72\*F72))  
 Cell G73:  $S_y$  (=SQRT(D72/5-G72\*G72))  
 Cell H73: Cov(x,y) (=E72/5-F72\*G72)  
 Cell H74: Correlation coefficient (=H73/(F73\*G73))

The above formulas are in line with formulas described earlier.

	A	B	C	D	E	F	G	H	I	
64	<b>CORRELATION</b>									
65										
66	<b>X</b>	<b>Y</b>	<b>X^2</b>	<b>Y^2</b>	<b>XY</b>	<b>Xm</b>	<b>Ym</b>	<b>Cov(X,Y)</b>		
67	2	60	4	3600	120					
68	5	100	25	10000	500					
69	4	70	16	4900	280					
70	6	90	36	8100	540					
71	3	80	9	6400	240					
72	<b>20</b>	<b>400</b>	<b>90</b>	<b>33000</b>	<b>1680</b>	<b>4</b>	<b>80</b>			
73	<b>Standard deviation S =</b>					<b>1.4</b>	<b>14.14</b>	<b>16</b>		
74						<b>0.8 r=</b>	<b>=H73/(F73*G73)</b>			
75										

- **CORREL**  
Returns the correlation coefficient of the array1 and array2 cell ranges. Use the correlation coefficient to determine the relationship between two properties. For example, you can examine the relationship between a location's average temperature and the use of air conditioners.

#### Syntax

**CORREL(array1,array2)**

Array1 is a cell range of values.

Array2 is a second cell range of values.

#### Remarks

- The arguments must be numbers, or they must be names, arrays, or references that contain numbers.
- If an array or reference argument contains text, logical values, or empty cells, those values are ignored; however, cells with the value zero are included.
- If array1 and array2 have a different number of data points, CORREL returns the #N/A error value.
- If either array1 or array2 is empty, or if s (the standard deviation) of their values equals zero, CORREL returns the #DIV/0! error value.
- The equation for the correlation coefficient is:

- The equation for the correlation coefficient is:

$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$$

where:

$$-1 \leq \rho_{XY} \leq 1$$

and:

$$\text{Cov}(X,Y) = \frac{1}{n} \sum_{j=1}^n (x_j - \mu_X)(y_j - \mu_Y)$$

### EXCEL Calculation

The X and Y arrays are in cells A79 to A83 and B79 to B83 respectively. The formula for correlation coefficient was entered in cell D84 as =CORREL(A79:A83;B79:B83). The value or r (0.8) is shown in cell C86.

The screenshot shows a Microsoft Excel spreadsheet titled "Lecture\_29". The spreadsheet contains the following data:

	A	B	C	D	E	F	G
75							
76	<b>CORREL(array1,array2)</b>						
77							
78	<b>X</b>	<b>Y</b>					
79	<b>2</b>	<b>60</b>					
80	<b>5</b>	<b>100</b>					
81	<b>4</b>	<b>70</b>					
82	<b>6</b>	<b>90</b>					
83	<b>3</b>	<b>80</b>					
84	<b>20</b>	<b>400</b>	<b>r=</b>	<b>=CORREL(A79:A83;</b>			
85				<b>B79:B83)</b>			
86			<b>0.8</b>				

### SAMPLE CORRELATION

The unknown value of r is estimated by the sample coefficient.

## Sample Correlation

The unknown value of  $r$  is estimated by the sample coefficient:  $r = s_{xy} / s_x s_y$

$$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

18

---

### EASY CALCULATION FORMULA

A simplified formula for the variance is given in the following slide.

#### Easy calculation formula

- For calculation, don't use the previous (definition) formula, but instead make a column of values of the squares of  $X$  then use the mean

$$s^2 = \frac{1}{n-1} \left[ \sum_{i=1}^n x_i^2 - n\bar{x}^2 \right]$$

20

### STANDARD DEVIATION

In practice the numerical statistic used to describe the “spread” of a sample is the square root of the variance which is called the “**standard deviation**”.

- $s = s^2 \text{ } ^{1/2}$  (for populations:  $\sigma = \sigma^2 \text{ } ^{1/2}$ )
- we say : “ $s$ ” estimates “ $\sigma$ ”
- If “range” = (max. value - min. value) then  $s = (\text{range}/4)$  approximately

#### Rules of thumb

- If the data are reasonably symmetric, and cluster near the mean:

- **About 70% of observations are included in an interval 1 standard deviation (s.d) either side of the mean**  
**about 95% .....2 s.d.'s.....**  
**about 99.7%.....3.....**

**Population Parameters**

- Sample -->(estimates) population
  - Statistic “ “ parameter
  - $\bar{x}$  “  $\mu$
  - $s^2$  “  $\sigma^2$  “
  - $s$  “  $\sigma$  “
- Rel.freq.polygon “ prob.distribution

**LECTURE 30**  
**Measures of Dispersion**  
**LINE FITTING**  
**PART 1**

**OBJECTIVES**

The objectives of the lecture are to learn about:

- Review Lecture 29
- Line Fitting

**EXCEL SUMMARY OF SAMPLE STATISTICS**

For summary of sample statistics, use:

**Tools > Data Analysis > Descriptive Statistics**

For individual sample statistics, use:

**Insert > Function > Statistical**

and select the function you need

**EXCEL STATISTICAL ANALYSIS TOOL**

You can use **EXCEL** to perform a statistical analysis:

- On the **Tools** menu, click **Data Analysis**. If **Data Analysis** is not available, load the **Analysis ToolPak**.
- In the **Data Analysis** dialog box, click the **name of the analysis tool** you want to use, and then click **OK**.
- In the **dialog box** for the **tool** you selected, set the **analysis options you want**.
- You can use the **Help** button on the dialog box to get more information about the options.

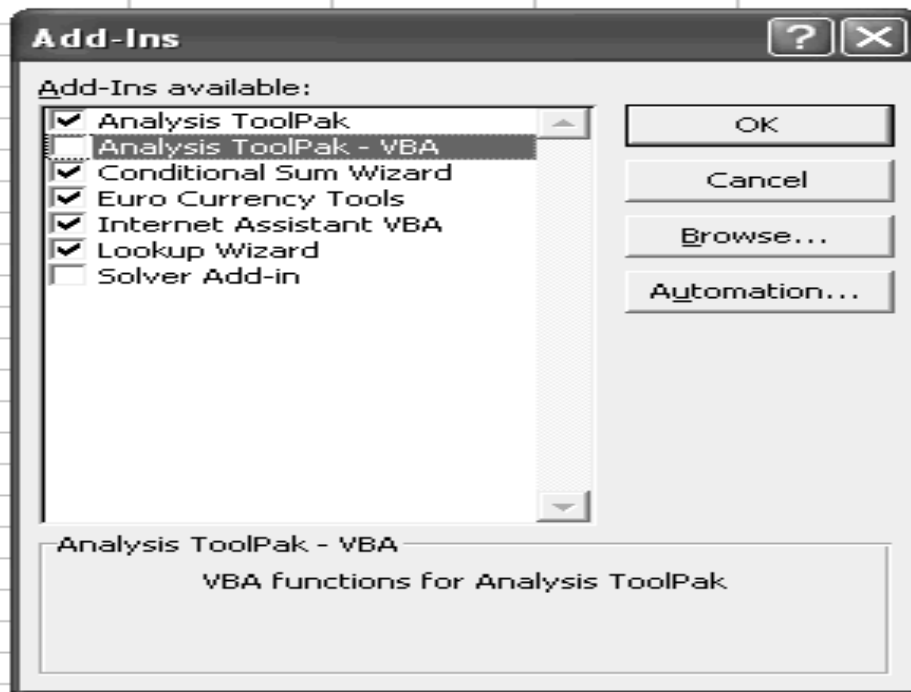
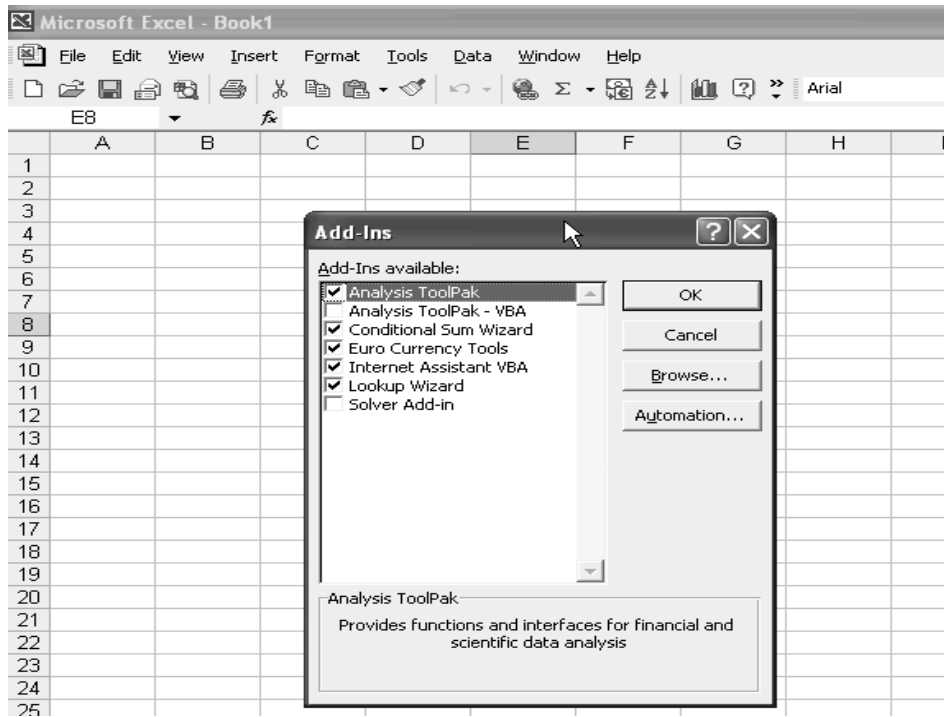
**LOAD THE ANALYSIS TOOLPAK**

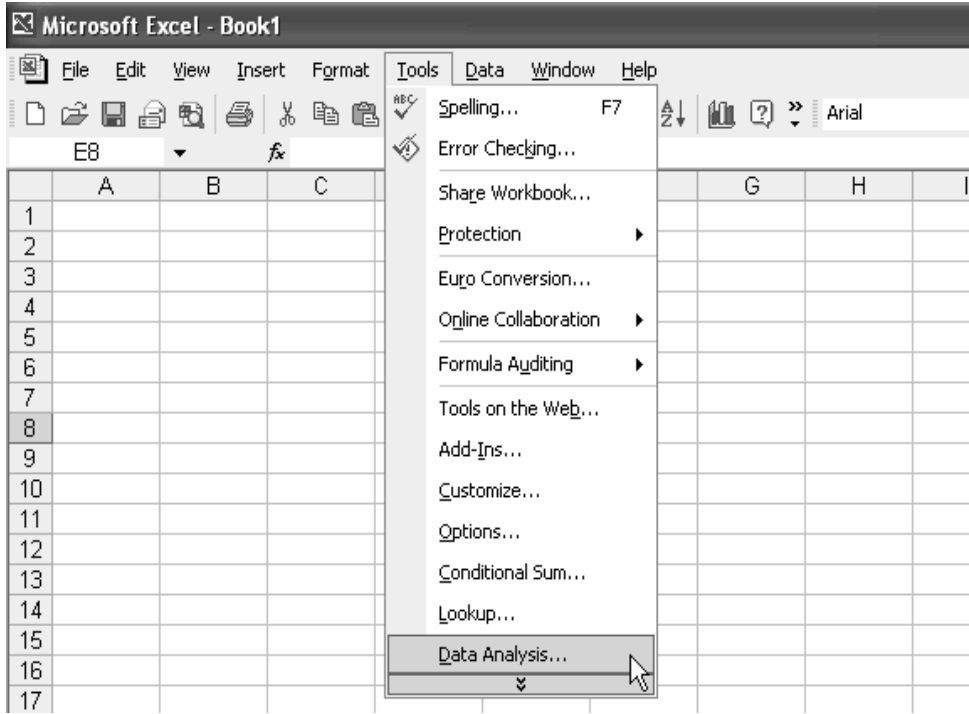
**You can load the EXCEL Analysis ToolPak as follows:**

**On the Tools menu, click Add-Ins.**

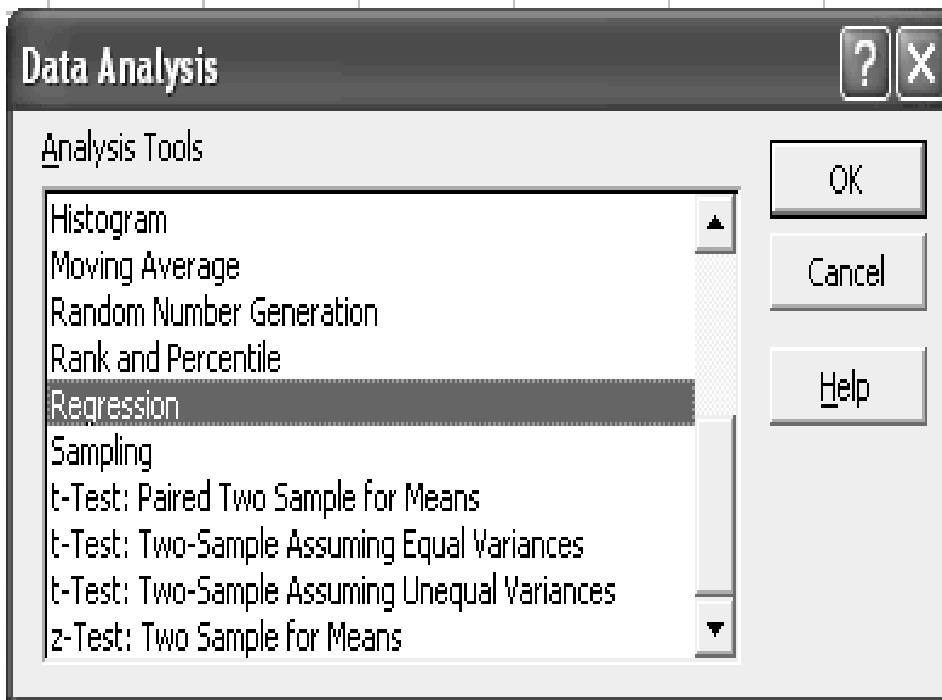
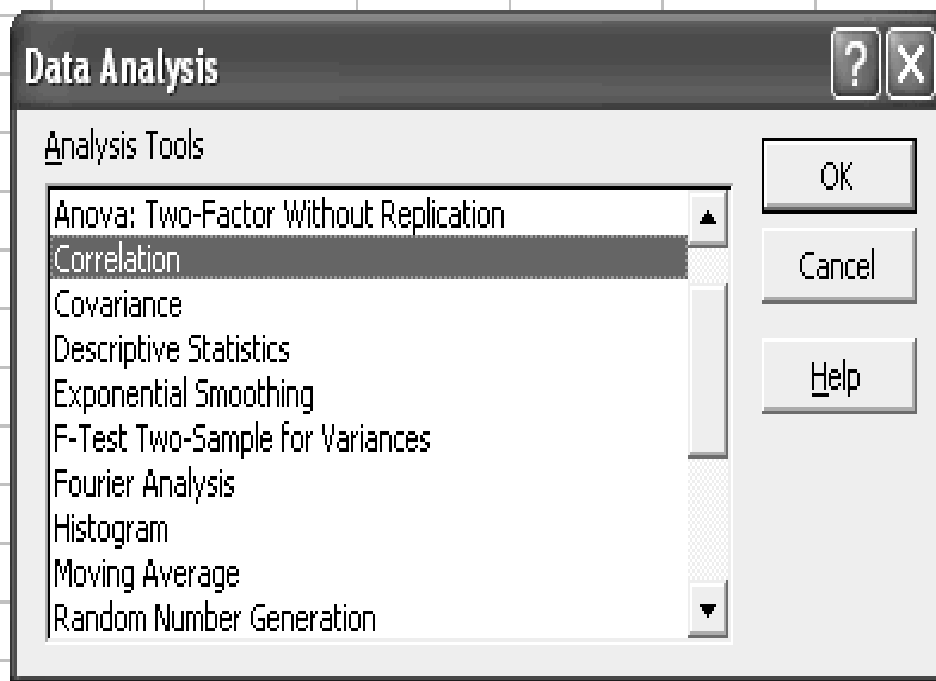
In the **Add-Ins** available **list**, select the **Analysis ToolPak** box, and then click **OK**.

If necessary, follow the instructions in the setup program









**SLOPE**

Returns the slope of the linear regression line through data points in known\_y's and known\_x's. The slope is the vertical distance divided by the horizontal distance between any two points on the line, which is the rate of change along the regression line.

**Syntax****SLOPE(known\_y's,known\_x's)**

Known\_y's is an array or cell range of numeric dependent data points.

Known\_x's is the set of independent data points.

**Remarks**

- The arguments must be numbers or names, arrays, or references that contain numbers.
- If an array or reference argument contains text, logical values, or empty cells, those values are ignored; however, cells with the value zero are included.
- If known\_y's and known\_x's are empty or have a different number of data points, SLOPE returns the #N/A error value.
- The equation for the slope of the regression line is:

$$b = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

**Example**

The known y-values and x-values were entered in cells A4 to A10 and B4 to B10 respectively. The formula =SLOPE(A4:A10;B4:B10) was entered in cell A11. The result 0.305556 is the value of slope in cell B12.

	A	B	C	D	E	F
2	<b>=SLOPE(known_y's,known_x's)</b>					
3	<b>Known y</b>	<b>Known x</b>				
4	2	6				
5	3	5				
6	9	11				
7	1	7				
8	8	5				
9	7	4				
10	5	4				
11	<b>=SLOPE(A4:A10; B4:B10)</b>					
12		<b>0.305556</b>				

### INTERCEPT

Calculates the point at which a line will intersect the y-axis by using existing x-values and y-values. The intercept point is based on a best-fit regression line plotted through the known x-values and known y-values. Use the INTERCEPT function when you want to determine the value of the dependent variable when the independent variable is 0 (zero). For example, you can use the INTERCEPT function to predict a metal's electrical resistance at 0°C when your data points were taken at room temperature and higher.

#### **Syntax**

**INTERCEPT(known\_y's,known\_x's)**

Known\_y's is the dependent set of observations or data.

Known\_x's is the independent set of observations or data.

#### **Remarks**

- The arguments should be either numbers or names, arrays, or references that contain numbers.
- If an array or reference argument contains text, logical values, or empty cells, those values are ignored; however, cells with the value zero are included.
- If known\_y's and known\_x's contain a different number of data points or contain no data points, INTERCEPT returns the #N/A error value.
- The equation for the intercept of the regression line is:

$$a = \bar{Y} - b\bar{X}$$

where the slope is calculated as:

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

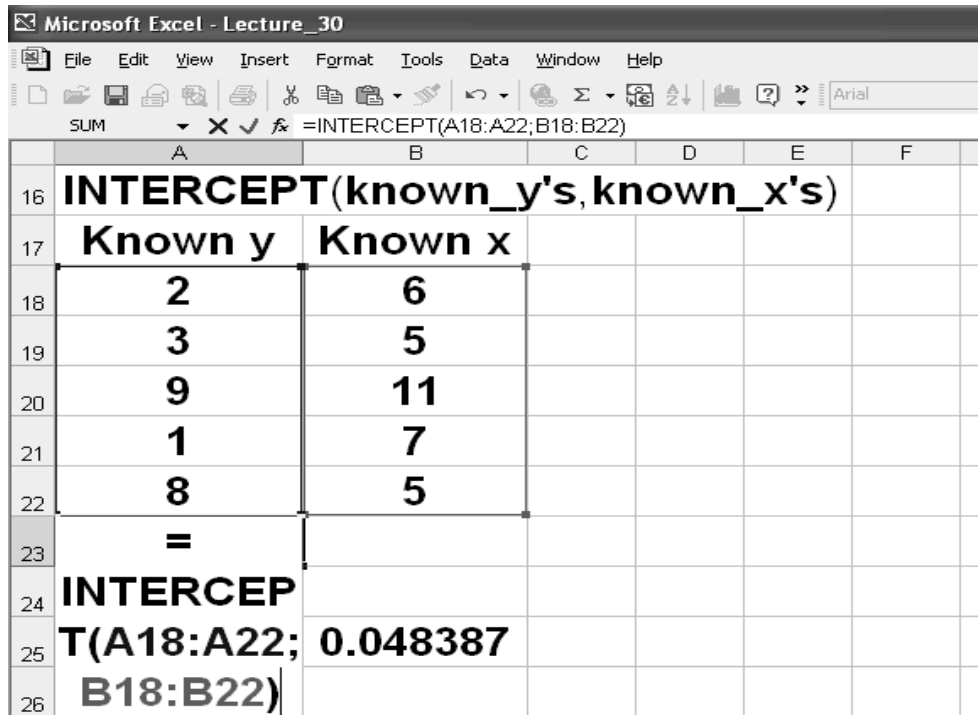
#### **Example**

The data for y-values was entered in cells A18 to A22.

The data for x-values was entered in cells B18 to B22.

The formula =INTERCEPT(A18:A22;B18:B22) was entered in cell A24.

The answer 0.048387 is shown in cell B25.



	A	B	C	D	E	F
16	<b>INTERCEPT(known_y's, known_x's)</b>					
17	<b>Known y</b>	<b>Known x</b>				
18	<b>2</b>	<b>6</b>				
19	<b>3</b>	<b>5</b>				
20	<b>9</b>	<b>11</b>				
21	<b>1</b>	<b>7</b>				
22	<b>8</b>	<b>5</b>				
23	<b>=</b>					
24	<b>INTERCEP</b>					
25	<b>T(A18:A22;</b>	<b>0.048387</b>				
26	<b>B18:B22)</b>					

**LECTURE 31**  
**LINE FITTING**  
**PART 2**

**OBJECTIVES**

The objectives of the lecture are to learn about:

- Review Lecture 30
- Line Fitting

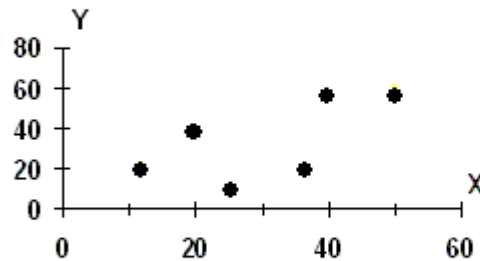
**Types of Regression Models**

There are different types of regression models. The simplest is the Simple Linear Regression Model or a relationship between variables that can be represented by a straight line equation.

To determine whether a linear relationship exists, a Scatter Diagram is developed first.

## The Scatter Diagram

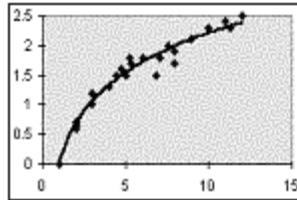
Plot of all  $(X_i, Y_i)$  pairs



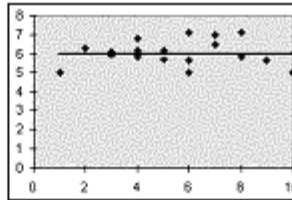
In linear regression two types of models are considered. The first one is the Population Linear Regression that represents the linear relationship between the variables of the entire population (i.e. all the data). It is quite customary to carry out sample surveys and determine linear relationship between two variables on the basis of sample data. Such regression analysis is called Sample Linear Regression.

# Types of Regression Models

Relationship NOT Linear



No Relationship



Relationship between Variables is described by a Linear Function. The change of one variable causes the other variable to change. The relationship describes the dependency of one variable on the other. The population dependent variable is  $Y_i$ . The regression equation for  $Y_i$  is shown in the slide along with explanations. The first and second terms give the population regression line. The third term is the random error.

## Population Linear Regression

Population Regression Line Is A Straight Line that Describes The Dependence of The Average Value of One Variable on The Other

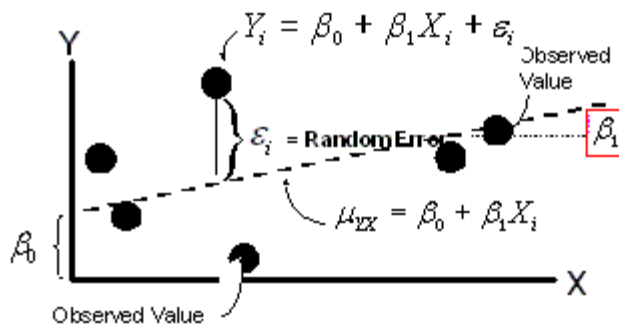
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Population Y intercept (points to  $\beta_0$ )  
 Population Slope Coefficient (points to  $\beta_1$ )  
 Population Random Error (points to  $\varepsilon_i$ )  
 Dependent (Response) Variable (points to  $Y_i$ )  
 Population Regression Line (bracketed under  $\beta_0 + \beta_1 X_i$ )  
 Independent (Explanatory) Variable (points to  $X_i$ )

The slide below shows the graphical representation of the population regression equation. It may be seen that the distance of the points from the regression line (obtained by inserting values of X in the equation) is the random error. The intercept is shown on the Y-axis.

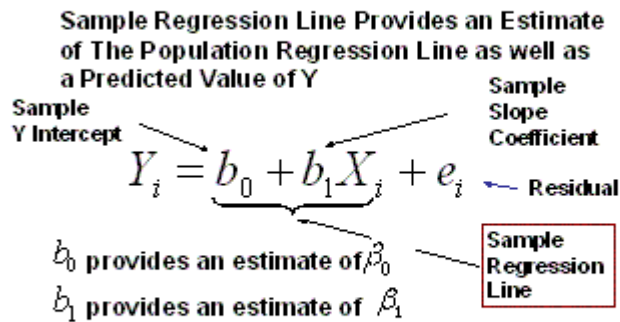
### Population Linear Regression

(continued)



The slide below shows the regression equation for the sample. Note that the intercept in this case has a notation  $b_0$ . The slope is  $b_1$ . The random error is  $e_1$ . Different notations are used to distinguish between population regression and sample regression.

### Sample Linear Regression



#### REGRESSION EQUATION

The formula for the regression equation is as under:

**Equation of Least Squares Regression line**

$$y - y_m = (r \cdot s(y) / s(x)) \cdot (x - x_m)$$

#### Example

Based on analysis of data the following values have been worked out:

$$x_m = 4;$$

$$y_m = 80;$$

$$s(x) = 2^{1/2};$$

$$s(y) = 200^{1/2};$$

$$r = 0.8$$

Find the regression equation  $Y = a + b \cdot X$

Using the formula given above:

$$y - y_m = (r \cdot s(y)/s(x)) \cdot (x - x_m)$$

$$y - 80 = \frac{0.8 \cdot 200^{1/2}}{2^{1/2}} \cdot (x - 4)$$

$$y - 80 = 8(x - 4)$$

$$y = 8x - 32 + 80$$

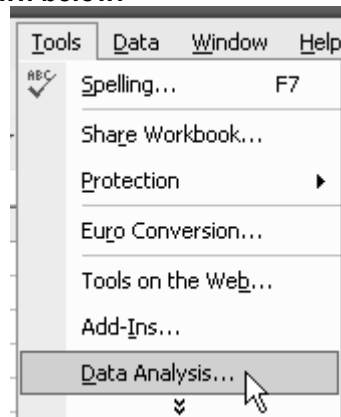
$$y = 8x + 48$$

**REGRESSION EXAMPLE 1**

Regression Analysis can be carried out easily using EXCEL Regression Tool. Let us see how it can be done. We chose to carry out regression on data given in the slide below. Y-values are 60, 100, 70, 90 and 80. X-values are 2, 5, 4, 6 and 3.

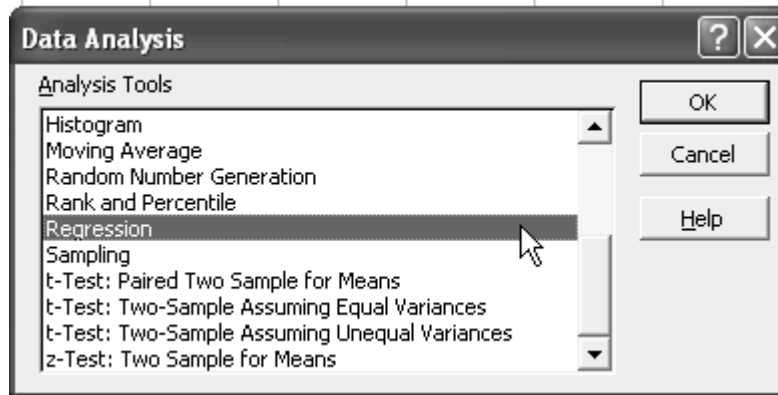
	A	B	C
37	<b>REGRESSION ANALYSIS</b>		
38	<b>EXAMPLE 1</b>		
39			
40	<b>Y</b>	<b>X</b>	
41	<b>60</b>	<b>2</b>	
42	<b>100</b>	<b>5</b>	
43	<b>70</b>	<b>4</b>	
44	<b>90</b>	<b>6</b>	
45	<b>80</b>	<b>3</b>	
46			

We start the regression analysis by going to the Tools menu and selecting the Data Analysis menu as shown below.

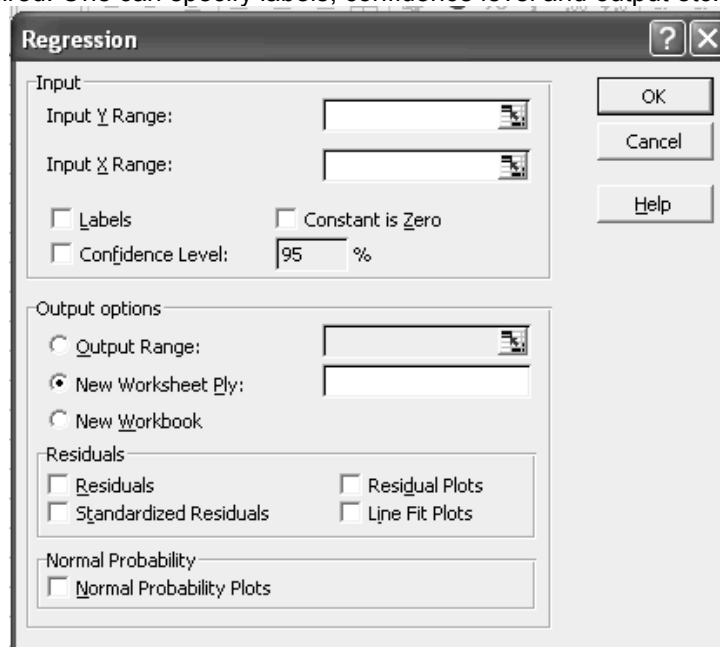


The Regression dialog box opens as shown in the following slide. You click the Regression analysis tool and then OK.

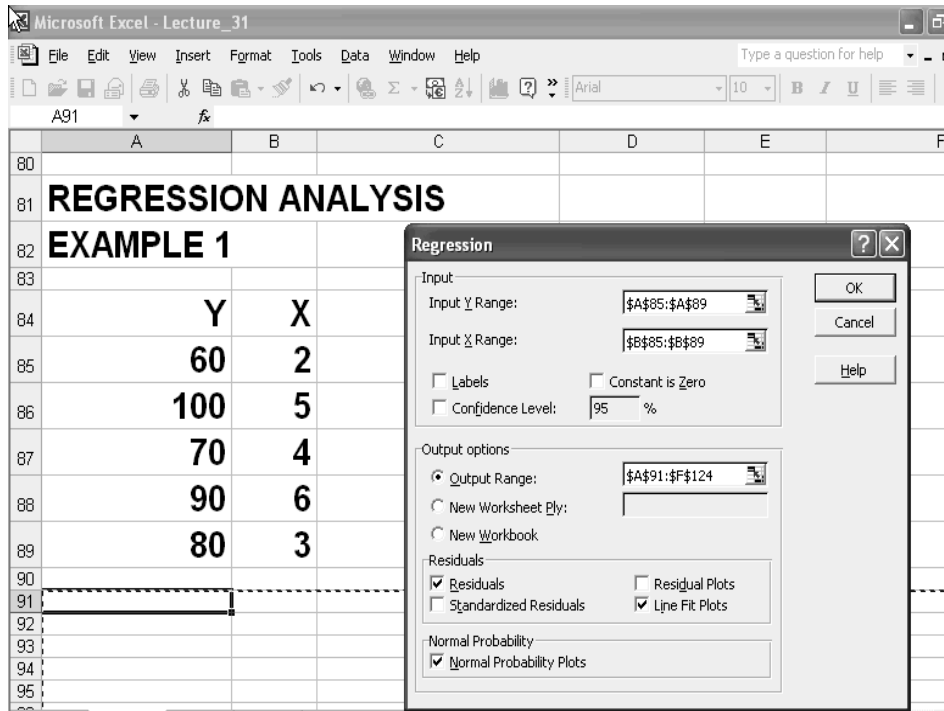




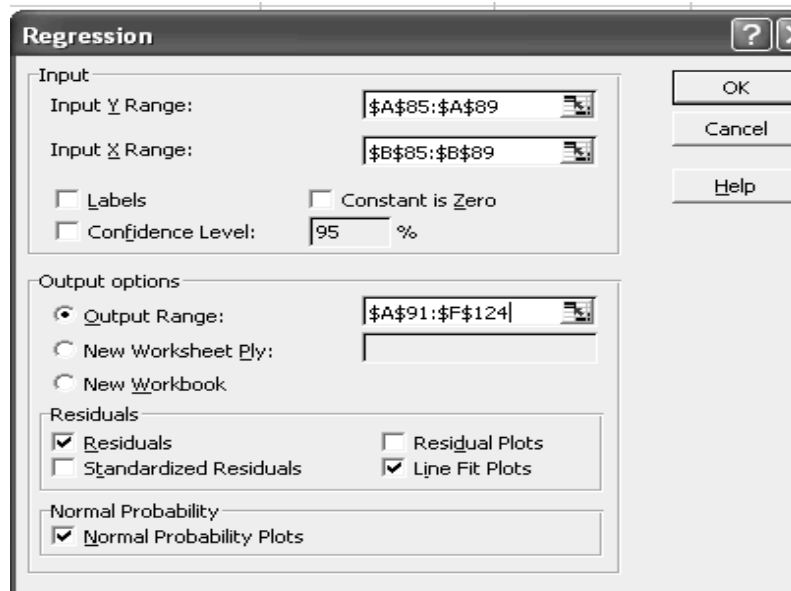
The regression dialog box opens as shown below. In this dialog box, Input range for X and Y is required. One can specify labels, confidence level and output etc.



For the sample data the input Y range was selected by clicking in the text box for input y range data first and then selecting the Y range (A85:A89). The regression tool adds the \$ sign in front of the column and row number to fix its location. The input range for X was specified in a similar fashion. No labels were chosen. The default value of 95% confidence interval was accepted. The output range was also selected in an arbitrary fashion. All you need to do is to select a range of cells for the output tables and the graphs. The range A91:F124 was selected as output range by selecting cell A91 and then dragging the mouse in such a manner that the last cell selected on the right was F124.



The Regression dialog box with data is shown below for clarity.



When you click OK on the Regression tool box a detailed SUMMARY OUTPUT is generated by the Regression Tool. This output is shown in parts below.

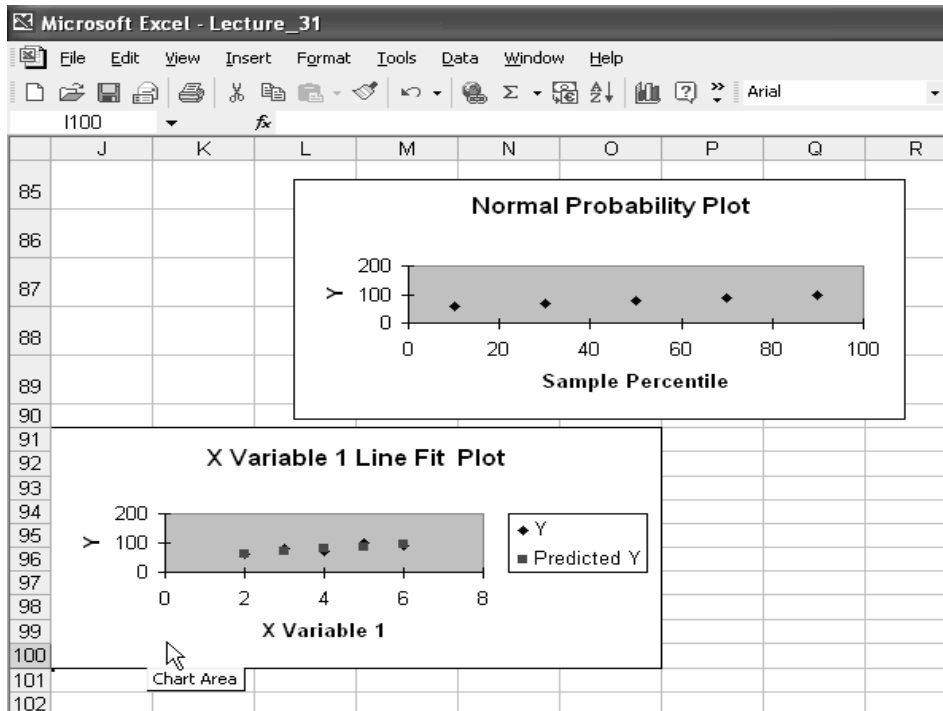
<b>SUMMARY OUTPUT</b>	
<i>Regression Statistics</i>	
<b>Multiple R</b>	<b>0.8</b>
<b>R Square</b>	<b>0.64</b>
<b>Adjusted R Square</b>	<b>0.52</b>
<b>Standard Error</b>	<b>10.9544512</b>
<b>Observations</b>	<b>5</b>

<b>ANOVA</b>			
	<i>df</i>	<i>SS</i>	<i>MS</i>
<b>Regression</b>	<b>1</b>	<b>640</b>	<b>640</b>
<b>Residual</b>	<b>3</b>	<b>360</b>	<b>120</b>
<b>Total</b>	<b>4</b>	<b>1000</b>	
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>
<b>Intercept</b>	<b>48</b>	<b>14.69693846</b>	<b>3.2659</b>
<b>X Variable 1</b>	<b>8</b>	<b>3.464101615</b>	<b>2.309</b>

57					
58	<b>F</b>	<b>Significance F</b>			
59	<b>5.3333</b>	<b>0.104088039</b>			
60					
61					
62					
63	<b>P-value</b>	<b>Lower 95%</b>	<b>Upper 95%</b>	<b>Lower 95.0%</b>	<b>Upper 95.0%</b>
64	<b>0.0469</b>	<b>1.227738632</b>	<b>94.77226</b>	<b>1.22773863</b>	<b>94.7722614</b>
65	<b>0.1041</b>	<b>-3.024327728</b>	<b>19.02433</b>	<b>-3.0243277</b>	<b>19.0243277</b>
66					

68			
69	<b>RESIDUAL OUTPUT</b>		
70			
71	<b>Observation</b>	<b>Predicted Y</b>	<b>Residuals</b>
72	<b>1</b>	<b>64</b>	<b>-4</b>
73	<b>2</b>	<b>88</b>	<b>12</b>
74	<b>3</b>	<b>80</b>	<b>-10</b>
75	<b>4</b>	<b>96</b>	<b>-6</b>
76	<b>5</b>	<b>72</b>	<b>8</b>
77			

The regression Tool also generates a normal probability plot and Line Fit Plot.



### **EXCEL REGRESSION TOOL OUTPUT**

In the regression Tool output there are a number of outputs for detailed analysis including Analysis of Variance (ANOVA) that is not part of this course. The main points of our interest for simple linear regression are:

#### **Multiple R**

Correlation Coefficient

#### **R Square**

Coefficient of determination

#### **STEM-Standard Error of mean:**

Standard deviation of population/sample size

#### **T-Statistic**

= (sample slope – population slope) / Standard error

#### **RSQ**

There is a separate function RSQ in EXCEL to calculate the coefficient of determination square of r. Description of this function is as follows:

Returns the square of the Pearson product moment correlation coefficient through data points in known\_y's and known\_x's. For more information, see PEARSON. The r-squared value can be interpreted as the proportion of the variance in y attributable to the variance in x.

#### **Syntax**

#### **RSQ(known\_y's,known\_x's)**

Known\_y's is an array or range of data points.

Known\_x's is an array or range of data points.

#### **Remarks**

- The arguments must be either numbers or names, arrays, or references that contain numbers.
- If an array or reference argument contains text, logical values, or empty cells, those values are ignored; however, cells with the value zero are included.
- If known\_y's and known\_x's are empty or have a different number of data points, RSQ returns the #N/A error value.
- The equation for the r value of the regression line is:

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

**Example**

	<b>A</b>	<b>B</b>
	<b>Known y</b>	<b>Known x</b>
1		
2	2	6
3	3	5
4	9	11
5	1	7
6	8	5
7	7	4
8	5	4
	<b>Formula</b>	<b>Description (Result)</b>
	=RSQ(A2:A8,B2:B8)	Square of the Pearson product moment correlation coefficient through data points above (0.05795)

**P-VALUE**

In the EXCEL regression Tool, the P-Value is defined as under:

P-value is the Probability of not getting a sample slope as high as the calculated value.

Smaller the value more significant the result. In our example

P-value=0.000133.

It means that slope is very significantly different from zero.

**Conclusion**

X and y are strongly associated

**SAMPLING DISTRIBUTION IN r**

It is possible to construct a sampling distribution for r similar to those for sampling distributions for means and percentages.

Tables at the end of books give minimum values of r (ignoring sign) for a given sample size to demonstrate a significant non-zero correlation at various significance levels (0.1, 0.05, 0.02, 0.01 and 0.001) and degrees of freedom (1 to 100).

It is to be noted that  $v = \text{degrees of freedom} = n - 2$  in all these calculations.

**SAMPLING DISTRIBUTION IN r-EXAMPLE 1**

*Look at a sample size  $n = 5$ .*

**Null hypothesis:**  $r = 0$ .

Calculated coefficient = 0.8.

Test the significance at 5% confidence level.

**Solution:**

Look in the table at row with  $v = 3$  and column headed by 0.05.

You will find the Tabulated value = 0.8783.

Sample value of 0.8 is less than 0.8783.

**Conclusion**

Correlation is not significantly different from zero at 5% level.

Variables are not strongly associated.

**SAMPLING DISTRIBUTION IN r-EXAMPLE 2**

*Look at a sample size  $n = 5$ .*

**Null hypothesis:**  $r = 0$

Calculated coefficient = -0.95

Test the significance at 5% confidence level.

**Solution:**

Look at row with  $v = 3$  and column headed by 0.05.

Tabulated value = 0.8783.

Sample value of 0.95 (ignoring sign) is greater than 0.8783

**Conclusion**

Correlation is significantly different from zero at 5% level.

Variables are strongly associated.

**LECTURE 32  
TIME SERIES AND  
EXPONENTIAL SMOOTHING  
PART 1**

**OBJECTIVES**

The objectives of the lecture are to learn about:

- Review Lecture 31
- Time Series and Exponential Smoothing

**SIMPLE LINEAR REGRESSION EQUATION. EXAMPLE**

The slide below shows the data from 7 stores covering square ft and annual sales. The question is whether there is a relationship between the area and the sale for these stores. It is required to find the regression equation that best fits the data.

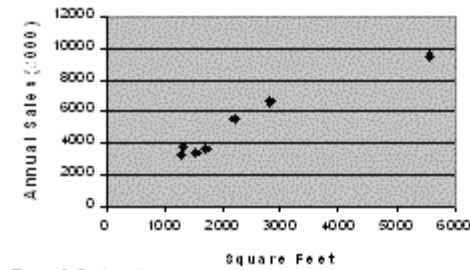
**Simple Linear Regression  
Equation: Example**

	Store	Square Feet	Annual Sales Rs. (000)
	1	1,726	3,681
You wish to examine the relationship between the square footage of produce stores and their annual sales. Sample data for 7 stores were obtained. Find the equation of the straight line that fits the data best	2	1,542	3,395
	3	2,816	6,653
	4	5,555	9,543
	5	1,292	3,318
	6	2,208	5,563
	7	1,313	3,760

First of all a scatter diagram is prepared using EXCEL Chart Wizard as shown below. The points on the scatter diagram clearly show a positive linear relationship between the annual sale and the area of store. It means that it will make sense to proceed further with regression analysis.



## Scatter Diagram Example



Using the EXCEL Regression Tool, the regression equation was derived as given below.

## Equation for Sample Regression Line

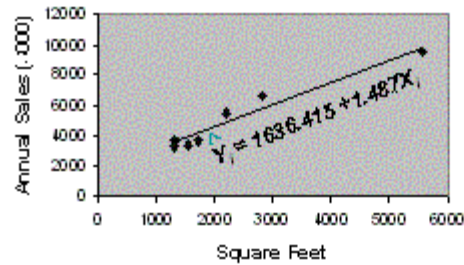
$$\begin{aligned} \hat{Y}_i &= b_0 + b_1 X_i \\ &= 1636.415 + 1.487 X_i \end{aligned}$$

From Excel Printout

	Coefficients
Intercept	1636.414726
X Variable	1.486633657

The graph of the regression line was prepared using the regression Tool. The result shows the data points, regression line and text showing the equation. As you see, it is possible to carry out linear regression very easily using Excel's Regression Tool.

## Graph of the Sample Regression Line



### Interpreting the Results

The slide below gives the main points, namely, that for every increase of 1 sq. ft. there is a sale of 1.487 units or 1407 Rs. As each unit was equal to 1,000. Now that the equation has been developed, we can estimate sale of stores of other sizes using this equation.

## Interpreting the Results

$$\hat{Y}_i = 1636.415 + 1.487X_i$$

*The slope of 1.487 means that each increase of one unit in  $X_i$  we predict the average of  $Y$  to increase by an estimated 1.487 units.*

*The model estimates that for each increase of 1 square foot in the size of the store, the expected annual sales are predicted to increase by Rs. 1487*

## Interpreting the Results

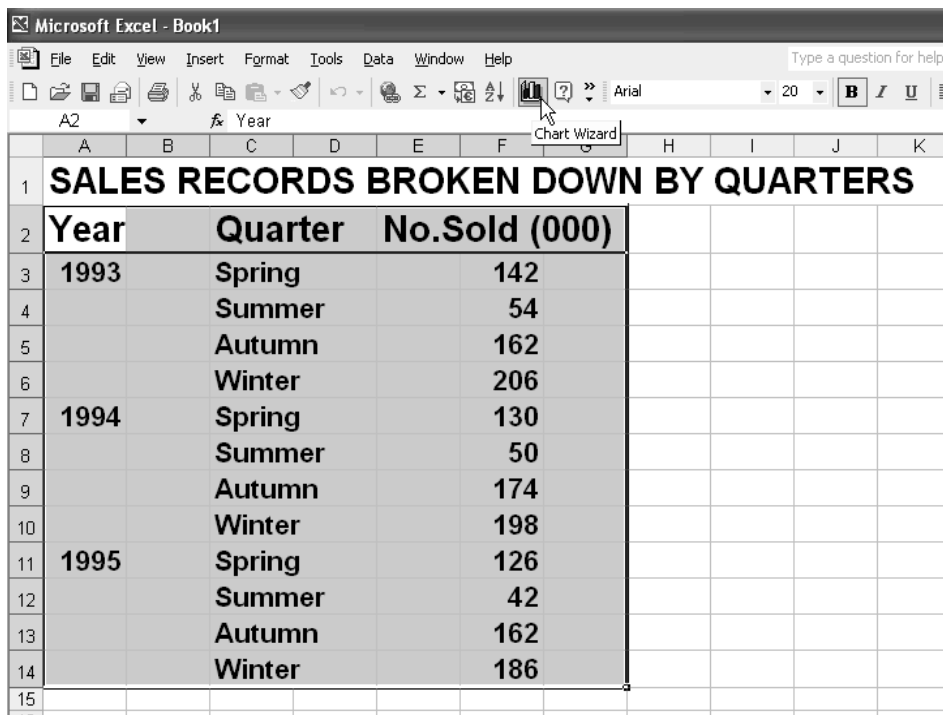
$$\hat{Y}_i = 1636.415 + 1.487X_i$$

*The slope of 1.487 means that each increase of one unit in  $X$ , we predict the average of  $Y$  to increase by an estimated 1.487 units.*

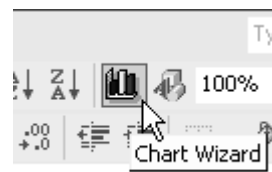
*The model estimates that for each increase of 1 square foot in the size of the store, the expected annual sales are predicted to increase by Rs.1487*

### CHART WIZARD

Let us look at how we can use the Chart Wizard. We wish to study the problem shown in the slide below.

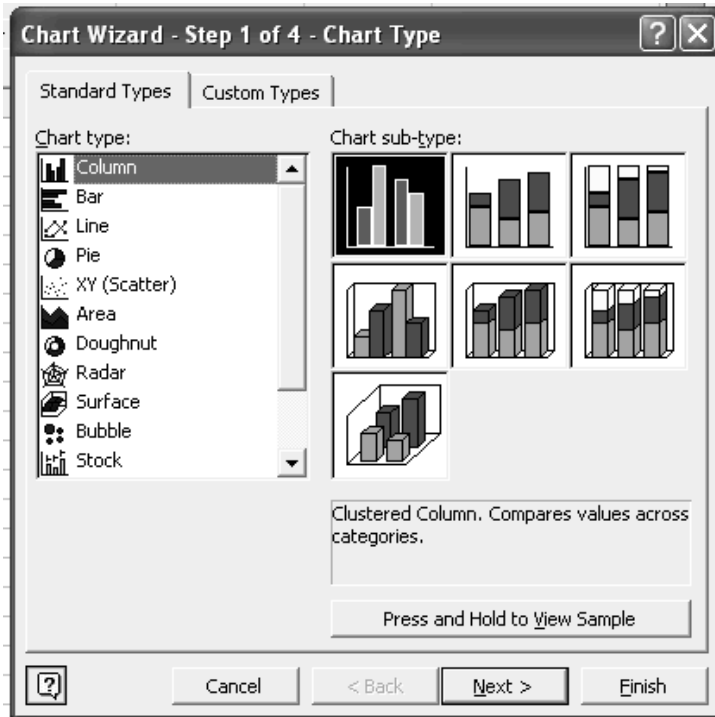


SALES RECORDS BROKEN DOWN BY QUARTERS			
Year	Quarter	No.Sold (000)	
1993	Spring	142	
	Summer	54	
	Autumn	162	
	Winter	206	
1994	Spring	130	
	Summer	50	
	Autumn	174	
	Winter	198	
1995	Spring	126	
	Summer	42	
	Autumn	162	
	Winter	186	



You can start with the Chart Icon as shown on the right.  
There are 4 steps in using the Chart wizard as shown below:

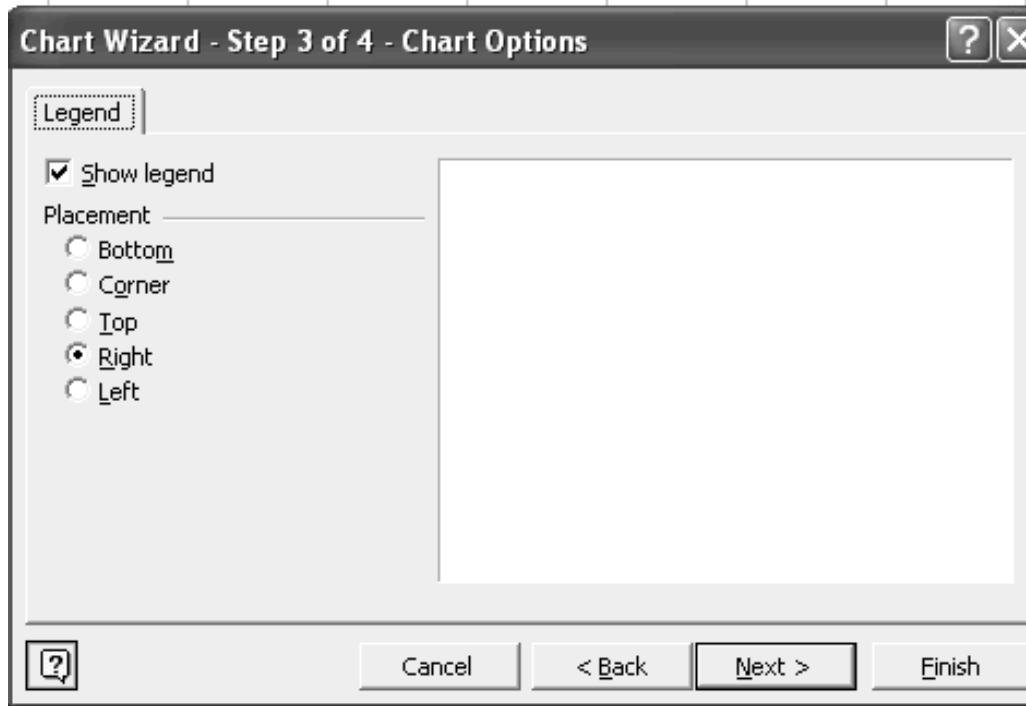
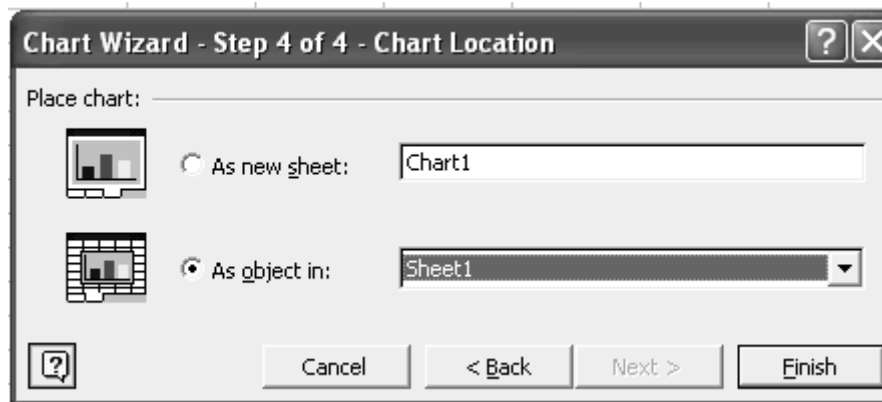
**Step 1**



**Step 2**



**Step 3**

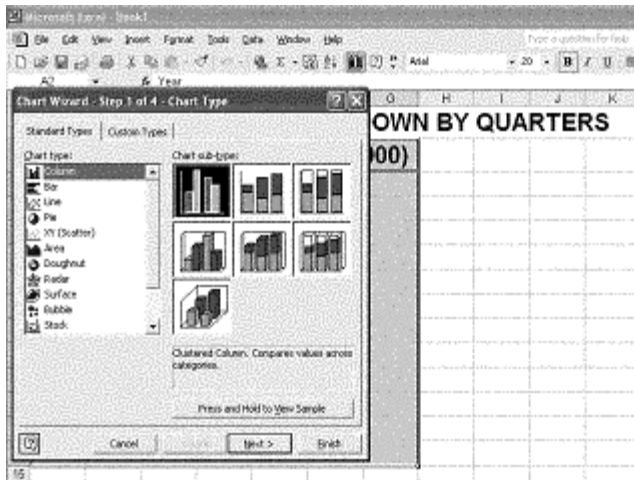
**Step 4**

The dialog boxes are self-explanatory. Let us look at the example above and see how Chart wizard was used.

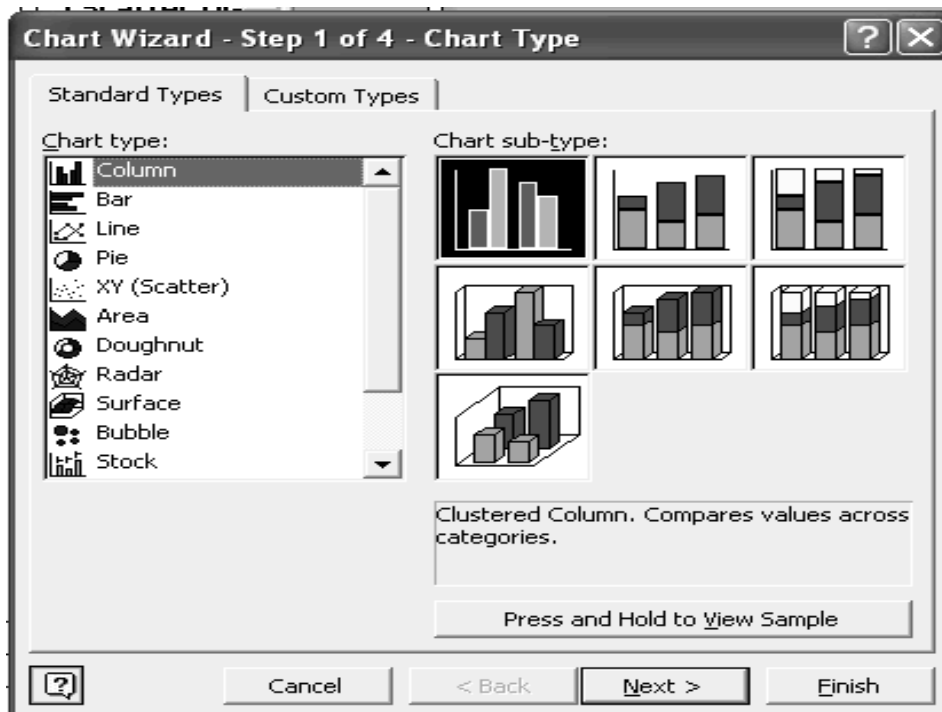
First the data was selected on the worksheet. Next the Chart Wizard was selected.

We chose Column Graph as the option as you can see in the slide below.

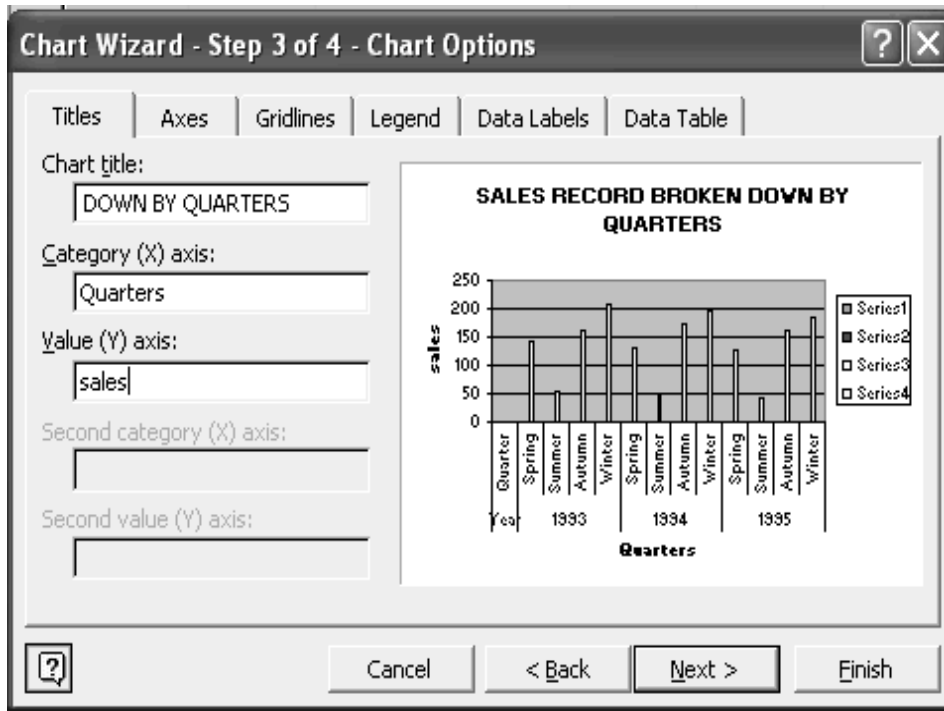
We clicked Next.



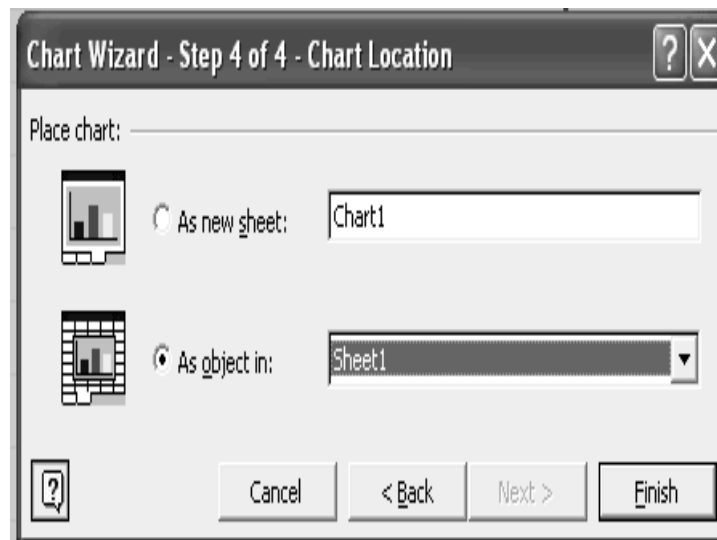
You can see the selection of Column graph in the slide below.



Under Step 2, the Chart Title, Category (X) axis and value (Y) were entered as shown in the slide. Then the button Next was clicked.



Under the 4<sup>th</sup> step, the default values Chart1 and Sheet1 were selected. Then the button Finish was clicked.



The result is shown below as a column graph.

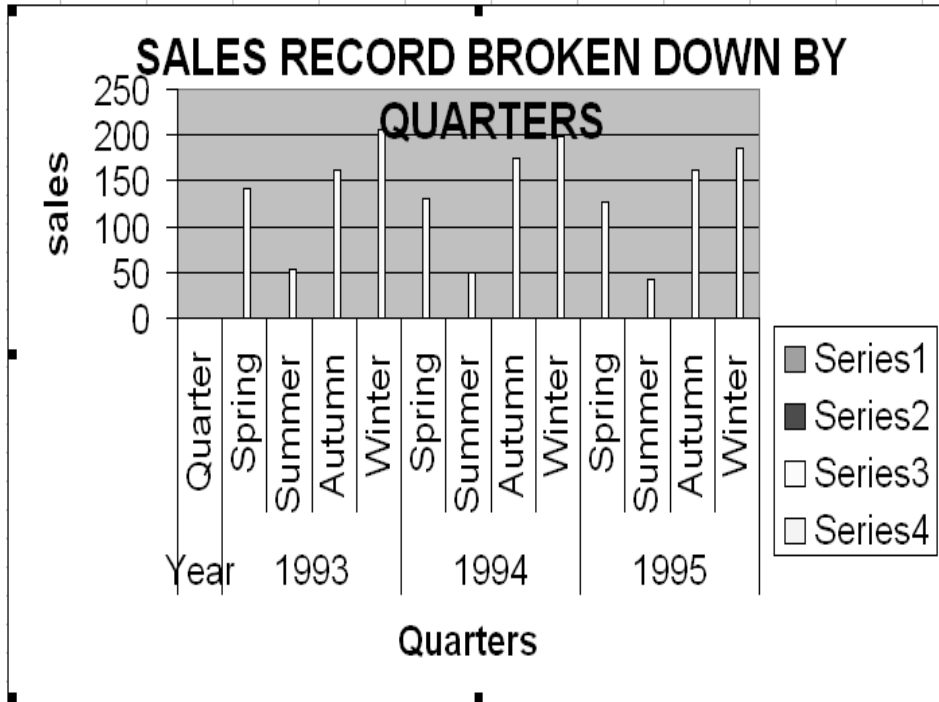
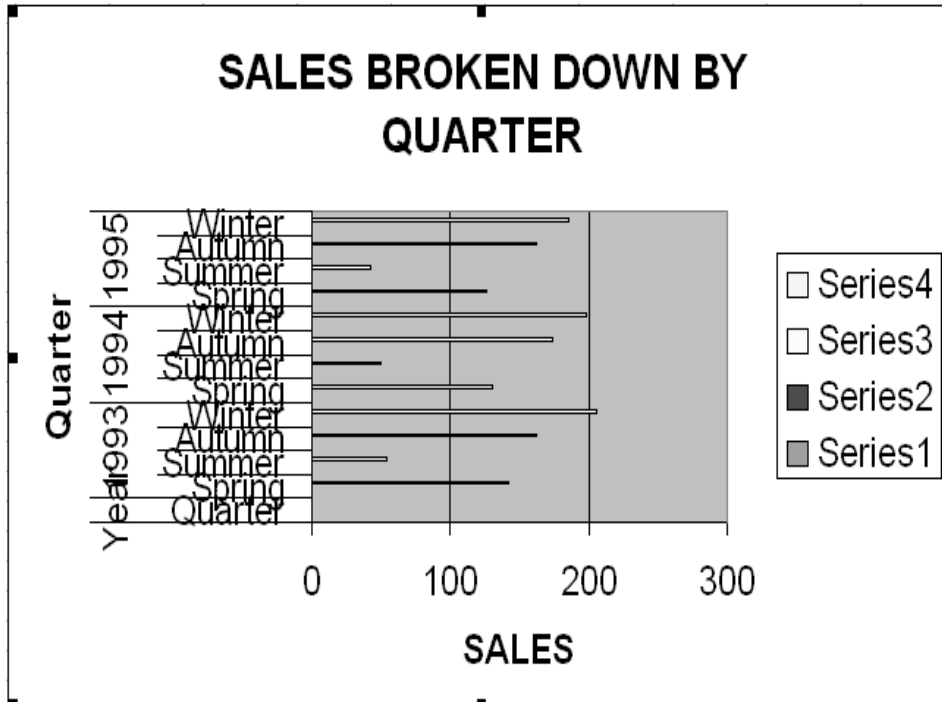
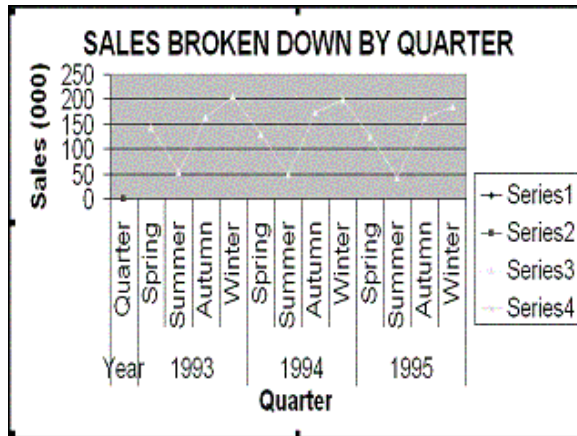


Chart Wizard was used again to draw a Side by Side chart using the same data. The result is shown below.



A line graph of the data was also prepared as shown below. This graph shows the seasonal variations in the values of sales.





**EXAMINATION OF GRAPH TREND**

The graph shows that there is a general upward or downward steady behaviour of figures. There are Seasonal Variations also. These are variations which repeat themselves regularly over short term, less than a year. There is also a random effect that is variations due to unpredictable situations. There are cyclical variations which appear as alternation of upward and downward movement.

**EXTRACTING THE TREND FROM DATA**

Look at the following data:

170, 140, 230, 176, 152, 233, 182, 161, 242

There is no explanation regarding time periods. What to do?

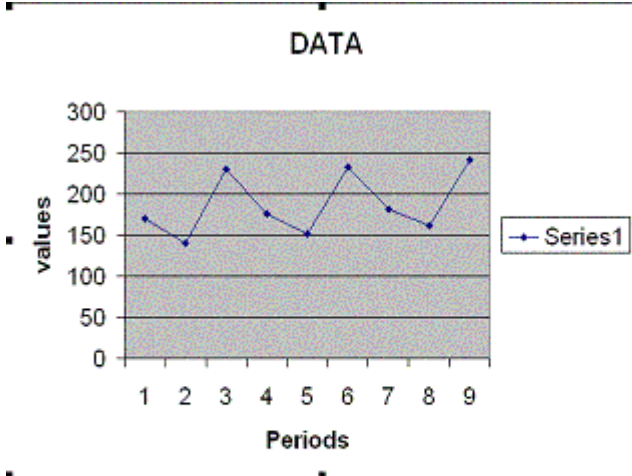
**First step**

- Plot figures on graph
- Horizontal as period 1
- Vertical as period 2

**Conclusion**

There is a marked pattern that repeats itself.

There is a well established method to extract trend with strong repeating pattern



**MOVING AVERAGES**

Look at the data in the slide below. There is sales data for morning, afternoon and evening for day 1, 2 and 3. We can calculate averages for each day as shown. These are simple averages for each day.

The screenshot shows an Excel worksheet titled "Lecture\_32". The data is organized as follows:

	B	C	D	E	F	G	H	I	J	K	L	M
122					<b>AVERAGES</b>							
123					<b>Data</b>	<b>Moving Average = Trend</b>						
124	<b>Day 1</b>		<b>Morning</b>		<b>170</b>							
125			<b>Afternoon</b>		<b>140</b>							
126			<b>Evening</b>		<b>230</b>							
127			<b>Average</b>						<b>180</b>	<b>=AVERAGE(F124:F126)</b>		
128	<b>Day 2</b>		<b>Morning</b>		<b>176</b>							
129			<b>Afternoon</b>		<b>152</b>							
130			<b>Evening</b>		<b>233</b>							
131			<b>Average</b>						<b>187</b>	<b>=AVERAGE(F128:F130)</b>		
132	<b>Day 3</b>		<b>Morning</b>		<b>182</b>							
133			<b>Afternoon</b>		<b>161</b>							
134			<b>Evening</b>		<b>242</b>							
135			<b>Average</b>						<b>195</b>	<b>=AVERAGE(F132:F134)</b>		
136												

Now let us look at the idea of moving averages.

**First Average- Day 1**

$$= (170 + 140 + 230)/3 = 540/3 = 180$$

**Next Average-Morning**

$$= (140 + 230 + 176)/3 = 546/3 = 182$$

**Next Average-Afternoon**

$$= (230 + 176 + 152)/3 = 558/3 = 186$$

**Another method**

$$\text{Drop } 170; \text{ Add } 176; = (176-170)/3 = 6/3 = 2$$

$$\text{Last average} + 2 = 180 + 2 = 182$$

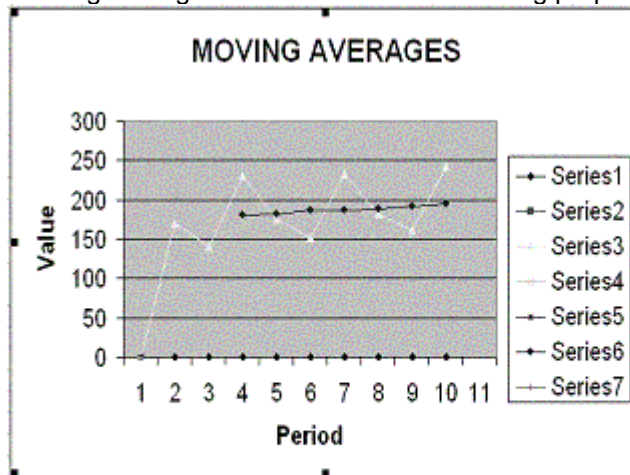
**Caution**

You may make a mistake

You saw how it is possible to start with the first 3 values 170, 140 and 230 for the first day and work out the average (180). Next we dropped 170 and added 152 the morning value from day 2. This gave us an average of 182. Similarly, the next value was calculated. Look at the worksheet below for the complete calculation. These averages are called moving averages. You could have used the alternative method but you may make a mistake in mental arithmetic. So let us only use EXCEL worksheets.

MOVING AVERAGES												
	Period	Data	Moving Average	= Trend								
Day 1	Morning	170										
	Afternoon	140		180	=AVERAGE(F140:F142)							
	Evening	230		182	=AVERAGE(F141:F143)							
Day 2	Morning	176		186	=AVERAGE(F142:F144)							
	Afternoon	152		187	=AVERAGE(F143:F145)							
	Evening	233		189	=AVERAGE(F144:F146)							
Day 3	Morning	182		192	=AVERAGE(F145:F147)							
	Afternoon	161		195	=AVERAGE(F146:F148)							
	Evening	242										

The moving averages were plotted as shown below. You can see that the seasonal variation has disappeared. Instead you see a clear trend of increase in sales. This plot shows that moving averages can be used for forecasting purposes.



**ANALYSING SEASONAL VARIATIONS**

Let us find out how much each period differs from trend

**Calculate Actual – trend for each period**

**Day 1, Afternoon**

Actual = 180, Trend = 140

Actual – Trend = 140 – 180 = -40

Here, -40 is the seasonal variation.

Similarly, other seasonal variations can be worked out.

### LECTURE 33

## TIME SERIES AND EXPONENTIAL SMOOTHING

### PART 2

#### OBJECTIVES

The objectives of the lecture are to learn about:

- Review Lecture 32
- Time Series and Exponential Smoothing.

#### TREND

As discussed briefly in the handout for lecture 32, the trend is given by the moving average minus the actual data. Look at the slide shown below. The average of the morning, afternoon and evening of the first day is 180. This value is written in cell I179, which is the middle value for first day. The next moving average is written in cell I180. This means that the last moving average will be written in cell I185 as the moving average of the morning, afternoon and evening of 3<sup>rd</sup> day will be written against the middle value in cell F185.

Now that all the moving averages have been worked out we can calculate the trend as difference of moving average and actual value.

ACTUAL MINUS TREND					
Day	Period	Data	Moving Average	= Trend	
1	Morning	170			
	Afternoon	140	180	-40 =F179-I179	
	Evening	230	182	48	
2	Morning	176	186	-10	
	Afternoon	152	187	-35	
	Evening	233	189	44	
3	Morning	182	192	-10	
	Afternoon	161	195	-34	
	Evening	242			

The actual trend figures are now written as shown in the slide below with M for morning, A for afternoon and E for evening. The titles Day 1, day 2 and Day 3 were written on the left hand side of the table. Further Total for each column was calculated. The total was divided by the non-zero values in the column. For example, in column M, there are 2 non-zero values. Hence, the total 20 was divided by 2 to obtain the average -10. Similarly, the averages in column A and E were calculated. This data is the seasonal variation and can now be used for estimating trend and random variations.

Microsoft Excel - Lecture_32								
Type a question for help								
A202								
A	B	C	D	E	F	G	H	I
191	<b>ACTUAL-TREND FIGURES TOGETHER</b>							
192			<b>M</b>	<b>A</b>	<b>E</b>			
193	<b>Day 1</b>		<b>0</b>	<b>-40</b>	<b>48</b>			
194	<b>Day 2</b>		<b>-10</b>	<b>-35</b>	<b>44</b>			
195	<b>Day 3</b>		<b>-10</b>	<b>-34</b>	<b>0</b>			
196								
197	<b>Total</b>		<b>-20</b>	<b>-109</b>	<b>92</b>			
198	<b>Average</b>		<b>-10</b>	<b>-36</b>	<b>46</b>			
199								
200								

### **EXTRACTING RANDOM VARIATIONS**

#### **Day 1**

Afternoon trend = 180

Afternoon seasonal variation = 36

Trend – variation = 180 – 36 = 144

Actual value = 140

Random variation = 140 – 144 = -4

#### **Conclusion**

**Expected = Trend + Seasonal**

**Random = Actual – expected**

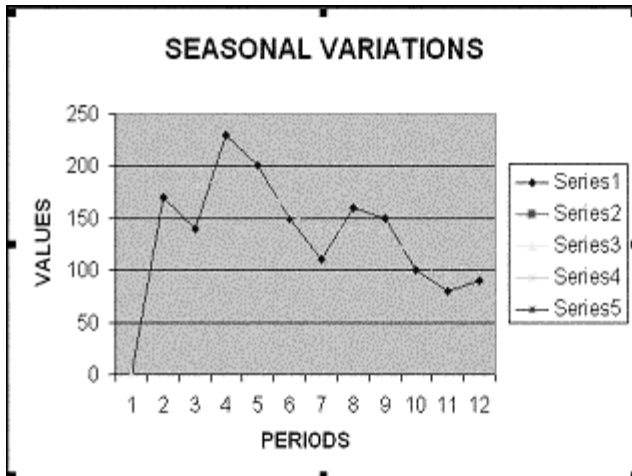
	A	B	C	D	E	F	G	H	I	J	K	L
202			<b>ACTUAL; EXPECTED AND RANDOM</b>									
203												
204	<b>ACTUAL</b>				140	230	176	152	233	182	161	
205	<b>Expected</b>											
206	<b>(trend + seasonal)</b>				144	228	176	151	235	182	159	
207	<b>Random</b>											
208	<b>(actual - expected)</b>				-4	2	0	1	-2	0	2	
209												

**Forecast for day 4**

=  
 Trend for afternoon of day 4  
 +  
 Seasonal adjustment for afternoon period  
 Trend = 180 to 195 (6 intervals)  
 = 15/6 = 2.5 per period  
 Figure for evening of day 3 = 195 + 2.5 = 197.5  
 Morning of day 4 = 197.5 + 2.5 = 200  
 Afternoon of day 4 = 200 + 2.5 = 202.5  
 After adjustment of seasonal variation = -36  
 = 202.5 - 36 = 166.5 or 166

**SEASONABLE VARIATIONS**

Seasonal Variations are regarded as constant amount added to or subtracted from the trends. This is a reasonable assumption as seasonal peaks and troughs are roughly of constant size. In practice Seasonal variations will not be constant. These will themselves vary as trend increases or decreases. Peaks and troughs can become less pronounced Seasonal variations as well as the trend are shown in the graph below. You can see that the trend clearly shows a downward slide in values.



In the following slide, the actual values are for 4 quarters per year. Here there is no middle value per year. The moving averages were therefore summarised against the 3<sup>rd</sup> quarter. As this does not reflect the correct position, the average of the first two moving averages was calculated and written as centred moving average in column H. The first centred moving average is the average of 141 and 138 or 139.5. This is used as the trend and the value Actual-Trend is the difference of Actual – Centred Moving Average. Here also the last row does not have a value as the moving average was shifted one position upwards.

TREND AND SEASONAL VARIATIONS						
Quarter	Actual	Moving Average	Centred M:Average	Actual - trend		
1	142					
2	54					
3	162	141	139.5	22.5		
4	206	138	137.5	68.5		
1	130	137	138.5	-8.5		
2	50	140	139.0	-89.0		
3	174	138	137.5	36.5		
4	198	137	136.0	62.0		
1	126	135	133.5	-7.5		
2	42	132	130.5	-88.5		
3	162	129				

The data from the previous slide was summarised as in the following slide using the approach described earlier. It may be seen that the average seasonal variation for Spring, Summer, Autumn and Winter is -8, -88.8, 29.5 and 65.3 respectively.

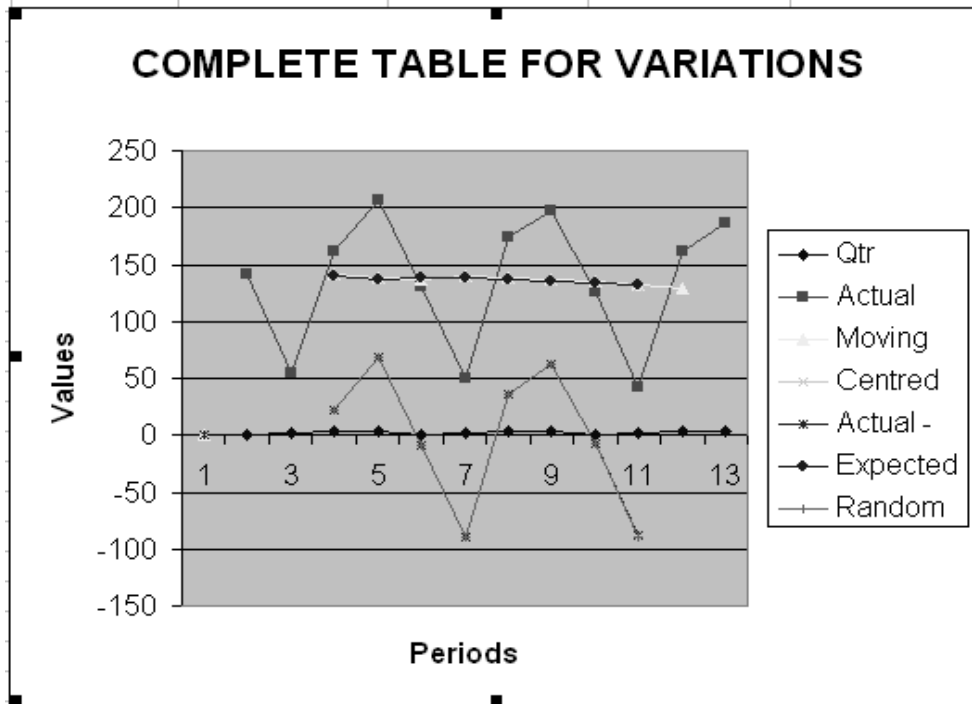
	Spring	Summer	Autumn	Winter
1993	0	0	22.5	68.5
1994	-8.5	-89	36.5	62
1995	-7.5	-88.5	0	0
<b>Total</b>	<b>-16.0</b>	<b>-177.5</b>	<b>59.0</b>	<b>130.5</b>
<b>Average</b>	<b>-8.0</b>	<b>-88.8</b>	<b>29.5</b>	<b>65.3</b>
<b>Rounded</b>	<b>-8</b>	<b>-89</b>	<b>30</b>	<b>65</b>

The expected value now is the sum of centred moving average and random

Qtr	Actual	Moving Average	Centred M:Average	Actual - trend	Expected	Random
1	142					
2	54				=E92+H85	=C92-G92
3	162	141	139.5	22.5	169.5	-7.5 =C92-G9
4	206	138	137.5	68.5	203.5	2.5
1	130	137	138.5	-8.5	130.5	-0.5
2	50	140	139.0	-89.0	50.0	0.0
3	174	138	137.5	36.5	167.5	6.5
4	198	137	136.0	62.0	202.0	-4.0
1	126	135	133.5	-7.5	125.5	0.5
2	42	132	130.5	-88.5	43	-1.0
3	162	129				
4	186					

variation. The random variation is the difference between the Actual and Expected value. This gives us a complete table with all the values. The values in this table were plotted using the EXCEL Chart Wizard as shown below. You can see that the different components can now be seen clearly.





### **FORECASTING APPLE PIE SALES**

#### **Forecast**

Sale steadily declined from 139.0 to 130.5.

Over 4 quarters, the sales declined by =  $139.0 - 130.5 = 8.5$

Trend in Spring 1995 was 133.5.

We can assume annual decrease as on the basis of decline over the last 4 quarters = 8.5

Trend in 1996 = trend in 1995 less decline =  $133.5 - 8.5 = 125$

Seasonal variation as already worked out = -8

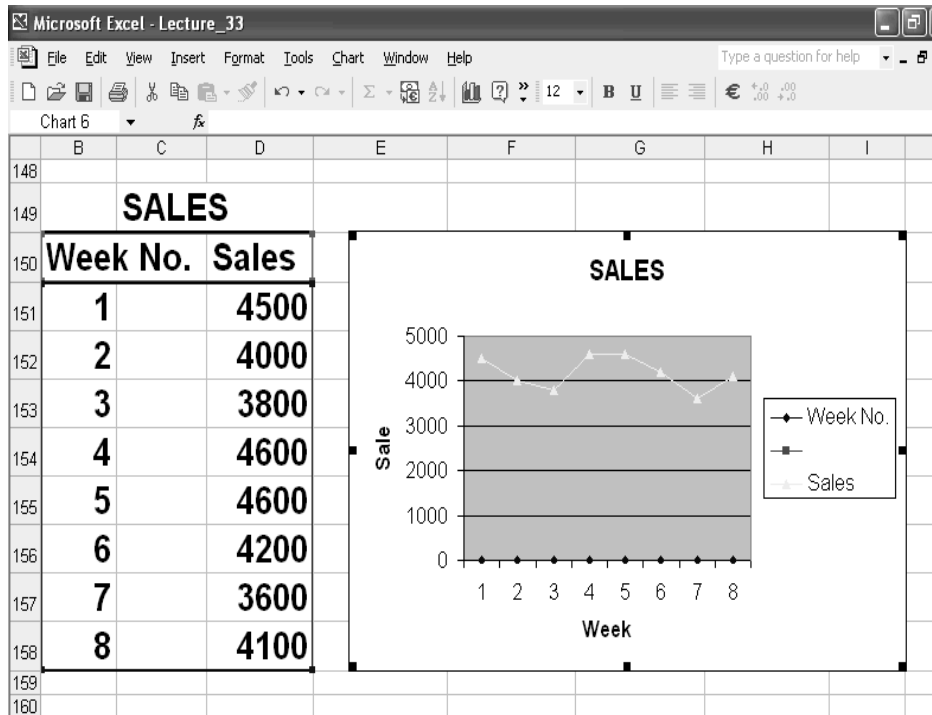
#### **Hence:**

Final forecast =  $125 - 8 = 117$

### **FORECASTING IN UNPREDICTABLE SITUATIONS**

Two methods were studied above. Each one has certain features. If there is steady increase in data and repeated seasonal variations, there are many cases that do not conform to these patterns. There may not be a trend. There may not be a short term pattern. Figures may hover around an average mark. How to forecast under such conditions?

Data for sales over a period of 8 weeks is summarized and plotted in the slide below. You may see that the values hover around an average value without any particular pattern. This problem requires a different solution.



**FORECAST**

Let us assume that the forecast for week 2 is the same as the actual data for week 1, that is 4500.

Week no.	Actual sales	Forecast
1	4500	-
2	4000	4500

The Actual sale was 4000. Thus, the Forecast is 500 too high.

Another approach would be to incorporate the proportion of error in the estimate as follows:

**new forecast = old forecast + proportion of error  $\alpha$**

Or

**new forecast = old forecast +  $\alpha$  x (old actual – old forecast)**

This method is called Exponential Smoothing. We shall learn more about this method in lecture 34.

**LECTURE 34**  
**FACTORIALS**  
**PERMUTATIONS AND COMBINATIONS**

**OBJECTIVES**

The objectives of the lecture are to learn about:

- Review Lecture 33
- Factorials
- Permutations and Combinations

**Module 7**

Module 7 covers the following:

Factorials

Permutations and Combinations

(Lecture 34)

Elementary Probability

(Lectures 35-36)

Chi-Square

(Lectures 37)

Binomial Distribution

(Lectures 38)

**FORECAST**

Please refer to the Example discussed in Handout 33.

Let  $\alpha = 0.3$

**Then:**

Forecast week 3 = week 2 forecast +  $\alpha$  x (week 2 actual sale – week 2 forecast)  
= 4500 – 0.3 x 500 = 4350

**Conclusion**

Overestimate is reduced by 30% of the error margin 500.

The slide below shows the calculation for normal error as well as alpha x error. You can see that the error is considerably reduced using this approach.

	B	C	D	E	F	G	H	I
162	<b>Week No.</b>	<b>Sales</b>	<b>Forecast</b>			<b>Error</b>	<b>alpha x Error</b>	
163	1	4500						
164	2	4000	4500			-500	-150	
165	3	3800	4350			-550	-165	
166	4	4600	4185			415	124.5	
167	5	4600	4309.5			290.5	87.2	
168	6	4200	4396.7			-196.7	-59	
169	7	3600	4337.7			-737.7	-221.3	
170	8	4100	4116.4			-16.4	-4.91	
171	9							
172	<b>Forecast =E164+H164</b>			<b>Error=D164-E164</b>				
173	<b>ErrorAlpha = 0.3</b>							

The forecast is now calculated by adding alpha x Error to the actual sales. The error is the difference between the actual sales and the forecast. The first value is the same as the sale last week. Use of alpha =0.3 is considered very common. This method is called Exponential Smoothing and alpha is Smoothing Constant.

**Rule for obtaining a forecast:**

Let A= Actual and F= Forecast.

Then:

$$F(t) = F(t-1) + \alpha(A(t-1) - F(t-1))$$

$$= \alpha A(t-1) + (1-\alpha) F(t-1)$$

$$F(t-1) = \alpha A(t-2) + (1-\alpha) F(t-2)$$

Substituting

$$F3 = \alpha A(t-1) + (1-\alpha) [\alpha A(t-2) + (1-\alpha) F(t-2)]$$

$$= \alpha [A(t-1) + (1-\alpha) A(t-2)] + (1-\alpha)^2 F(t-2)$$

Replacing F(t-2) by  $\alpha A(t-3) + (1-\alpha) F(t-3)$

$$F(t) = \alpha [A(t-1) + (1-\alpha) A(t-2) + (1-\alpha)^2 A(t-3)] + (1-\alpha) F(t-3)$$

**WHERE TO APPLY EXPONENTIAL SMOOTHING**

What kinds of situations require the application of Exponential Smoothing?

What are good values of  $\alpha$ ?

**The accepted Criterion is Mean Square Error (MSE).**

You can find MSE for by squaring all and including the present one and dividing by the number of periods included.

Sign of good forecast is when MSE stabilizes.

Generally alpha between 0.1 and 0.3 performs best.

**Example**

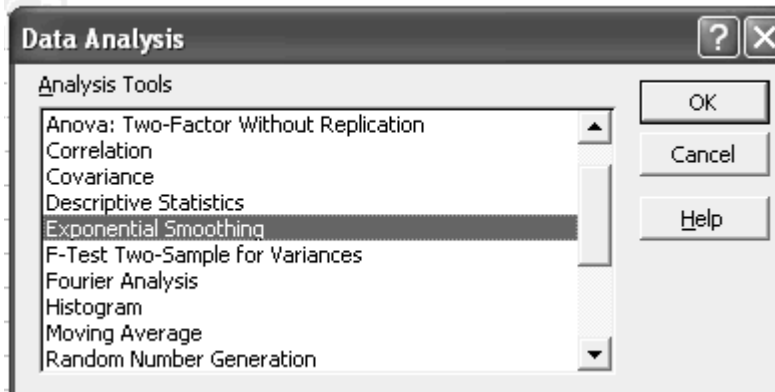
The slide below shows the calculation of MSE. Detailed formulas can be seen in the Worksheet for Lecture 34.

Week	Actual	Forecast	Error	0.3xError	Error <sup>2</sup>	MSE
22	2200	2200	0	0	0	0
23	2400	2200	200	60	40000	20000
24	2600	2260	340	102	115600	51867
25	2800	2362	438	131.4	191844	86861
26	3000	2493.4	506.6	152.0	256644	120818
Forecast =D179+F179		MSE =SUM(H176:H180)/5				

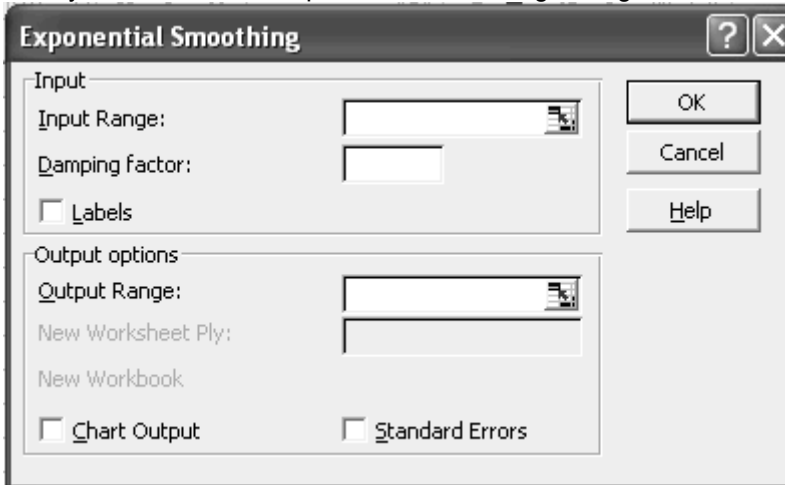
EXCEL

**EXPONENTIAL SMOOTHING TOOL**

It is possible to use the Exponential Smoothing Tool included in the EXCEL Tools.



After you click OK, the Exponential Smoothing Dialog Box is shown as below:



Different items in the Dialog Box are described below:

#### Input Range

Enter the cell reference for the range of data you want to analyze. The range must contain a single column or row with four or more cells of data.

#### Damping factor

Enter the damping factor you want to use as the exponential smoothing constant. The damping factor is a corrective factor that minimizes the instability of data collected across a population. The default damping factor is 0.3.

Note Values of 0.2 to 0.3 are reasonable smoothing constants. These values indicate that the current forecast should be adjusted 20 to 30 percent for error in the prior forecast. Larger constants yield a faster response but can produce erratic projections. Smaller constants can result in long lags for forecast values

#### Labels

Select if the first row and column of your input range contain labels. Clear this check box if your input range has no labels; Microsoft Excel generates appropriate data labels for the output table.

#### Output Range

Enter the reference for the upper-left cell of the output table. If you select the Standard Errors check box, Excel generates a two-column output table with

standard error values in the right column. If there are insufficient historical values to project a forecast or calculate a standard error, Excel returns the #N/A error value.

**Note** The output range must be on the same worksheet as the data used in the input range. For this reason, the New Worksheet Ply and New Workbook options are unavailable.

### Chart Output

Select to generate an embedded chart for the actual and forecast values in the output table.

### Standard Errors

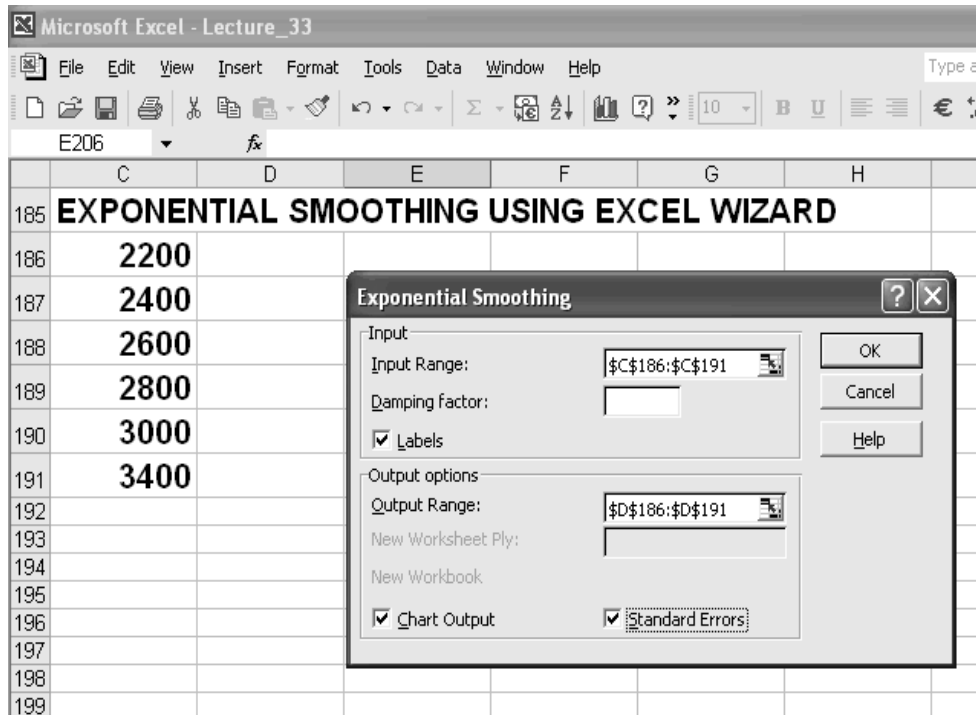
Select if you want to include a column that contains standard error values in the output table. Clear if you want a single-column output table without standard error values.

### Example

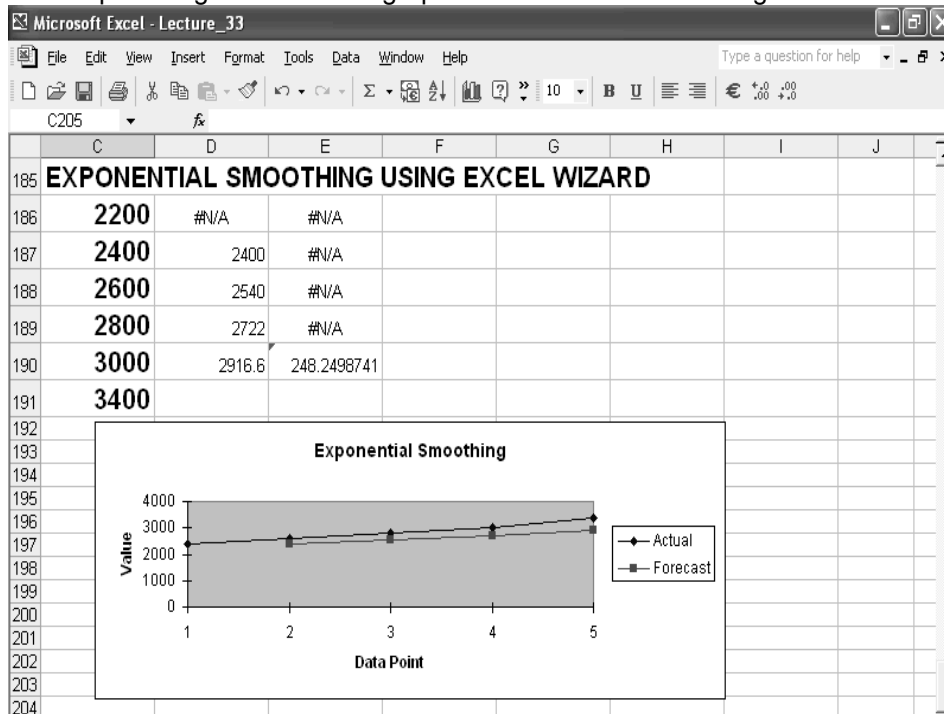
Use of the Exponential Smoothing Tool is shown in the following slides. First the Exponential Tool was selected.

	C	D	E	F	G	H	I	J	
185	<b>EXponential Smoothing Using Excel Wizard</b>								
186	2200	#N/A							
187	2400	2200							
188	2600	2340							
189	2800	2522							
190	3000	2716.6							
191	3400	2914.98							
192									
193									
194									
195									
196									
197									

Next the Input and Output Range were specified. Labels, Chart Output and Standard Errors were ticked as options in check boxes.



The output along with standard graphs is shown on the following slide.



**FACTORIAL**

Let us look at natural numbers.

**Natural Numbers**

1, 2, 3,...

Let us now define a factorial of natural numbers, say factorial of 5.

**Five Factorial**

$5! = 5.4.3.2.1$  or  $1.2.3.4.5$

Similarly factorial of 10 is:

**Ten Factorial**

$= 1.2.3.4.5.6.7.8.9.10 = 10.9.8.7.6.5.4.3.2.1$

**In general**

$n! = n(n-1)(n-2)...3.2.1$  or

$n! = n(n-1)(n-2)!$

$= n(n-1)!$

**FACTORIAL EXAMPLES**

$10! = 10.9.8.7.6.5.4.3.2.1 = 3,628,800$

$8!/5! = 8.7.6.5! = 8.7.6 = 336$

$12!/9! = 12.11.10.9!/9! = 12.11.10 = 1320$

$10!8!/9!5! = 10.9!8.7.6.5!/9!5! =$

$10.8.7.6 = 3360$

**WAYS**

If operation A can be performed in m ways and B in n ways, then the two operations can be performed together in m.n ways.

**Example**

A coin can be tossed in 2 ways. A die can be thrown in 6 ways. A coin and a die together can be thrown in  $2.6 = 12$  ways

**PERMUTATIONS**

An arrangement of all or some of a set of objects in a definite order is called permutation.

**Example 1**

There are 4 objects A, B, C and D

Permutations of 2 objects A & B: AB, BA

Permutations in three objects A, B and C:

ABC, ACB, BCA, BAC, CAB, CBA

**Example 2**

Number of permutations of 3 objects taken 2 at a time =  $3P_2$

$= 3!/(3-2)! = 3.2 = 6$

= AB, BA, AC, CA, BC, CB

Number of permutations of n objects taken r at a time =

$nPr = n!/(n-r)!$

**PERMUTATIONS OF n OBJECTS**

Number of n permutations of n different objects taken n at a time is n!

$nP_n = n(n-1)(n-2)...3.2.1$

Number of permutations of n objects of which  $n_1$  are alike of one kind,  $n_2$  are alike of one kind and  $n_k$  are alike.

$n!/n_1!n_2!...n_k!$

**Example 3**

How many possible permutations can be formed from the word STATISTICS?

S=3, A=1, T=3, I=2, C=1

**Formula**

$nPr = n!/n_1!n_2!...n_k!$

$= 10!/3!1!3!2!1! = 10.9.8.7.6.5.4.3!/3!3!2!$

= 50400

**PERMUT**

**EXCEL function PERMUT can be used to calculate number of permutations.**



Returns the number of permutations for a given number of objects that can be selected from number objects. A permutation is any set or subset of objects or events where internal order is significant. Permutations are different from combinations, for which the internal order is not significant. Use this function for lottery-style probability calculations.

**Syntax**

**PERMUT(number,number\_chosen)**

**Number** is an integer that describes the number of objects.

**Number\_chosen** is an integer that describes the number of objects in each permutation.

**Remarks**

- Both arguments are truncated to integers.
- If number or number\_chosen is nonnumeric, PERMUT returns the #VALUE! error value.
- If number  $\leq 0$  or if number\_chosen  $< 0$ , PERMUT returns the #NUM! error value.
- If number  $<$  number\_chosen, PERMUT returns the #NUM! error value.
- The equation for the number of permutations is:

$$P_{k,n} = \frac{n!}{(n-k)!}$$

**Example**

Suppose you want to calculate the odds of selecting a winning lottery number. Each lottery number contains three numbers, each of which can be between 0 (zero) and 99, inclusive. The following function calculates the number of possible permutations:

	A	B	C	D	E	F	G
1	PERMUT(number,number_chosen)						
2							
3	Data	Description					
4	100	Number of objects					
5	3	Number of objects in each permutation					
6	970200						
7	=PERMUT(A4;A5)						
8							
9							

**LECTURE 35**  
**COMBINATIONS**  
**ELEMENTARY PROBABILITY**  
**PART 1**

**OBJECTIVES**

The objectives of the lecture are to learn about:

- Review Lecture 34
- Combinations
- Elementary Probability

**COMBINATIONS**

Arrangements of objects without caring for the order in which they are arranged are called Combinations.

Number of  $n$  objects taken  $r$  at a time, denoted by  $nCr$  or  $(n)$  given by

$$nCr = \frac{n!}{r!(n-r)!}$$

**Example**

Number of combinations of 3 different objects A, B, C taken two at a time  
 $= \frac{3!}{2!(3-2)!} = \frac{6}{2} = 3$ .

These combinations are: AB, AC, and BC.

**COMBINATIONS EXAMPLES**

Here are a few examples of combinations which are based on the above formula.

**Example 1**

$$4C2 = \frac{4!}{2!(4-2)!} = \frac{4 \cdot 3 \cdot 2!}{2 \cdot 2!} = 6$$

**Example 2**

$$5C2 = \frac{5!}{2!(5-2)!} = \frac{5 \cdot 4 \cdot 3!}{2 \cdot 3!} = 10$$

**Example 3**

In how many ways a team of 11 players be chosen from a total of 15 players?

$$n = 15, r = 11$$

$$15C11 = \frac{15!}{11!(15-11)!} = \frac{15 \cdot 14 \cdot 13 \cdot 12 \cdot 11!}{11! \cdot 4!} = \frac{15 \cdot 7 \cdot 13}{4 \cdot 3 \cdot 2} = 1365 \text{ ways}$$

**Example 4**

There are 5 white balls and 4 black balls. In how many ways can we select 3 white and 2 black balls?

$$5C3 \times 4C2 = \frac{5!}{3!(5-3)!} \cdot \frac{4!}{2!(4-2)!} = 10 \cdot 6 = 60$$

**RESULTS OF SOME COMBINATIONS**

Here are some important combinations that can simplify the process of calculations for Binomial Expansion.

1.  $nC0 = nCn = 1$   
e.g.,  $4C0 = 4C4 = 1$
2.  $nC1 = nCn-1 = n$   
e.g.,  $4C1 = 4C3 = 4$
3.  $nCr = nCn-r$
4. e.g.,  $5C2 = 5C3$

**BINOMIAL EXPANSION**

An expression consisting of two terms joined by + or – sign is called a Binomial Expression. Expressions such as (a+b), (a-b), (x+y)<sup>2</sup> are examples of Binomial Expressions

We can verify that:

$$(x+y)^1 = x + y$$

$$(x+y)^2 = x^2 + 2xy + y^2$$

$$(x+y)^3 = x^3 + 3x^2y + 3xy^2 + y^3$$

$$(x+y)^4 = x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + y^4$$

Expressions on the right hand side are called Binomial Expansions.

**COEFFICIENTS OF BINOMIAL EXPANSION**

*The coefficients of the binomial expansion for any binomial expression can be written in combinatorial notation:*

$$(x+y)^5 = {}^5C_0x^5 + {}^5C_1x^4y + {}^5C_2x^3y^2 + {}^5C_3x^2y^3 + {}^5C_4xy^4 + {}^5C_5y^5$$

**Solving:**

$$= x^5 + 5x^4y + 10x^3y^2 + 10x^2y^3 + 5xy^4 + y^5$$

**CALCULATION OF BINOMIAL EXPANSION COEFFICIENTS**

Coefficient of first and last term is always 1

Coefficient of any other term = (coefficient of previous term).(power of x from previous term)/number of that term

**Example**

First term =  $x^5$

Last term =  $y^5$

Second coefficient =  $5/1 = 5$

Third coefficient =  $5 \cdot 4/2 = 10$

Fourth coefficient =  $10 \cdot 3/3 = 10$

Fifth coefficient =  $10 \cdot 2/4 = 5$

**PROJECT DEVELOPMENT MANAGER'S PROBLEM**

A toys manufacturer intends to start development of new product lines. A new toy is to be developed. Development of this toy is tied with a new TV series with the same name. There is 40% chance of TV series. The production in such a case is estimated at 12,000 units. The Profit per toy would be Rs. 2.

Without TV series-sale there may be demand for 2,000 units.

Already 500,000 Rs. has been invested.

A rival may bring to the market a similar toy. If so the sale may be 8000 units. The chance of rival bringing this toy to the market is 50%.

**Choices:**

The company has two choices:

- Abandon new product
- Risk new development

**How should the company tie it all to financial results?**

**PROBABILITY EXAMPLE 1**

How can we make assessment of chances? Look at a simple example.

A worker out of 600 gets a prize by lottery.

What is the chance of any one individual say Rashid being selected?

**Solution:**

Chance of any one individual say Rashid being selected =  $1/600$

The probability of the event "Rashid is selected" is the probability of an event occurring =  $p(\text{Rashid}) = 1/600$

This is a priori method of finding probability as we can assess the probability before the event occurred

**PROBABILITY EXAMPLE 2**

When all outcomes are equally likely a priori probability is defined as:

$p(\text{event}) = \text{Number of ways that event can occur} / \text{Total number of possible outcomes}$

If out of 600 persons 250 are women, then the chance of a woman being selected =  $p(\text{woman}) = 250/600$

**PROBABILITY - EMPIRICAL APPROACH**

In many situations, there is no prior knowledge to calculate probabilities.

What is the probability of a machine being defective?

**Method:**

1. Monitor the machine over a period of time.
2. Find out how many times it becomes defective.

This experimental or empirical approach

**EXPERIMENTAL AND THEORETICAL PROBABILITY**

$p(\text{event}) = \text{Number of times event occurs} / \text{Total number of experiments}$ .

Larger the number of experiments, more accurate the estimate.

Experimental probability approaches theoretical probability as the number of experiments becomes very large.

**OR RULE**

Consider two events A and B.

What is the probability of either A or B happening?

What is the probability of A and B happening?

What is the number of possibilities?

Probability of A or B happening =  $\text{Number of ways A or B can happen} / \text{Total number of possibilities}$

=  $\text{Number of ways A can happen} + \text{number of ways B can happen} / \text{Total number of possibilities}$

Or

=  $\text{Number of ways A can happen} / \text{Total number of possibilities} + \text{Number of ways B can happen} / \text{Total number of possibilities}$

=  $\text{Probability of A happening} + \text{Probability of B happening}$

**Condition for Or Rule**

A and B must be mutually exclusive.

When A and B are mutually exclusive:

$p(A \text{ or } B) = p(A) + p(B)$

**OR RULE EXAMPLE**

If a dice is thrown what is the chance of getting an **even number** or a number **divisible by three**?

$p(\text{even}) = 3/6$

$p(\text{div by } 3) = 2/6$

$p(\text{even or div by } 3) = 3/6 + 2/6 = 5/6$

The number 6 is not mutually exclusive.

**Hence:**

Correct answer =  $4/6$

**AND RULE**

Probability of A and B happening = Probability of A x Probability of B

**Example**

In a factory 40% workforce are women. Twenty five percent females are in management grade. Thirty percent males are in management grade. What is the probability that a worker selected is a women from management grade?

**Solution**

$$p(\text{woman chosen}) = 2/5$$

25% females = management grade

30% of males = management grade

$$p(\text{woman \& Management grade}) = p(\text{woman}) \times p(\text{management})$$

Assume that the total workforce = 100

$$p(\text{woman}) = 0.4$$

$$p(\text{ management}) = 0.25$$

$$p(\text{woman}) \times p(\text{ management}) = 0.4 \times 0.25 = 0.1 \text{ or } 10\%$$

**SET OF MUTUALLY EXCLUSIVE EVENTS**

To cover all possibilities between mutually exclusive events add up all the probabilities.

Probabilities of all these events together add up to 1.

$$p(A) + p(B) + p(C) + \dots p(N) = 1$$

**EXHAUSTIVE EVENTS**

A happens or A does not happen then A and B are Exhaustive Events.

$$p(A \text{ happens}) + p(A \text{ does not happen}) = 1$$

**Example 1**

$$p(\text{you pass}) = 0.9$$

$$p(\text{you fail}) = 1 - 0.9 = 0.1$$

**EXAMPLE1 - EXHAUSTIVE EVENTS**

A production line uses 3 machines. The Chance that 1<sup>st</sup> machine breaks down in any week is 1/10. The Chance for 2<sup>nd</sup> machine is 1/20. Chance of 3<sup>rd</sup> machine is 1/40. What is the chance that at least one machine breaks down in any week?

**Solution**

$$p(\text{at least one not working}) + p(\text{all three working}) = 1$$

$$p(\text{at least one not working}) = 1 - p(\text{all three working})$$

$$p(\text{all three working}) = p(\text{1st working}) \times p(\text{2nd working}) \times p(\text{3rd working})$$

$$p(\text{1st working}) = 1 - p(\text{1st not working}) = 1 - 1/10 = 9/10$$

$$p(\text{2nd working}) = 19/20$$

$$p(\text{3rd working}) = 39/40$$

$$p(\text{all working}) = 9/10 \times 19/20 \times 39/40 = 6669/8000$$

$$p(\text{at least 1 working}) = 1 - 6669/8000 = 1331/8000$$

**APPLICATION OF RULES**

A firm has the following rules:

When a worker comes late there is 1/4 chance that he is caught.

First time he is given a warning.

Second time he is dismissed.

What is the probability that a worker is late three times is not dismissed?

**Solution**

Let us use the denominations:

1C: Probability of being Caught first time

1NC: Probability of being Not Caught first time

2C: Probability of being Caught 2nd time

2NC: Probability of being Not Caught 2nd time

3C: Probability of being Caught 3rd time

3NC: Probability of being Not Caught 3<sup>rd</sup> time

Probabilities of different events can be calculated by applying the AND Rule.

$$1C(1/4) \& 2C(1/4) \text{ (Dismissed 1)} = (1/16 = 4/64)$$

$$1C(1/4) \& 2NC(3/4) \& 3C(1/4) \text{ (Dismissed 2)} (3/64)$$

$$1C(1/4) \& 2NC(3/4) \& 3NC(3/4) \text{ (Not dismissed 1)} (9/64)$$

$$1NC(3/4) \& 2C(1/4) \& 3C(1/4) \text{ (Dismissed 3)} (3/64)$$

$$1NC(3/4) \& 2C(1/4) \& 3NC(3/4) \text{ (Not dismissed 2)} (9/64)$$

$$1NC(3/4) \& 2NC(3/4) \& 3C(1/4) \text{ (Not dismissed 3)} (9/64)$$

$$1NC(3/4) \& 2NC(3/4) \& 3NC(3/4) \text{ (Not dismissed 4)} (27/64)$$

$$p(\text{caught first time but not the second or third time}) = \frac{1}{4} \times \frac{3}{4} \times \frac{3}{4} = 9/64$$

$$p(\text{caught only on second occasion}) = \frac{3}{4} \times \frac{1}{4} \times \frac{3}{4} = 9/64$$

$$p(\text{late three times but not dismissed}) = p(\text{not dismissed 1}) + p(\text{not dismissed 2}) +$$

$$p(\text{not dismissed 3}) + p(\text{not dismissed 4}) = 9/64 + 9/64 + 9/64 + 27/64 = 54/64$$

**p(caught) using OR Rule**

$$p(\text{caught}) =$$

$$p(\text{dismissed 1}) + p(\text{dismissed 2}) + p(\text{dismissed 3}) = 4/64 + 3/64 + 3/64$$

$$= 10/64$$

**p(caught) and p(not caught) using rule about Exhaustive events**

$$p(\text{not caught}) = 1 - p(\text{not caught})$$

$$= 1 - 10/64$$

$$= 54/64$$

**LECTURE 36**  
**ELEMENTARY PROBABILITY**  
**PART 2**

**OBJECTIVES**

The objectives of the lecture are to learn about:

- Review Lecture 35
- Elementary Probability

**PROBABILITY CONCEPTS REVIEW**

Most of the material on probability theory along with examples was included in the handout for lecture 35. You are advised to refer to handout 35. Some of the concepts and examples have been further elaborated in this handout.

Probability means making assessment of chances. The simplest example was the probability of Rashid getting the lottery when he was one of 600. The probability of the event was  $1/600$ .

**PERMUT EXAMPLE**

In handout for lecture 35, we looked at the function PERMUT, that can be used for calculations of permutations. An example is shown in the slide

	A	B	C	D	E	F	G
1	<b>PERMUT(number,number_chosen)</b>						
2							
3	<b>Data</b>	<b>Description</b>					
4	<b>100</b>	<b>Number of objects</b>					
5	<b>3</b>	<b>Number of objects in each permutation</b>					
6	<b>970200</b>						
7	<b>=PERMUT(A4;A5)</b>						
8							
9							

below.

**OR RULE REVIEW**

When two events are mutually exclusive, the probability of either one of those occurring is the sum of individual probabilities. This is the OR Rule. This is a very extensively used rule.

A and B must be mutually exclusive. The formula for the OR rule is as under.

$$p(A \text{ or } B) = p(A) + p(B)$$

**Example**

If a dice is thrown what is the chance of getting an odd number or a number divisible by two?

$$P(\text{odd}) = 3/6$$

$$p(\text{div by 3}) = 2/6$$

$$p(\text{odd or div by 3}) = 3/6 + 2/6 = 5/6$$

The number 6 is not mutually exclusive

Hence correct answer = 4/6

**AND RULE REVIEW**

The AND Rule requires that the events occur simultaneously.

**Example**

60% workforce are men.

$$p(\text{man chosen}) = 3/5$$

25% females = management grade

30% of males = management grade

What is the probability that a worker selected is a man from management grade?

**Example**

$$p(\text{man \& management grade}) = p(\text{man}) \times p(\text{management})$$

Total workforce = 100

$$p(\text{man}) = 0.6$$

$$p(\text{management}) = 0.3$$

$$p(\text{man}) \times p(\text{management}) = 0.6 \times 0.3 = 0.18 \text{ or } 18\%$$

**SET OF MUTUALLY EXCLUSIVE EVENTS REVIEW**

Between them they cover all possibilities. Probabilities of all these events together add up to 1. Exhaustive Events are events that happen or do not happen.

$$p(\text{it rains}) = 0.9$$

$$p(\text{it does not rain}) = 1 - 0.9 = 0.1$$

**Example**

In Handout for lecture 35 we studied the problem of the three machines.

A production line uses 3 machines.

Chance that 1<sup>st</sup> machine breaks down in any week was 1/10. Chance for 2<sup>nd</sup> machine was 1/20. Chance of 3<sup>rd</sup> machine was 1/40. What is the chance that at least one machine breaks down in any week?

What are the probabilities?

Probability that one or two or three machines are not working (in other words at least one not working) and that all three are working add up to 1 as exhaustive events.

$$P(\text{at least one not working}) + p(\text{all three working}) = 1$$

From the above, the probability that at least one is not working is worked out.

$$P(\text{at least one not working}) = 1 - p(\text{all three working})$$

Now to work out the probability that all three are working, we need to think in terms of machine 1 and machine 2 and machine 3 working. This means application of the AND Rule.

$p(\text{all three working}) = p(\text{1st working}) \times p(\text{2nd working}) \times p(\text{3rd working})$  Now the probability of machine 1 working is not known. The probability that machine 1 is not working is given. These two events (working and not working) are exhaustive events and add up to 1. Thus, the event that machine 1 is working,  $p(\text{1st working})$ , can be calculated as:

$$= 1 - p(\text{1st not working}) = 1 - 1/10 = 9/10$$

The calculations for the other machines are:



$$p(\text{2nd working}) = 1 - 1/20 = 19/20$$

$$p(\text{3rd working}) = 1 - 1/40 = 39/40$$

Now the combined probability of  $p(\text{all working})$  is a product of their individual probabilities using the AND Rule:

$$= 9/10 \times 19/20 \times 39/40 = 6669/8000$$

$$\text{Finally } P(\text{at least 1 working or } ) = 1 - 6669/8000 = 1331/8000$$

**LECTURE 37**  
**PATTERNS OF PROBABILITY: BINOMIAL, POISSON AND NORMAL**  
**DISTRIBUTIONS**  
**PART 1**

**OBJECTIVES**

The objectives of the lecture are to learn about:

- Review Lecture 36
- Patterns of Probability: Binomial, Poisson and Normal Distributions

**MODULE 7**

Module 7 covers the following:

Factorials

Permutations and Combinations

(Lecture 34)

Elementary Probability

(Lectures 35- 36)

Patterns of probability: Binomial, Poisson and Normal Distributions

Part 1- 4

(Lectures 37- 40)

**MODULE 8**

Module 8 covers the following.

Estimating from Samples: Inference

(Lectures 41- 42)

Hypothesis Testing: Chi-Square Distribution (Lectures 43 - 44)

Planning Production Levels: Linear Programming (Lecture 45)

Assignment Module 7- 8

End-Term Examination

**EXAMPLE 1**

We covered in the past two lectures Elementary Probability. Most of the material was included in Handout 35. Some questions were discussed in detail in handout 36. In lecture 37, the example where the employee was warned on coming late and dismissed if late twice will be discussed. The material for this example is given in handout 35. Here we shall cover the main points and the method.

A firm has the following rules:

When a worker comes late there is  $\frac{1}{4}$  chance that he is caught First time he is given a warning. Second time he is dismissed.

What is the probability that a worker is late three times is not dismissed?

**Solution**

How do we solve a problem of this nature? The answer is to develop the different options first. Let us see how it can be done.

**First time**

There are two options:

Caught: 1C

Not Caught: 1NC

**2nd time**

Caught: 2C

Not Caught: 2NC

**3<sup>rd</sup> time**

Caught: 3C

Not Caught: 3NC

**Look at combinations up to 2<sup>nd</sup> stage****1C > 2C**

1C &gt; 2NC

1NC &gt; 2C

1NC &gt; 2NC

**Look at combinations up to 3<sup>rd</sup> stage****1C & 2C**

1C &amp; 2NC &amp; 3C

1C &amp; 2NC &amp; 3NC

1NC &amp; 2C &amp; 3C

1NC &amp; 2C &amp; 3NC

1NC &amp; 2NC &amp; 3C

1NC &amp; 2NC &amp; 3NC

You saw that the first case is 1C & 2C. Here the employee was caught twice and was dismissed. He can not continue. Hence this case was closed here.

In other cases, the combinations were as given above.

Now the probability of being caught was  $\frac{1}{4}$ . As an exhaustive event the probability of not being caught was  $1 - \frac{1}{4} = \frac{3}{4}$ .

Now the probabilities can be calculated as follows:

1C & 2C ( $\frac{1}{4} \times \frac{1}{4} = \frac{1}{16}$ )1C & 2NC & 3C ( $\frac{1}{4} \times \frac{3}{4} \times \frac{1}{4} = \frac{3}{64}$ )1C & 2NC & 3NC ( $\frac{1}{4} \times \frac{3}{4} \times \frac{3}{4} = \frac{9}{64}$ )1NC & 2C & 3C ( $\frac{3}{4} \times \frac{1}{4} \times \frac{1}{4} = \frac{3}{64}$ )1NC & 2C & 3NC ( $\frac{3}{4} \times \frac{1}{4} \times \frac{3}{4} = \frac{9}{64}$ )1NC & 2NC & 3C ( $\frac{3}{4} \times \frac{3}{4} \times \frac{1}{4} = \frac{9}{64}$ )1NC & 2NC & 3NC ( $\frac{3}{4} \times \frac{3}{4} \times \frac{3}{4} = \frac{27}{64}$ )

The probabilities for each combination of events are now summarized below:

First Caught, Second Caught, Dismissed:

**1C ( $\frac{1}{4}$ ) & 2C ( $\frac{1}{4}$ ) (Dismissed 1) ( $\frac{1}{16} = \frac{4}{64}$ )**

First caught, Second Not Caught, 3<sup>rd</sup> Caught, Dismissed:

**1C ( $\frac{1}{4}$ ) & 2NC ( $\frac{3}{4}$ ) & 3C ( $\frac{1}{4}$ ) (Dismissed 2) ( $\frac{3}{64}$ )**

First caught, Second Not Caught, 3<sup>rd</sup> Not Caught, Not Dismissed

**1C ( $\frac{1}{4}$ ) & 2NC ( $\frac{3}{4}$ ) & 3NC ( $\frac{3}{4}$ ) (Not dismissed 1) ( $\frac{9}{64}$ )**

First Not Caught, Second Caught, 3<sup>rd</sup> Caught, Dismissed

**1NC ( $\frac{3}{4}$ ) & 2C ( $\frac{1}{4}$ ) & 3C ( $\frac{1}{4}$ ) (Dismissed 3) ( $\frac{3}{64}$ )**

First Not caught, Second Caught, 3<sup>rd</sup> Not Caught, Not Dismissed

**1NC ( $\frac{3}{4}$ ) & 2C ( $\frac{1}{4}$ ) & 3NC ( $\frac{3}{4}$ ) (Not dismissed 2) ( $\frac{9}{64}$ )**

First caught, Second Not Caught, 3<sup>rd</sup> Caught, Not Dismissed

**1NC ( $\frac{3}{4}$ ) & 2NC ( $\frac{3}{4}$ ) & 3C ( $\frac{1}{4}$ ) (Not dismissed 3) ( $\frac{9}{64}$ )**

First caught, Second Not Caught, 3<sup>rd</sup> Not Caught, Not Dismissed

**1NC ( $\frac{3}{4}$ ) & 2NC ( $\frac{3}{4}$ ) & 3NC ( $\frac{3}{4}$ ) (Not dismissed 4) ( $\frac{27}{64}$ )****Probabilities**p(caught) =

The probability of being caught can be calculated by thinking that these are mutually events. All situations where there was a dismissal can be considered.

Probability(caught) =

$$p(\text{dismissed 1}) + p(\text{dismissed 2}) + p(\text{dismissed 3}) = \frac{4}{64} + \frac{3}{64} + \frac{3}{64} \\ = \frac{10}{64}$$

$p(\text{not caught}) =$

Once we have the probability of being caught we can find out the probability of not being caught as an exhaustive event. Thus:

$$\begin{aligned} p(\text{not caught}) &= 1 - p(\text{caught}) \\ &= 1 - 10/64 \\ &= 54/64 \end{aligned}$$

### **EXAMPLE 2**

Two firms compete for contracts.

A has probability of  $\frac{3}{4}$  of obtaining one contract.

B has probability of  $\frac{1}{4}$ .

What is the probability that when they bid for two contracts, firm A will obtain either the first or second contract?

**Solution:**

$$P(\text{A gets first or A gets second}) = \frac{3}{4} + \frac{3}{4} = \frac{6}{4}$$

Wrong! Probability greater than 1!

We ignored the restriction: events must be mutually exclusive.

We are looking for probability that A gains the first or second or both.

We are not interested in B getting both the contracts

$$p(\text{B gets first}) \times p(\text{B gets both}) = \frac{1}{4} \times \frac{1}{4} = \frac{1}{16}$$

$$p(\text{A gets one or both}) = 1 - \frac{1}{16} = \frac{15}{16}$$

### **Alternative Method**

Split "A gets first or the second or both" into 3 parts

$$\text{A gets first but not second} = \frac{3}{4} \times \frac{1}{4} = \frac{3}{16}$$

$$\text{A does not get first but gets second} = \frac{1}{4} \times \frac{3}{4} = \frac{3}{16}$$

$$\text{A gets both} = \frac{3}{4} \times \frac{3}{4} = \frac{9}{16}$$

$$P(\text{A gets first or second or both}) = \frac{3}{16} + \frac{3}{16} + \frac{9}{16} = \frac{15}{16}$$

### **EXAMPLE 3**

In a factory 40% workforce is female.

25% females belong to the management cadre.

30% males are from management cadre.

If management grade worker is selected, what is the probability that it is a female?

Draw up a table first.

	Male	Female	Total
Management	?	?	?
Non-Management	?	?	?
Total	?	40	100

**Calculate**

$$\text{Total male} = 100 - 40 = 60$$

$$\text{Management female} = 0.25 \times 40 = 10$$

$$\text{Non-Management female} = 40 - 10 = 30$$

$$\text{Management male} = 0.3 \times 60 = 18$$

$$\text{Non-Management male} = 60 - 18 = 42$$

$$\text{Management total} = 18 + 10 = 28$$

$$\text{Non-Management total} = 42 + 30 = 72$$

**Summary**

	Male	Female	Total
<b>Management</b>	<b>18</b>	<b>10</b>	<b>28</b>
<b>Non-Management</b>	<b>42</b>	<b>30</b>	<b>72</b>
<b>Total</b>	<b>60</b>	<b>40</b>	<b>100</b>

$$p(\text{management grade worker is female}) = 10/28$$

**EXAMPLE 4**

A pie vendor has collected data over sale of pies. This data is organized as follows:

No. Pies sold	Income(X)	% Days(f)	fX Rs.
40	$x 35 = 1400$	20	28000
50	1750	20	35000
60	2100	30	63000
70	2450	20	49000
80	2800	10	28000
Total		100	203000
Mean/day = $203000/100 = 2030$			

The selling price per pie was Rs. 35. What was the mean sale per day?

Such a question can be solved by calculating the sale in each slab and then dividing the total sale by number of pies.

% days is the probability. If multiplied with the income from each pie, the expected sale from all pies can be calculated. The overall expected value was 203,000. When divided by the number of days (100) an average of 2,030 Rs. Per day was obtained as average sale per day.

**EXPECTED VALUE**

$$EMV = \sum (\text{probability of outcome} \times \text{financial result of outcome})$$

**Example**

In an insurance company 80% of the policies have no claim.

In 15% cases the Claim is 5000 Rs.

For the remaining 5% the Claim is 50000 Rs.

What is the Expected value of claim per policy?

Applying the formula above:

$$EMV = 0.8 \times 0 + 0.15 \times 5000 + 0.05 \times 50000$$

$$= 0 + 750 + 2500$$

$$= 3250 \text{ Rs.}$$

**TYPICAL PRODUCTION PROBLEM**

In a factory producing biscuits, the packing machine breaks 1 biscuit out of twenty ( $p = 1/20 = 0.05$ ).

What proportion of boxes will contain more than 3 broken biscuits?

This is a typical Binomial probability situation!

The individual biscuit is broken or not

= two possible outcomes

**Conditions for Binomial Situation**

1. Either or situation
2. Number of trials ( $n$ ) known and fixed
3. Probability for success on each trial ( $p$ ) is known and fixed

**CUMULATIVE BINOMIAL PROBABILITIES**

The Cumulative Probability table gives the probability of  $r$  or more successes in  $n$  trials, with the probability  $p$  of success in one trial

In the table:

The total number of trials  $n = 1$  to 10

The number of successes  $r = 1$  to 10

The probability  $p = 0.05, 0.1, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5$

**LECTURE 38**  
**PATTERNS OF PROBABILITY: BINOMIAL, POISSON AND NORMAL**  
**DISTRIBUTIONS**  
**PART 2**

**OBJECTIVES**

The objectives of the lecture are to learn about:

- Review Lecture 37
- Patterns of Probability: Binomial, Poisson and Normal Distributions

**CUMULATIVE BINOMIAL PROBABILITIES**

Probability of r or more successes in n trials with the probability of success in each trial

- Look in column for n
- Look in column for r
- Look at column for value of p(0.05 to 0.5)

**Example**

n = 5; r = 4; p = 0.5

p( 4 or more successes in 5 trials)

= 0.1874 = 18.74 %

**BINOMDIST**

Returns the individual term binomial distribution probability. Use BINOMDIST in problems with a fixed number of tests or trials, when the outcomes of any trial are only success or failure, when trials are independent, and when the probability of success is constant throughout the experiment. For example, BINOMDIST can calculate the probability that two of the next three babies born are male.

**Syntax**

**BINOMDIST(number\_s, trials, probability\_s, cumulative)**

**Number\_s** is the number of successes in trials.

**Trials** is the number of independent trials.

**Probability\_s** is the probability of success on each trial.

**Cumulative** is a logical value that determines the form of the function. If cumulative is TRUE, then BINOMDIST returns the cumulative distribution function, which is the probability that there are at most number\_s successes; if FALSE, it returns the probability mass function, which is the probability that there are number\_s successes.

**Remarks**

- Number\_s and trials are truncated to integers.
- If number\_s, trials, or probability\_s is nonnumeric, BINOMDIST returns the #VALUE! error value.
- If number\_s < 0 or number\_s > trials, BINOMDIST returns the #NUM! error value.
- If probability\_s < 0 or probability\_s > 1, BINOMDIST returns the #NUM! error value.
- The binomial probability mass function is:

$$b(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$

where:

$$\binom{n}{x},$$

The cumulative binomial distribution is:

$$B(x; n, p) = \sum_{y=0}^x b(y; n, p)$$

Microsoft Excel - Lecture\_38

File Edit View Insert Format Tools Data Window Help

€ +.0 -.00

SUM X ✓ ✕ =BINOMDIST(A3;A4;A5;FALSE)

	A	B	C	D	E
1	<b>BINOMDIST(number_s, trials, probability_s, cumulative)</b>				
2	<b>Data</b>	<b>Description</b>			
3	6	Number of successes in trials			
4	10	Number of independent trials			
5	0.5	Probability of success on each trial			
6					
7	<b>=BINOMDIST(A3;A4;A5;FALSE)</b>				
8	<b>Probability of exactly 6 of 10 trials</b>				
9	<b>being successful (0.205078)</b>				
10					

**Example**

In the above example, the BINOMDIST function was used to calculate the probability of exact 6 out of 10 trials being successful. Here the value of Cumulative was set as False. The following example also shows a similar calculation.

Microsoft Excel - Lecture\_38

File Edit View Insert Format Tools Data Window Help

€ +.0 -.00

B33 X ✓ ✕

	A	B	C
19	<b>EXAMPLE</b>		
20	<b>Five coins are tossed simultaneously</b>		
21	<b>What is the chance of obtaining 3 heads?</b>		
22			
23	<b>Data</b>	<b>Description</b>	
24	3	Number of successes in trials	
25	5	Number of independent trials	
26	0.5	Probability of success on each trial	
27			
28	<b>0.3125</b>		
29	<b>Probability of exactly 3 of 5 trials</b>		
30	<b>being successful (0.3125)</b>		
31			

**EXAMPLE USING TABLES**



We have the probability of 3 or more dry days in a week. What is the chance of getting 5 or more wet days next week?

$$n = 7; r = 3; p = 0.4$$

From the tables, the probability of 3 or more in a sample of 7 was found as 0.5800.

$$p(3 \text{ or more dry days}) = 0.5800$$

Now:

$$p(2 \text{ or less dry days}) + p(3 \text{ or more dry days}) = 1$$

$$p(2 \text{ or less dry days}) = 1 - p(3 \text{ or more dry days})$$

$$p(2 \text{ or less dry days}) = 1 - 0.5800 = 0.4200$$

= Chance of 5 or more wet days next week.

Note that we thought in terms of 2 or less dry days. In reality, it means 5 or more wet days which we wanted to find out.

### **EXAMPLE 1**

The probability of wet days is 60%. Note that the figure 0.6 is beyond the maximum value 0.5 as given in the tables. Let us first convert our problem to  $p(\text{dry}) = 1 - 0.6 = 0.4$ . Now  $p(5 \text{ or more wet days})$  can be restated as  $p(2 \text{ or less dry days})$ . The BINOMDIST function is for  $p(r \text{ or more})$ . Let us convert  $p(2 \text{ or less dry days})$  to  $1 - p(3 \text{ or more days})$ . Now the value of  $n = 7$ ,  $r = 3$  and  $p = 0.4$ .

Using BINOMDIST, the answer is 0.4199. Note that the value of cumulative was TRUE.

	A	B	C
35	<b>EXAMPLE</b>		
36	<b>Probability of wet days in current month = 60%</b>		
37	<b>Probability of 5 or more wet days next week?</b>		
38	<b>p = 0.6. The tables are for p upto 0.5.</b>		
39	<b>Turn the question. P(dry) = 1 - 0.6 = 0.4</b>		
40	<b>p( 5 or more wet days) = (7 - 5 = 2 or less dry days)</b>		
41	<b>p( 2 or less dry days) = 1 -p( 3 or more dry days)</b>		
42	<b>Data</b>	<b>Description</b>	
43	<b>2</b>	<b>Number of successes in trials</b>	
44	<b>7</b>	<b>Number of independent trials</b>	
45	<b>0.4</b>	<b>Probability of success on each trial</b>	
46	<b>0.4199</b>	<b>=BINOMDIST(A43;A44;A45;TRUE)</b>	
47	<b>At most 2 successes: 1, 2 (0.4199)</b>		

**EXAMPLE 2**

In a transmission where 8 bit message is transmitted electronically there is 10% probability of one bit being transmitted erroneously? What is the chance that entire message is transmitted correctly?

We can state that the probability required is for 0 successes (errors) in 8 trials (bits).

$p(\text{one bit transmitted erroneously}) = 0.1$

$n = 8; r = 8, p = 0.1; p(\text{exactly 0 errors})?$

$p(0 \text{ errors}) + p(1 \text{ or more errors}) = 1$

$p(0 \text{ errors}) = 1 - p(1 \text{ or more errors})$

**From the Tables**

$p(1 \text{ or more})$  is 0.5695.

Hence  $p(0) = 1 - 0.5695 = 0.4305$

**Using BINOMDIST**

The data was for 0 or more successes. BINOMDIST function gives the value for at most r successes. Hence the answer was obtained directly.

	A	B	C	D
51	<b>EXAMPLE</b>			
52	<b>Probability of 1 erroneous bit = 0.1</b>			
53	<b>Probability of 8 correct bits? = 0 erroneous bits</b>			
54	<b>Data</b>	<b>Description</b>		
55		<b>0</b>	<b>Number of successes in trials</b>	
56		<b>8</b>	<b>Number of independent trials</b>	
57		<b>0.1</b>	<b>Probability of success on each trial</b>	
58	<b>0.4305</b>	<b>=BINOMDIST(B55;B56;B57;TRUE)</b>		
59	<b>At most 0 successes: (0.4305)</b>			
60				

**EXAMPLE 3**

A surgery is successful for 75% patients. What is the probability of its success in at least 7 cases out of randomly selected 9 patients?

$p(\text{success in at least 7 cases in randomly selected 9 patients})?$

Here

$n = 9; p(\text{success}) = 0.75; p(\text{at least 7 cases})?$

$p = 0.75$  is outside the table

Let us invert the problem.

$p(\text{failure}) = 1 - 0.75 = 0.25$

Success at least 7 = Failure 2 or less

$P(\text{failure 2 or less}) = 1 - p(\text{failure 3 or more})$

$= 1 - 0.3993 = 0.6007 = 60\%$

**Calculation using BINOMDIST**

Here the question was inverted.

We had to find 7 successes out of 9. The probability was 75% for success. It becomes 25% for failure. Now let us restate the problem in terms of failure.

We are interested in 7 or more successes. It means 2 or less failures.

Now the BINOMDIST function gives us at most  $r$  successes. In other words 2 or less. Hence if we specify  $r = 2$ , we get the answer 0.6007 directly.

	A	B	C	D	E
76	<b>EXAMPLE</b>				
77	<b>Probability of success = 0.75. Failure = 0.25</b>				
78	<b>Probability of 7 successes out of 9? Or 2 or less failures</b>				
79					
80	<b>Data</b>	<b>Description</b>			
81		<b>2</b>	<b>Number of successes in trials</b>		
82		<b>9</b>	<b>Number of independent trials</b>		
83		<b>0.25</b>	<b>Probability of success on each trial</b>		
84		<b>0.6007</b>	<b>=BINOMDIST(B81;B82;B83;TRUE)</b>		
85	<b>At most 2 successes: (0.6007)</b>				
86					

**NEGBINOMDIST**

Returns the negative binomial distribution. NEGBINOMDIST returns the probability that there will be number\_f failures before the number\_s-th success, when the constant probability of a success is probability\_s. This function is similar to the binomial distribution, except that the number of successes is fixed, and the number of trials is variable. Like the binomial, trials are assumed to be independent.

For example, you need to find 10 people with excellent reflexes, and you know the probability that a candidate has these qualifications is 0.3. NEGBINOMDIST calculates the probability that you will interview a certain number of unqualified candidates before finding all 10 qualified candidates.

**Syntax**

**NEGBINOMDIST(number\_f,number\_s,probability\_s)**

**Number\_f** is the number of failures.

**Number\_s** is the threshold number of successes.

**Probability\_s** is the probability of a success.

**Remarks**

- **Number\_f** and **number\_s** are truncated to integers.
- If any argument is nonnumeric, NEGBINOMDIST returns the #VALUE! error value.
- If probability\_s < 0 or if probability > 1, NEGBINOMDIST returns the #NUM! error value.
- If (number\_f + number\_s - 1) ≤ 0, NEGBINOMDIST returns the #NUM! error value.
- The equation for the negative binomial distribution is:

$$nb(x; r, p) = \binom{x+r-1}{r-1} p^r (1-p)^x$$

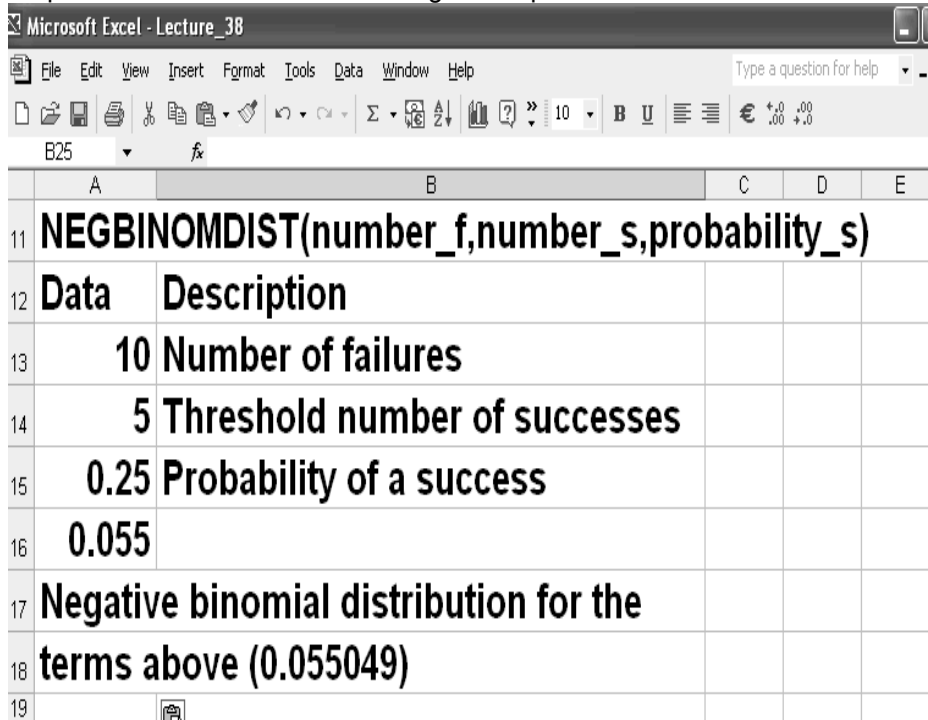
where:

x is number\_f, r is number\_s, and p is probability\_s.

### **NEGBINOMDIST- EXAMPLE**

You need to find 10 people with excellent reflexes, and you know the probability that a candidate has these qualifications is 0.3

NEGBINOMDIST calculates the probability that you will interview a certain number of unqualified candidates before finding all 10 qualified candidates.



The screenshot shows a Microsoft Excel spreadsheet with the following content:

	A	B	C	D	E
11	<b>NEGBINOMDIST(number_f,number_s,probability_s)</b>				
12	<b>Data</b>	<b>Description</b>			
13	<b>10</b>	<b>Number of failures</b>			
14	<b>5</b>	<b>Threshold number of successes</b>			
15	<b>0.25</b>	<b>Probability of a success</b>			
16	<b>0.055</b>				
17	<b>Negative binomial distribution for the</b>				
18	<b>terms above (0.055049)</b>				
19					

### **CRITBINOM**

Returns the smallest value for which the cumulative binomial distribution is greater than or equal to a criterion value. Use this function for quality assurance applications. For example, use CRITBINOM to determine the greatest number of defective parts that are allowed to come off an assembly line run without rejecting the entire lot.

#### **Syntax**

**CRITBINOM(trials,probability\_s,alpha)**

**Trials** is the number of Bernoulli trials.

**Probability\_s** is the probability of a success on each trial.

**Alpha** is the criterion value.

#### **Remarks**

- If any argument is nonnumeric, CRITBINOM returns the #VALUE! error value.
- If trials is not an integer, it is truncated.
- If trials < 0, CRITBINOM returns the #NUM! error value.
- If probability\_s is < 0 or probability\_s > 1, CRITBINOM returns the #NUM! error value.
- If alpha < 0 or alpha > 1, CRITBINOM returns the #NUM! error value.

### **Example**

---

	<b>A</b>	<b>B</b>
1	<b>Data</b>	<b>Description</b>
2	6	Number of Bernoulli trials
3	0.5	Probability of a success on each trial
4	0.75	Criterion value
	<b>Formula</b>	<b>Description (Result)</b>
	=CRITBINOM(A2,A3,A4)	Smallest value for which the cumulative binomial distribution is greater than or equal to a criterion value (4)

**LECTURE 39**  
**PATTERNS OF PROBABILITY: BINOMIAL, POISSON AND NORMAL**  
**DISTRIBUTIONS**  
**PART 3**

**OBJECTIVES**

The objectives of the lecture are to learn about:

- Review Lecture 38
- Patterns of Probability: Binomial, Poisson and Normal Distributions

**CRITBINOM EXAMPLE**

The example shown under CRITBINOM in Handout 38 is shown below.

	A	B	C	D
65	<b>CRITBINOM(trials, probability_s, alpha)</b>			
66				
67	<b>Data</b>	<b>Description</b>		
68	<b>6</b>	<b>Number of Bernoulli trials</b>		
69	<b>0.5</b>	<b>Probability of a success on each trial</b>		
70	<b>0.75</b>	<b>Criterion value</b>		
71	<b>=</b>	<b>Smallest value for which the cumulative</b>		
72	<b>CRITBINOM</b>	<b>binomial distribution is greater than</b>		
73	<b>M(A68;</b>	<b>or equal to a criterion value (4)</b>		
74	<b>A69;A70)</b>			
75				

**EXPECTED VALUE EXAMPLE**

A lottery has 100 Rs. Payout on average 20 turns.

Is it worthwhile to buy the lottery if the ticket price is 10 Rs.?

Expected win per turn =  $p(\text{winning}) \times \text{gain per win} + p(\text{losing}) \times \text{loss if you loose}$

$$= \frac{1}{20} \times (100 - 10) + \frac{19}{20} \times (-10) \text{ Rs.}$$

$$= \frac{90}{20} - \frac{190}{20} \text{ Rs.}$$

$$= 4.5 - 9.5 = -5 \text{ Rs.}$$

So on an average you stand to loose 5 Rs.

**DECISION TABLES**

Look at the data in the table below:

No. of Pies demanded	% Occasions
25	10
30	20
35	25
40	20
45	15
50	10

Price per pie = Rs. 15

Refund on return = Rs. 5

Sale price = Rs. 25

Profit per pie = Rs. 25 – 15 = Rs. 10

Loss on each return = Rs. 15 – 5 = Rs. 10

How many pies should be bought for best profit?

To solve such a problem, a decision table is set up as shown below. The values in the first column are number of pies to be purchased. Figures in columns are the sale with % share of sale within brackets. If the number of pies bought is less than the number that can be sold, the number of pies sold remains constant. If the number of pies bought exceeds the number of pies sold then the remaining are returned. This means a loss. For every value the sum of profit for sale and loss for pies returned is calculated.

The average sale for each row is calculated by multiplying the profit for each sale with % sale in the column. An example calculation is given as a guide for 30 pies.

### **DECISION TABLES**

	25(0.1)	30(0.2)	35(0.25)	40(0.2)	45(0.15)	50(0.1)	EMV
25	250	250	250	250	250	250	250
30	200	300	300	300	300	300	290
35	150	250	350	350	350	350	310
Buy							
40	100	200	300	400	400	400	305
45	50	150	250	350	450	450	280
–	0	100	200	300	400	500	240

### **Expected profit 30 pies**

$$= 0.1 \times 200 + 0.2 \times 300 + 0.25 \times 300 + 0.2 \times 300 + 0.15 \times 300 + 0.1 \times 300$$

$$= 20 + 60 + 75 + 60 + 45 + 30$$

$$= 290 \text{ Rs.}$$

### **Best Profit**

It may be noted that the best profit is for 35 Pies = Rs. 310

### **DECISION TREE TOY MANUFACTURING CASE**

The problem of the manufacturer intending to start manufacturing a new toy under the conditions that the TV series may or may not appear, that the rival may or may not sell a similar toy is now solved below.

Here a Decision tree has been developed with the possible branches as shown below. Each sequence represents an application of the AND rule.

1A Abandon

1B Go ahead >2A: Series appears (60%)

>2B: No series (40%)

>2A>3A: Rival markets (50%)

>2A>3B: No Rival (50%)

### **Production**

Series, no rival = 12000 units

Series, rival = 8000 units

No series = 2000 units

Investment = Rs. 500000

Profit per unit = Rs. 200

Loss if abandon = Rs. 500000

What is the best course of action?

**Decision Tree**

Profit if rival markets, series appears =  $8000 \times 200 - 500000 = 1600000 - 500000 = 1100000$  Rs.

Profit if no rivals =  $12000 \times 200 - 500000 = 2400000 - 500000 = 1900000$  Rs.

Profit/Loss if no series =  $2000 \times 200 - 500000 = 400000 - 500000 = -100000$  Rs. (No series)

EMV = Rival markets and no rivals =  $0.5 \times 1100000 + 0.5 \times 1900000 = 1500000$  (Series)

EMV =  $0.6 \times 1500000 + 0.4 \times -100000 = 900000 - 40000 = 860000$  Rs.

**Conclusion**

It is clear that in spite of the uncertainty, there is a likelihood of a reasonable profit.

Hence the conclusion is:

Go ahead

**THE POISSON DISTRIBUTION**

The POISSON Distribution has the following characteristics:

- Either or situation
- No data on trials
- No data on successes
- Average or mean value of successes or failures

This is a typical Poisson Situation.

**Characteristics**

- Either/or situation
- Mean number of successes per unit,  $m$ , known and fixed
- $p$ , chance, unknown but small, (event is unusual)

**THE POISSON TABLES OF PROBABILITIES**

Gives cumulative probability of  $r$  or more successes

Knowledge of  $m$  is required.

Table gives the probability of that  $r$  or more random events are contained in an interval when the average number of events per interval is  $m$

**Example 1**

$m = 7$ ;  $r = 9$ ;

$P(r \text{ or more successes}) = 0.2709$

Values given in 4 decimals

**Example 2**

Attendance in a factory shows 7 absences.

What is the probability that on a given day there will be more than 8 people absent?

**Solution**

$m = 7$

$r = \text{More than } 8 = 9 \text{ or more}$

$p(9 \text{ or more successes}) = 0.2709$

**Example 3**

An automatic production line breaks down every 2 hours.

Special production requires uninterrupted operation for 8 hours.

What is the probability that this can be achieved?

**Solution**

$m = 8/2 = 4$

$r = 0$  (no breakdown)

$p(0 \text{ breakdown}) = 1 - p(1 \text{ or more breakdowns})$

$= 1 - 0.9817 = 0.0183 = 1.83\%$



**Example 4**

An automatic packing machine produces on an average one in 100 underweight bags. What is the probability that 500 bags contain less than three underweight bags?

**Solution**

$$\begin{aligned}m &= 1 \times 500/100 = 5 \\p(r = \text{less than three}) &= 1 - p(r = 3 \text{ or more}) \\&= 1 - 0.8753 \\&= 0.1247 \\&= 12.47\%\end{aligned}$$

**Example 5**

Faulty apple toffees in a production line average out at 6 per box. The management is willing to replace one box in a hundred. What is the number of faulty toffees that this probability corresponds to?

**Solution**

$$\begin{aligned}p &= 1/100 = 0.1 \\m &= 6\end{aligned}$$

**Look for value of p close to 0.1**

$$p(r = 12) = 0.0201$$

$$p(r = 13) = 0.0088$$

Hence 13 or more faulty toffees correspond to this probability.

**LECTURE 40**  
**PATTERNS OF PROBABILITY: BINOMIAL, POISSON AND NORMAL**  
**DISTRIBUTIONS**  
**PART 4**

**OBJECTIVES**

The objectives of the lecture are to learn about:

- Review Lecture 39
- Patterns of Probability: Binomial, Poisson and Normal Distributions Part 4

**POISSON WORKSHEET FUNCTION**

Returns the Poisson distribution. A common application of the Poisson distribution is predicting the number of events over a specific time, such as the number of cars arriving at a toll plaza in 1 minute.

**Syntax**

**POISSON(x,mean,cumulative)**

X is the number of events.

Mean is the expected numeric value.

Cumulative is a logical value that determines the form of the probability distribution returned. If cumulative is TRUE, POISSON returns the cumulative Poisson probability that the number of random events occurring will be between zero and x inclusive; if FALSE, it returns the Poisson probability mass function that the number of events occurring will be exactly x.

**Remarks**

- If x is not an integer, it is truncated.
- If x or mean is nonnumeric, POISSON returns the #VALUE! error value.
- If  $x \leq 0$ , POISSON returns the #NUM! error value.
- If  $\text{mean} \leq 0$ , POISSON returns the #NUM! error value.
- POISSON is calculated as follows.

For cumulative = FALSE:

$$POISSON = \frac{e^{-\lambda} \lambda^x}{x!}$$

For cumulative = TRUE:

$$CUMPOISSON = \sum_{k=0}^x \frac{e^{-\lambda} \lambda^k}{k!}$$

**Example**

An application of the POISSON function is shown below. In this slide the value of Cumulative was TRUE. It means that the probability is for at the most case.

The screenshot shows an Excel spreadsheet titled "Microsoft Excel - Lecture\_39". The formula bar displays `=POISSON(A3;A4;TRUE)`. The spreadsheet content is as follows:

	A	B	C	D	E
1	<b>POISSON(x, mean, cumulative)</b>				
2	<b>Data</b>	<b>Description</b>			
3	<b>2</b>	<b>Number of events</b>			
4	<b>5</b>	<b>Expected mean</b>			
5	<b>=</b>				
6	<b>POIS</b>	<b>Cumulative Poisson probability with</b>			
7	<b>SON(</b>	<b>the terms above (0.124652)</b>			
8	<b>A3;</b>				
9	<b>A4;</b>				
10	<b>TRUE</b>				
11	<b>)</b>				

In the slide below the Cumulative is FALSE, which means that the probability is for exactly 2 events.

The screenshot shows an Excel spreadsheet titled "Microsoft Excel - Lecture\_39". The formula bar displays `=POISSON(A17;A18;FALSE)`. The spreadsheet content is as follows:

	A	B	C	D	E
15	<b>POISSON(x, mean, cumulative)</b>				
16	<b>Data</b>	<b>Description</b>			
17	<b>2</b>	<b>Number of events</b>			
18	<b>5</b>	<b>Expected mean</b>			
19	<b>=</b>				
20	<b>POISSON(</b>	<b>Cumulative Poisson probability with</b>			
21	<b>A17;A18;</b>	<b>the terms above (0.084224337)</b>			
22	<b>FALSE)</b>				

**THE PATTERN**

In Binomial and Poisson the situations are: either/or  
 Number of times could be counted.

In the Candy problem with underweight boxes, there is measurement of weight.

Binomial and Poisson are discrete probability distributions.

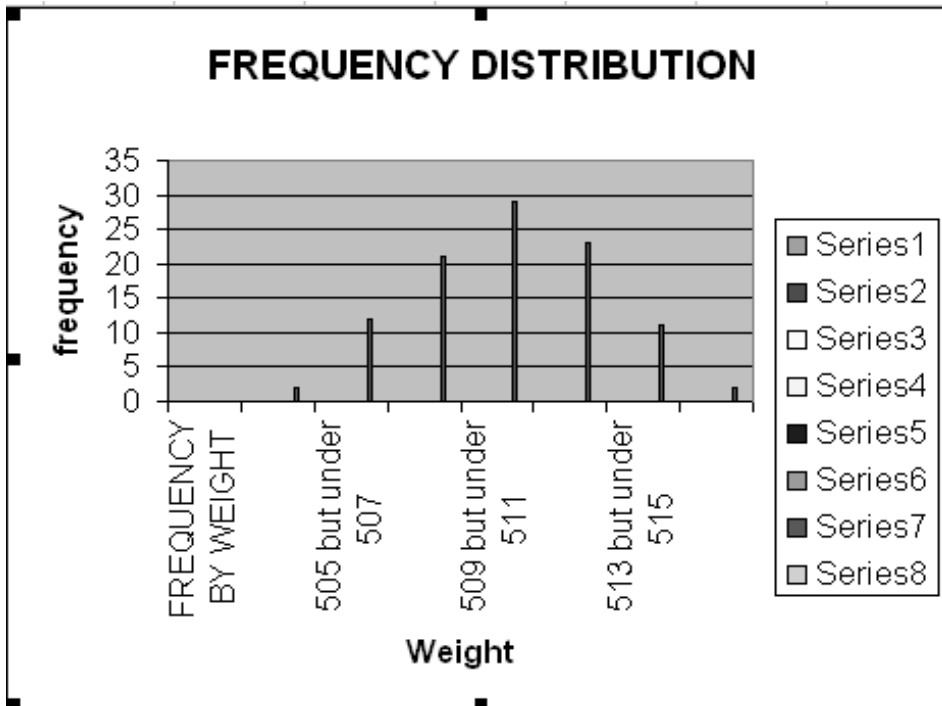
Candy problem is a Continuous probability distribution. Such problems need a different treatment.

**FREQUENCY BY WEIGHT**

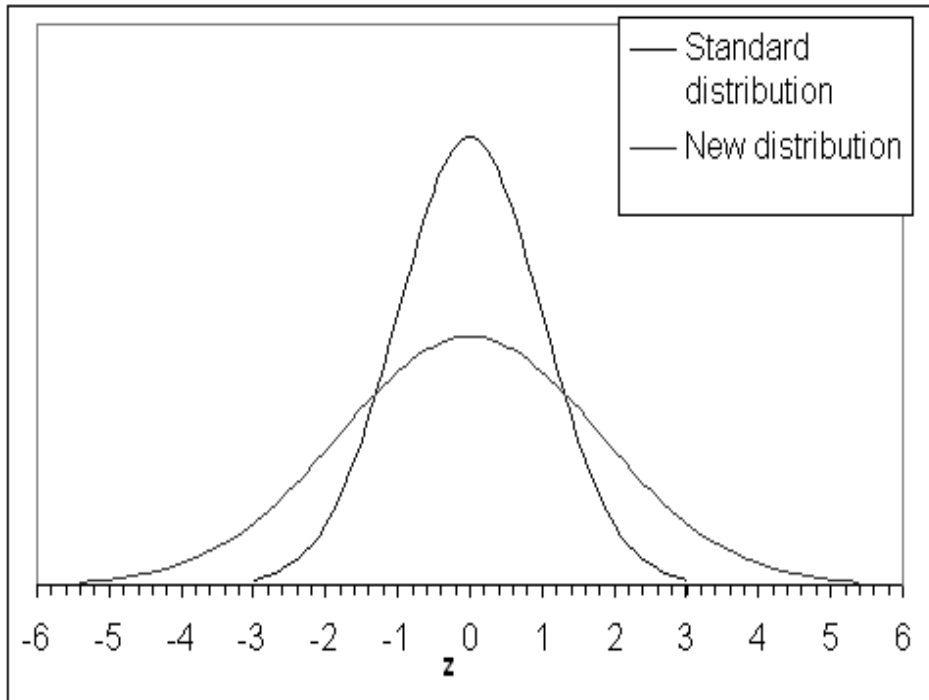
Look at the frequency distribution of weight of sample bags.

FREQUENCY BY WEIGHT		No. of bags
503 but under 505		2
505 but under 507		12
507 but under 509		21
509 but under 511		29
511 but under 513		23
513 but under 515		11
515 but under 517		2

Frequency distribution graph of the sample is shown below. You may see a distinct shape in the graph. It appears to be symmetrical.



The shape of the distribution is that of a Normal Distribution as shown as New distribution in the slide below. On this slide you also see a Standard Normal Distribution with 0 mean and standard deviations 1, 2, 3, 4 etc.



### **NORMAL DISTRIBUTION**

The blue Curve is a typical Normal Distribution.

A standard normal distribution is a distribution with mean = 0 and standard deviation = 1.

The Y-axis gives the probability values.

The X-axis gives the z (measurement) values.

Each point on the curve corresponds to the probability p that a measurement will yield a particular z value (value on the x-axis.).

Probability is a number from 0 to 1.

Percentage probabilities –multiply p by 100.

Area under the curve must be one.

Note how the probability is essentially zero for any value z that is greater than 3 standard deviations away from the mean on either side.

Mean gives the peak of the curve.

Standard deviation gives the spread.

### **Weight distribution case**

Mean = 510 g

StDev = 2.5 g

What proportion of bags weigh more than 515 g?

Proportion of area under the curve to the right of 515 g gives this probability

### **AREA UNDER THE STANDARD NORMAL CURVE**

The normal distribution table gives the area under one tail only.

z-value

Ranges between 0 and 4 in first column.

Ranges between 0 and 0.09 in other columns.

### **Example**

Find area under one tail for z-value of 2.05.

- Look in column 1. Find 2.0.
- Look in column 0.05 and go to intersection of 2.0 and 0.05.
- The area (cumulative probability of a value greater than 2.05) is the value at the intersection = 0.02018 or 2.018%

**CALCULATING Z-VALUES**

$$z = (\text{Value } x - \text{Mean})/\text{StDev}$$

Process of calculating z from x is called Standardisation.

z indicates how many standard deviations the point is from the mean

**Example 1**

Find proportion of bags which have weight in excess of 515 g.

Mean = 510. StDev = 2.5 g

**Solution**

$$z = (515 - 510)/2.5 = 2$$

From tables: Area under tail = 0.02275 or 2.28%

**Example 2**

What percentage of bags filled by the machine will weigh less than 507.5 g?

Mean = 510 g; StDev = 2.5 g

**Solution**

$$z = (507.5 - 510)/2.5 = -1$$

Look at value of z = +1

Area = 0.158

Hence:

15.8% bags weigh less than 507.5 g

**Example 3**

What is the probability that a bag filled by the machine weighs less than 512 g?

$$z = (512 - 510)/2.5 = 0.8$$

**Solution**

Area under right tail = 0.2119

= p(weighs more than 512)

p(weighs less than 512) = 1 - p(weighs more than 512)

= 1 - 0.2119

= 0.7881

**Example 4**

What percentage of bags weigh between 512 and 515?

$$z_1 = (512 - 510)/2.5 = 0.8$$

**Solution**

Area 1 = 0.2119

$$z_2 = (515 - 510)/2.5 = 2$$

Area 2 = 0.02275

p(bags weighs between 512 and 515) =

Area 1 - Area 2

= 0.2119 - 0.02275

= 0.18915 = 18.9%

**LECTURE 41**  
**ESTIMATING FROM SAMPLES: INFERENCE**  
**PART 1**

**OBJECTIVES**

The objectives of the lecture are to learn about:

- Review Lecture 40
- Estimating from Samples: Inference

**NORMDIST**

Returns the normal distribution for the specified mean and standard deviation. This function has a very wide range of applications in statistics, including hypothesis testing.

**Syntax**

**NORMDIST(x,mean,standard\_dev,cumulative)**

**X** is the value for which you want the distribution.

**Mean** is the arithmetic mean of the distribution.

**Standard\_dev** is the standard deviation of the distribution.

**Cumulative** is a logical value that determines the form of the function. If cumulative is TRUE, NORMDIST returns the cumulative distribution function; if FALSE, it returns the probability mass function.

**Remarks**

- If mean or standard\_dev is nonnumeric, NORMDIST returns the #VALUE! error value.
- If standard\_dev ≤ 0, NORMDIST returns the #NUM! error value.
- If mean = 0, standard\_dev = 1, and cumulative = TRUE, NORMDIST returns the standard normal distribution, NORMSDIST.
- The equation for the normal density function (cumulative = FALSE) is:

$$f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

- When cumulative = TRUE, the formula is the integral from negative infinity to x of the given formula.

**Example**

In the slide the x value is 42. Arithmetic mean is 40. Standard deviation is 1.5. The cumulative distribution is 0.9.



The screenshot shows an Excel spreadsheet with the following content:

	A	B	C
1	<b>NORMDIST(x,mean,standard_dev,cumulative)</b>		
2	<b>Data Description</b>		
3	<b>42</b>	<b>Value for which you want the distribution</b>	
4	<b>40</b>	<b>Arithmetic mean of the distribution</b>	
5	<b>1.5</b>	<b>Standard deviation of the distribution</b>	
6	<b>0.9</b>		
7	<b>Cumulative distribution function for</b>		
8	<b>the terms above (0.908789)</b>		
9			

**NORMSDIST**

Returns the standard normal cumulative distribution function. The distribution has a mean of 0 (zero) and a standard deviation of one. Use this function in place of a table of standard normal curve areas.

**Syntax****NORMSDIST(z)**

**z** is the value for which you want the distribution.

**Remarks**

- If **z** is nonnumeric, NORMSDIST returns the #VALUE! error value.
- The equation for the standard normal density function is:

**Example**

The input to the NORMSDIST function is the z-value. The output is the cumulative probability distribution. In the example  $z = 1.333333$ . The normal cumulative probability function is 0.908789.

	A	B
11	<b>NORMSDIST(z)</b>	
12	<b>A</b>	<b>B</b>
13	<b>Formula Description (Result)</b>	
14	<b>=NORMSDIST(1.333333)</b>	
15	<b>Normal cumulative distribution function</b>	
16	<b>at 1.333333 (0.908789)</b>	
17		

**NORMINV**

Returns the inverse of the normal cumulative distribution for the specified mean and standard deviation.

**Syntax**

**NORMINV(probability,mean,standard\_dev)**

**Probability** is a probability corresponding to the normal distribution.

**Mean** is the arithmetic mean of the distribution.

**Standard\_dev** is the standard deviation of the distribution.

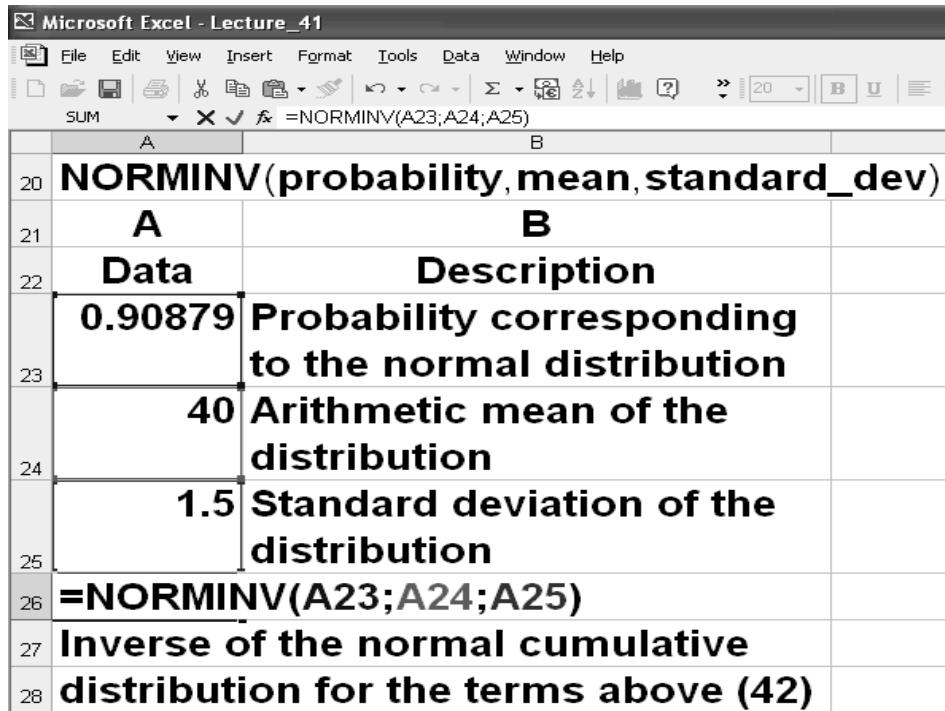
**Remarks**

- If any argument is nonnumeric, NORMINV returns the #VALUE! error value.
- If probability < 0 or if probability > 1, NORMINV returns the #NUM! error value.
- If standard\_dev ≤ 0, NORMINV returns the #NUM! error value.
- If mean = 0 and standard\_dev = 1, NORMINV uses the standard normal distribution (see NORMSINV).

NORMINV uses an iterative technique for calculating the function. Given a probability value, NORMINV iterates until the result is accurate to within  $\pm 3 \times 10^{-7}$ . If NORMINV does not converge after 100 iterations, the function returns the #N/A error value.

**Example**

Here the probability value, arithmetic mean and standard deviation are given. The answer is the x-value.



	A	B
20	<b>NORMINV(probability, mean, standard_dev)</b>	
21	<b>A</b>	<b>B</b>
22	<b>Data</b>	<b>Description</b>
23	<b>0.90879</b>	<b>Probability corresponding to the normal distribution</b>
24	<b>40</b>	<b>Arithmetic mean of the distribution</b>
25	<b>1.5</b>	<b>Standard deviation of the distribution</b>
26	<b>=NORMINV(A23;A24;A25)</b>	
27	<b>Inverse of the normal cumulative</b>	
28	<b>distribution for the terms above (42)</b>	

**NORMSINV**

Returns the inverse of the standard normal cumulative distribution. The distribution has a mean of zero and a standard deviation of one.

**Syntax****NORMSINV(probability)**

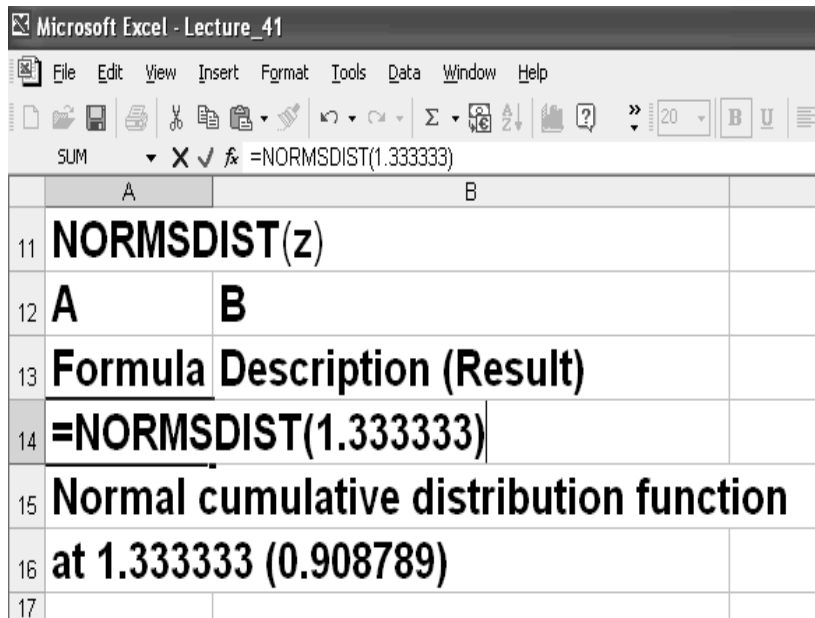
Probability is a probability corresponding to the normal distribution.

**Remarks**

- If probability is nonnumeric, NORMSINV returns the #VALUE! error value.
  - If probability < 0 or if probability > 1, NORMSINV returns the #NUM! error value.
- NORMSINV uses an iterative technique for calculating the function. Given a probability value, NORMSINV iterates until the result is accurate to within  $\pm 3 \times 10^{-7}$ . If NORMSINV does not converge after 100 iterations, the function returns the #N/A error value.

**Example**

In this case, the input is the z-value. The corresponding cumulative distribution is calculated.



**SAMPLING VARIATIONS**

Electronic components are despatched by a manufacturer in boxes of 500.

A small number of faulty components are unavoidable.

Customers have agreed to a defect rate of 2%.

One customer recently found 25 faulty components (5%) in a box.

Was this box representative of production as a whole?

The box represents a sample from the whole output. In such a case sampling variations are expected

If overall proportion of defective items has not increased, just how likely is it that a box of 500 with 25 defective components will occur?

**SAMPLING VARIATIONS EXAMPLE 1**

In a section of a residential colony there are 6 households say Household A, B, C, D, E and F. A survey is to be carried out to determine % of households who use corn flakes (cf) in breakfast.

Survey data exists and the following information is available:

**Households A, B, C and D:** Use corn flakes

**Households E and F:** Do not

It was decided to take random samples of 3 households

The first task is to list all possible samples and find % of each sample using corn flakes.

Possible Samples

<b>Sample</b>	<b>% cf users</b>	<b>Sample</b>	<b>% cf users</b>
ABC	100	BCD	100
ABD	100	BCE	67
ABE	67	BCF	67
ABF	67	BDE	67
ACD	100	BDF	67
ACE	67	BEF	33
ACF	67	CDE	67
ADE	67	CDF	67
ADF	67	CEF	33
AEF	33	DEF	33

**Percentage In Sample**

Out of 20 samples:

4 contain 100% of users,

12 contain 67% of users,

4 contain 33% of users,

with required characteristic

If the samples are selected randomly, then each sample is likely to arise.

The probability of getting a sample

with 100% of users is:  $4/20$  or 0.2

with 67% :  $12/20$  or 0.6

with 33% :  $4/20$  or 0.2

This is a **Sampling Distribution**.

### **SAMPLING DISTRIBUTION**

The sampling distribution of percentages is the distribution obtained by taking all possible samples of fixed size  $n$  from a population, noting the percentage in each sample with a certain characteristic and classifying these into percentages

#### **Mean of the Sampling Distribution**

Using the above data:

Mean =  $100\% \times 0.2 + 67\% \times 0.6 + 33\% \times 0.2 = 67\%$

Mean of the sampling distribution is the true percentage for the population as a whole.

You must make allowance for variability in samples.

#### **Conditions For Sample Selection**

- Number of items in the sample,  $n$ , is fixed and known in advance
- Each item either has or has not the desired characteristic
- The probability of selecting an item with the characteristic remains constant and is known to be  $P$  percent

If  $n$  is large ( $>30$ ) then the distribution can be approximated to a normal distribution

#### **STANDARD ERROR OF PERCENTAGES**

Standard deviation of the sampling distribution tells us how the sample values differ from the mean  $P$ .

It gives us an idea of error we might make if we were to use a sample value instead of the population value.

For this reason it is called Standard Error of Percentages or STEP.

#### **STEP**

The sampling distribution of percentages in samples of  $n$  items ( $n > 30$ ) taken at random from an infinite population in which  $P$  percent of items have characteristic  $X$  will be:

**A Normal Distribution**

with **mean  $P\%$**

and **standard deviation (STEP) =  $[P(100-P)/n]^{1/2} \%$**

The mean and StDev of the sampling distribution of percentages will also be percentages.

**LECTURE 42**  
**Estimating from Samples: Inference**  
**Part 2**

**OBJECTIVES**

The objectives of the lecture are to learn about:

- Review Lecture 41
- Estimating from Samples: Inference

**EXAMPLE 1**

In a factory 25% workforce is women.

How likely is it that a random sample of 80 workers contains 25 or more women?

**Solution:**

Mean = 25%

$$\text{STEP} = [25(100 - 25)/80]^{1/2}$$

$$= [(25 \times 75)/80]^{1/2}$$

$$= 4.84\%$$

$$\% \text{ women in sample} = (25/80) \times 100 = 31.25\%$$

$$z = (31.25 - 25)/4.84$$

$$= 1.29$$

We need to find  $p(\text{sample contains 25 women})$ .

Look for  $p$  against  $z = 1.29$ .

$p(\text{sample contains 25 women}) = 0.985$  or about 10%.

**APPLICATIONS OF STEP**

Some important issues are:

- What is the probability that such a sample will arise?
- How to estimate the percentage  $P$  from information obtained from a single sample?
- How large a sample will be required in order to estimate a population percentage with a given degree of accuracy?

To obtain answers to these questions, let us solve some typical problems.

**CONFIDENCE LIMITS**

A market researcher wishes to conduct a survey to determine % consumers buying the company's products.

He selects a sample of 400 consumers at random.

He finds that 280 of these (70%) are purchasers of the product.

What can he conclude about % of all consumers buying the product?

First let us decide some limits.

It is common to use 95% confidence limits.

These will be symmetrically placed around the 70% buyers.

In a normal sampling distribution 2.5% corresponds to a  $z$ -value of 1.96 on either side of 70%.

Now the sample percentage of 70% can be used as an approximation for population percentage  $P$ .

**Hence:**

$$\text{STEP} = [(70(100 - 70)/400)]^{1/2} = 2.29\%$$

**Confidence Limits**

Estimate for population percentage =  $70 \pm 1.96 \times \text{STEP}$

Or

$$70 \pm 1.96 \times 2.29$$

= 65.515% and 74.49% as the two limits for 95% confidence interval.

We can round off 1.96 to 2

Then with 95% confidence we estimate the population percentage with that characteristic as lying in the interval  
 $P \pm 2 \times \text{STEP}$

**EXAMPLE 2**

A sample of 60 students contains 12 (20%) who are left handed.  
 Find the range with 95% confidence in which the entire left handed students fall.  
 $\text{Range} = 20 \pm 2 \times \text{STEP}$   
 $= 20 \pm 2 \times [(20 \times 80)/60]^{1/2}$   
 $= 9.67\% \text{ and } 30.33\%$

**ESTIMATING PROCESS SUMMARY**

1. Identify n and P (the sample size and percentage) in the sample.
2. Calculate STEP using these values.

•The 95% confidence interval is approximately  $P \pm 2 \text{ STEP}$ .

**99% confidence**

For 99% confidence limits:  
 z-value = 2.58.

**FINDING A SAMPLE SIZE**

To satisfy 95% confidence:  
 $2 \times \text{STEP} = 5$   
 $\text{STEP} = 2.5$   
 Pilot survey value of P = 30%.  
 $\text{STEP} = [(30 \times 70)/n]^{1/2} = 2.5$

**Solving**

$n = 336$

We must interview 336 persons to be 95% confident that our estimate is within 5% of the true answer.

**DISTRIBUTION OF SAMPLE MEANS**

The standard deviation of the Sampling Distribution of means is called Standard Error of the Mean STEM.

$$\text{STEM} = \text{s.d}/(n)^{1/2}$$

s.d denotes standard deviation of the population.  
 n is the size of the sample.

**EXAMPLE 3**

What is the probability that if we take a random sample of 64 children from a population whose mean IQ is 100 with a StDev of 15, the mean IQ of the sample will be below 95?

**Solution:**

$$s = 15; n = 64; \text{population mean} = 100$$

$$\text{STEM} = 15/(64)^{1/2} = 15/(64)^{1/2} = 15/8 = 1.875$$

$$z = 100 - 95 / \text{STEM}$$

$$= 5/1.875 = 2.67$$

This gives a probability of 0.0038.

So the chance that the average IQ of the sample is below 95 is very small.

**LECTURE 43**  
**HYPOTHESIS TESTING: CHI-SQUARE DISTRIBUTION**  
**PART 1**

**OBJECTIVES**

The objectives of the lecture are to learn about:

- Review Lecture 42
- Hypothesis testing: Chi-Square Distribution

**EXAMPLE 1**

An inspector took a sample of 100 tins of beans. The sample weight is 225 g. Standard deviation is 5 g.

Calculate with 95% confidence the range of the population mean.

**Solution:**

$$\text{STEM} = s.d/(n)^{1/2}$$

s.d is not known

**Use s.d of sample as an approximation.**

$$\text{STEM} = 5/(100)^{1/2} = 0.5$$

$$95\% \text{ confidence interval} = 225 \pm 2 \times 0.5 \text{ or From } 224 \text{ to } 226 \text{ g}$$

**PROBLEM OF FAULTY COMPONENTS REVISITED**

Box of 500 components may have 25 or 5% faulty components.

Overall faulty items = 2%

P = 2%; n = 500;

$$\text{STEP} = [(2 \times 98)/500]^{1/2}$$

$$= 0.626$$

**To find the probability that the sample percentage is 5% or over:**

$$z = (5 - 2)/\text{STEP} = 3/0.626 = 4.79$$

Area against z = 4.79 is negligible.

Chance of such a sample is very small

**FINITE POPULATION CORRECTION FACTOR**

If population is very large compared to the sample then multiply STEM and STEP by the:

$$\text{Finite Population Correction Factor} = [1 - (n/N)]^{1/2}$$

Where

N = Size of the population

n = Size of the sample

n = less than 0.1N

**TRAINING MANAGER'S PROBLEM**

New refresher course for training of workers was completed.

The Training Manager would like to assess the effect of retraining if any.

**Particular questions:**

- Is quality of product better than produced before retraining?
- Has the speed of machines increased?
- Do some classes of workers respond better to retraining than others?

**Training Manager hopes to:**

- Compare the new position with established
- Test a theory or hypothesis about the course



Case Study**Before the course:**

Worker X produced 4% rejects.

**After the course:**

Out of 400 items 14 were defective = 3.5%

**An improvement?**

The 3.5% figure may not demonstrate overall improvement.

It does not follow that every single sample of 400 items contains exactly 4% rejects.

**To draw a sound conclusion:**

Sampling variations must be taken into account.

We do not begin by assuming what we are trying to prove NULL HYPOTHESIS.

We must begin with the assumption that there is no change at all.

This initial assumption is called

**NULL HYPOTHESIS****Implication of Null Hypothesis:**

That the sample of 400 items taken after the course was drawn from a population in which the percentage of reject items is still 4%.

**NULL HYPOTHESIS EXAMPLE****Data:**

$P = 4\%$ ;  $n = 400$

**STEP**

$$= [P(100 - P)/n]^{1/2}$$

$$= [4(100 - 4)/400]^{1/2} = 0.98\%$$

**At 95% confidence limit:**

$$\text{Range} = 4 \pm 2 \times 0.98 = 2.04 \text{ to } 5.96 \%$$

**Conclusion:**

Sample with 3.5% rejects is not inconsistent.

No ground to assume that % rejects has changed at all.

On the strength of sample there were no grounds for rejecting Null Hypothesis.

**ANOTHER EXAMPLE****Before the course:**

5% rejects

**After the course:**

2.5% rejects (10 out of 400)

$P = 5$

$$\text{STEP} = [5(100 - 5)/400]^{1/2}$$

$$= [5 \times 95/400]^{1/2}$$

$$= 1.09$$

Range at 95 % Confidence Limits

$$= 5 \pm 2 \times 1.09$$

$$= 2.82 \% \text{ to } 7.18 \%$$

**Conclusion:**

Doubt about Null Hypothesis most of the time

Null hypothesis to be rejected

**PROCEDURE FOR CARRYING OUT HYPOTHESIS TEST**

1. Formulate null hypothesis
2. Calculate STEP &  $P \pm 2 \times \text{STEP}$
3. Compare the sample % with this interval to see whether it is inside or outside  
If the sample falls outside the interval, reject the null hypothesis (sample differs significantly from the population %)  
If the sample falls inside the interval,

do not reject the null hypothesis (sample does not differ significantly from the population % at 5% level)

**HOW THE RULE WORKS?**

Bigger the difference between the sample and population percentages, less likely it is that the population percentages will be applicable.

- When the difference is so big that the sample falls outside the 95% interval, then the population percentages cannot be applied.  
Null Hypothesis must be rejected
- If sample belongs to majority and it falls within 95% interval, then there are no grounds for doubting the Null Hypothesis

**FURTHER POINTS ABOUT HYPOTHESIS TESTING**

- 99% interval requires 2.58 x STEP. Interval becomes wider. It is less likely to conclude that something is significant.
- (A) We might conclude there is a significant difference when there is none.  
Chance of error = 5% (type 1 Error)  
(B) We might decide that there is no significant difference when there is one  
(Type 2 Error)

**LECTURE 44**  
**HYPOTHESIS TESTING : CHI-SQUARE DISTRIBUTION**  
**PART 2**

**OBJECTIVES**

The objectives of the lecture are to learn about:

- Review Lecture 43
- Hypothesis Testing : Chi-Square Distribution

**FURTHER POINTS ABOUT HYPOTHESIS TESTING**

This is a continuation of the points covered under Handout 43.

3. We cannot draw any conclusion regarding the direction the difference is in

**(A) Possible to do 1-tailed test**

Null Hypothesis:  $P \geq 4\%$  against the alternative  $P > 4\%$

$z = 1.64$  for 5% significance level

Range =  $P - 1.64 \times \text{STEP}$  (0.98%)

**Example**

Range =  $4 - 1.64 \times 0.98 = 2.39\%$

New figure = 3.5%.

**Hence:**

There is no reason to conclude that things have improved.

4. We cannot draw any conclusion regarding the direction the difference is in.

**(B) Possible to do 2-tailed test**

Null Hypothesis:  $P \geq 4\%$  against the alternative  $P > 4\%$

$z = 1.96$  for 5% significance level

Range =  $P \pm 1.96 \times \text{STEP}$  (0.98%)

**Example**

Range =  $4 \pm 2 \times 1.96 \times 0.98 = 2.08\%$  to  $5.92\%$

New figure = 3.5%

There is no reason to conclude that things have improved

**HYPOTHESES ABOUT MEANS**

Let us go back to the problem of retraining course discussed earlier.

**Before the course:**

Worker X took 2.5 minutes to produce 1 item.

StDev = 0.5 min

**After the course:**

For a sample of 64 items, mean time = 2.58 min

**Null hypothesis**

No change after the course.

**STEM**

$= s.d./\sqrt{n} = 0.5/\sqrt{64} = 0.0625$

**Range**

$= 2.5 \pm 2 \times 0.0625 = 2.375$  to  $2.625$  min

**Conclusion:**

No grounds for rejecting the Null Hypothesis.

There is no change significant at 5% level.

**ALTERNATIVE HYPOTHESIS TESTING USING Z-VALUE**

$$z = (\text{sample percentage} - \text{population mean})/\text{STEP}$$

$$= (3.5 - 4)/0.98 = 0.51$$

Compare it with z-value which would be needed to ensure that our sample falls in the 5% tails of distribution (1.96 or about 2).

z is much less than 2.

We conclude that the probability of getting by random chance a sample which differs from the mean of 4% or more is quite high.

Certainly it is greater than the 5% significance level.

Sample is quite consistent with null hypothesis.

Null hypothesis should not be rejected.

**PROCESS SUMMARY**

1. State Null Hypothesis (1-tailed or 2-tailed)
2. Decide on a significance level and find corresponding critical value of z
3. Calculate sample z (sample value – population value divided by STEP or STEM as appropriate)
4. Compare sample z with critical value of z
5. If sample z is smaller, do not reject the Null Hypothesis
6. If sample z is greater than critical value of z, sample provides ground for rejecting the Null Hypothesis.

**TESTING HYPOTHESES ABOUT SMALL SAMPLES**

Whatever the form of the underlying distribution the means of large samples will be normally distributed.

This does not apply to small samples.

We can carry out hypothesis testing using the methods discussed only if the underlying distribution is normal.

If we only know the Standard Deviation of sample and have to approximate population Standard Deviation then we use **Student's t-distribution**.

**STUDENT'S t-DISTRIBUTION**

Student's T-Distribution is very much like normal distribution.

In fact it is a whole family of t-distributions.

As n gets bigger, t-distribution approximates to normal distribution.

t-distribution is wider than normal distribution.

95% confidence interval reflects greater degree of uncertainty in having to approximate the population Standard Deviation by that of the sample.

**EXAMPLE**

Mean training time for population = 10 days.

Sample mean for 8 women = 9 days.

Sample Standard Deviation = 2 days.

To approximate population Standard Deviation by a sample divide the sum of squares by  $n - 1$ :

$$\text{STEM} = 2/(8)^{1/2} = 0.71$$

**Null Hypothesis:**

There is no difference in overall training time between men and women.

**t-value = (sample mean – population mean)/STEM**

$$= (9 - 10)/0.71 = - 1.41$$

For  $n = 8$ ,  $v = 8 - 1 = 7$ ;

For 5% (.05) significance level looking at 0.025 (2-tailed):

$t = 2.365$  (Calculated table value)

**Decision:**

Do not reject the Null Hypothesis

**SUMMARY - I**

If underlying population is normal and we know the Standard Deviation  
Then

Distribution of sample means is normal  
with

**Standard Deviation = STEM = population s.d./ $(n)^{1/2}$**

and

**we can use a z-test.**

**SUMMARY - II**

If underlying population is unknown but the sample is large  
Then

Distribution of sample means is approximately normal  
With

**StDev = STEM = population s.d./ $(n)^{1/2}$**

and again

**we can use a z-test.**

**SUMMARY - III**

If underlying population is normal but we do not know its StDev and the sample is small  
Then We can use the sample s.d to approximate that of the population with  $n - 1$  divisor  
in the calculation of s.d.

Distribution of sample means is a t-distribution with  $n - 1$  degrees of freedom  
With

**Standard Deviation = STEM = sample s.d./ $(n)^{1/2}$**

And we can use a t-test.

**SUMMARY - IV**

If underlying **population is not normal** and we have a **small sample**  
Then **none of the hypothesis testing procedures can be safely used.**

**TESTING DIFFERENCE BETWEEN TWO SAMPLE MEANS**

A group of 30 from production has a mean wage of 120 Rs. per day with  
Standard Deviation = Rs. 10.

50 Workers from Maintenance had a mean of Rs. 130 with  
Standard Deviation = 12

Is there a difference in wages between workers?

**Difference of two sample means =  $s[(1/n_1) + (1/n_2)]^{1/2}$**

$s = [(n_1.s_1^2 + n_2.s_2^2)/(n_1 + n_2)]^{1/2}$

$N_1 = 30; n_2 = 50; s_1 = 10; s_2 = 12$

$s = [(30 \times 100 + 50 \times 144)/(30 + 50)]^{1/2} = 11.29$

**Standard Error of Difference in Sample Means (STEDM)**

$= 11.29(1/30 + 1/50)^{1/2} = 2.60$

**$z = (\text{difference in sample means} - 0)/\text{STEDM}$**

$= 120 - 130/2.60 = - 3.85$

This is well outside the critical z for 5% significance.

There are grounds for rejecting Null Hypothesis (There is difference in the two samples).

**PROCEDURE SUMMARY**

1. State Null Hypothesis and decide significance level
2. Identify information (no. of samples, large or small, mean or proportion) and decide what standard error and what distribution are required
3. Calculate standard error
4. Calculate z or t as difference between sample and population values divided by standard error
5. Compare your z or t with critical value from tables for the selected significance level; if z or t is greater than critical value, reject the Null Hypothesis

**MORE THAN ONE PROPORTION**

Look at a problem, where after the course some in different age groups shows improvement while others did not.

Let us assume that the expected improvement was uniform. An improvement of 40%, if applied to 21, 24 and 15 would give 14, 16 and 10 respectively, who improved. Let us write these values within brackets. Subtracting 14, 16 and 10 from the totals 21, 24 and 15 gives us 7, 8 and 5 respectively, who did not improve. This is the estimate if every person was affected in a uniform manner.

Let us write the observations as O, in one line (17 17 6 4 7 9).

Let us write down the expected as E, in the next line as (14 16 10 7 8 8).

Calculate O-E.

Next calculate  $(O-E)^2$ .

Now standardize  $(O-E)^2$  by dividing by E.

Calculate the total and call it  $\chi^2$ .

Age	Improved	Did not improve	Total			
Under 35	17(14)	4(7)	21			
35 – 50	17(16)	7(8)	24			
Over 50	6(10)	9(5)	15			
Total	40	20	60			
O	17	17	6	4	7	9
E	14	16	10	7	8	8
O-E	3	1	-4	-3	-1	4
$(O-E)^2$ :	9	1	16	9	1	16
$(O-E)^2/E$ :	0.643	0.0625	1.6	1.286	0.125	3.2 = 6.92

Measurement of disagreement = Sum  $[(O-E)^2/E]$

is known as **Chi-squared ( $\chi^2$ )**

**Degrees of freedom  $v = (r-1) \times (c-1) = (3-1)(2-1) = 2$**

There are tables that give Critical value of chi-squared at different confidence limits and degrees of freedom  $v$  (columns-1) x (rows-1). In the above case

$$v = 2-1 \times 3-1 = 2$$

In the present case, the Critical value of chi-squared at 5% (and  $v = 2$ ) = 5.991.

The value 6.92 is greater than 5.991.

This means that the Sample falls outside of 95% interval.

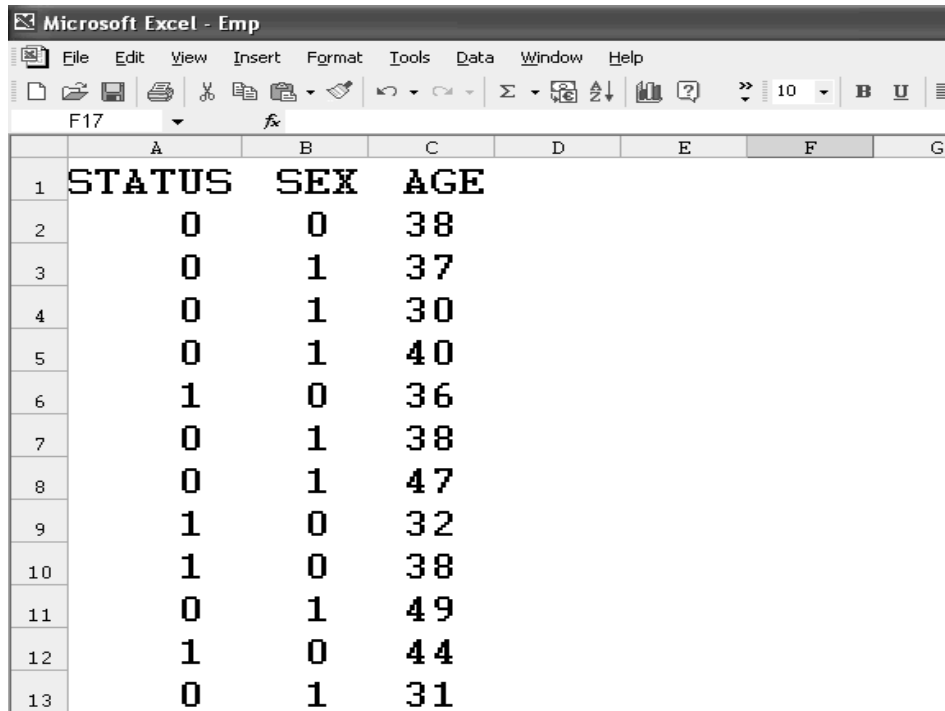
Null hypothesis should be rejected.

**CHI-SQUARED SUMMARY**

1. Formulate null hypothesis (no association form)
2. Calculate expected frequencies
3. Calculate  $\chi^2$
4. Calculate degrees of freedom (rows minus 1) x (columns minus 1); look up the critical  $\chi^2$  under the selected significance level
5. Compare the calculated value of  $\chi^2$  from the sample with value from the table; if the sample  $\chi^2$  is smaller (within the interval) don't reject the null hypothesis; if it is bigger (outside) reject the null hypothesis

**Example**

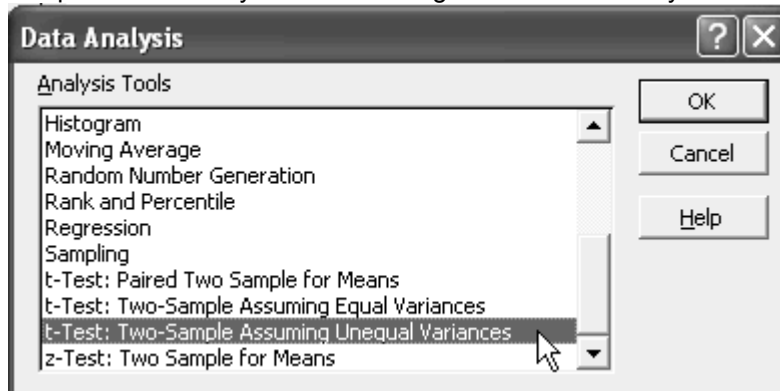
Look at the data in the slide below.



The screenshot shows a Microsoft Excel window titled "Emp" with a menu bar (File, Edit, View, Insert, Format, Tools, Data, Window, Help) and a toolbar. The active cell is F17. The data is as follows:

	A	B	C	D	E	F	G
1	<b>STATUS</b>	<b>SEX</b>	<b>AGE</b>				
2	0	0	38				
3	0	1	37				
4	0	1	30				
5	0	1	40				
6	1	0	36				
7	0	1	38				
8	0	1	47				
9	1	0	32				
10	1	0	38				
11	0	1	49				
12	1	0	44				
13	0	1	31				

It is possible to carry out t-tests using EXCEL Data Analysis tools.



When you select the tool and press OK, the t-test dialog box is opened as below.

**t-Test: Two-Sample Assuming Unequal Variances**

Input

Variable 1 Range:

Variable 2 Range:

Hypothesized Mean Difference:

Labels

Alpha:

Output options

Output Range:

New Worksheet Ply:

New Workbook

OK  
Cancel  
Help

The ranges for the two variables, labels and output options are specified. For the above data the output was as follows:

Microsoft Excel - Emp

File Edit View Insert Format Tools Data Window Help

Type a question for help

D15

	A	B	C	D	E
1	<b>t-Test: Two-Sample Assuming Unequal Variances</b>				
2					
3		<i>38</i>	<i>42</i>		
4	<b>Mean</b>	37.913	37.571		
5	<b>Variance</b>	37.356	75.252		
6	<b>Observations</b>	23	35		
7	<b>Hypothesized Mean Di</b>	0			
8	<b>df</b>	56			
9	<b>t Stat</b>	0.1758			
10	<b>P(T&lt;=t) one-tail</b>	0.4305			
11	<b>t Critical one-tail</b>	1.6725			
12	<b>P(T&lt;=t) two-tail</b>	0.8611			
13	<b>t Critical two-tail</b>	2.0032			
14					

#### CHITEST

Returns the test for independence. CHITEST returns the value from the chi-squared ( $\chi^2$ ) distribution for the statistic and the appropriate degrees of freedom. You can use  $\chi^2$  tests to determine whether hypothesized results are verified by an experiment.

#### **Syntax**

**CHITEST(actual\_range,expected\_range)**

**Actual\_range** is the range of data that contains observations to test against expected values.

**Expected\_range** is the range of data that contains the ratio of the product of row totals and column totals to the grand total.

#### **Remarks**



- If actual\_range and expected\_range have a different number of data points, CHITEST returns the #N/A error value.
- The  $\chi^2$  test first calculates a  $\chi^2$  statistic and then sums the differences of actual values from the expected values. The equation for this function is  $\text{CHITEST} = p(X > \chi^2)$ , where:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(A_{ij} - E_{ij})^2}{E_{ij}}$$

and where:

$A_{ij}$  = actual frequency in the i-th row, j-th column

$E_{ij}$  = expected frequency in the i-th row, j-th column

r = number of rows

c = number of columns

CHITEST returns the probability for a  $\chi^2$  statistic and degrees of freedom, df, where  $df = (r - 1)(c - 1)$ .

### Example

	A	B	C
2	<b>Men (Actual)</b>	<b>Women (Actual)</b>	<b>Description</b>
3	58	35	Agree
4	11	25	Neutral
5	10	23	Disagree
6	<b>Men (Expected)</b>	<b>Women (Expected)</b>	<b>Description</b>
7	45.35	47.65	Agree
8	17.56	18.44	Neutral
9	16.09	16.91	Disagree
10	A3:B5;A7:	<b>The <math>\chi^2</math> statistic for the data above is 16.16957</b>	
11	B9)	<b>with 2 degrees of freedom (0.000308)</b>	

The above example shows two different groups. The calculation shows that the probability for chi-squared 16.16957 with 2 degrees of freedom was 0.000308, which is negligible.

## LECTURE 45

### Planning Production Levels: Linear Programming

#### **OBJECTIVES**

The objectives of the lecture are to learn about:

- Review Lecture 44
- Planning Production Levels: Linear Programming

#### **INTRODUCTION TO LINEAR PROGRAMMING**

A Linear Programming model seeks to maximize or minimize a linear function, subject to a set of linear constraints.

The linear model consists of the following components:

1. A set of decision variables,  $x_j$ .
2. An objective function,  $\sum C_j x_j$ .
3. A set of constraints,  $\sum a_{ij} x_j \leq b_i$ .

#### **THE FORMAT FOR AN LP MODEL**

Maximize or minimize  $\sum C_j x_j = C_1 x_1 + C_2 x_2 + \dots + C_n x_n$

Subject to

$a_{ij} x_j \leq b_i, i = 1, \dots, m$

Non-negativity conditions: all  $x_j \geq 0, j = 1, \dots, n$

Here  $n$  is the number of decision variables.

Here  $m$  is the number of constraints.

(There is no relation between  $n$  and  $m$ )

#### **THE METHODOLOGY OF LINEAR PROGRAMMING**

1. Define decision variables
2. Hand-write objective
3. Formulate math model of objective function
4. Hand-write each constraint
5. Formulate math model for each constraint
6. Add non-negativity conditions

#### **THE IMPORTANCE OF LINEAR PROGRAMMING**

Many real world problems lend themselves to linear programming modeling.

Many real world problems can be approximated by linear models.

There are well-known successful applications in:

- Operations
- Marketing
- Finance (investment)
- Advertising
- Agriculture

There are efficient solution techniques that solve linear programming models.

The output generated from linear programming packages provides useful "what if" analysis.

#### **ASSUMPTIONS OF THE LINEAR PROGRAMMING MODEL**

1. The parameter values are known with certainty
  2. The objective function and constraints exhibit constant returns to scale
  3. There are no interactions between the decision variables (the additivity assumption)
- The Continuity assumption: Variables can take on any value within a given feasible range.

#### **A PRODUCTION PROBLEM – A PROTOTYPE EXAMPLE**

A company manufactures two toy doll models:

Doll A

Doll B

**Resources are limited to:**

1000 kg of special plastic.

40 hours of production time per week.

**Marketing requirement:**

Total production cannot exceed 700 dozens.

Number of dozens of Model A cannot exceed number of dozens of Model B by more than 350.

The current production plan calls for:

- Producing as much as possible of the more profitable product, Model A (Rs. 800 profit per dozen).
- Use resources left over to produce Model B (Rs. 500 profit per dozen), while remaining within the marketing guidelines.

**Management is seeking:**

a production schedule that will increase the company's profit

A linear programming model

can provide:

- an insight and
- an intelligent solution to this problem

**Decisions variables:**

$X_1$  = Weekly production level of Model A (in dozens)

$X_2$  = Weekly production level of Model B (in dozens).

**Objective Function:**

Weekly profit, to be maximized

Maximize  $800X_1 + 500X_2$  (Weekly profit)

subject to

$2X_1 + 1X_2 \leq 1000$  (Plastic)

$3X_1 + 4X_2 \leq 2400$  (Production Time)

$X_1 + X_2 \leq 700$  (Total production)

$X_1 - X_2 \leq 350$  (Mix)

$X_j \geq 0, j = 1, 2$  (Nonnegativity)

**ANOTHER EXAMPLE**

A dentist is faced with deciding:

how best to split his practice

between the two services he offers—general dentistry

and pedodontics?

(children's dental care)

Given his resources,

how much of each service should he provide

to maximize his profits?

The dentist employs three assistants and uses two operatories.

Each pedodontic service requires .75 hours of operator time, 1.5 hours of an assistant's

time and .25 hours of the dentist's time

A general dentistry service requires .75 hours of an operator, 1 hour of an assistant's

time and .5 hours of the dentist's time.

Net profit for each service is Rs. 1000 for each pedodontic service and Rs. 750 for each general dental service.

Time each day is: eight hours of dentist's, 16 hours of operator time, and 24 hours of assistants' time.

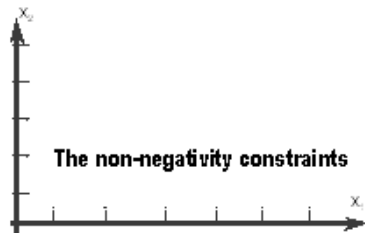
**THE GRAPHICAL ANALYSIS OF LINEAR PROGRAMMING**

Using a graphical presentation,  
we can represent:

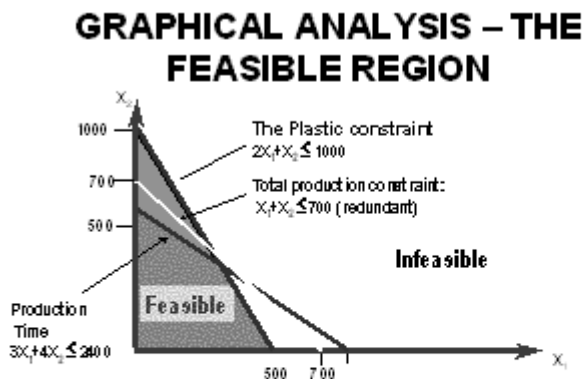
all the constraints, the objective function, and the three types of feasible points.

**GRAPHICAL ANALYSIS – THE FEASIBLE REGION**

The slide shows how a feasible region is defined with non-negativity constraints.

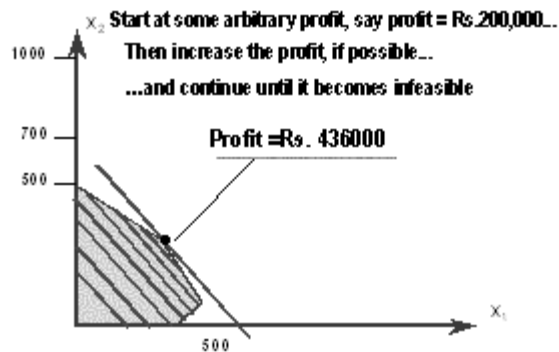
**GRAPHICAL ANALYSIS – THE FEASIBLE REGION****THE SEARCH FOR AN OPTIMAL SOLUTION**

The figure shows how different constraints can be represented by straight lines to define a feasible region. There is an area outside the feasible region that is infeasible.



It may be seen that each of the constraints is a straight line. The constraints intersect to form a point that represents the optimal solution. This is the point that results in maximum profit of 436,000 Rs. As shown in the slide below. The procedure is to start with a point that is the starting point say 200,000 Rs. Then move the line upwards till the last point on the feasible region is reached. This region is bounded by the lines representing the constraints.

## THE SEARCH FOR AN OPTIMAL SOLUTION



### SUMMARY OF THE OPTIMAL SOLUTION

Model A = 320 dozen

Model B = 360 dozen

Profit = Rs. 436000

This solution utilizes all the plastic and all the production hours.

Total production is only 680 (not 700).

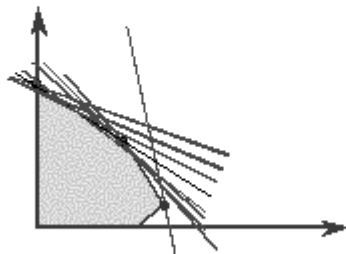
Model a production does not exceed Model B production at all.

### EXTREME POINTS AND OPTIMAL SOLUTIONS

If a linear programming problem has an optimal solution, an extreme point is optimal.

## EXTREME POINTS AND OPTIMAL SOLUTIONS

– If a linear programming problem has an optimal solution, an extreme point is optimal.



**MULTIPLE OPTIMAL SOLUTIONS**

There may be more than one optimal solutions. However, the condition is that the objective function must be parallel to one of the constraints. If a weighted average of different optimal solutions is obtained, it is also an optimal solution.

**MULTIPLE OPTIMAL SOLUTIONS**

- For multiple optimal solutions to exist, the objective function must be parallel to one of the constraints

