

BUSINESS MATHEMATICS & STATISTICS

Module 5

Statistical data representation
(Lecture 23)

Measures of central tendency
(Lectures 24-25)

Measures of dispersion and skewness
(Lectures 26-27)

Module 6

Correlation

(Lecture 28-29)

Line Fitting

(Lectures 30-31)

Time Series and Exponential Smoothing

(Lectures 32-33)

Module 7

Factorials

Permutations and Combinations

(Lecture 34)

Elementary Probability

(Lectures 35-36)

Chi-Square

(Lectures 37)

Binomial Distribution

(Lectures 38)

Module 8

Patterns of probability: Binomial, Poisson and Normal
Distributions

(Lecture 39-41)

Estimating from Samples: Inference

(Lectures 42-43)

Hypothesis testing : Chi-Square Distribution

(Lectures 44-45)

End-Term Examination

LECTURE 23

Statistical Data Representation

STATISTICAL DATA

Information is collected by government departments, market researchers, opinion pollsters and others

Information then has to be organised and presented in a way that is easy to understand

BASIS FOR CLASSIFICATION

1. Qualitative

Attributes: sex, religion

2. Quantitative

Characteristics: Heights, weights, incomes etc.

3. Geographical

Geographical regions: Provinces, divisions etc.

4. Chronological or Temporal

By time of occurrence: Time series

TYPES OF CLASSIFICATION

One-way

One characteristic: Population

Two-way

Two characteristics at a time

Three-way

Three characteristics at a time

METHODS OF PRESENTATION

Text

"The majority of population of Punjab is located in rural areas."

Semitabular

Data in rows

Tabular

Tables with rows and columns

Graphic

Charts and graphs

TYPES OF GRAPHS

Column Graphs

Line Graphs

Circle Graphs (Sector Graphs)

Conversion Graphs

Travel Graphs

Statistical Graphs

Frequency Tables

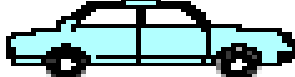
Histograms

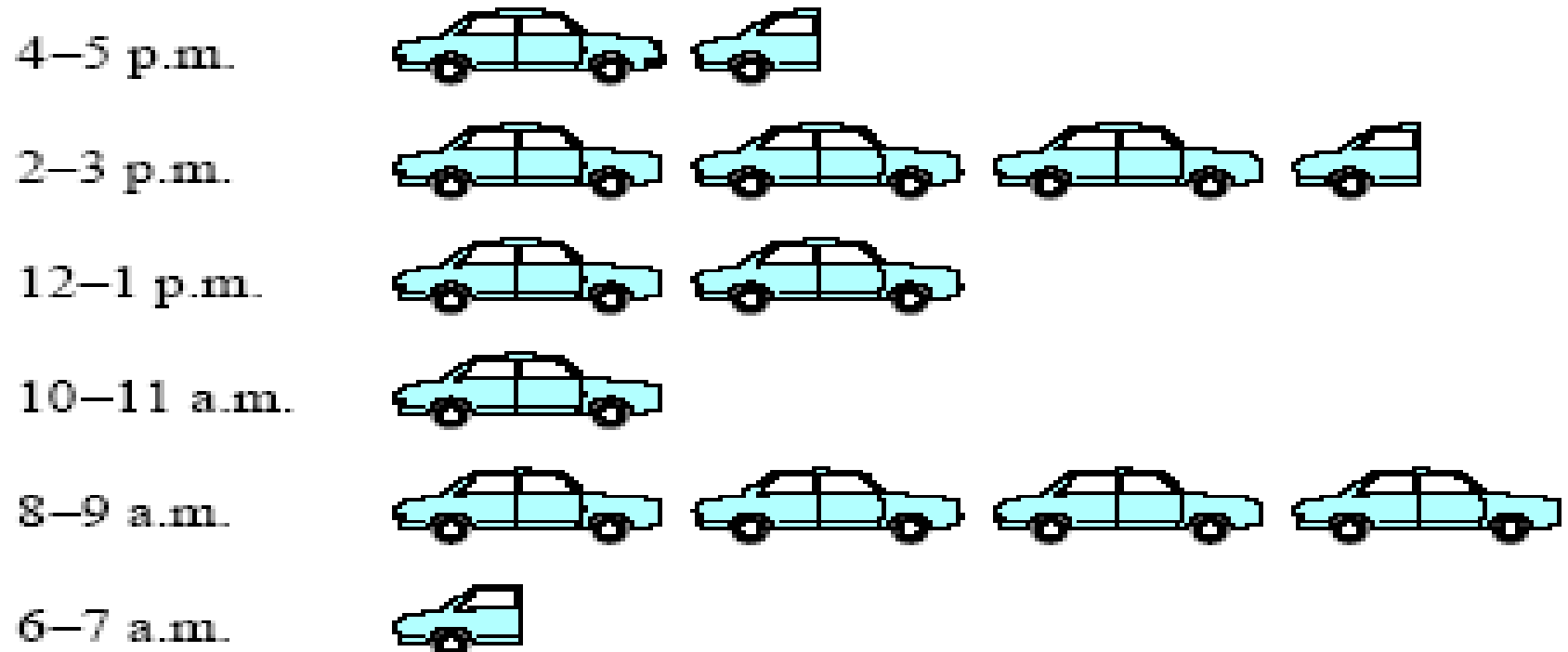
Frequency distributions

Cumulative Distributions

Commulated Distributions

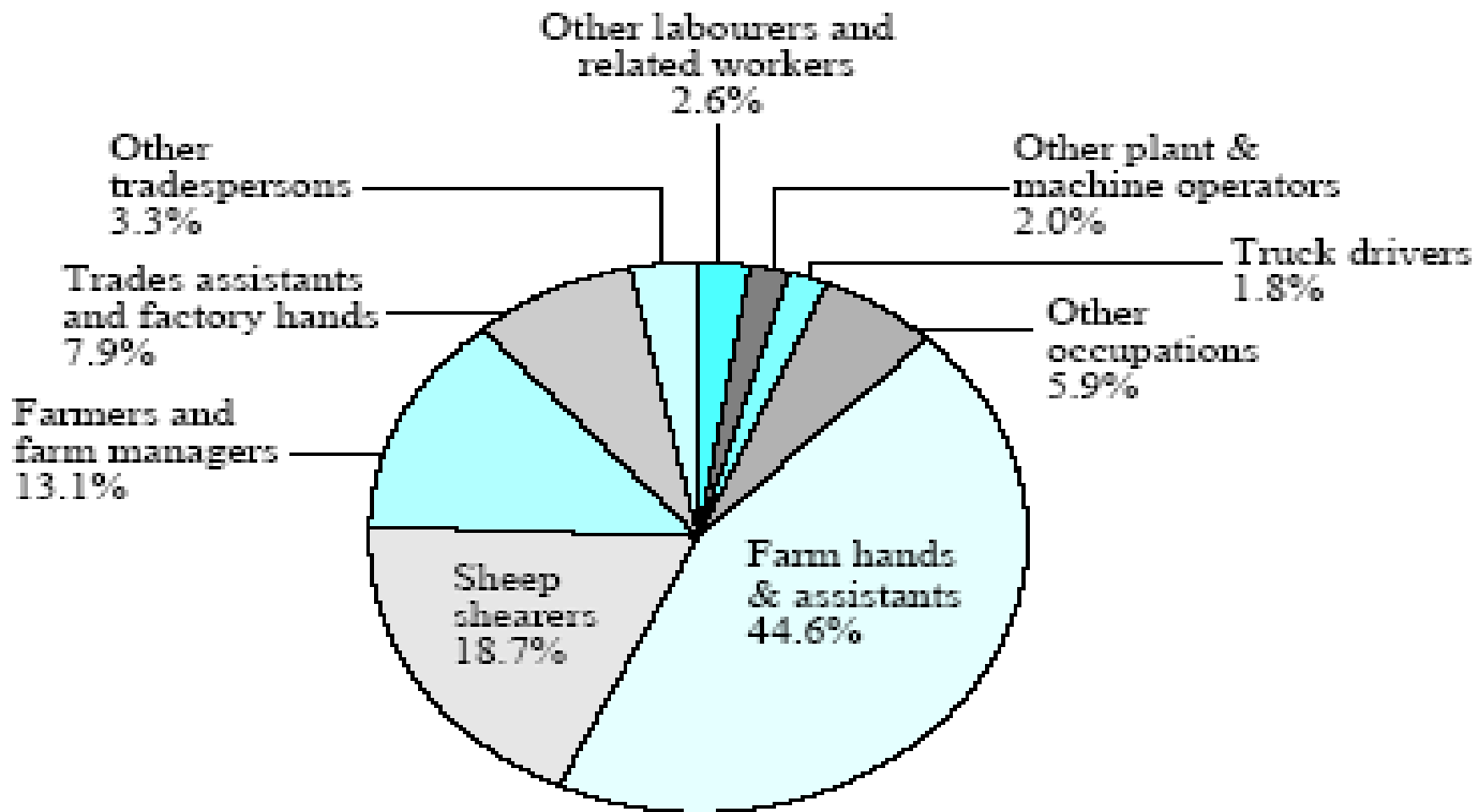
Cars passing school gate during the day

Each  represents 10 cars.



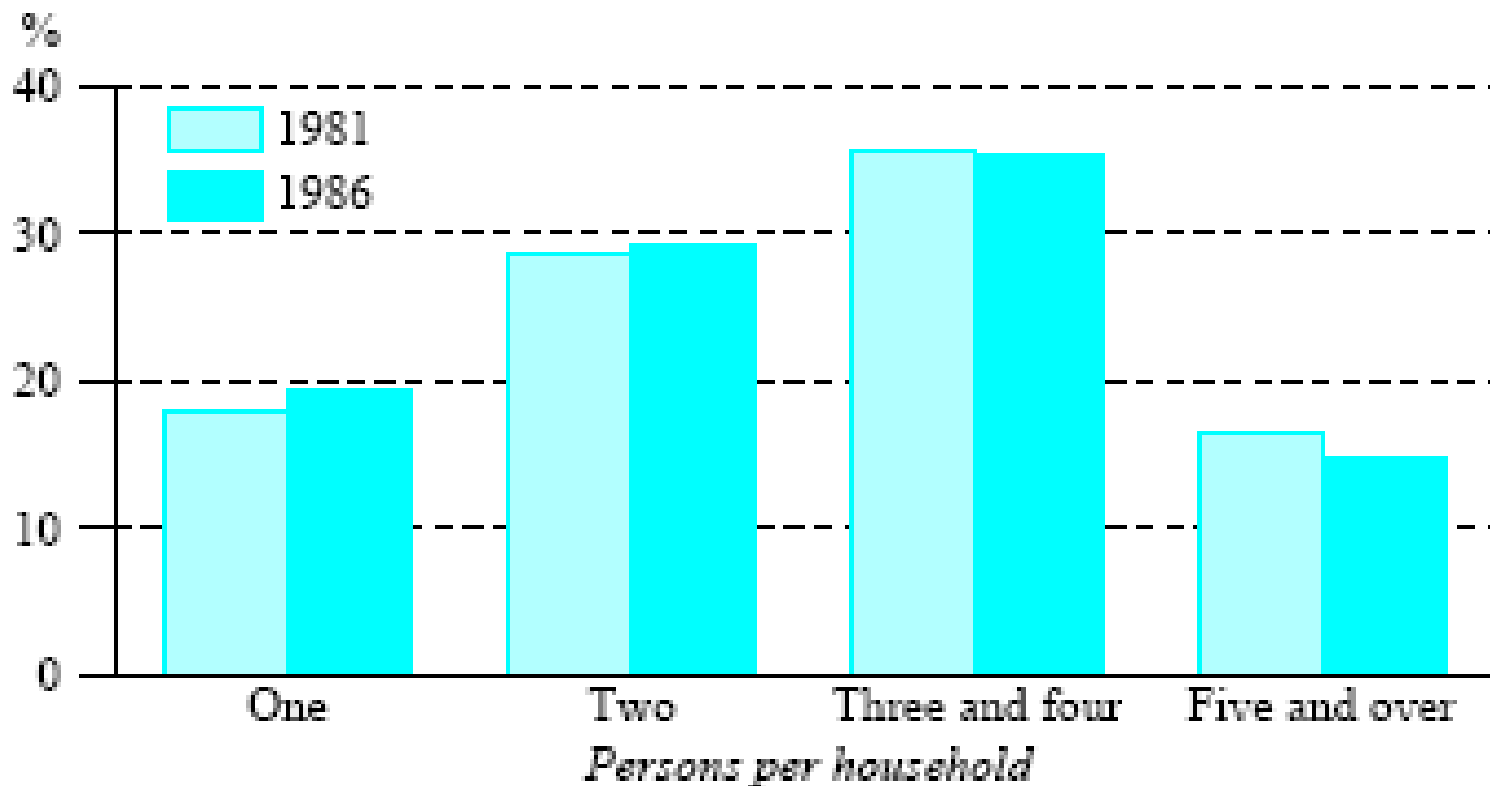
SECTOR GRAPHS

Distribution of employment injuries by occupation



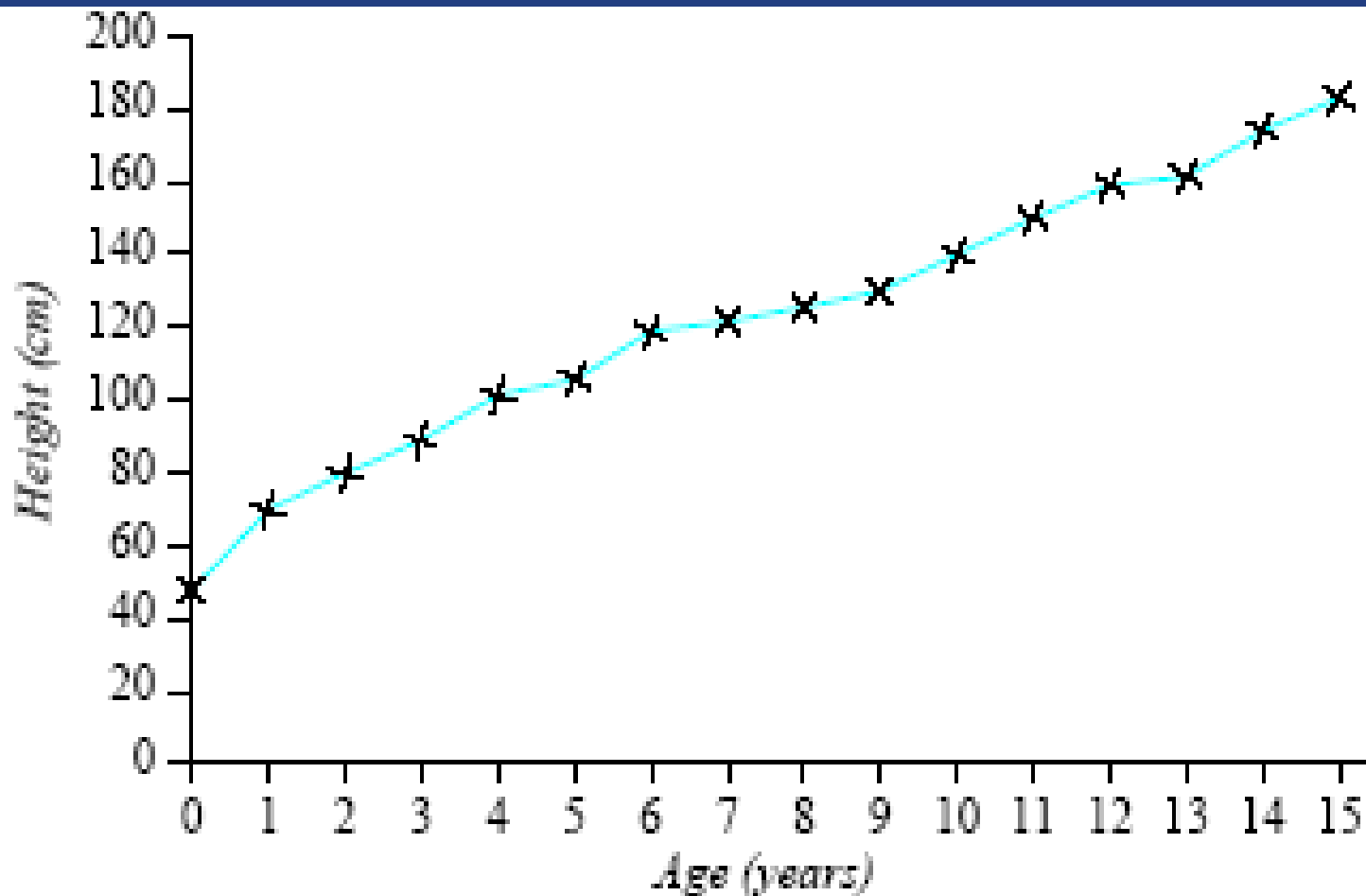
COLUMN AND BAR GRAPHS

Proportion of households by size



LINE GRAPHS

A student's height with age



BUSINESS MATHEMATICS & STATISTICS

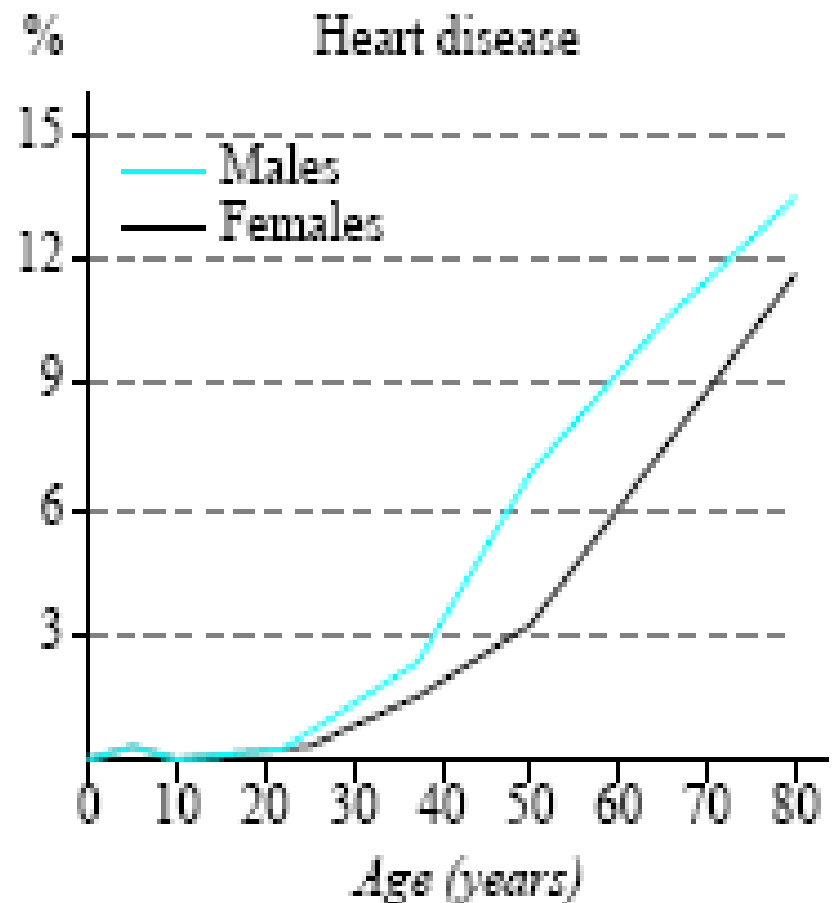
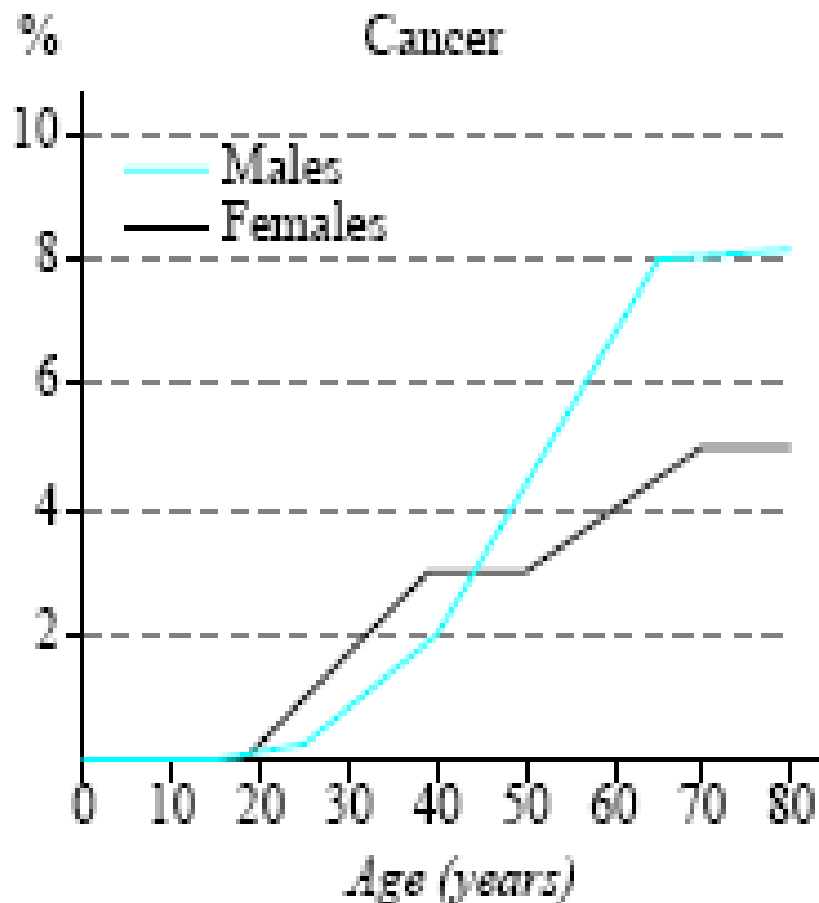
BUSINESS MATHEMATICS & STATISTICS

LECTURE 24

Statistical Representation Measures of Central Tendency Part 1

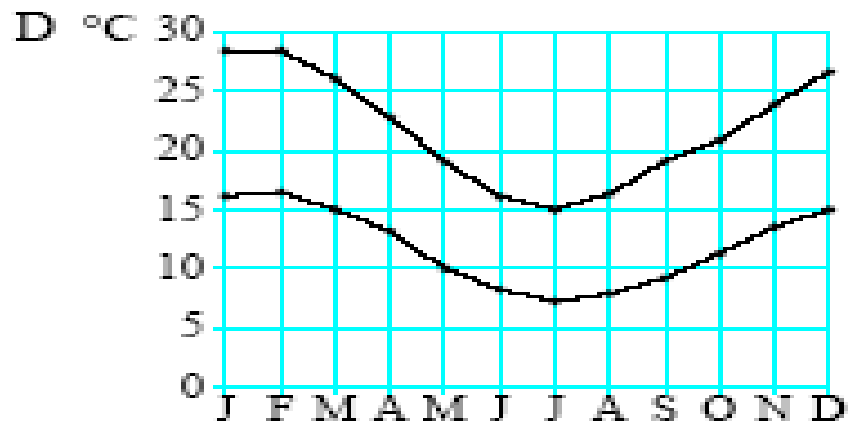
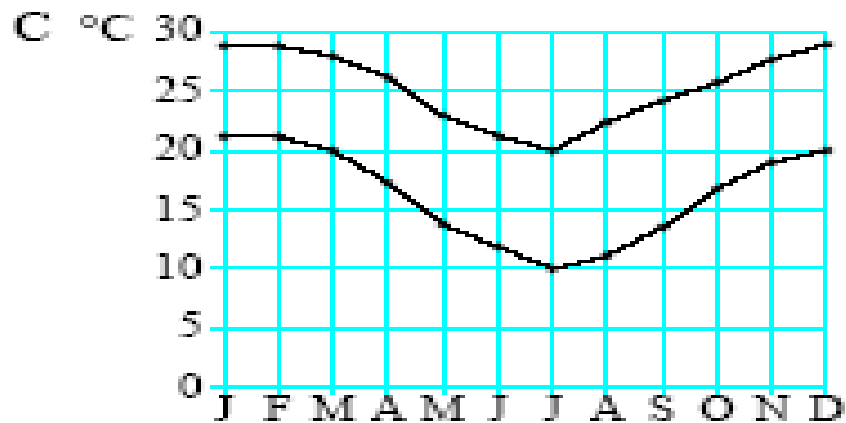
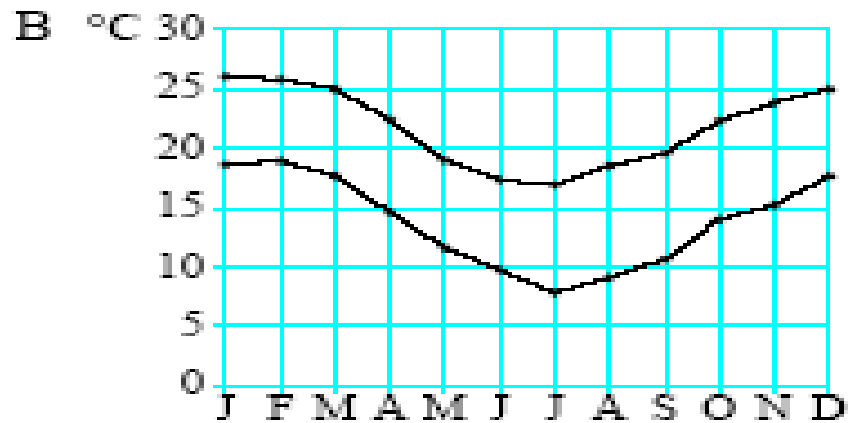
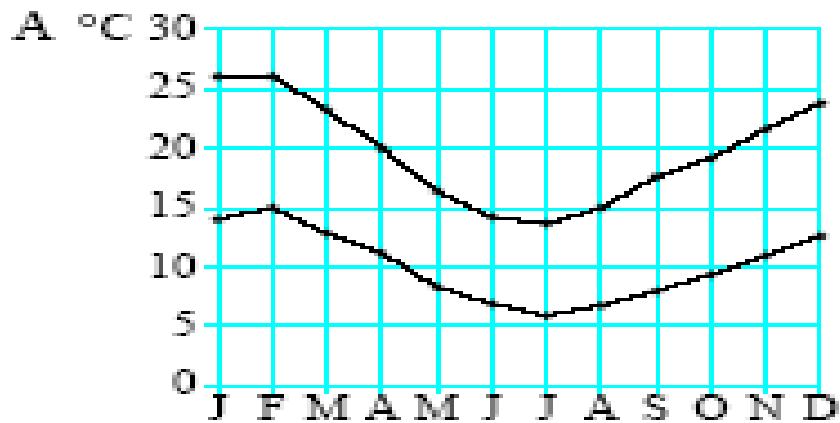
LINE GRAPHS

Causes of death



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 = $596/8 = 74.5$

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

3 6 11 14 19 19 21 24 31

The median is the middle score, or the **mean of the two middle scores**, when the scores are placed in order

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

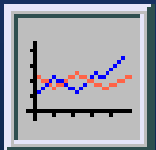
Organising Numerical Data

The Ordered Array and Stem-leaf Display

Tabulating and Graphing Numerical Data

Frequency Distributions: Tables, Histograms, Polygons

Cumulative Distributions: Tables, the Ogive

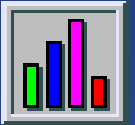


ORGANISING DATA

Tabulating and Graphing Univariate Categorical Data

The Summary Table

Bar and Pie Charts, the Pareto Diagram



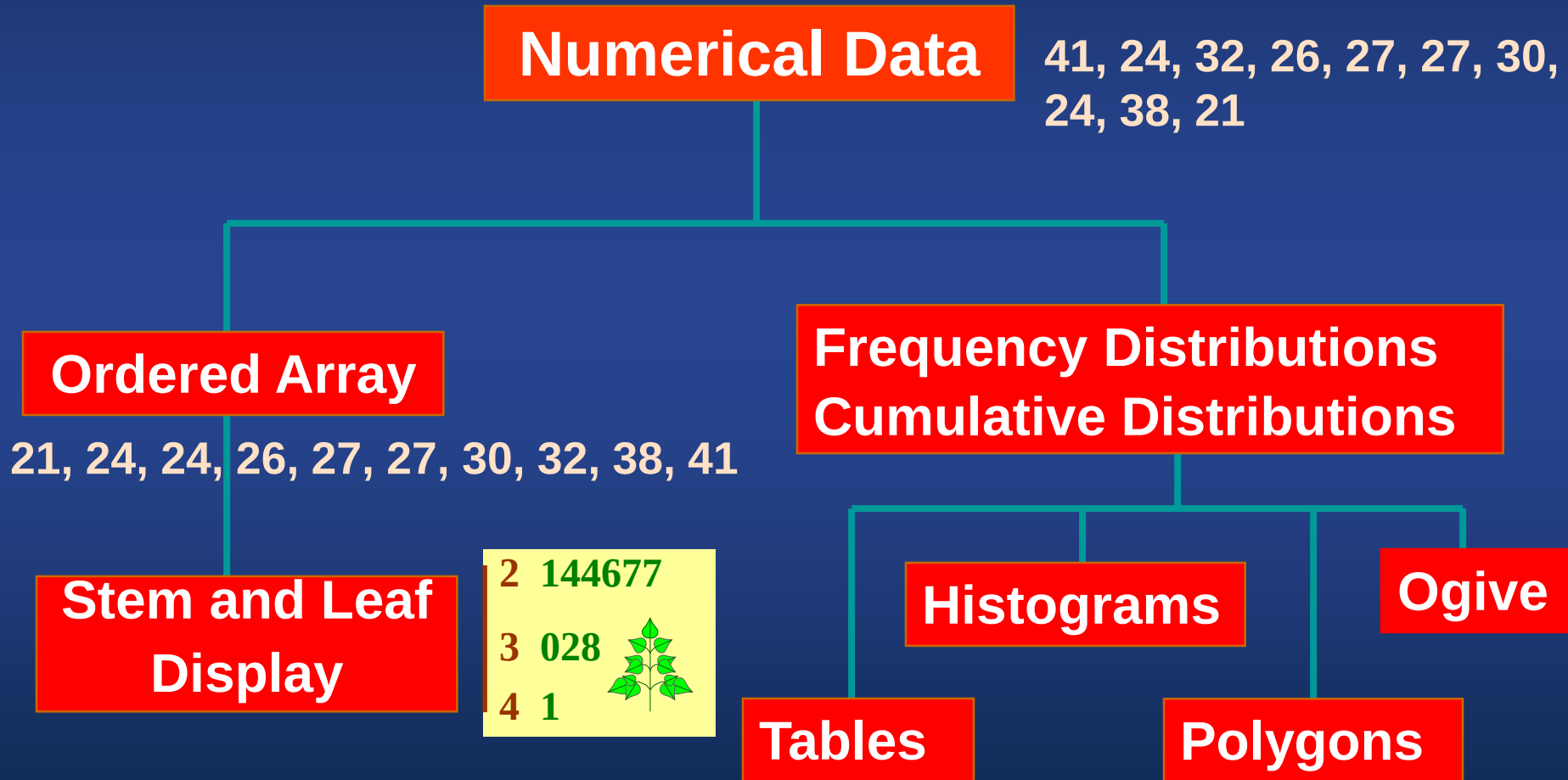
Tabulating and Graphing Bivariate Categorical Data

Contingency Tables

Side by Side Bar charts

Graphical Excellence and Common Errors in Presenting Data

Organizing Numerical Data



ORGANISING NUMERICAL DATA

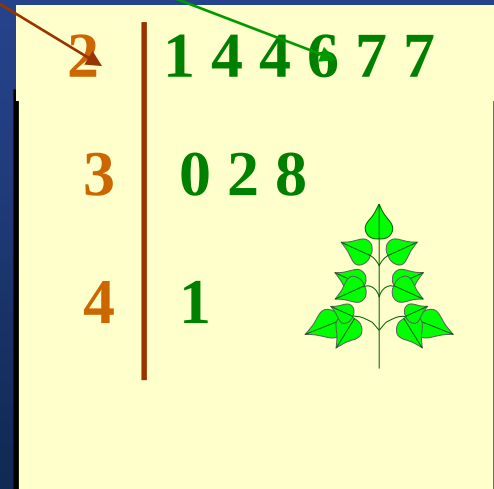
Data in Raw form (as collected)

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 Numerical Data

Numerical Data

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

Ordered Array

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

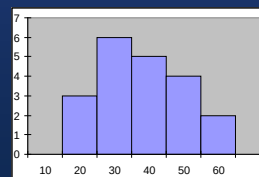
Stem and Leaf Display

2	144677
3	028
4	1

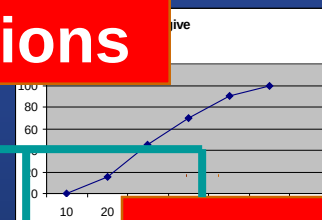


Frequency Distributions Cumulative Distributions

Histograms



Ogive



Tables

Polygons

Tabulating Numerical Data: Frequency Distributions

Sort Raw Data in Ascending Order

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

Find Range

$$58 - 12 = 46$$

Select Number of Classes

5 (usually between 5 and 15)

Tabulating Numerical Data: Frequency Distributions

Compute Class Interval (width)

10 ($46/5$ then round up)

Determine Class Boundaries (limits)

10, 20, 30, 40, 50

Compute Class Midpoints

15, 25, 35, 45, 55

Count Observations & Assign to Classes

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
Total	20	1	100

BUSINESS MATHEMATICS & STATISTICS

BUSINESS MATHEMATICS & STATISTICS

BUSINESS MATHEMATICS & STATISTICS

BUSINESS MATHEMATICS & STATISTICS

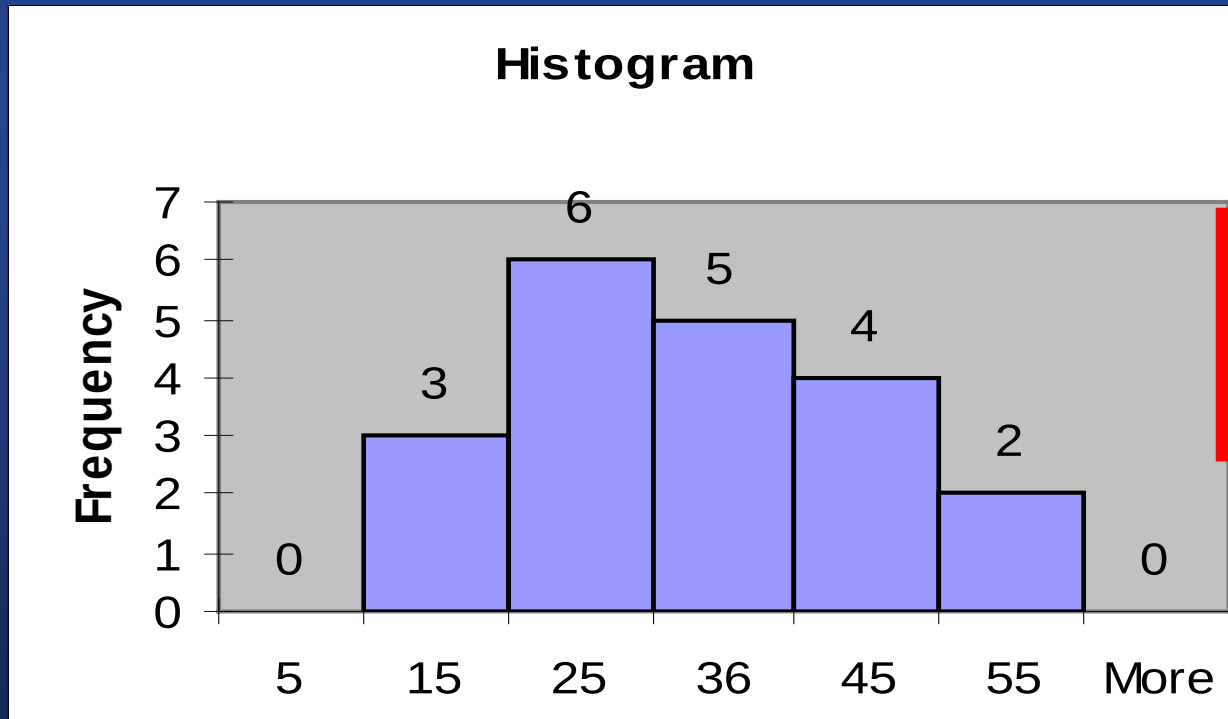
LECTURE 25

Statistical Representation
Measures of Central Tendency
Part 2

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,
46, 53, 58



**No Gaps
Between
Bars**

Class Midpoints

MEASURES OF CENTRAL TENDENCY

Arithmetic Mean

Arithmetic Mean for Grouped Data

Weighted Mean

Median

Median for Grouped Data

Median for Discrete Data

Graphic Location of Median

MEASURES OF CENTRAL TENDENCY

Quantiles(Quartiles, Deciles, Percentiles)

Quantiles from Grouped Data

Quantiles from Discrete Data

Graphic Location of Quantiles

Mode

Mode from Grouped Data

Mode from Discrete Data

Empirical Relation Between mean, Median and Mode

AVERAGE & AVERAGEA

AVERAGE(number1,number2, ...)

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

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.

AVERAGEA(value1,value2,...)



=AVERAGE(A4:A8)

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

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.

E16		=						
	A	B	C	D	E	F	G	
12	MEDIAN(number1,number2, ...)							
13								
14	1							
15	2							
16	3							
17	4							
18	5							
19	6							
20	3	=MEDIAN(1, 2, 3, 4, 5)						
21	3,5	=MEDIAN(A14:A19))						
22								

MODE

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.

File Edit View Insert Format Tools Data Window Help Adobe PDF

File Edit View Insert Format Tools Data Window Help Adobe PDF

10 B

B35		=						
	A	B	C	D	E	F	G	
24								
25	MODE(number1,number2, ...)							
26								
27	5.6							
28	4							
29	4							
30	3							
31	2							
32	4							
33	4	=MODE(A27:A32)						
34								

COUNT FUNCTION

Counts the number of cells that contain numbers and numbers within the list of arguments. Use COUNT to get the number of entries in a number field 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

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.

FREQUENCY

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**.

ARITHMETIC MEAN GROUPED DATA

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



H63 ▼ =

[illegible]

BUSINESS MATHEMATICS & STATISTICS

BUSINESS MATHEMATICS & STATISTICS

LECTURE 26

Review Lecture 25

Statistical Representation

Measures of Dispersion and Skewness

Part 1

1

Score Bins

79	70	Step 1: After copying the example to a blank					
85	79	worksheet, select the range A13:A16 starting					
78	89	with the formula cell.					

Step 2: Press F2

Step 3. Pess CTRL+SHIFT+ENTER

81

95

88

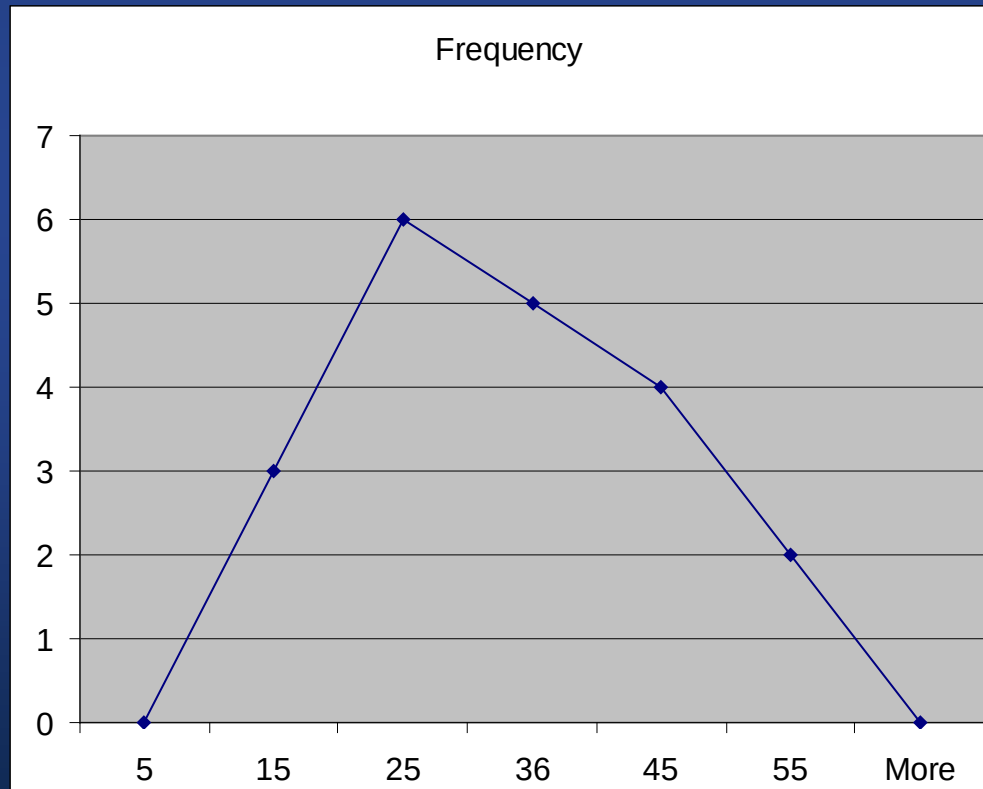
97

[illegible]

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



Class Midpoints

TABULATING NUMERICAL DATA: CUMULATIVE FREQUENCY

Data in ordered array

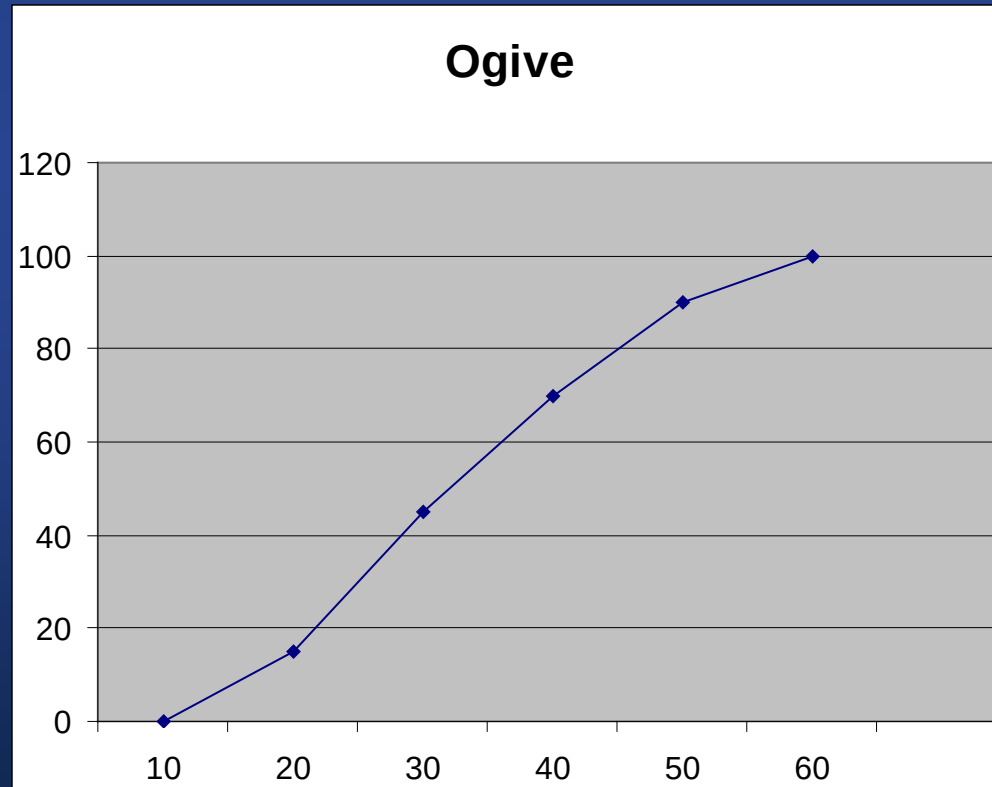
12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

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

GRAPHING NUMERICAL DATA: THE OGIVE (CUMULATIVE % POLYGON)

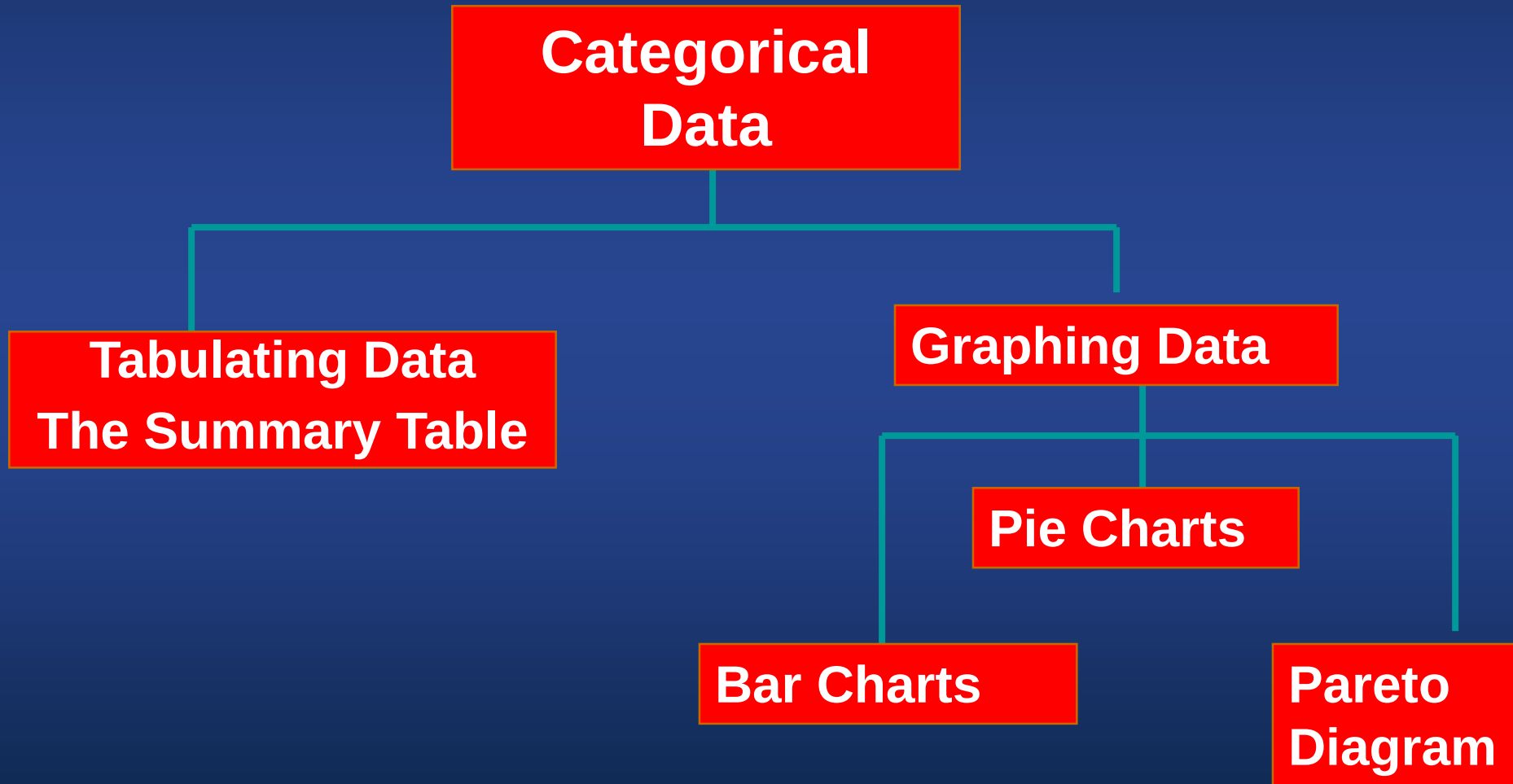
Data in ordered array

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



Class Boundaries (*Not Midpoints*)

TABULATING AND GRAPHING CATEGORICAL DATA: UNIVARIATE DATA



SUMMARY TABLE

(FOR AN INVESTOR'S PORTFOLIO)

Investment Category	Amount (in thousands Rs.)	Percentage
Stocks	46.5	42.27
Bonds	32.0	29.09
CD	15.5	14.09
Savings	16	14.55
Total	110	100

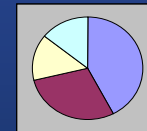
Variables are Categorical

GRAPHING CATEGORICAL DATA: UNIVARIATE DATA

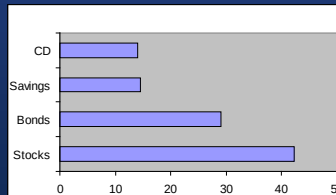
Categorical Data

Tabulating Data
The Summary Table

Graphing Data

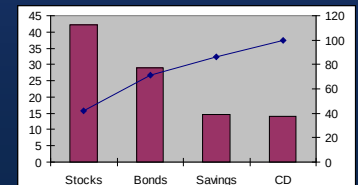


**Pie
Charts**



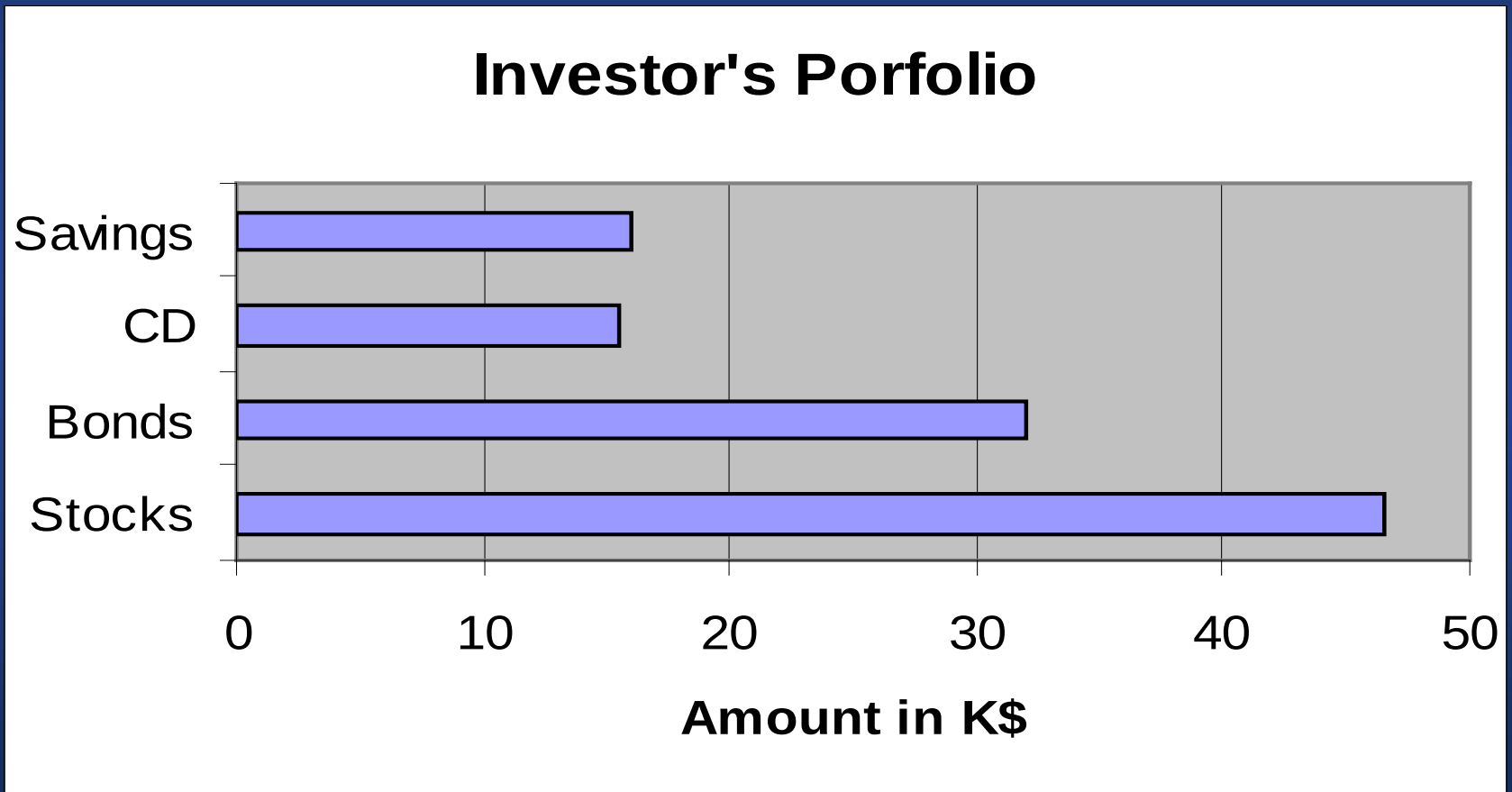
**Bar
Charts**

ParetoDiagram



BAR CHART

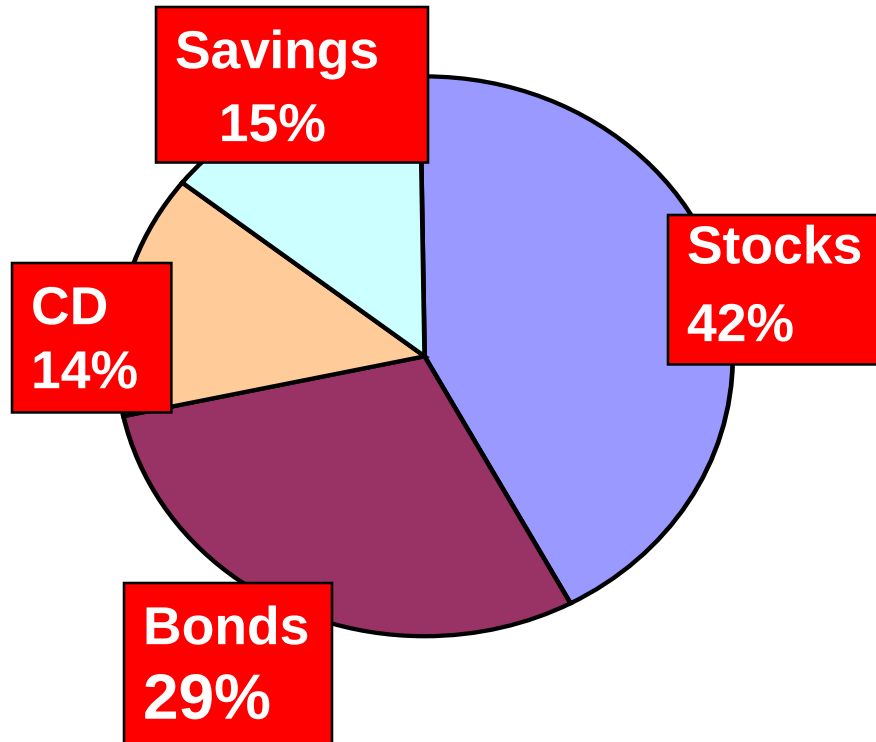
(FOR AN INVESTOR'S PORTFOLIO)



PIE CHART

(FOR AN INVESTOR'S PORTFOLIO)

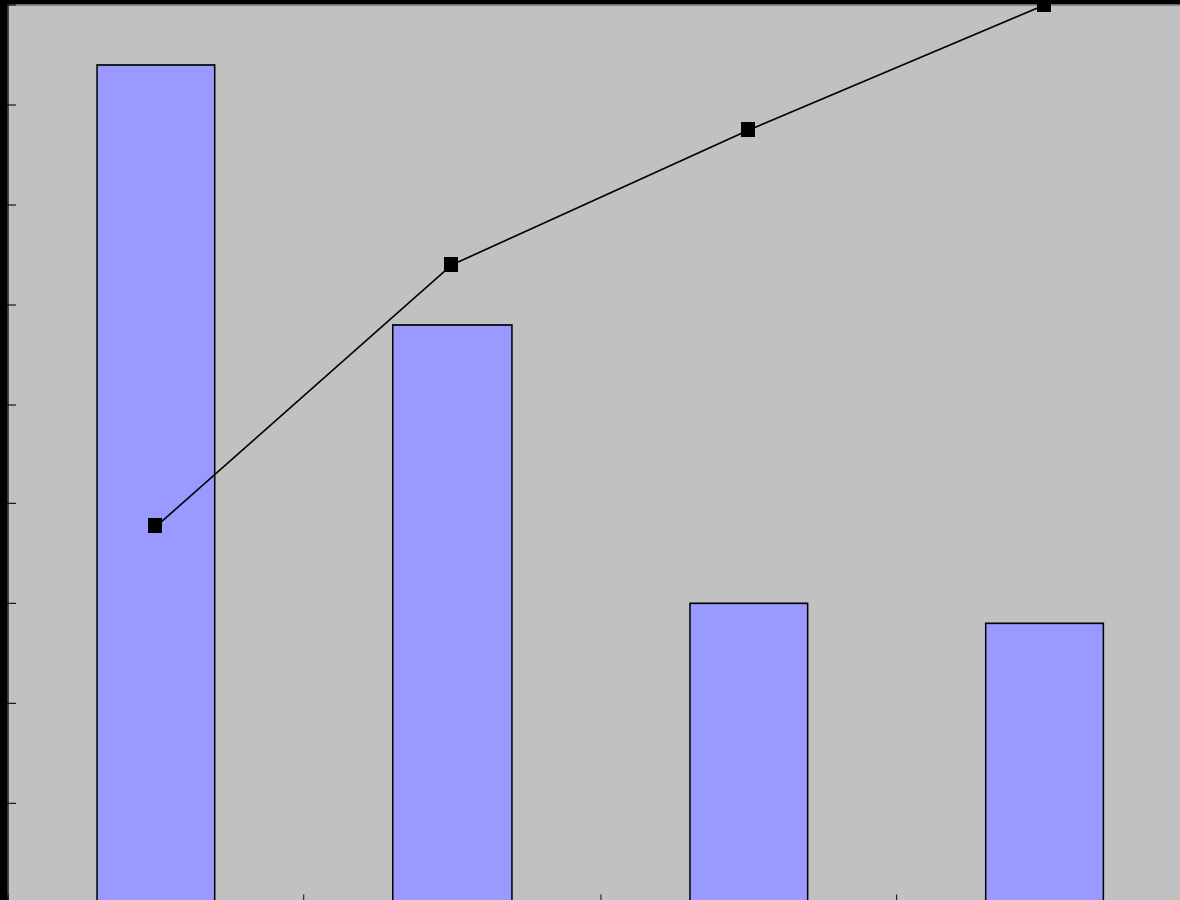
Amount Invested in Rs.



**Percentages
are rounded
to the nearest
percent.**

PARETO DIAGRAM

Axis for
bar chart
shows %
invested
in each
category



Axis for
line graph
shows
cumulative
% invested

TABULATING AND GRAPHING BIVARIATE CATEGORICAL DATA

Contingency Tables

Side by Side Charts

TABULATING CATEGORICAL DATA: BIVARIATE DATA

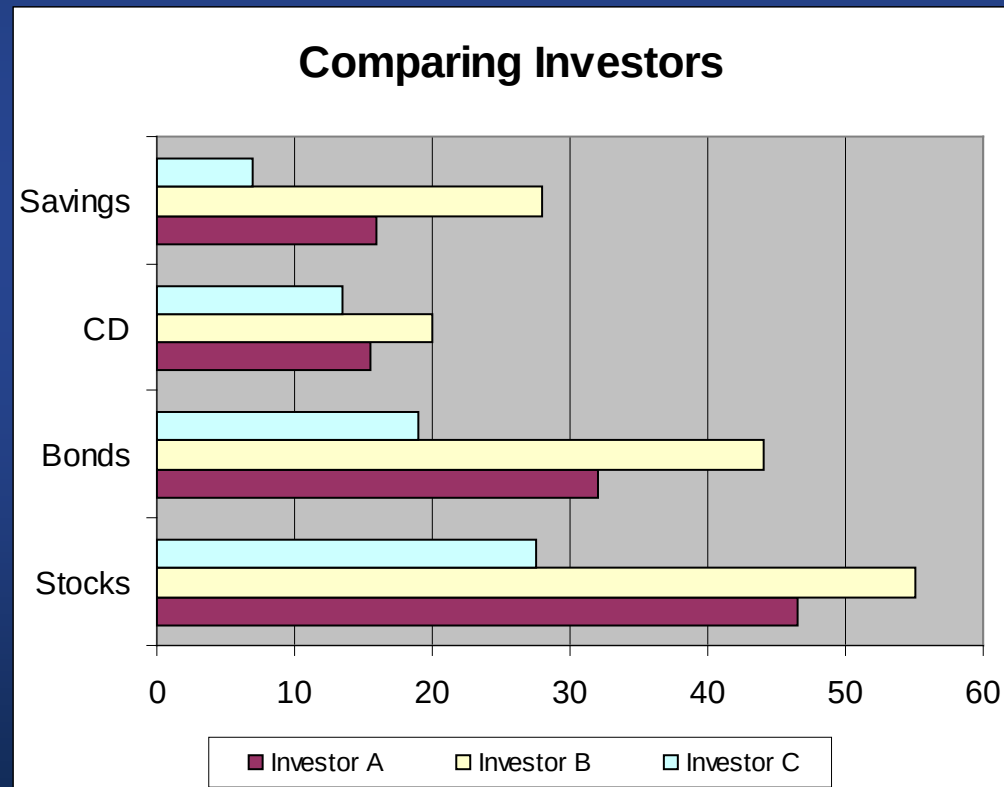
Contingency Table

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

GRAPHING CATEGORICAL DATA: BIVARIATE DATA

Side by
Side
Chart



GEOMETRIC MEAN

Syntax

$$G=(x1.x2.x3.....xn)^{1/n}$$

Example

Find GM of 130, 140, 160

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

MARMONIC MEAN

Syntax

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

Example

Find HM of 10, 8, 6

$$\begin{aligned} \text{HM} &= 3 / (1/10 + 1/8 + 1/6) \\ &= 7.66 \end{aligned}$$

BUSINESS MATHEMATICS

&

STATISTICS

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

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

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

$\text{mean} = \text{median} = \text{mode}$

Positively Skewed Distribution

(Tilted to left)

$\text{mean} > \text{median} > \text{mode}$

Negatively Skewed Distribution

$\text{mode} > \text{median} > \text{mean}$

(Tilted to right)

EMPIRICAL RELATIONSHIPS

Moderately Skewed and Unimodal Distribution

$$\text{Mean} - \text{Mode} = 3(\text{Mean} - \text{Median})$$

Example

mode = 15, mean = 18, median = ?

$$\text{Median} = 1/3[\text{mode} + 2 \text{ mean}]$$

$$= 1/3[15 + 2(18)]$$

$$= [15+36]/3 = 51/3 = 17$$

MODIFIED MEANS

TRIMMED MEAN

Remove all observations below 1st quartile and above 3rd Quartile

Winsorized MEAN

Replace each observation below first quartile with value of first quartile

Replace each observation above the third quartile with value of 3rd 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

$$Q1 = (6+1)/4 = 1.75 \text{ (2nd value)} = 9.2$$

$$Q3 = 3(6+1)/4 = 5.25 \text{ (6th value)} = 9.3$$

$$TM = (9.2+9.2+9.2+9.2+9.3)/5 = 9.22$$

$$WM = (9.2+9.2+9.2+9.2+9.2+9.3+9.3)/7$$

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

1. Range
2. Quartile Deviation
3. Mean Deviation
4. Standard Deviation or Variance

TYPES OF RELATIVE MEASURES

1. Coefficient of Range
2. Coefficient of Quartile Deviation
3. Coefficient of Mean Deviation
4. Coefficient of Variation

BUSINESS MATHEMATICS & STATISTICS

BUSINESS MATHEMATICS & STATISTICS

BUSINESS MATHEMATICS & STATISTICS

LECTURE 27

Measures of Dispersion and Skewness

Part 2

Measures of Central Tendency, Variation and Shape for a Sample

Mean, Median, Mode, Midrange, Quartiles, Midhinge

Range, Interquartile Range, Variance, Standard
Deviation, Coefficient of Variation

Right-skewed, Left-skewed, Symmetrical

Measures of Central Tendency, Variation and Shape



Exploratory Data Analysis

Five-Number Summary

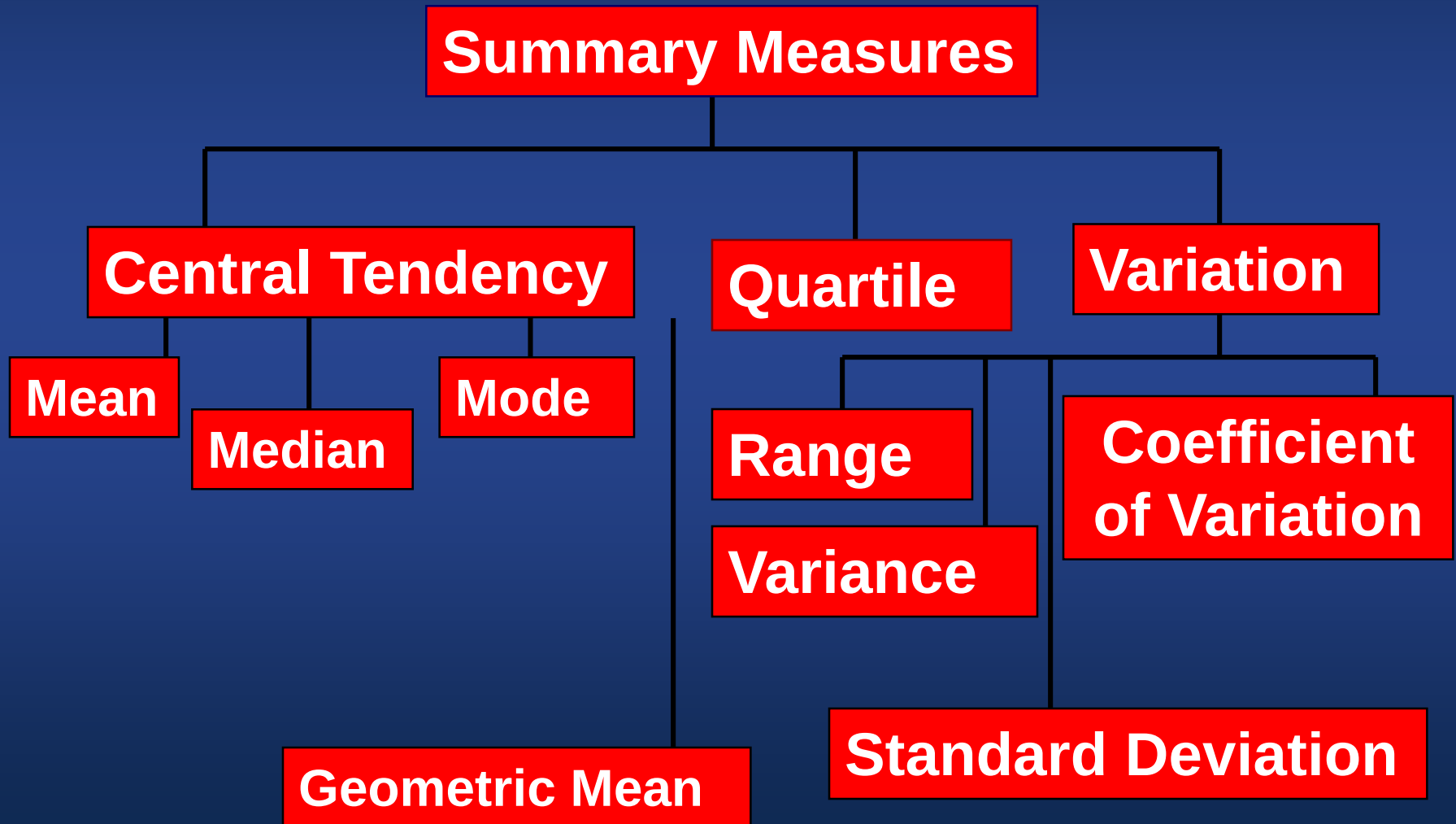
Box-and-Whisker Plot

Proper Descriptive Summarization

Exploring Ethical Issues

Coefficient of Correlation

Summary Measures



Measures of Central Tendency

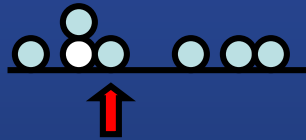
Central Tendency

Mean

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

$$\mu = \frac{\sum_{i=1}^N X_i}{N}$$

Median



Mode

Geometric Mean

$$\bar{X}_G = (X_1 \times X_2 \times \cdots \times X_n)^{1/n}$$

The Mean (Arithmetic Average)

The Arithmetic Average of data values:

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

Sample Mean

Sample Size

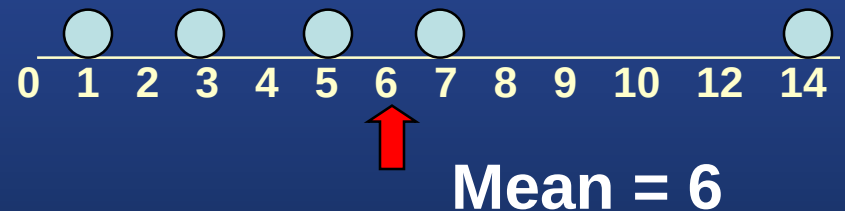
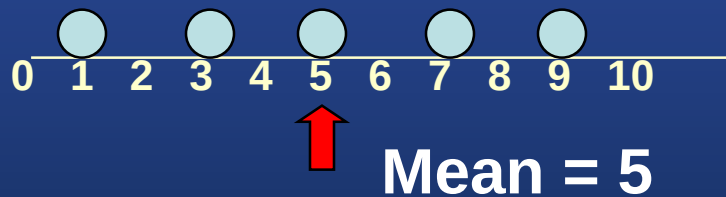
$$\mu = \sum_{i=1}^N X_i / N = \frac{X_1 + X_2 + \cdots + X_N}{N}$$

Population Mean

Population Size

The Mean

The Most Common Measure of Central Tendency
Affected by Extreme Values (Outliers)



The Median

Important Measure of Central Tendency

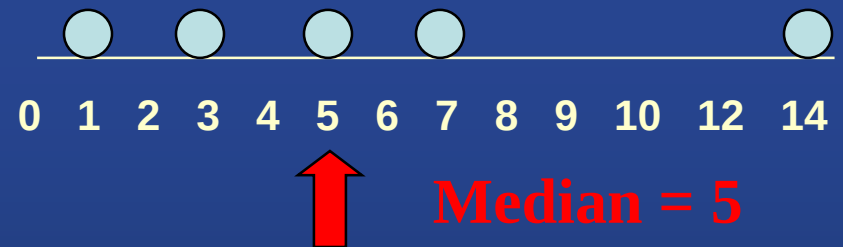
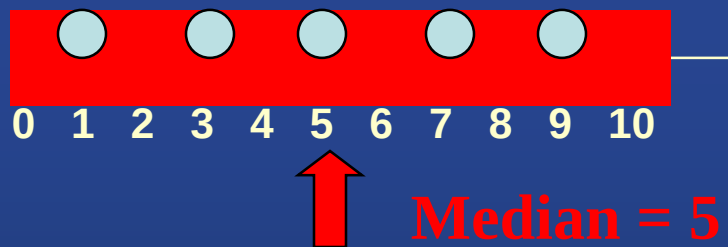
In an ordered array, the median is the
“middle” number.

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

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

The Median (continued)

Not Affected by Extreme Values

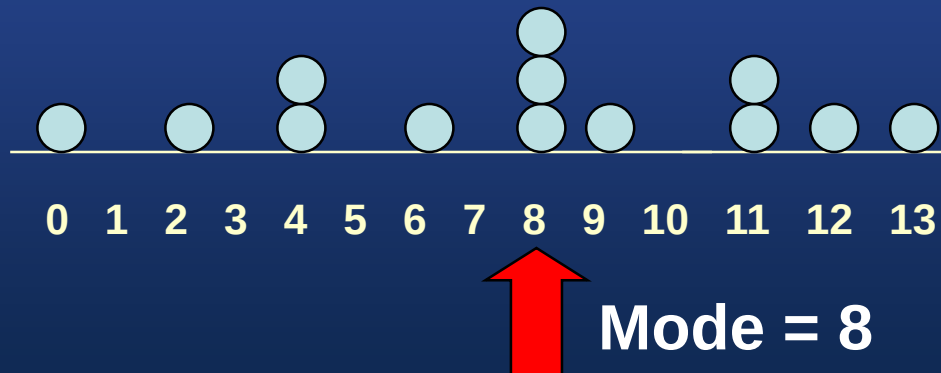


The Mode

A Measure of Central Tendency

Value that Occurs Most Often

Not Affected by Extreme Values

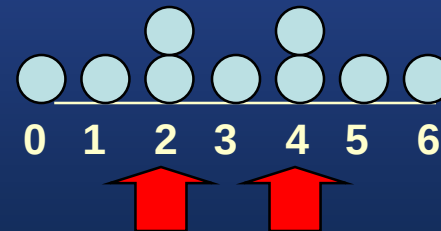
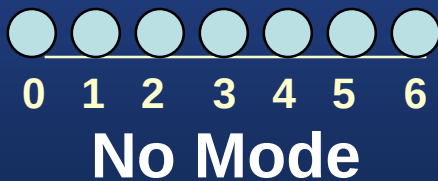


The Mode

There May Not be a Mode

There May be Several Modes

Used for Either Numerical or Categorical Data



Two Modes

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 = $43 - 12 = 31$

Midrange

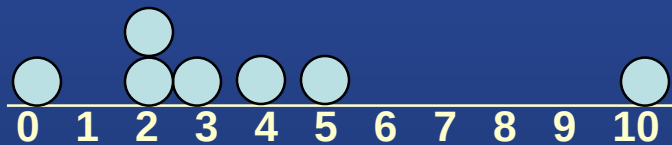
A Measure of Central Tendency

Average of Smallest and Largest Observation:

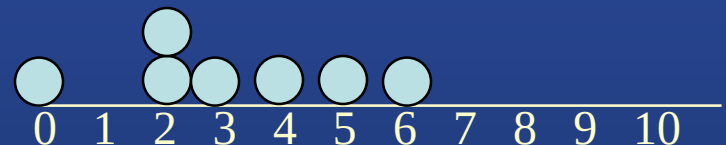
$$\text{Midrange} = \frac{X_{\text{largest}} + X_{\text{smallest}}}{2}$$

Midrange

Affected by Extreme Value



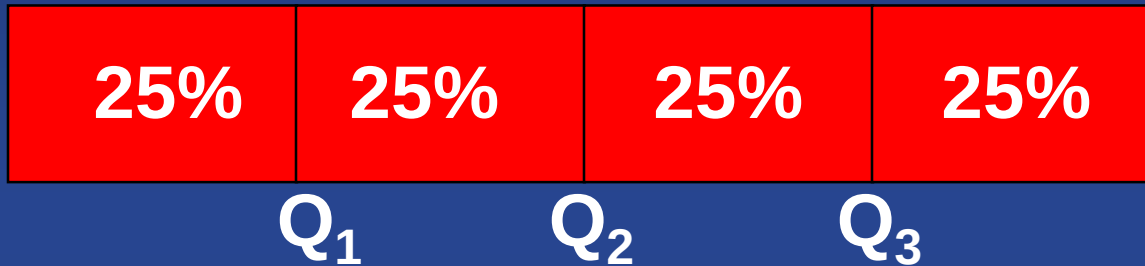
Midrange = 5



Midrange = 3

Quartiles

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



Position of i-th quartile:

$$Q_i = \frac{i(n+1)}{4}$$

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

A red arrow points upwards from the space between the numbers 12 and 13 in the ordered array.

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

$Q_1 = 12.5$

DISPERSION OF DATA

Quartile Deviation

$$Q.D = (Q3 - Q1)/2$$

Example

Find Q.D

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

$$\begin{aligned} Q1 &= (n+1)/4^{\text{th}} \text{ value} = (10+1)/4 = 2.75^{\text{th}} \\ &= 8 + 0.75(9 - 8) = 8 + 0.75 \times 1 = 8.75 \end{aligned}$$

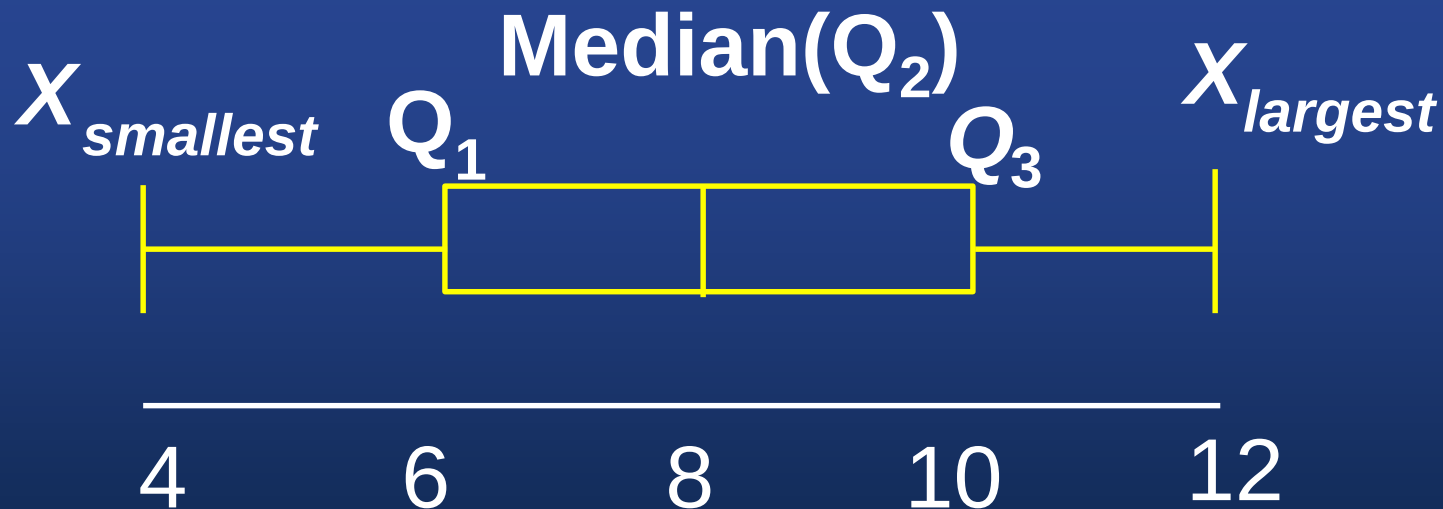
$$\begin{aligned} Q3 &= 3(2.75) = 7.75^{\text{th}} \text{ value} \\ &= 8^{\text{th}} \text{ value} + 0.25(8^{\text{th}} \text{ value} - 7^{\text{th}} \text{ value}) \\ &= 17 + 0.75 (20 - 17) = 19.25 \end{aligned}$$

$$Q.D = (19.25 - 8.75)/2 = 5.25$$

Exploratory Data Analysis

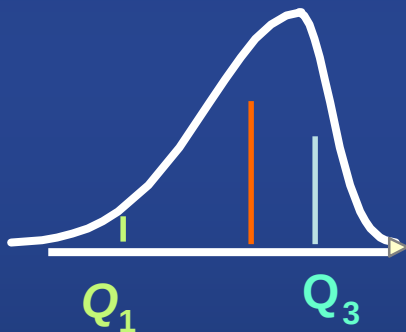
Box-and-whisker:

Graphical display of data using 5-number summary



Distribution Shape & Box-and-whisker Plots

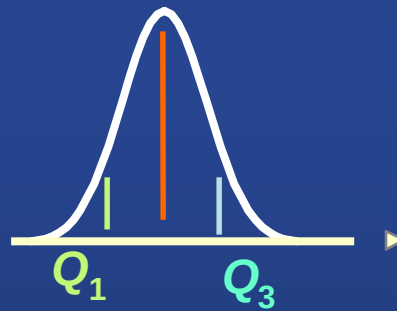
Left-Skewed



Median



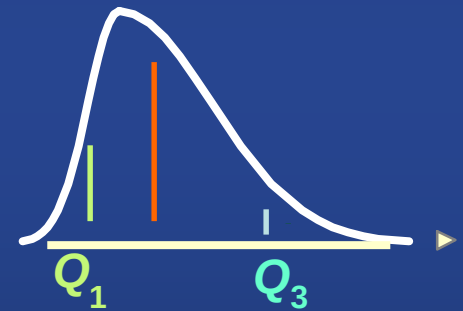
Symmetric



Median



Right-Skewed



Median



Five-Number Summary (Symmetrical Distribution)

Smallest Q_1 Median Q_3 X_{largest}

Data perfectly symmetrical if:

Distance from Q_1 to Median = Distance from Median to Q_3

Distance from X_{smallest} to Q_1 = Distance from Q_3 to X_{largest}

Median = Midhinge = Midrange

Five-Number Summary (Nonsymmetrical Distribution)

X_{smallest} Q_1 Median Q_3 X_{largest}

Right-skewed distribution

Median < Midhinge < Midrange

Distance from X_{largest} to Q_3 greatly exceeds distance from Q_1 to X_{smallest}

Left-skewed distribution

Median > Midhinge > Midrange

Distance from Q_1 to X_{smallest} greatly exceeds distance from X_{largest} to Q_3

Determining the 5-Number Summary

Example

Example

For the sample representing annual costs (in Rs.000) for attending 10 Conferences, the ordered array is:

Ordered: 13.0 14.3 14.9 15.2 15.2 15.4 15.6 16.4
 17.0 23.1

State the 5-number summaries:

$$X_{\text{smallest}} = 13.0$$

$$Q_1 = 14.9$$

$$\text{Median} = 15.3$$

$$Q_3 = 16.4$$

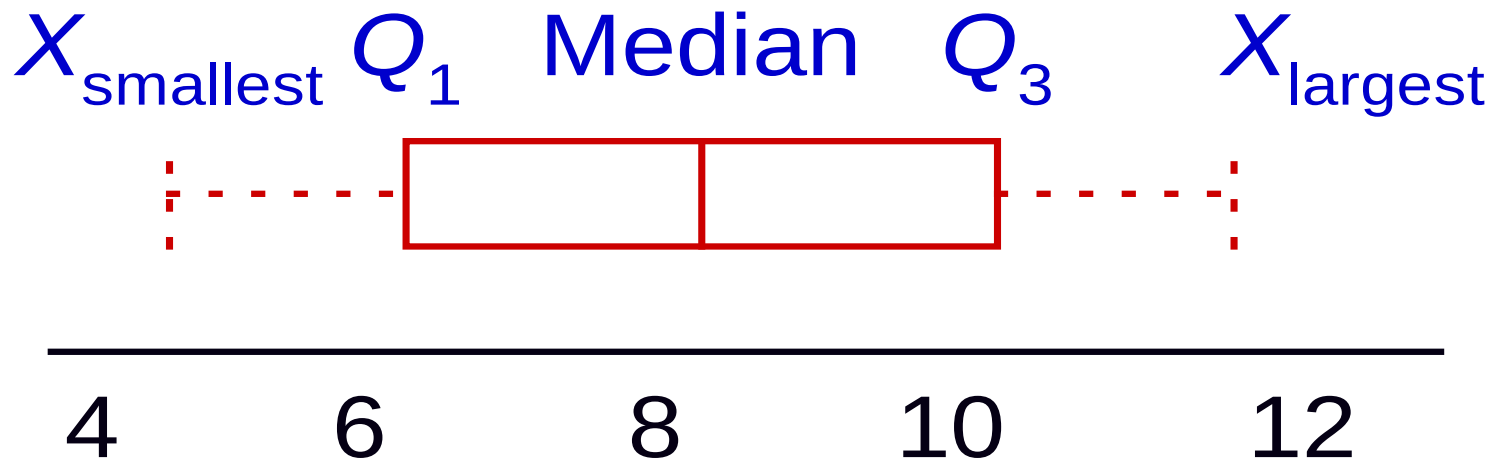
$$X_{\text{largest}} = 23.1$$

Data appear right-skewed

Median < Midhinge < Midrange

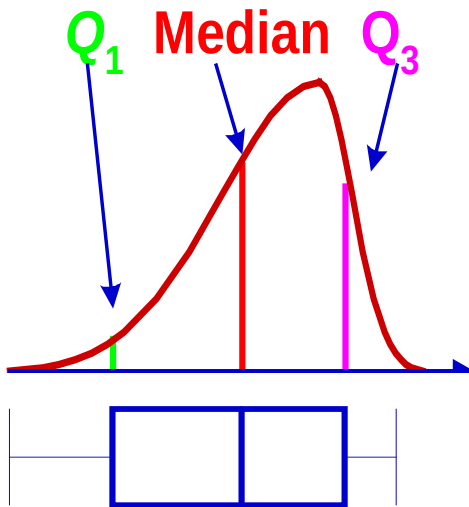
Box-and-Whisker Plot

**Graphical display of data using
5-number summary**

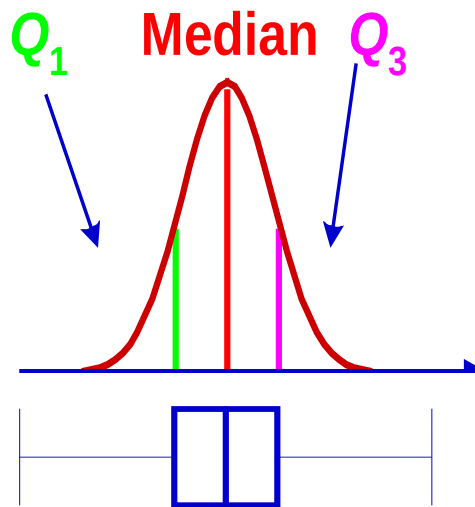


Shape & Box-and-Whisker Plot

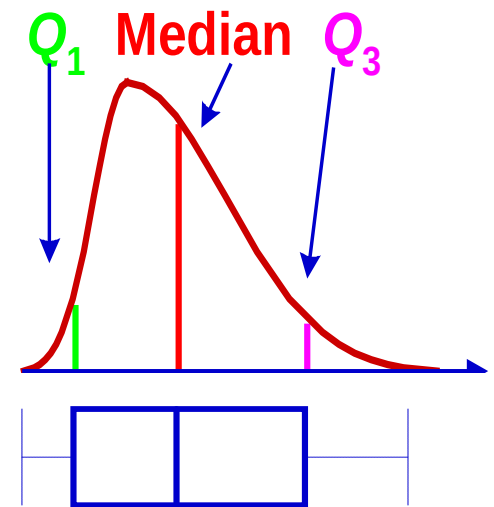
Left-Skewed



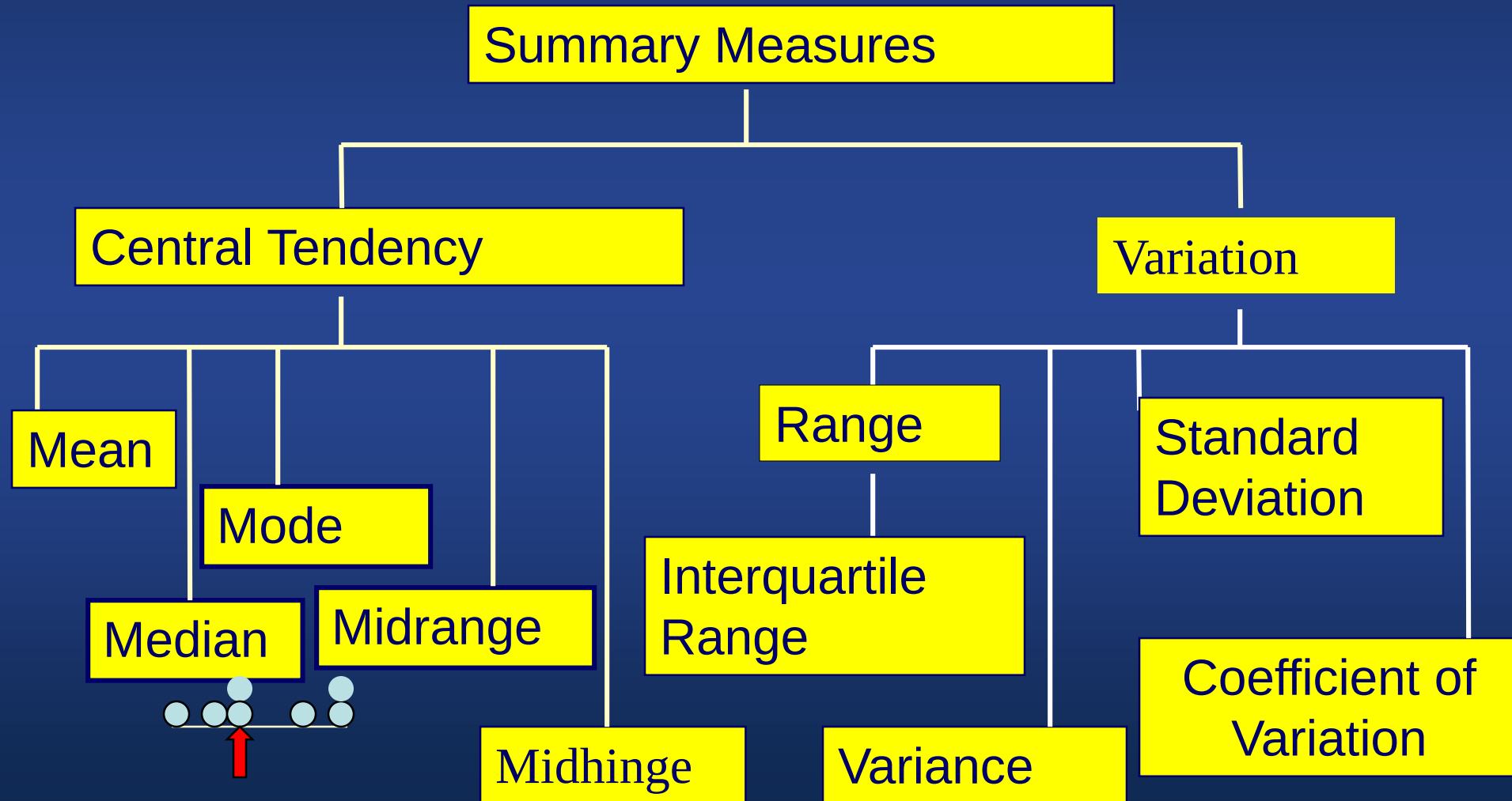
Symmetric



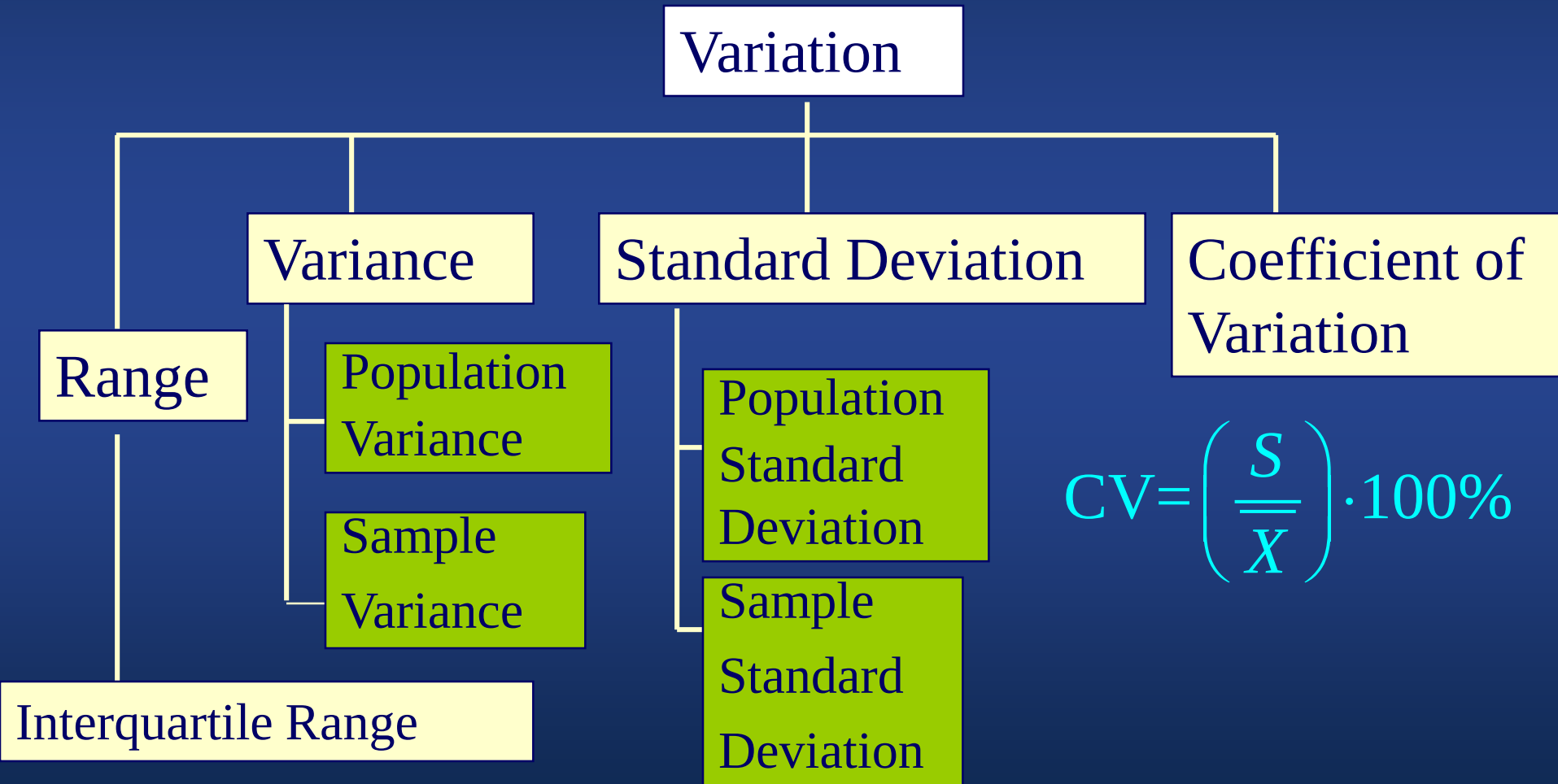
Right-Skewed



Summary Measures



Measures of Variation



The Range

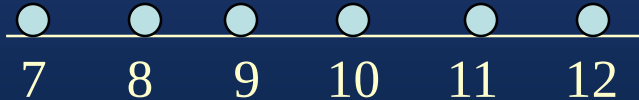
Measure of Variation Difference Between Largest & Smallest Observations:

Range =

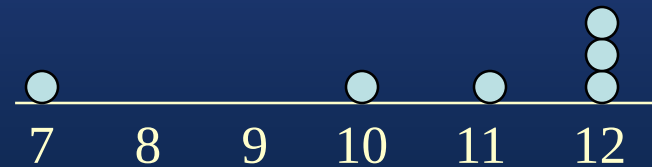
$$X_{\text{Largest}} - X_{\text{Smallest}}$$

Ignores How Data Are Distributed:

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



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



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

BUSINESS MATHEMATICS

&

STATISTICS

BUSINESS MATHEMATICS & STATISTICS

Module 6

Correlation

(Lecture 28-29)

Line Fitting

(Lectures 30-31)

Time Series and Exponential Smoothing

(Lectures 32-33)

LECTURE 28

Review Lecture 27

Measures of Dispersion

Correlation

Part 1

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.

Standard Deviation

Most Important Measure of Variation

Shows Variation About the Mean

Same unit of measurement as the observations

For the Population:

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

For the Sample:

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

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

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

Sample Standard Deviation

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

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

Data: X_i : 10 12 14 15 17 18 18 24

$n = 8$

Mean = 16

$$s = \sqrt{\frac{(10-16)^2 + (12-16)^2 + (14-16)^2 + (15-16)^2 + (17-16)^2 + (18-16)^2 + (18-16)^2 + (24-16)^2}{8-1}}$$

= 4.2426

Comparing Standard Deviations

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

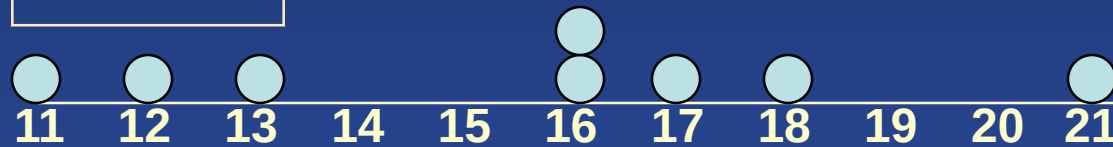
$$s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}} = 4.2426$$

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

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

Comparing Standard Deviations

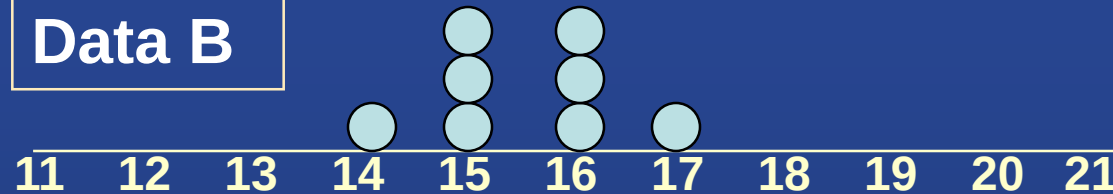
Data A



Mean = 15.5

$s = 3.338$

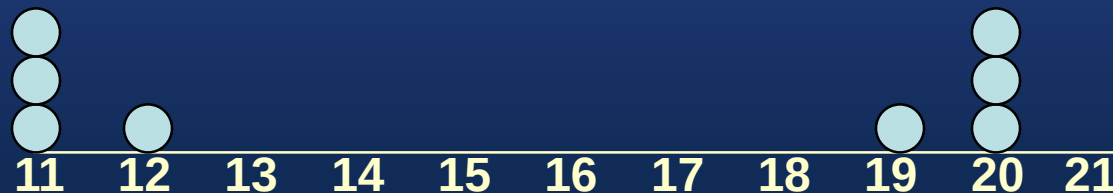
Data B



Mean = 15.5

$s = .9258$

Data C



Mean = 15.5

$s = 4.57$

Coefficient of Variation

Measure of **relative variation**

Always a %

Shows variation **relative to** mean

Used to **compare 2 or more groups**

Formula (for sample):

$$CV = \left(\frac{S}{\bar{X}} \right) \cdot 100\%$$

Comparing Coefficient of Variation

- Stock A: average price last year = Rs. 50
- Standard deviation = Rs. 5
- Stock B: average price last year = Rs. 100
- Standard deviation = Rs. 5

$$CV = \left(\frac{S}{\overline{X}} \right) \cdot 100\%$$

Coefficient of Variation:

Stock A: CV = 10%

Stock B: CV = 5%

Shape

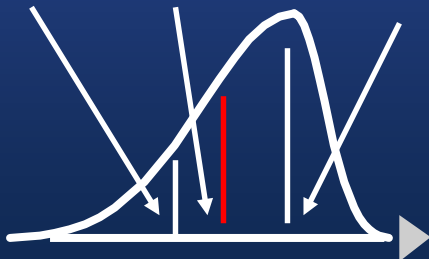
Describes how data are distributed

Measures of shape:

Symmetric or skewed

Left-Skewed

$\text{Mean} < \text{Median} < \text{Mode}$



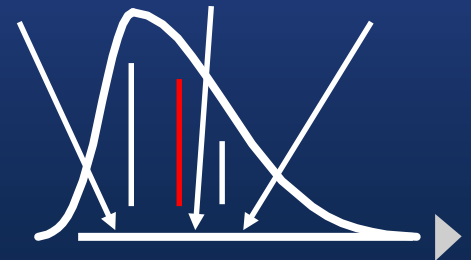
Symmetric

$\text{Mean} = \text{Median} = \text{Mode}$



Right-Skewed

$\text{Mode} < \text{Median} < \text{Mean}$



DISPERSION OF DATA

Mean Deviation About Mean

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

For Grouped data

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

Mean Deviation About Median

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

For Grouped data

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

CORRELATION

Types of regression models

Determining the simple linear regression equation

Measures of variation in regression and correlation

Assumptions of regression and correlation

CORRELATION

Residual analysis

Inferences about the slope

Estimation of predicted values

Pitfalls in regression and ethical issues

Correlation - measuring the strength of the association

Purpose of Regression Analysis

Regression Analysis Is Used Primarily to Model Causality
And Provide Prediction

Predict the values of a dependent (response) variable
based on values of at least one independent (explanatory
variable)

Explain the effect of the independent variables on the
dependent variable

Purpose of Regression Analysis

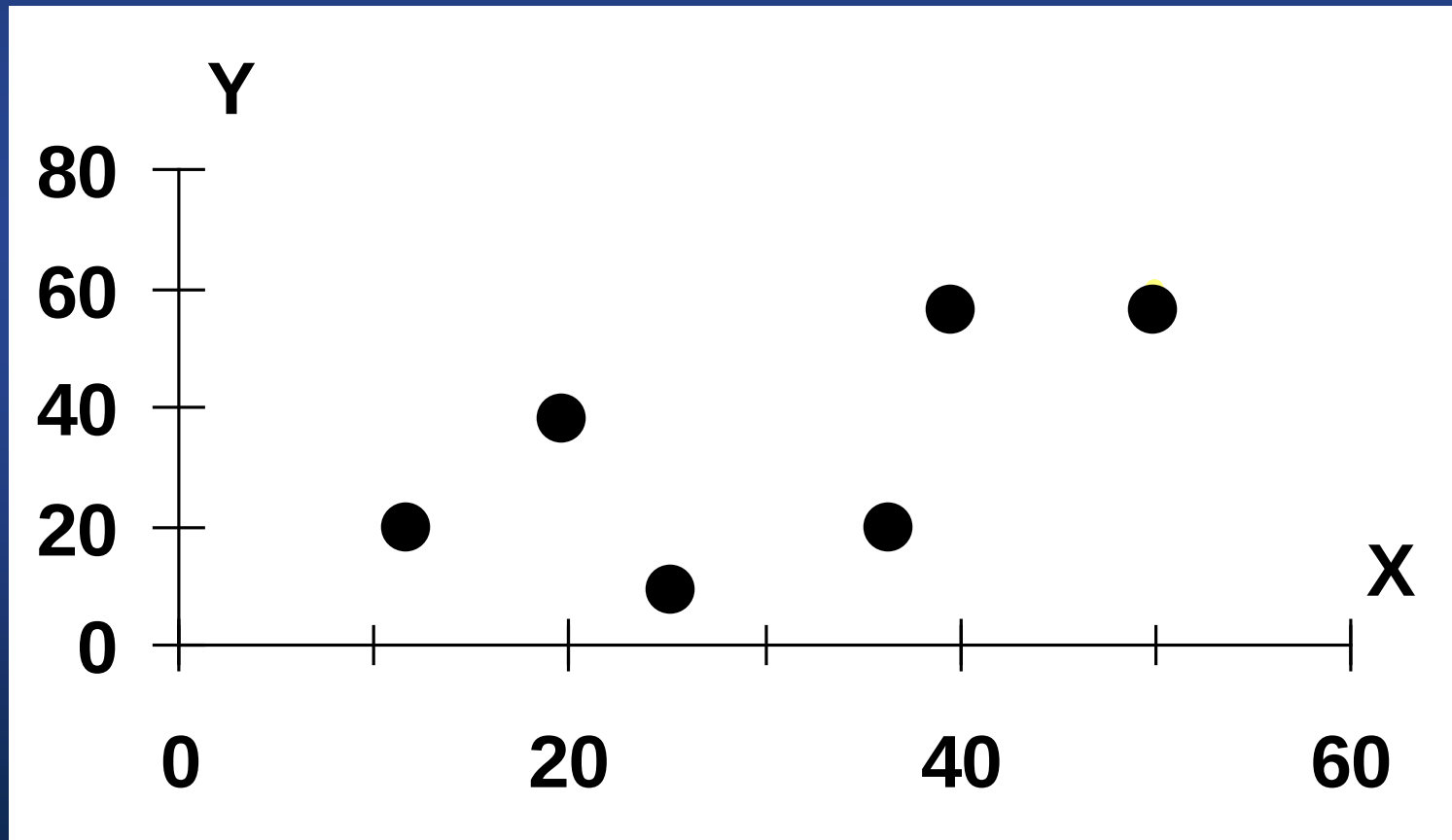
Correlation Analysis Is Used to Measure Strength of Association Between Numerical Variables

Only Strength of the Relationship is Concerned

No Causal Effect is Implied

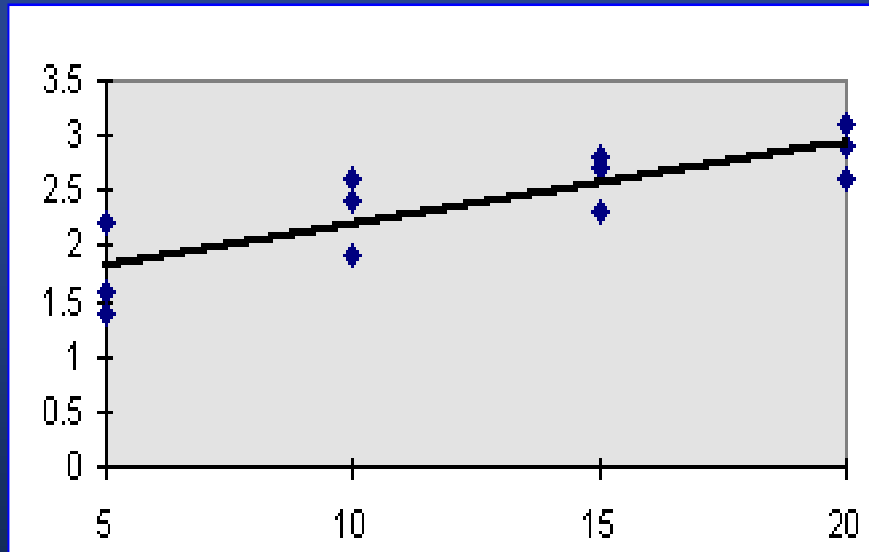
The Scatter Diagram

Plot of all (X_i, Y_i) pairs

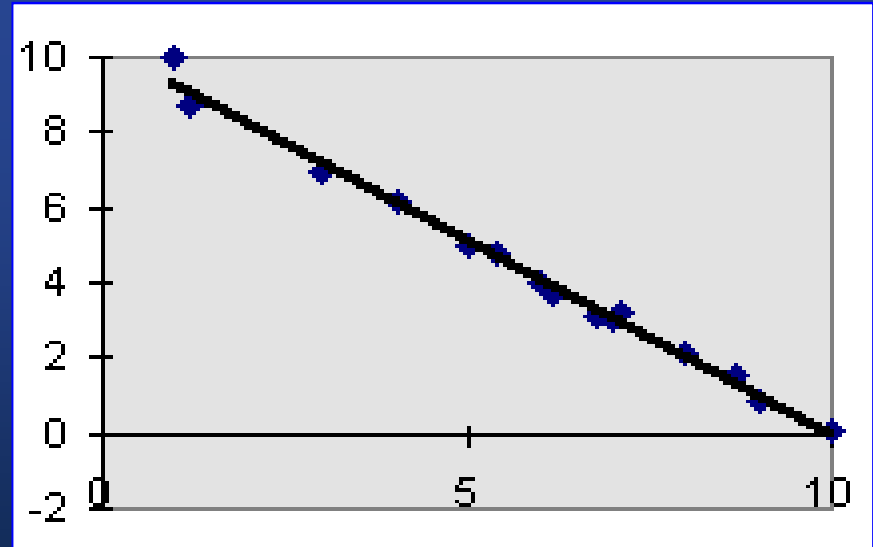


Types of Regression Models

Positive Linear Relationship



Negative Linear Relationship



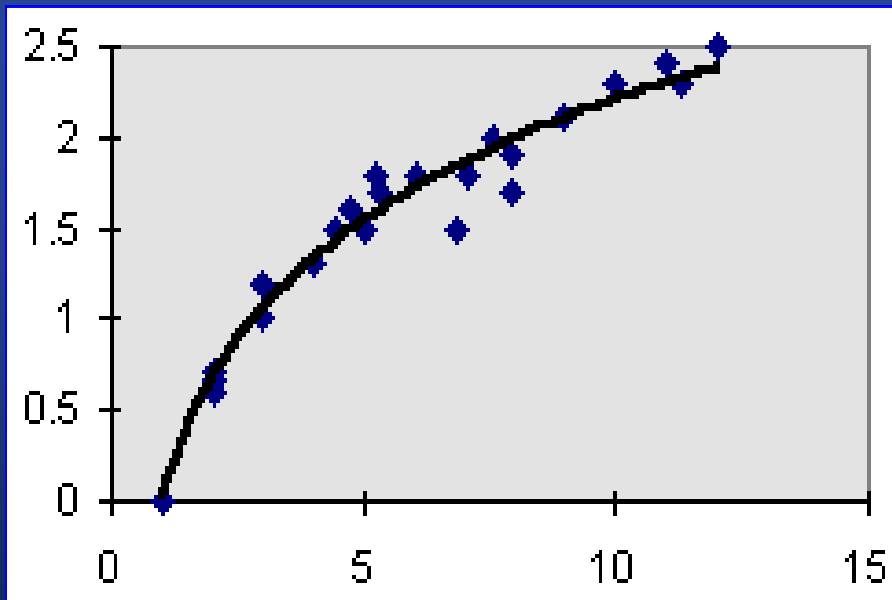
BUSINESS MATHEMATICS

&

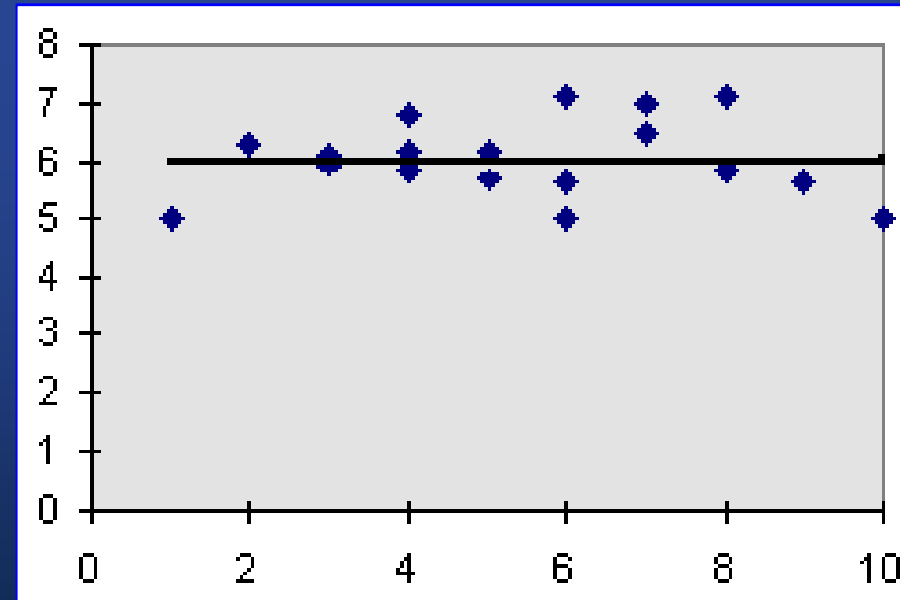
STATISTICS

Types of Regression Models

Relationship NOT Linear



No Relationship



Simple Linear Regression Model

**Relationship Between Variables Is
Described by A Linear
Function**

**The Change of One Variable Causes
The Other Variable to Change
A Dependency of One Variable on
the Other**

Population Linear Regression

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

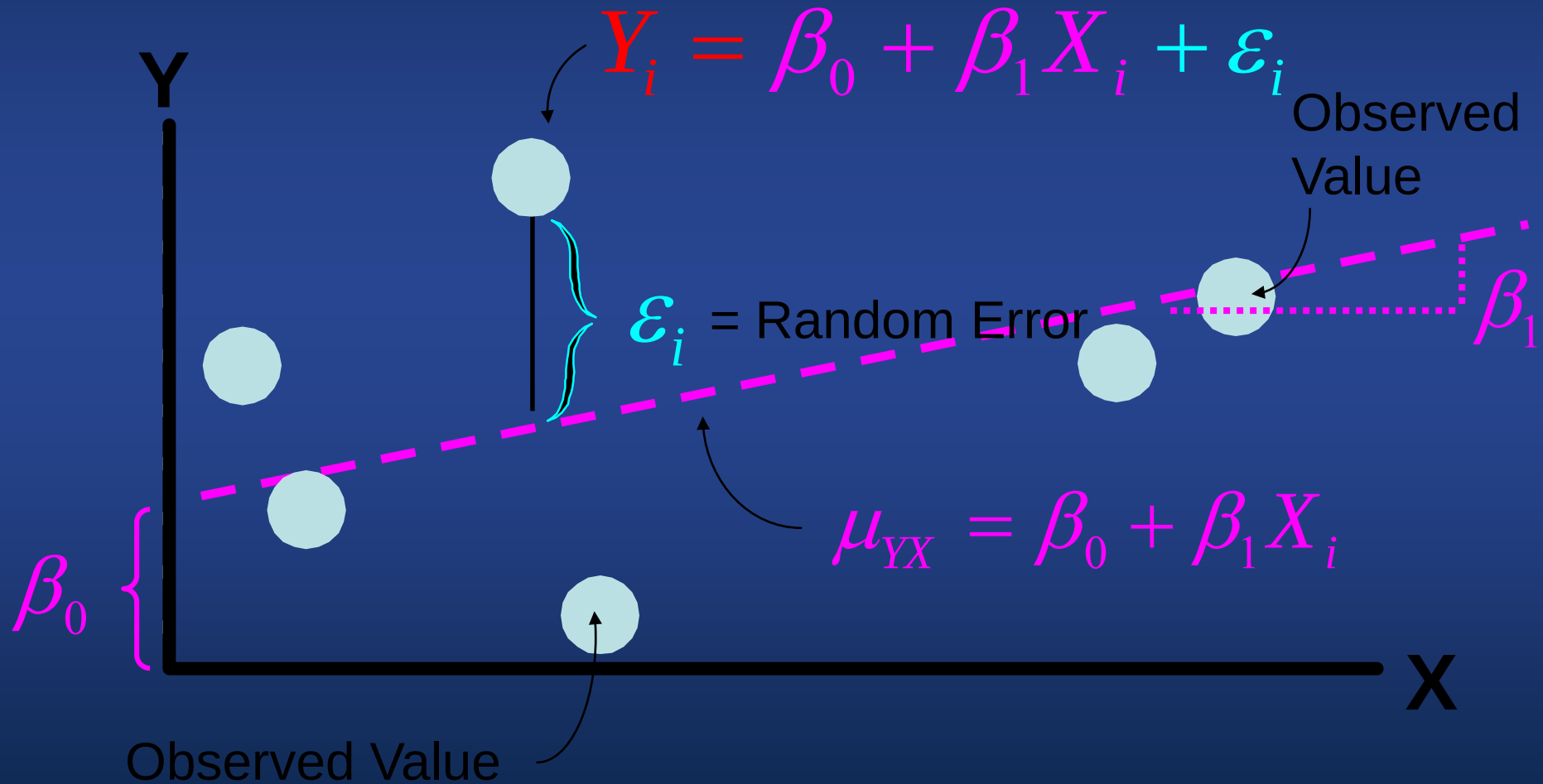
The diagram illustrates the Population Linear Regression equation, $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, with labels for each component:

- Population Y intercept**: Points to β_0 .
- Population Slope Coefficient**: Points to β_1 .
- Random Error**: Points to ε_i .
- Dependent (Response) Variable**: Points to Y_i .
- Population Regression Line**: Points to the entire equation.
- Independent (Explanatory) Variable**: Points to X_i .

The equation is displayed in a white box with the following terms highlighted in color: Y_i (red), β_0 (magenta), β_1 (magenta), X_i (magenta), and ε_i (cyan). A magenta bracket is placed under the $\beta_0 + \beta_1 X_i$ portion of the equation.

Population Linear Regression

(continued)



Sample Linear Regression

- Sample Regression Line Provides an Estimate of The Population Regression Line as well as a Predicted Value of Y

Sample
Y Intercept

Sample
Slope
Coefficient

$$Y_i = b_0 + b_1 X_i + e_i$$

Residual

b_0 provides an estimate of β_0

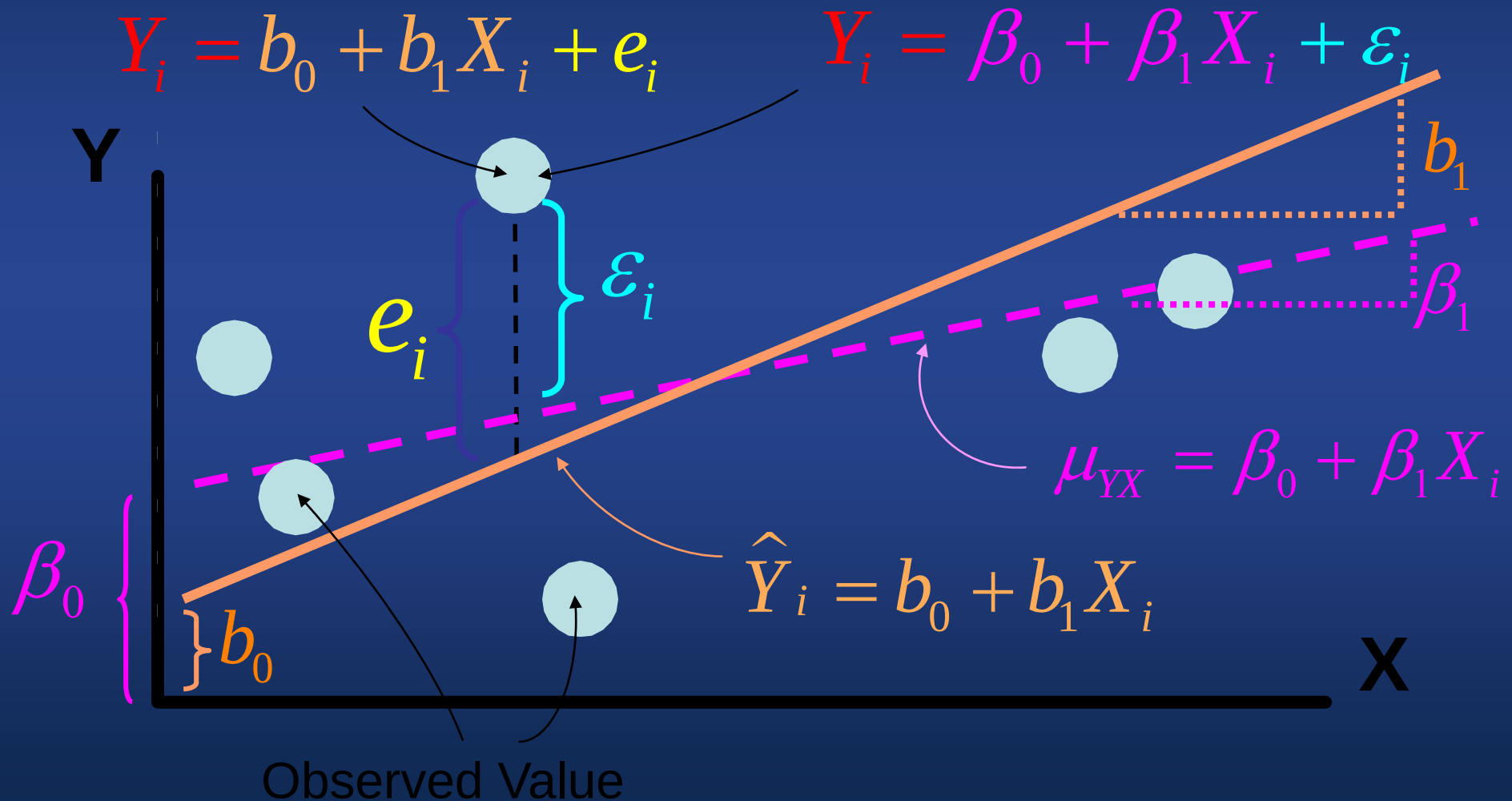
b_1 provides an estimate of β_1

Sample
Regression
Line

$\hat{Y}$

Sample Linear Regression

(continued)



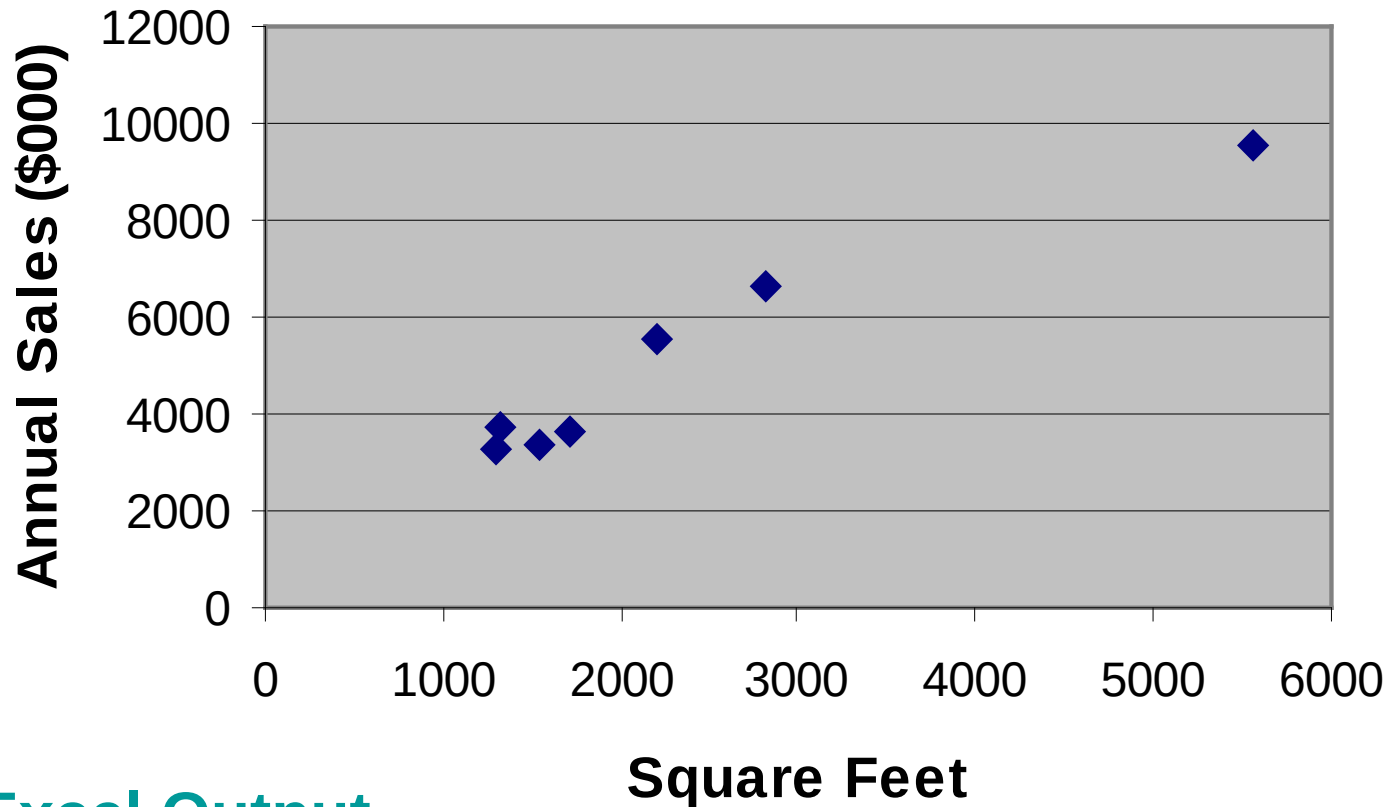
Simple Linear Regression

Equation: Example

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**

Store	Square Feet	Annual Sales (1000)
1	1,726	3,681
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

Scatter Diagram Example



Excel Output

Equation for Sample Regression Line

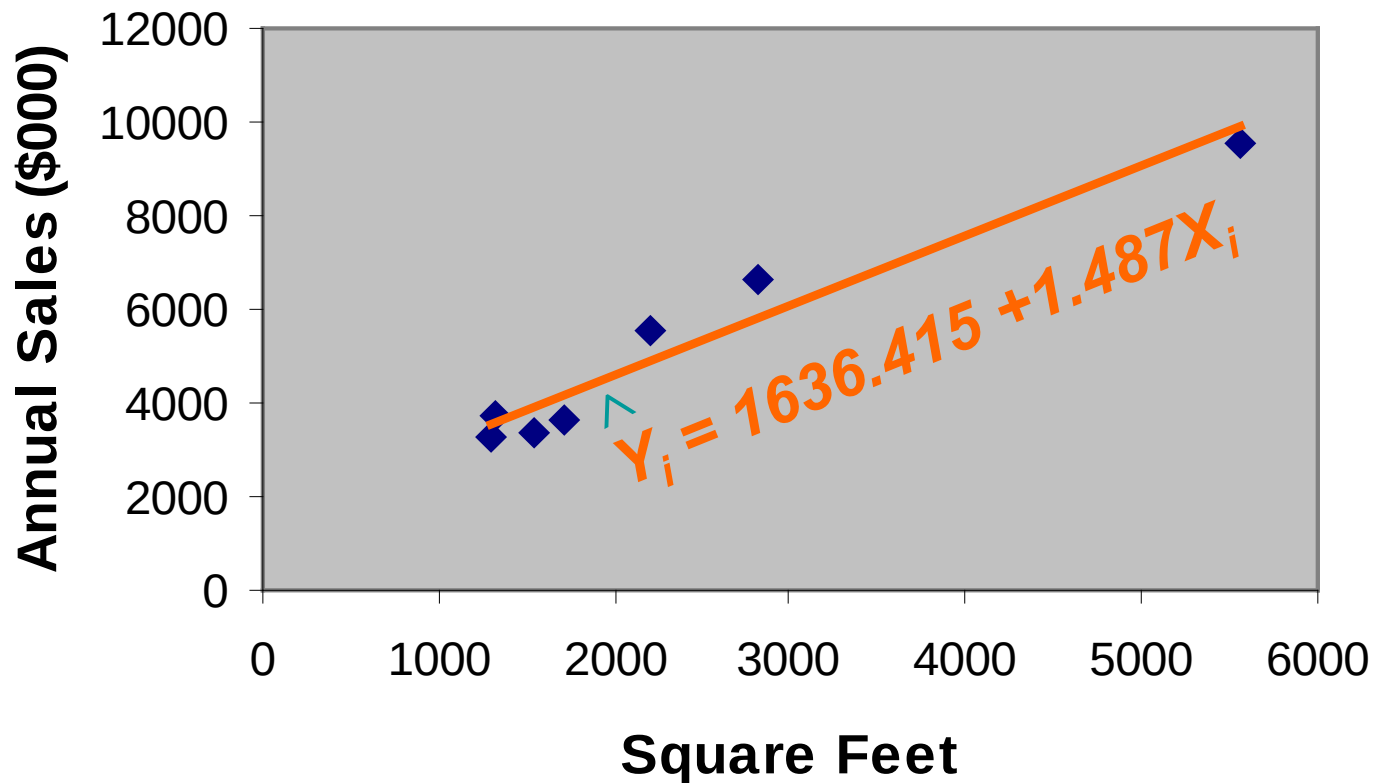
$$\begin{aligned}\hat{Y}_i &= b_0 + b_1 X_i \\ &= 1636.415 + 1.487 X_i\end{aligned}$$

From Excel Printout:

	<i>Coefficients</i>
Intercept	1636.414726
X Variable	1.486633657



Graph of the Sample Regression Line



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 \$1487.

BUSINESS MATHEMATICS & STATISTICS

BUSINESS MATHEMATICS & STATISTICS

LECTURE 29

Correlation

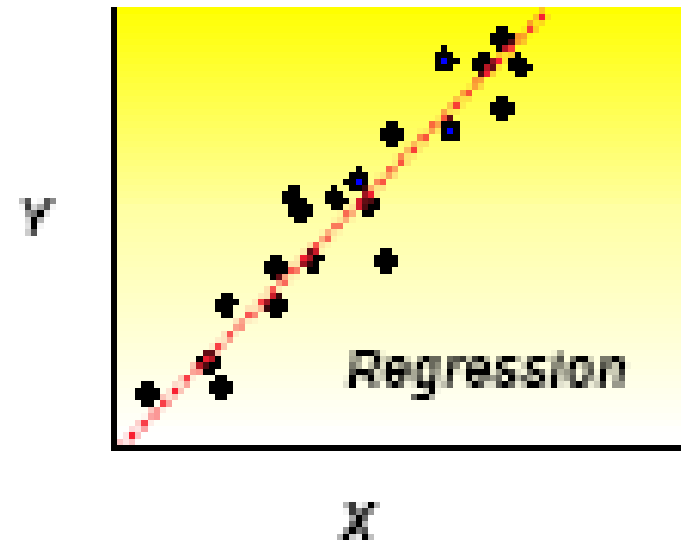
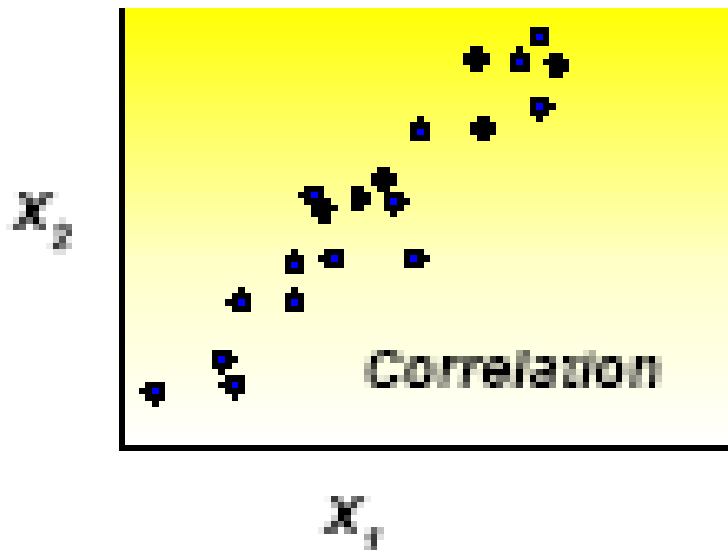
Part 2

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

Calculations are the same.

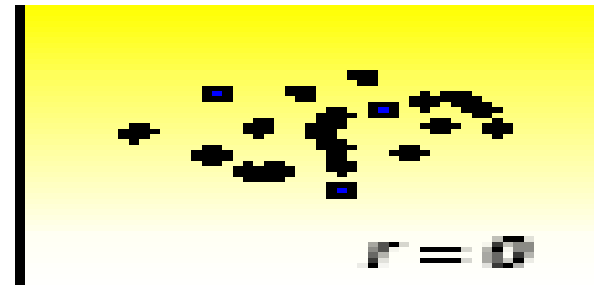
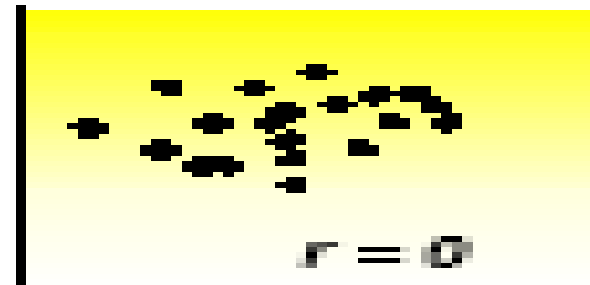
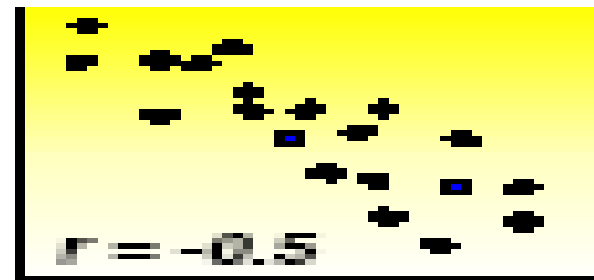
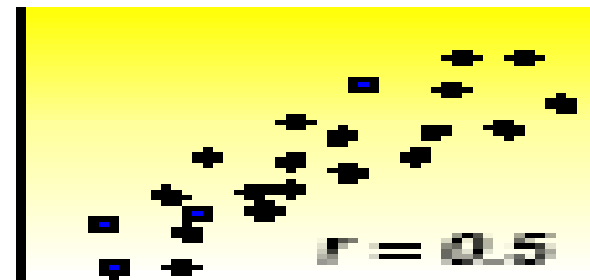
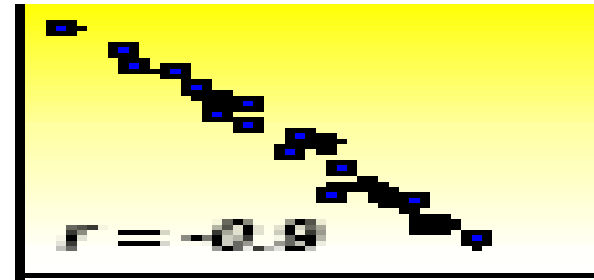
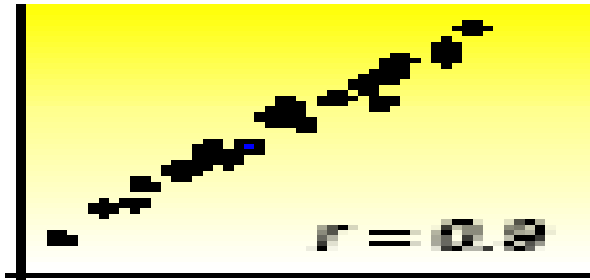
In correlation analysis, one must sample
randomly both X and Y.

Correlation deals with association
(importance)

Regression deals with prediction (intensity)

TYPICAL CORRELATION

y vs x



Correlation coefficient

- A standardized transform of the covariance is calculated by dividing it by the product of the standard deviations of X & Y.
- It is called the population correlation coefficient is defined as :

$$r = s_{XY}/s_X s_Y$$

Properties

- For the population :
- $-1 \leq r \leq +1$
- $r = 0$ no linear relationship
- $r = -1$ perfect negative relationship
- $r = +1$ perfect positive relationship

CORRELATION

- r always lies between -1 and 1
- r^2 is the coefficient of determination, which measures the proportion of the variance in X_1 (or X_2) “explained” by variation in X_2 or X_1
- 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
- There can exist spurious correlations, and correlations can arise because of the action of a third unmeasured or unknown variable

Measuring the strength of a correlation

Test statistic is the
product-moment
correlation coefficient r

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

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

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

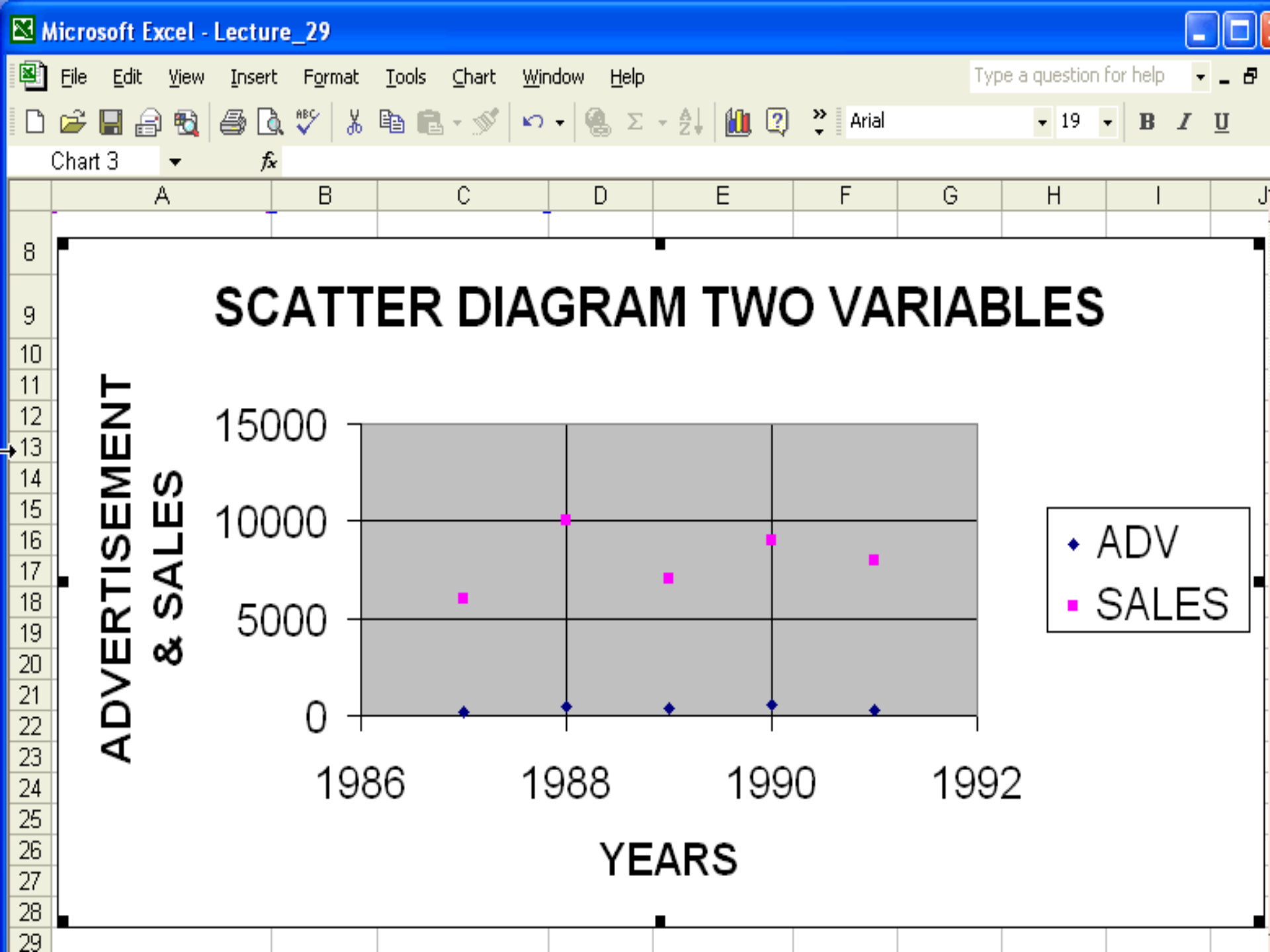
$$s(y) = \left(\frac{\sum (y^2) - (y_m^2)}{n} \right)^{1/2}$$

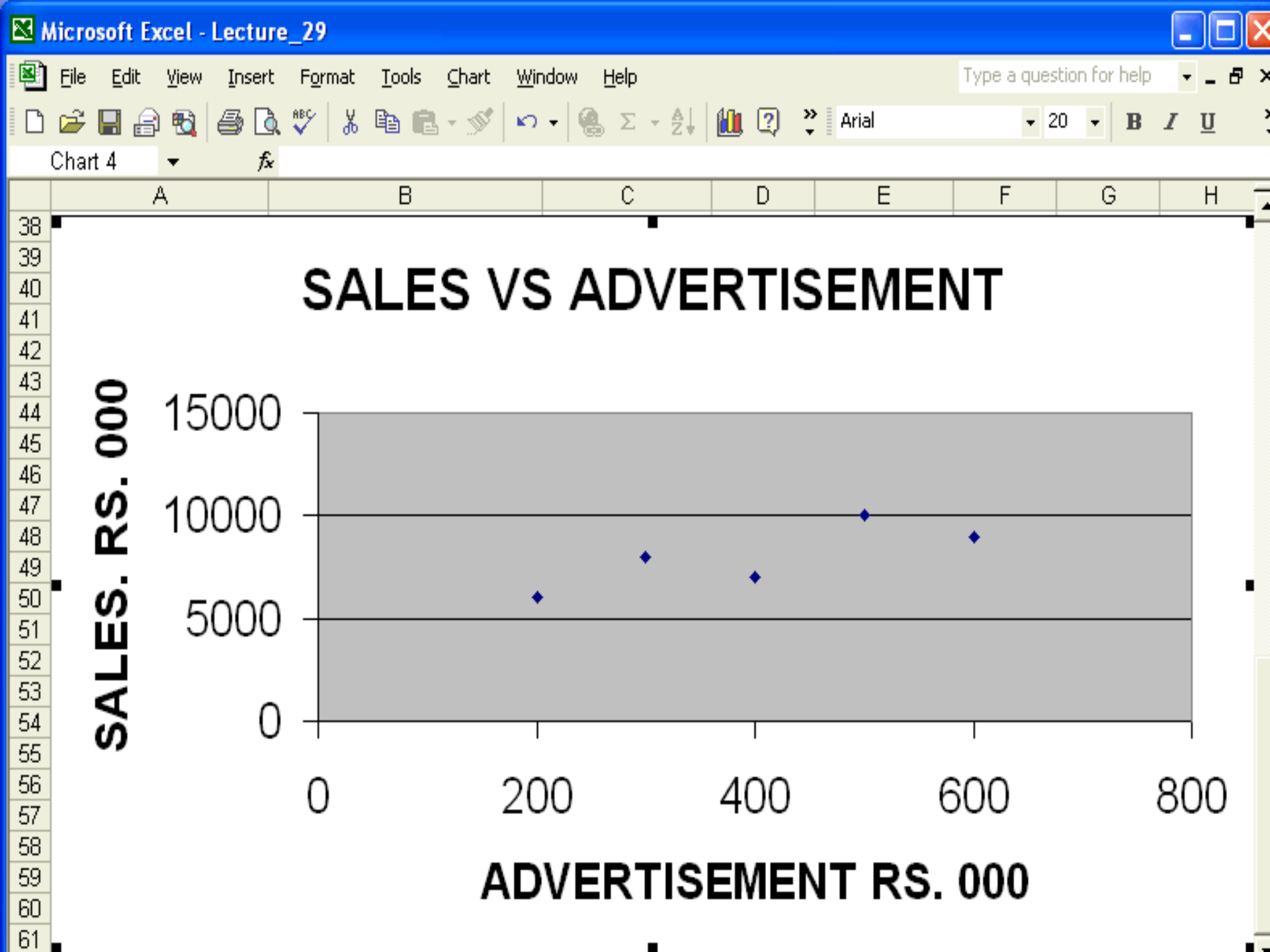
Excel

- 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

- 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





File Edit View Insert Format Tools Data Window Help

Σ $\frac{\Delta}{2}$? >>

SUM ✖ ✔ *fx* =CORREL(A79:A83;B79:B83)

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

Sample Correlation

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

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

Sample Variance (ungrouped)

- The sample variance(s^2) measures spread & is the average of the squares of all the deviations from the sample mean
- for a population, use σ^2 instead of s^2

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

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

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

Standard deviation

- In practice the numerical statistic used to describe the “spread” of a sample is the square root of the variance , & is called the “standard deviation”
- $s = \sqrt{s^2}$ (for populations: $\sigma = \sqrt{\sigma^2}$)
- we say : “s” estimates “ σ ”
- 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 either side of the mean
- about 95% 2 s.d.'s.....
- about 99.7%.....3.....

Population parameters

- sample -->>(estimates) population
- statistic “ “ parameter
- $\bar{x}$ “ “ μ
- s^2 “ “ σ^2
- s “ “ σ
- Rel.freq.polygon “ prob.distribution

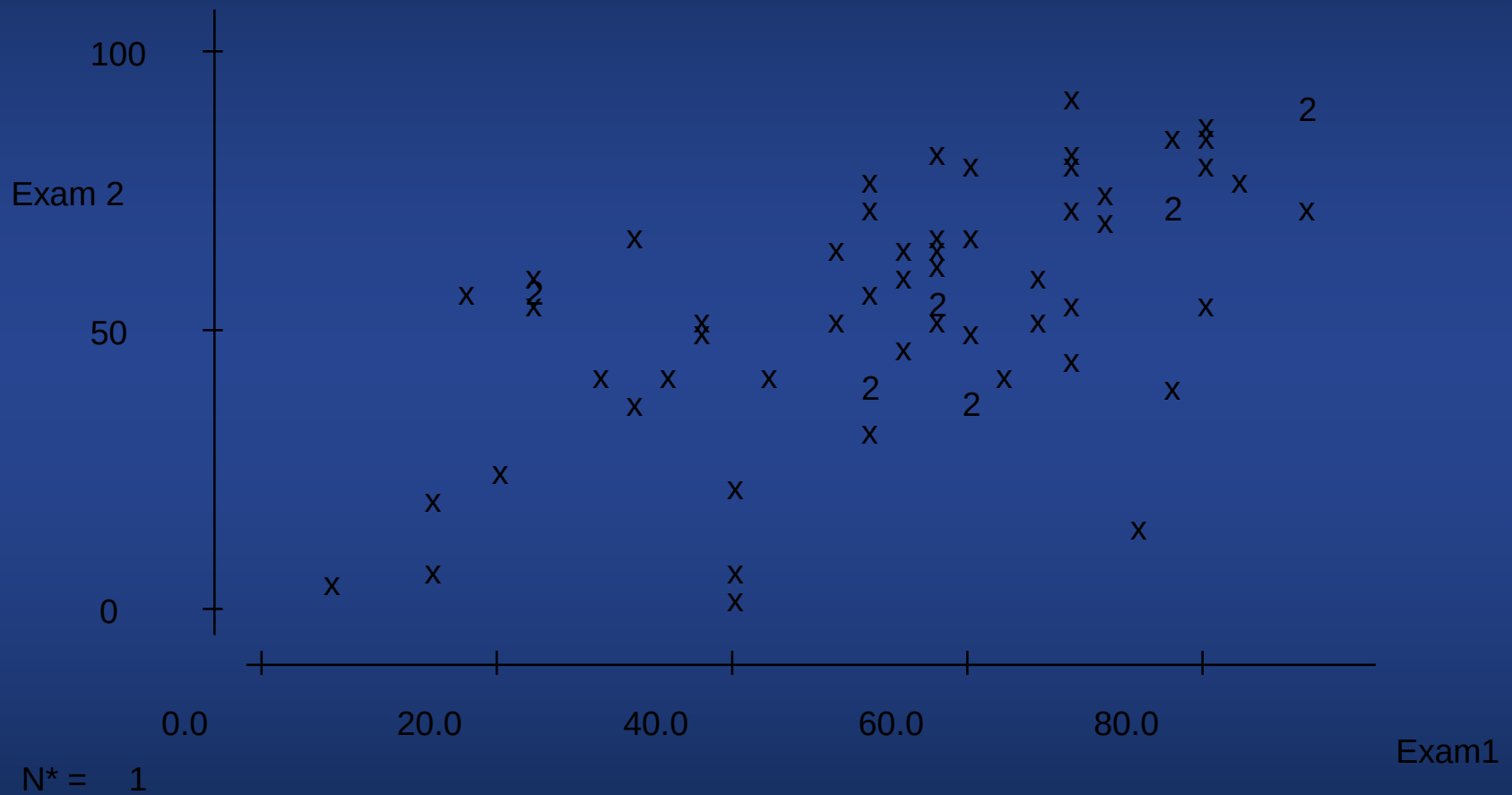
Output: Descriptive Statistics

	I	J
1	<i>April 1993</i>	
2		
3	Mean	219.392857
4	Standard Error	12.0526356
5	Median	189
6	Mode	185
7	Standard Deviation	90.1936661
8	Variance	8134.8974
9	Kurtosis	3.95428336
10	Skewness	1.77639095
11	Range	460
12	Minimum	115
13	Maximum	575
14	Sum	12286
15	Count	56

Example

- In 2nd Semester, 66 students sat for both the final exam for Bizmath & the midterm exam.
- How closely related are the marks on these two exams?
- Can I predict the next final exam scores from 3rd midterm scores using a regression model based on 2nd semester data?

Graph of relationship



Interpretation

- The graph shows that the marks in the two tests are fairly closely related
- The correlation coefficient measures the strength of the linear relationship between two variables

Example

- So, find if the correlation coefficient, r , or its estimate, r , is big or small
- For the 2nd semester exams data, $r=0.593$. This value is moderate, so the relation is neither strong nor weak.

Significance test on r

- We can test if the correlation between X & Y is significantly different from 0.
- The sampling distribution of $[(r-r)/sr]$ is “t”, with $n-2$ degrees of freedom.
- This means testing if $r = 0$ by using exactly the same t-statistic as before
- e.g., is the correlation between the marks in the 2 exams in 2nd semester non-zero?

Hypothesis test on ρ

- To test $H_0: r = 0$, vs $H_1: r \neq 0$
- we reject H_0 if $|t| > t_{\alpha/2} \text{ (dof=n-2)}$

$$t = \frac{r - \rho}{s_r} \text{ where } s_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

Example

- For the 2nd semester exams data we found that the correlation coefficient is 0.593. Also $n = 66$.
- We reject H_0 if $|t| > t_{\alpha/2} \Rightarrow 2.0$ for $\alpha = .05$
- We have $sr = 0.101$, & $t = 5.89$ so reject H_0
- The marks in the two tests are correlated

Shape : Skewness

- The housing data are positively-skewed or right-skewed



Skewness

- For positive skew, mode is to the left; the mean & the tail are to the right , example : scores on “difficult” exam.
- For negative skew, mode is to the right; the mean & the tail are to the left , example : scores on an “easy” exam.
- For skewness near zero: the distribution is symmetric , the mean, median and mode coincide

Skewness coefficient

- Ungrouped observations, sample : $b = \frac{1}{n} \sum (x_i - \bar{x})^3 / s^3$ b
 - $b = 0$ -->> no skewness
 - $b > 0$ -->> positive skew
 - $b < 0$ -->> negative skew
 - $b > 3$ or $b < -3$ -->> large skew
- b estimates the population skewness

kurtosis coefficient

- ungrouped, sample :
$$k = (1/n)[\sum (x_i - \bar{x})^4 / s^4]$$

 $k = 3$ for a normal distribution
 $k < 3$ flatter than normal
 $k > 3$ more peaked than normal
- k estimates flatness of the population distribution?

Grouped data-Mean

- Some data are only available in grouped format (e.g. freq.tables)
- Suppose the groups are intervals with midpoints $m_1, m_2, \dots, m_k$ & that there are f_i observations in group i
- The approximate sample mean is
 - $\bar{x} = [S (f_i m_i)] / n$
 - where $n = \sum f_i$

Grouped data : s²,b,k

- approximate calculation formulae using the group midpoints (m_i) are
- Variance :

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

Skewness : $b = [S f_i (m_i - \bar{x})^3 / s^3] / n$

Kurtosis : $k = [S f_i (m_i - \bar{x})^4 / s^4] / n$

Population Covariance

- The strength of association between 2 jointly distributed random variables X & Y , is measured by the Covariance between X & Y :

$$\text{Cov}(X,Y) = s_{xy} = E(XY) - m_x m_y$$

- If X, Y are independent $\rightarrow \text{Cov}(X,Y)=0$

Sample Covariance

- Sample covariance is an estimate of the population covariance , obtained from a sample of values of X & Y. It gives a measure of the strength of association between them in the original units .

$\text{Cov}(x,y) = s_{xy} = ([S_{xy}] - nxy)/n$: biased

or $s_{xy} = ([S_{xy}] - nxy)/(n-1)$: unbiased

--

BUSINESS MATHEMATICS & STATISTICS

LECTURE 30

Line Fitting

Part 1

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

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.

CORREL(array1,array2)

Array1 is a cell range of values.

Array2 is a second cell range of values.

Remarks

- 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.

Excel

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

STATISTICAL ANALYSIS TOOL

Perform a statistical analysis

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

LOAD THE ANALYSIS TOOLPAK

How?

1. On the **Tools** menu, click **Add-Ins**.
2. In the **Add-Ins available list**, select the **Analysis ToolPak box**, and then click **OK**.
3. If necessary, follow the instructions in the setup program

E8

fx

	A	B	C	D	E	F	G	H	I
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									
16									
17									
18									
19									
20									
21									
22									
23									
24									
25									

Add-Ins

Add-Ins available:

- ☒ Analysis ToolPak
- ☐ Analysis ToolPak - VBA
- ☒ Conditional Sum Wizard
- ☒ Euro Currency Tools
- ☒ Internet Assistant VBA
- ☒ Lookup Wizard
- ☐ Solver Add-in

OK

Cancel

Browse...

Automation...

Analysis ToolPak

Provides functions and interfaces for financial and scientific data analysis

Add-Ins



Add-Ins available:

- ☒ Analysis ToolPak
- ☐ Analysis ToolPak - VBA
- ☒ Conditional Sum Wizard
- ☒ Euro Currency Tools
- ☒ Internet Assistant VBA
- ☒ Lookup Wizard
- ☐ Solver Add-in

OK

Cancel

Browse...

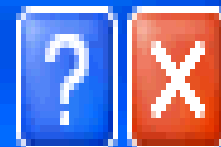
Automation...

Analysis ToolPak - VBA

VBA functions for Analysis ToolPak

	A	B	C
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			

Data Analysis



Analysis Tools

Anova: Two-Factor Without Replication

Correlation

Covariance

Descriptive Statistics

Exponential Smoothing

F-Test Two-Sample for Variances

Fourier Analysis

Histogram

Moving Average

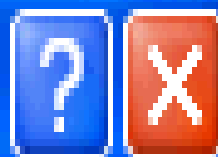
Random Number Generation

OK

Cancel

Help

Data Analysis



Analysis Tools

Histogram
Moving Average
Random Number Generation
Rank and Percentile
Regression
Sampling
t-Test: Paired Two Sample for Means
t-Test: Two-Sample Assuming Equal Variances
t-Test: Two-Sample Assuming Unequal Variances
z-Test: Two Sample for Means

OK

Cancel

Help

Regression



Input

Input Y Range:

Input X Range:

☐ Labels

☐ Constant is Zero

☐ Confidence Level:

95 %

OK

Cancel

Help

Output options

☐ Output Range:

☒ New Worksheet Ely:

☐ New Workbook

Residuals

☐ Residuals

☐ Residual Plots

☐ S_tandardized Residuals

☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

SLOPE

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

The equation for the slope of the regression line is:

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

 Microsoft Excel - Lecture 30

File Edit View Insert Format Tools Data Window Help





















SUM =SLOPE(A4:A10;B4:B10)

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

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.

INTERCEPT

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

Microsoft Excel - Lecture_30

File Edit View Insert Format Tools Data Window Help

SUM =INTERCEPT(A18:A22;B18:B22)

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

BUSINESS MATHEMATICS & STATISTICS

BUSINESS MATHEMATICS & STATISTICS

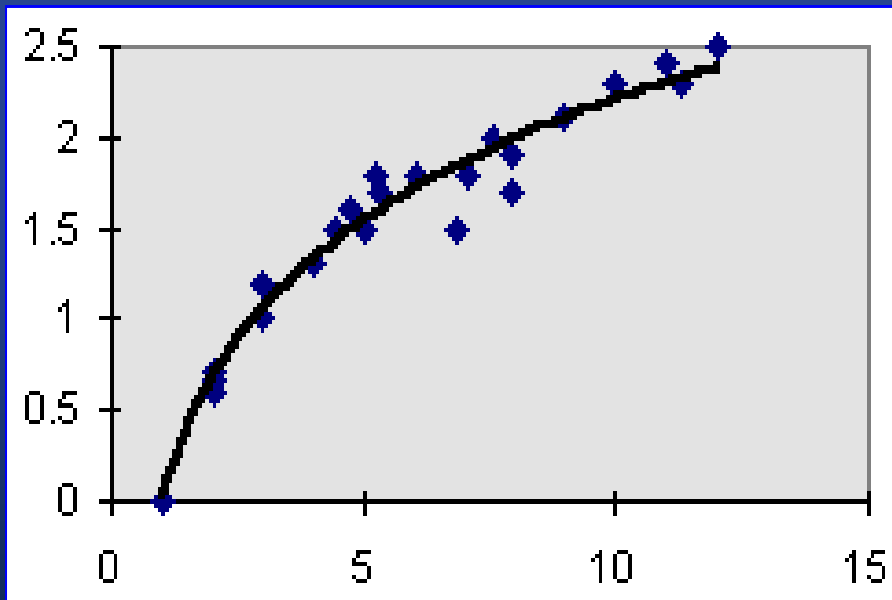
LECTURE 31

Line Fitting

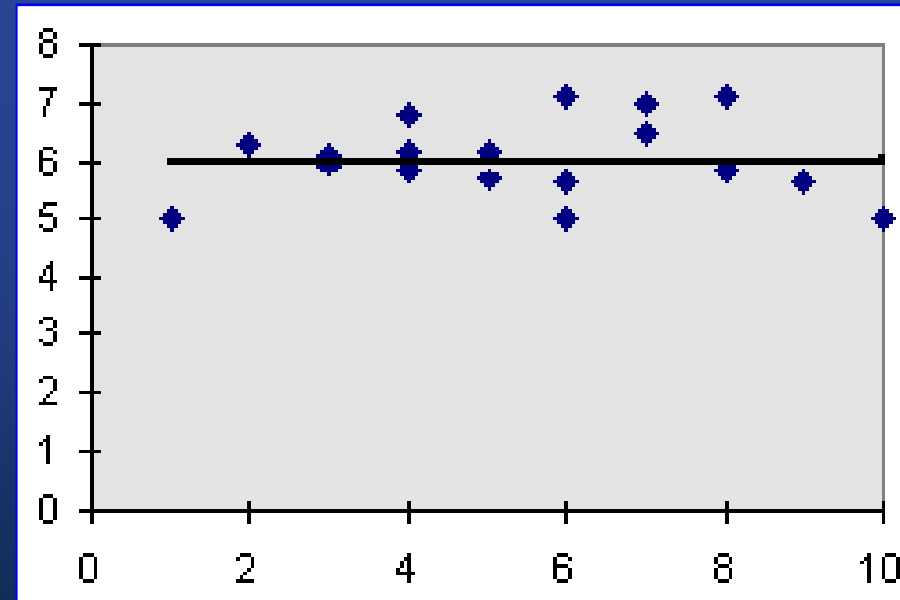
Part 2

Types of Regression Models

Relationship NOT Linear



No Relationship



Simple Linear Regression Model

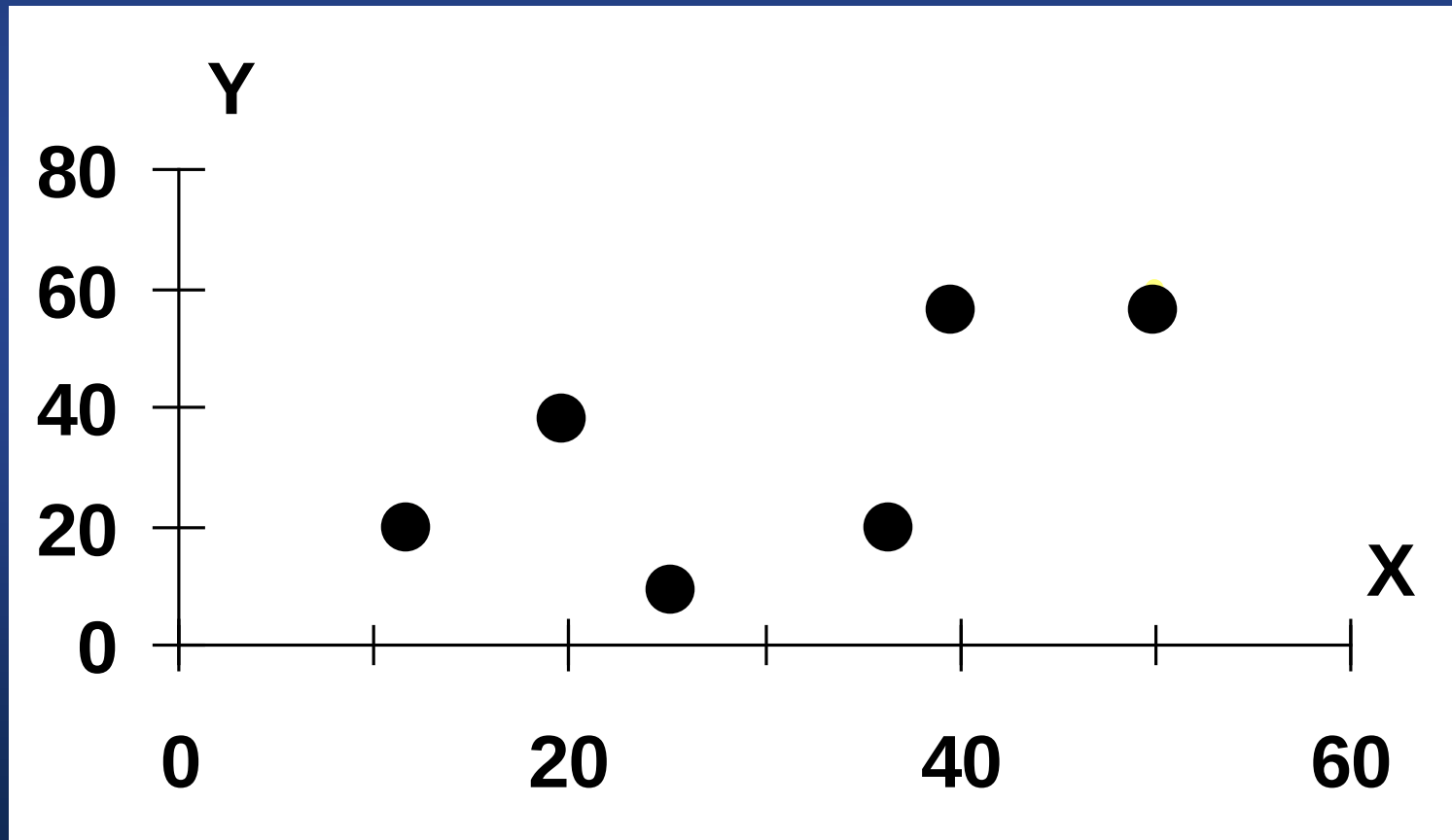
Relationship Between Variables Is Described by A Linear
Function

The Change of One Variable Causes The Other Variable to
Change

A Dependency of One Variable on the Other

The Scatter Diagram

Plot of all (X_i, Y_i) pairs



Population Linear Regression

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

The diagram illustrates the Population Linear Regression equation, $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, with labels for each component:

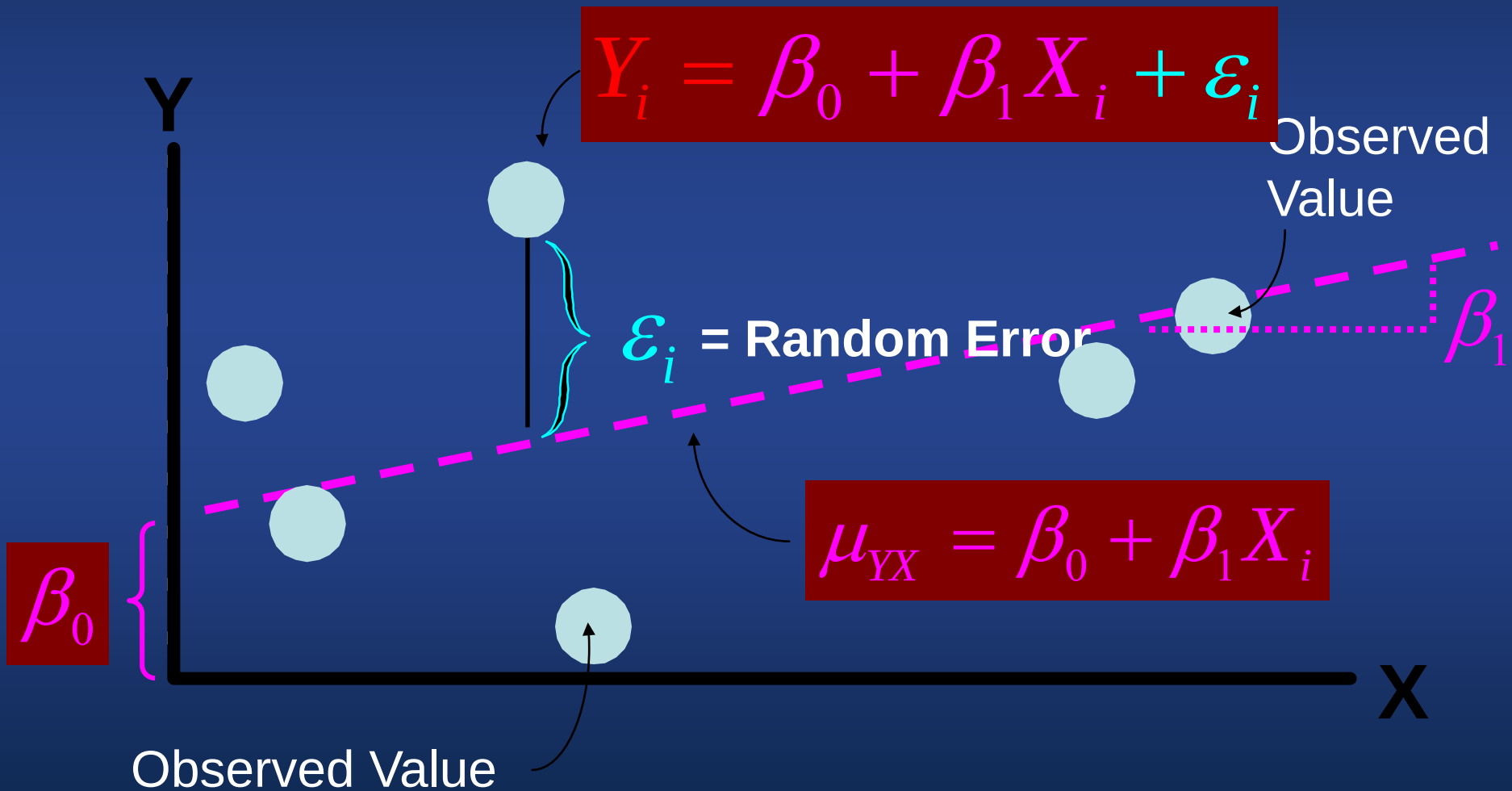
- Population Y intercept**: Points to β_0 .
- Population Slope Coefficient**: Points to β_1 .
- Random Error**: Points to ε_i .
- Dependent (Response) Variable**: Points to Y_i .
- Population Regression Line**: Points to the entire equation.
- Independent (Explanatory) Variable**: Points to X_i .

The equation is displayed as:

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

Population Linear Regression

(continued)



Sample Linear Regression

Sample Regression Line Provides an Estimate of The Population Regression Line as well as a Predicted Value of Y

Sample
Y Intercept

$$Y_i = b_0 + b_1 X_i + e_i$$

Sample
Slope
Coefficient

Residual

b_0 provides an estimate of
 b_1 provides an estimate of

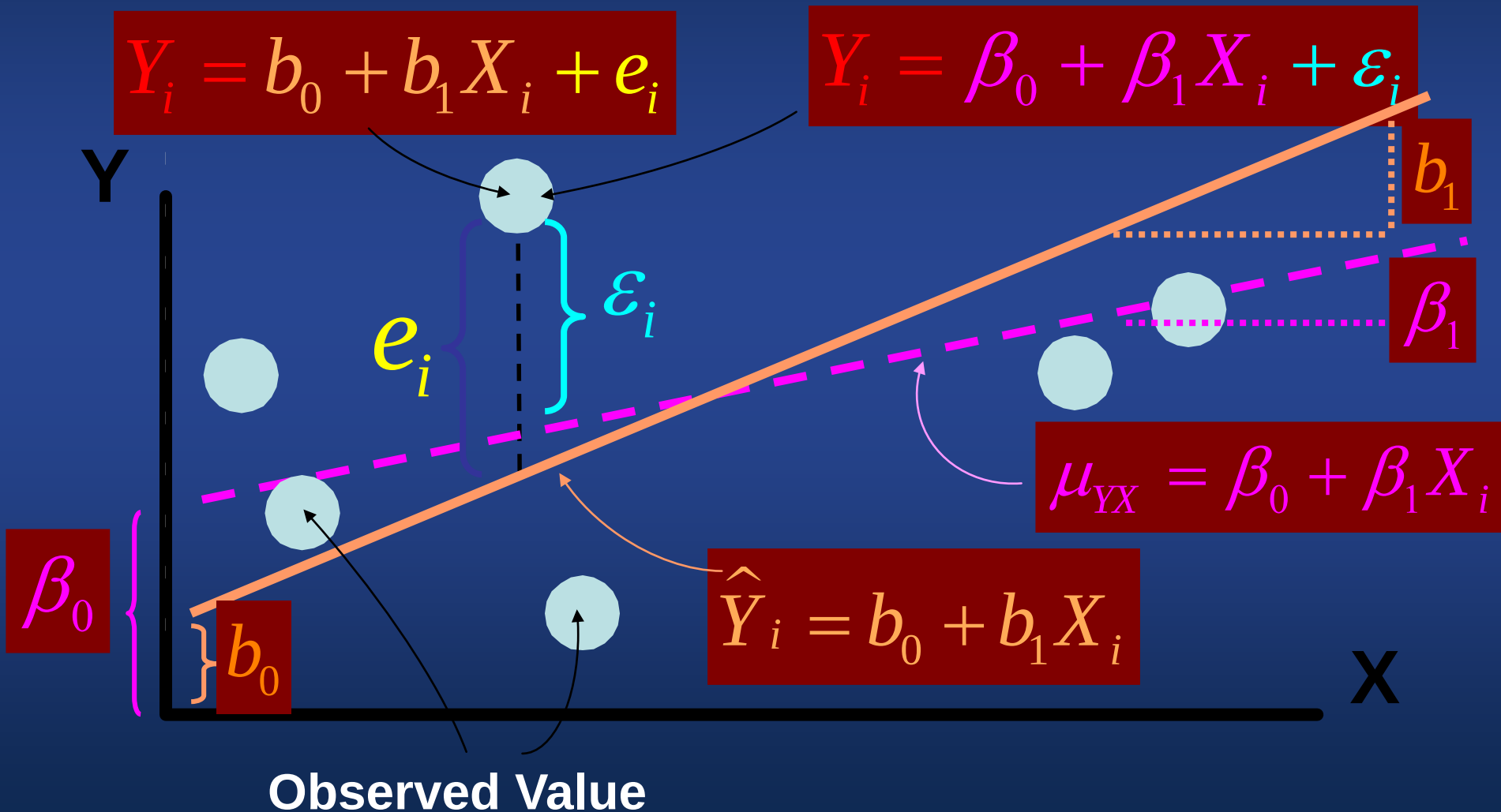
β_0

β_1

**Sample
Regression
Line**

Sample Linear Regression

(continued)



REGRESSION EQUATION

Equation of Least Squares Regression line

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

$$x_m = 4; y_m = 80; s(x) = 2^{1/2};$$

$$s(y) = 200^{1/2}; r = 0.8$$

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

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

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

$$y = 8x + 48$$



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

80

81

82

83

84

85

86

87

88

89

90

91

92

93

REGRESSION ANALYSIS

EXAMPLE 1

Y	X
60	2
100	5
70	4
90	6
80	3

Regression

Input

Input Y Range:

\$A\$85:\$A\$89

Input X Range:

\$B\$85:\$B\$89

☐ Labels☐ Constant is Zero☐ Confidence Level:

95 %

OK

Cancel

Help

Output options

☒ Output Range:

\$A\$91:\$F\$124

☐ New Worksheet Ply:☐ New Workbook

Residuals

☒ Residuals☐ Residual Plots☐ Standardized Residuals☒ Line Fit Plots

Normal Probability

☒ Normal Probability Plots

Regression



Input

Input Y Range:

\$A\$85:\$A\$89



Input X Range:

\$B\$85:\$B\$89



☐ Labels

☐ Constant is Zero

☐ Confidence Level:

95

%

OK

Cancel

Help

Output options

☒ Output Range:

\$A\$91:\$F\$124



☐ New Worksheet Ply:

☐ New Workbook

Residuals

☒ Residuals

☐ Residual Plots

☐ Standardized Residuals

☒ Line Fit Plots

Normal Probability

☒ Normal Probability Plots

SUMMARY OUTPUT		
Regression Statistics		
Multiple R	0.8	
R Square	0.64	
Adjusted R Square	0.52	
Standard Error	10.9544512	
Observations	5	

ANOVA			
	<i>df</i>	<i>SS</i>	<i>MS</i>
Regression	1	640	640
Residual	3	360	120
Total	4	1000	
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>
Intercept	48	14.69693846	3.2659
X Variable 1	8	3.464101615	2.3094



168

fx

	E	F	G	H	I
57					
58	<i>F</i>	<i>Significance F</i>			
59	5.3333	0.104088039			
60					
61					
62					
63	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
64	0.0469	1.227738632	94.77226	1.22773863	94.772261
65	0.1041	-3.024327728	19.02433	-3.0243277	19.024327
66					

 f_x

D

77

RESIDUAL OUTPUT

<i>Observation</i>	<i>Predicted Y</i>	<i>Residuals</i>

1	64	-4

2 88 12

3	80	-10

4	96	-6

5	72	8



1100

J K L M N O P Q R

85

86

87

88

89

90

91

92

93

94

95

96

97

98

99

100

101

102

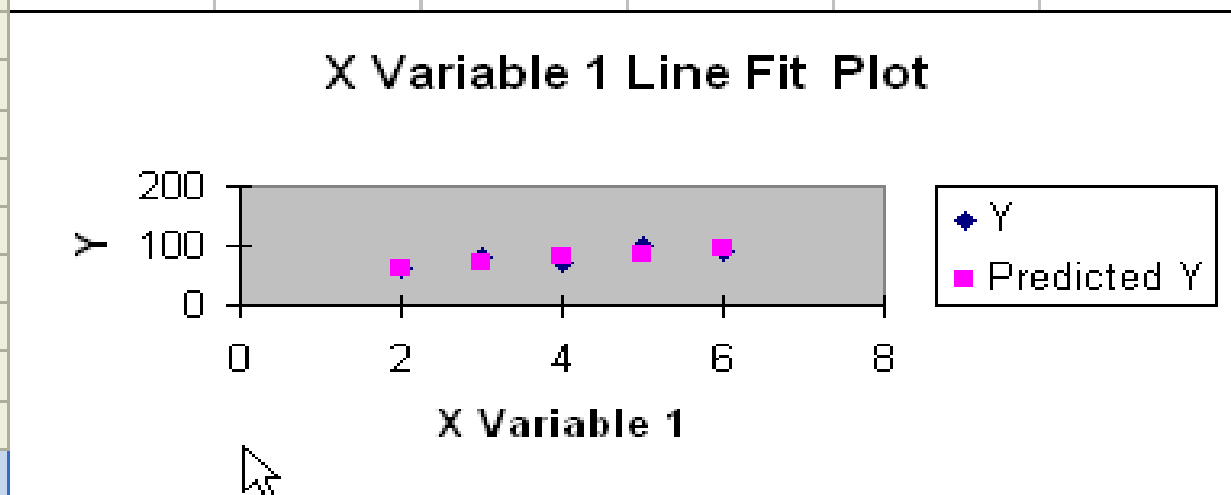
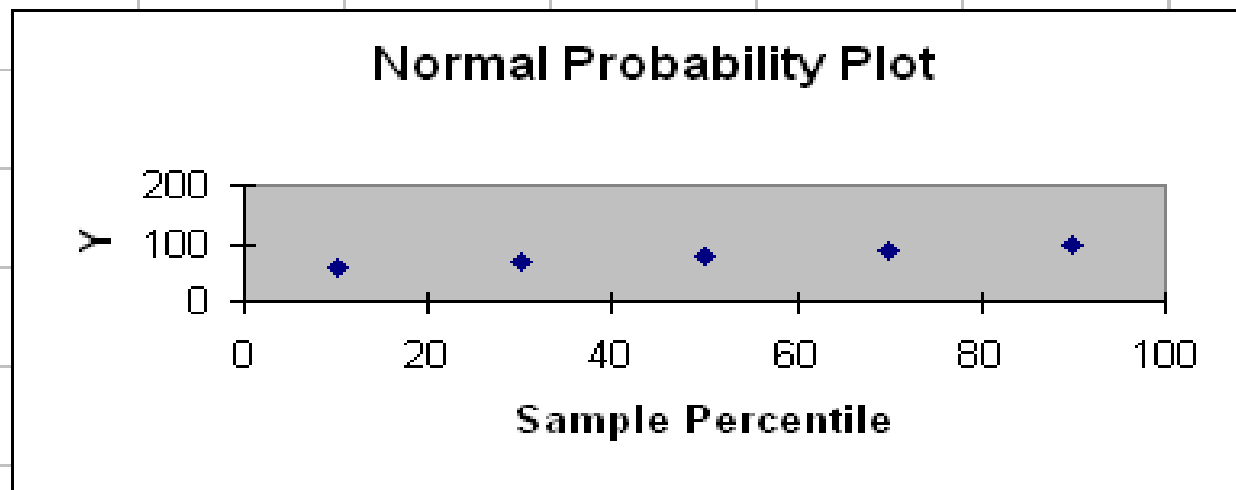


Chart Area

EXCEL REGRESSION TOOL

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

Returns the square of the Pearson product moment correlation coefficient through data points in known_y's and known_x's

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

EXCEL REGRESSION TOOL

P-Value

Probability of not getting a sample slope as high as the calculated value

Smaller the value more significant the result

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

Possible to construct a sampling distribution for r similar to those for sampling distributions for means and percentages

Tables 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 81 to 100)

$v = \text{degrees of freedom} = n - 2$

SAMPLING DISTRIBUTION IN r

Example 1

$n=5$

Null hypothesis: $r = 0$

Calculated coefficient = 0.8

Test the significance at 5% confidence level

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

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

$n=5$

Null hypothesis: $r = 0$

Calculated coefficient = -0.95

Test the significance at 5% confidence level

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

BUSINESS MATHEMATICS & STATISTICS

BUSINESS MATHEMATICS & STATISTICS

LECTURE 32

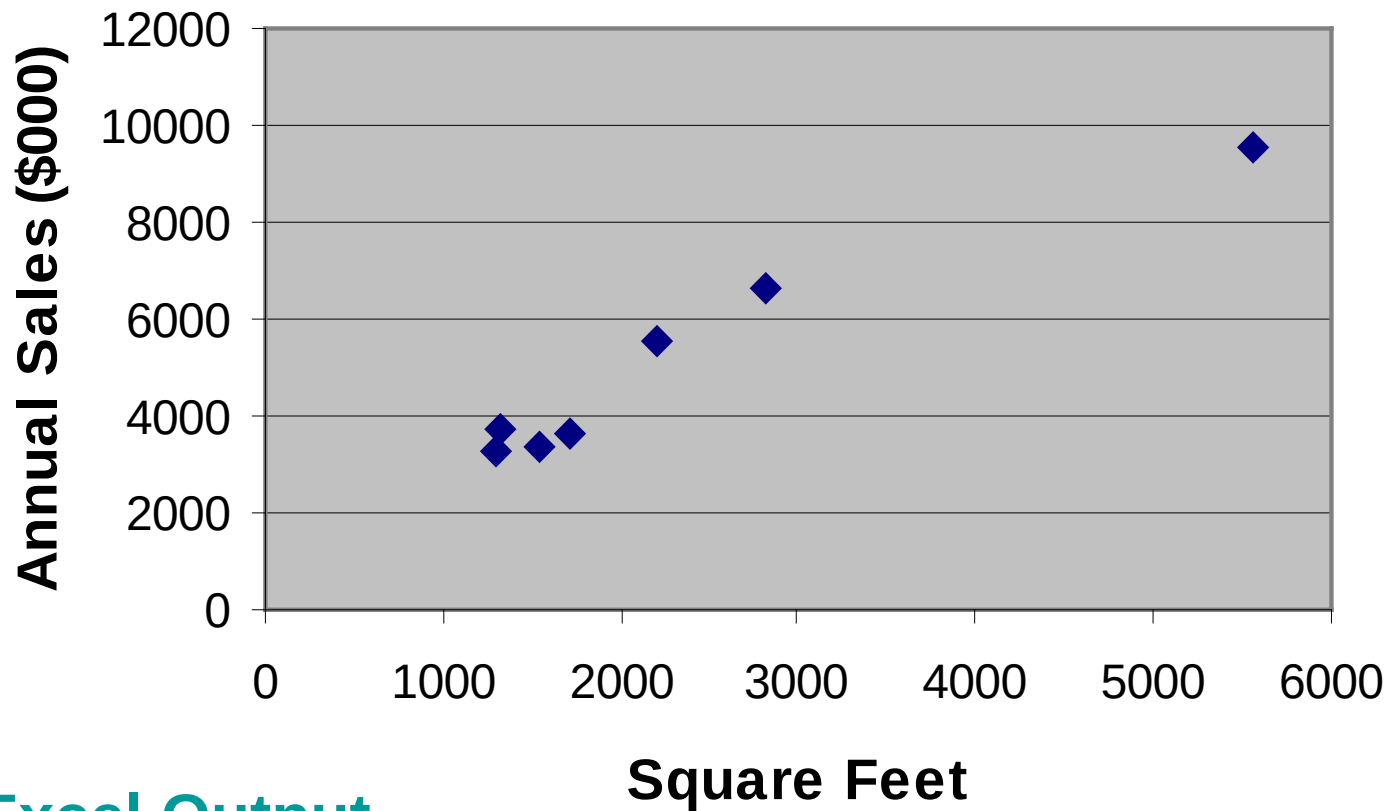
Time Series and
Exponential Smoothing
Part 1

Simple Linear Regression Equation: Example

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

Store	Square Feet	Annual Sales Rs. (000)
1	1,726	3,681
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

Scatter Diagram Example



Excel Output

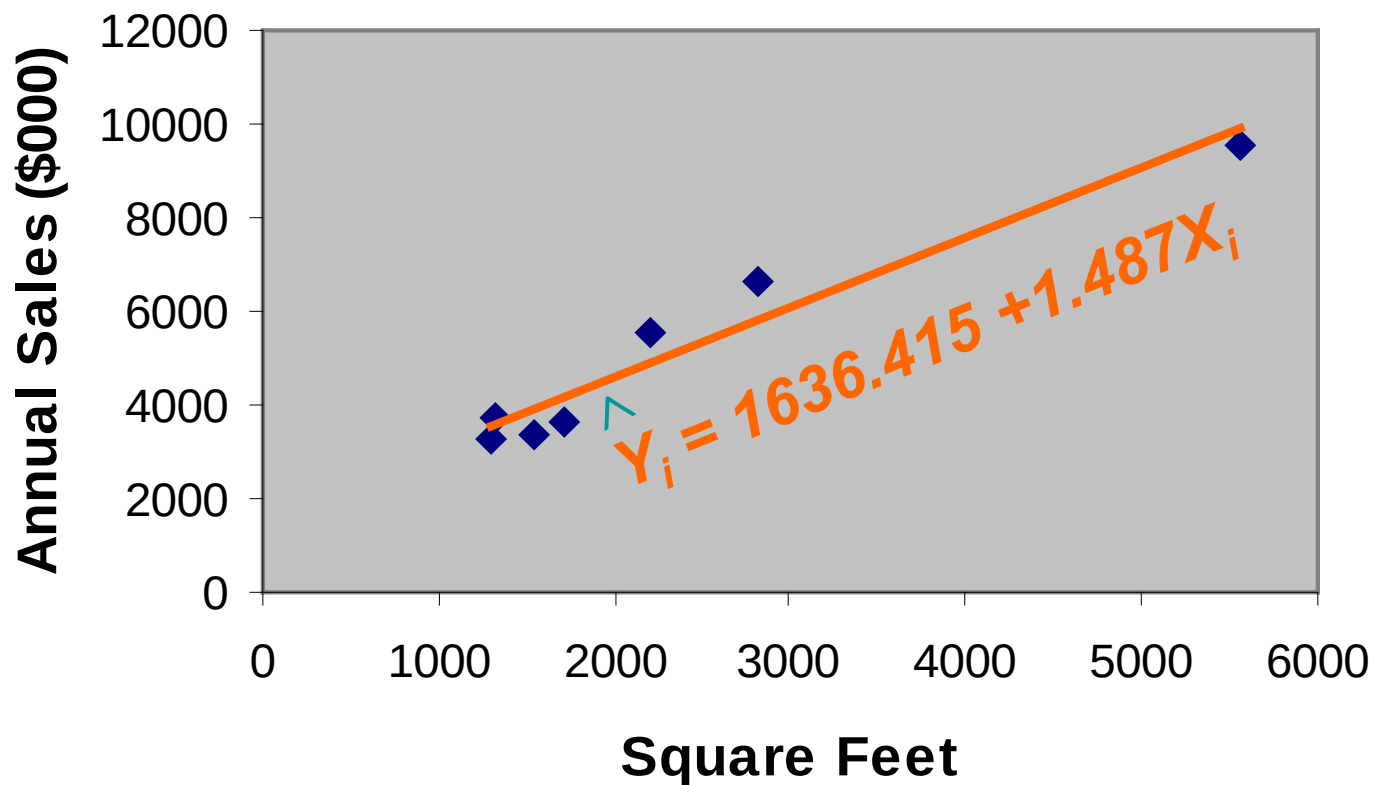
Equation for Sample Regression Line

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

From Excel Printout:

	<i>Coefficients</i>
Intercept	1636.414726
X Variable	1.486633657

Graph of the Sample Regression Line



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

	A	B	C	D	E	F	G	H	I	J	K

2	Year		Quarter	No.Sold (000)				

3	1993	Spring	142				
4		Summer	54				
5		Autumn	162				
6		Winter	206				
7	1994	Spring	130				
8		Summer	50				
9		Autumn	174				
10		Winter	198				
11	1995	Spring	126				
12		Summer	42				
13		Autumn	162				
14		Winter	186				
15							

1	SALES RECORDS BROKEN DOWN BY QUARTERS

2	Year	Quarter	No.Sold (000)				
3	1993	Spring	142				
4		Summer	54				
5		Autumn	162				
6		Winter	206				
7	1994	Spring	130				
8		Summer	50				
9		Autumn	174				
10		Winter	198				
11	1995	Spring	126				
12		Summer	42				
13		Autumn	162				
14		Winter	186				
15							
16							

A2 Year

Chart Wizard - Step 1 of 4 - Chart Type

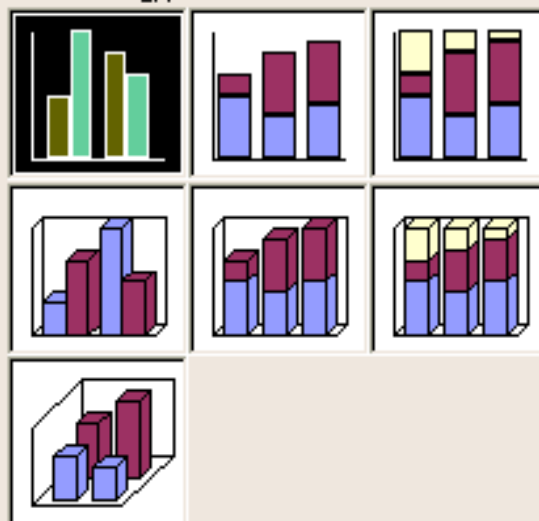
Standard Types

Custom Types

Chart type:

- Column
- Bar
- Line
- Pie
- XY (Scatter)
- Area
- Doughnut
- Radar
- Surface
- Bubble
- Stock

Chart sub-type:



Clustered Column. Compares values across categories.

Press and Hold to View Sample

Cancel

< Back

Next >

Finish

DOWN BY QUARTERS

000)

Chart Wizard - Step 1 of 4 - Chart Type



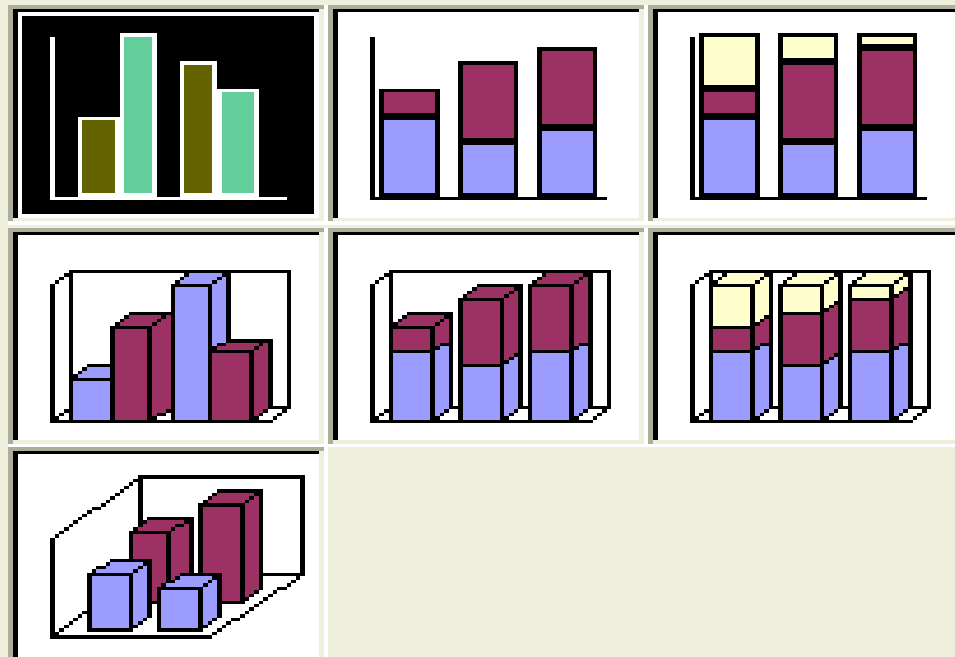
Standard Types

Custom Types

Chart type:

-  Column
-  Bar
-  Line
-  Pie
-  XY (Scatter)
-  Area
-  Doughnut
-  Radar
-  Surface
-  Bubble
-  Stock

Chart sub-type:



Clustered Column. Compares values across categories.

Press and Hold to view Sample



Cancel

< Back

Next >

Finish

Chart Wizard - Step 3 of 4 - Chart Options



Titles

Axes

Gridlines

Legend

Data Labels

Data Table

Chart title:

DOWN BY QUARTERS

Category (X) axis:

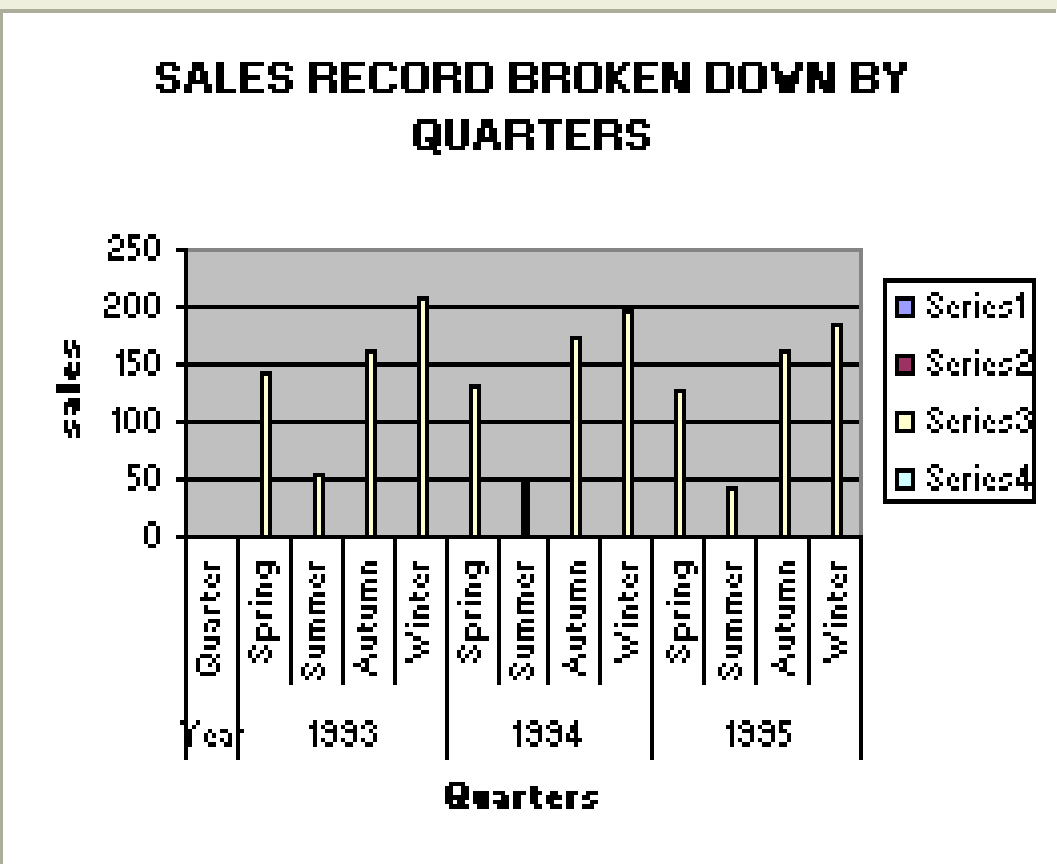
Quarters

Value (Y) axis:

sales

Second category (X) axis:

Second value (Y) axis:



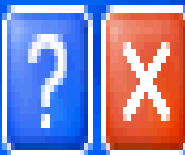
Cancel

< Back

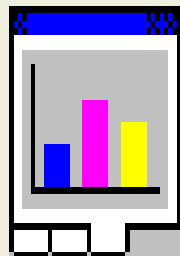
Next >

Finish

Chart Wizard - Step 4 of 4 - Chart Location

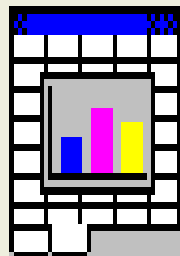


Place chart:



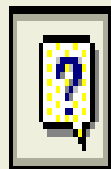
☐ As new sheet:

Chart1



☒ As object in:

Sheet1



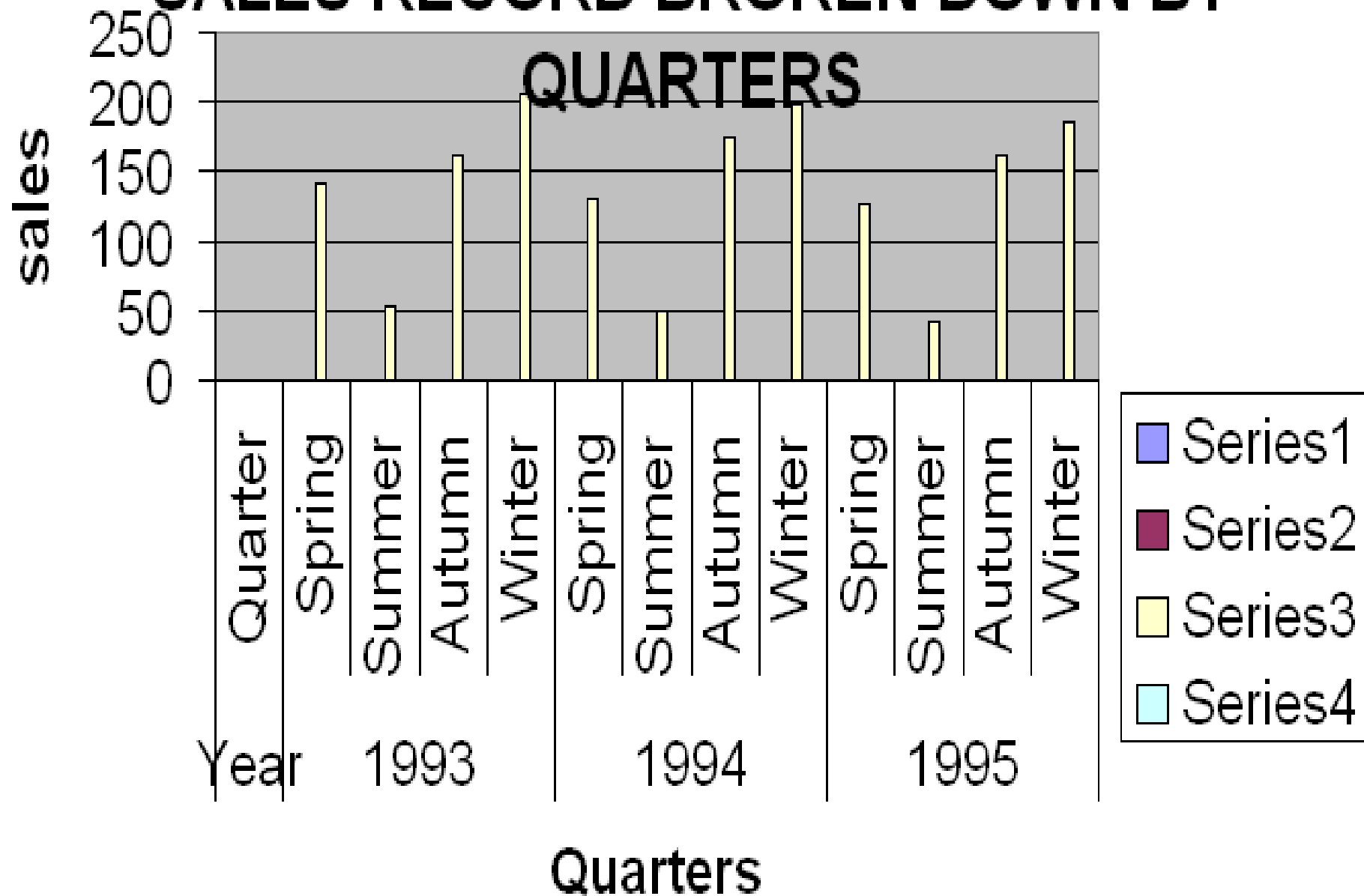
Cancel

< Back

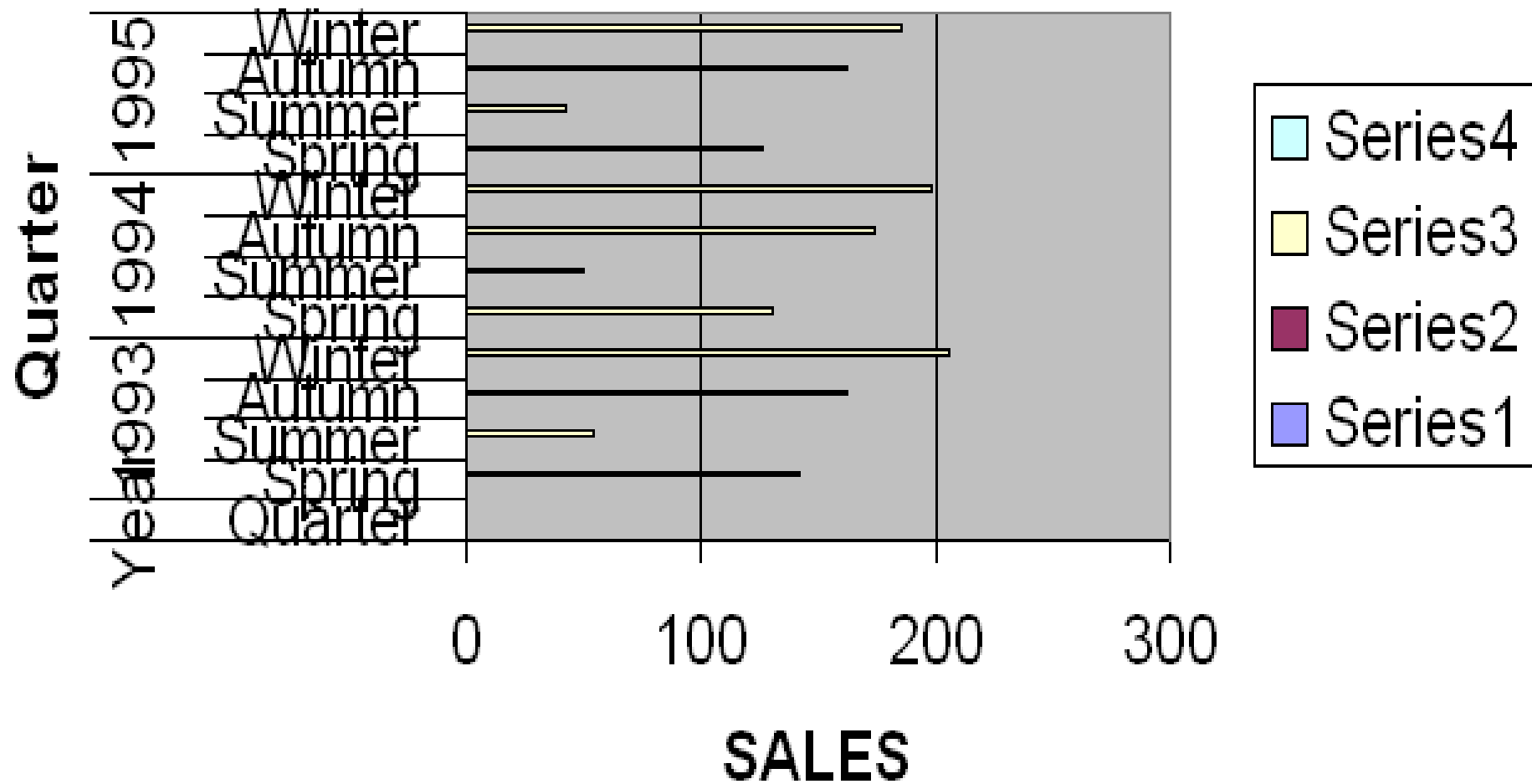
Next >

Finish

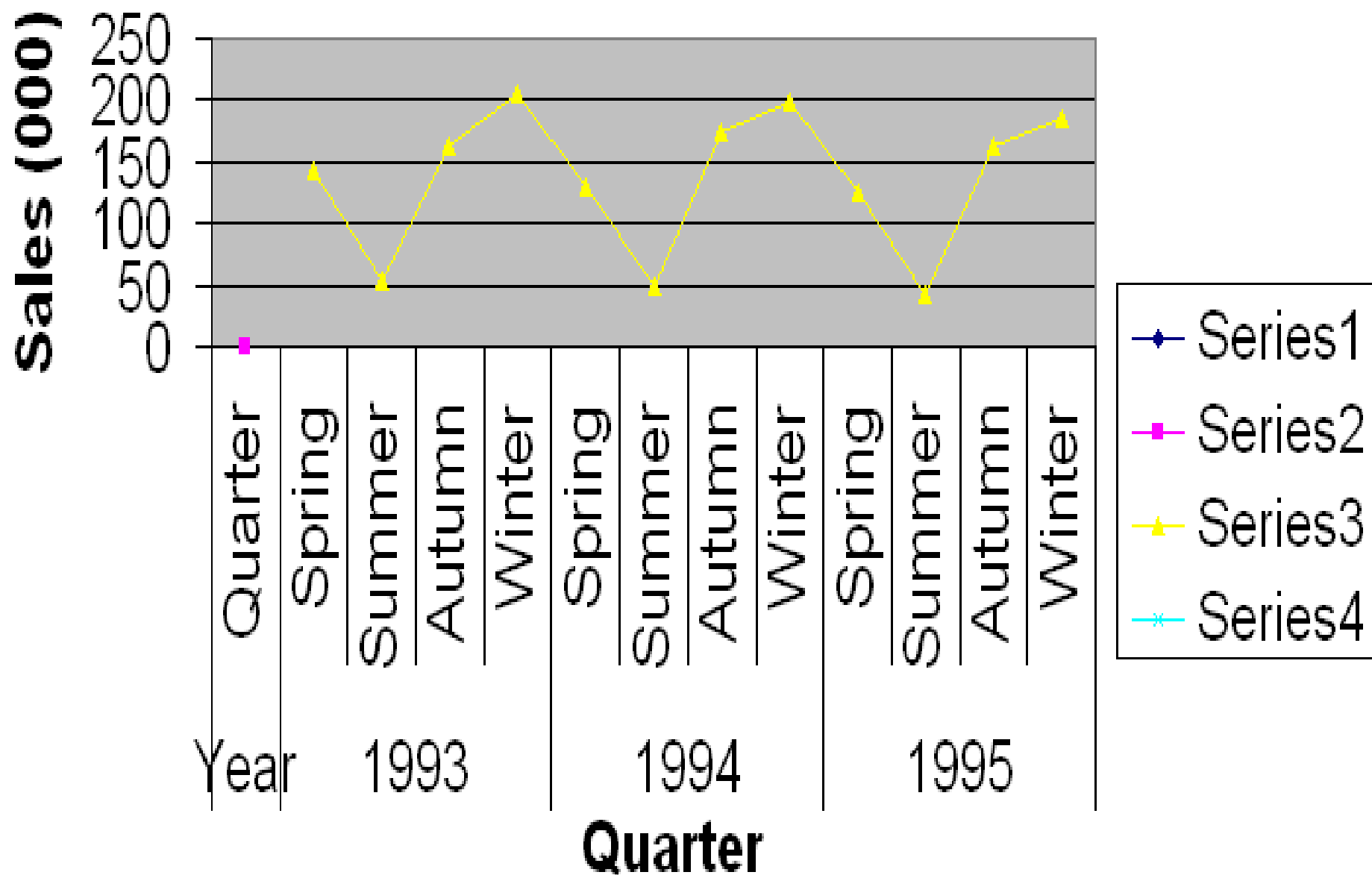
SALES RECORD BROKEN DOWN BY QUARTERS



SALES BROKEN DOWN BY QUARTER



SALES BROKEN DOWN BY QUARTER



EXAMINATION OF GRAPH

TREND

General upward or downward steady behaviour of figures

Seasonal Variation

Variations which repeat themselves regularly over short term, less than a year

Random Effect

Variations due to unpredictable situations

Cyclical Variations

Alternation of upward and downward movement

EXTRACTING THE TREND

DATA

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

No explanation regarding time periods

First step

Plot figures on graph

Horizontal as period 1

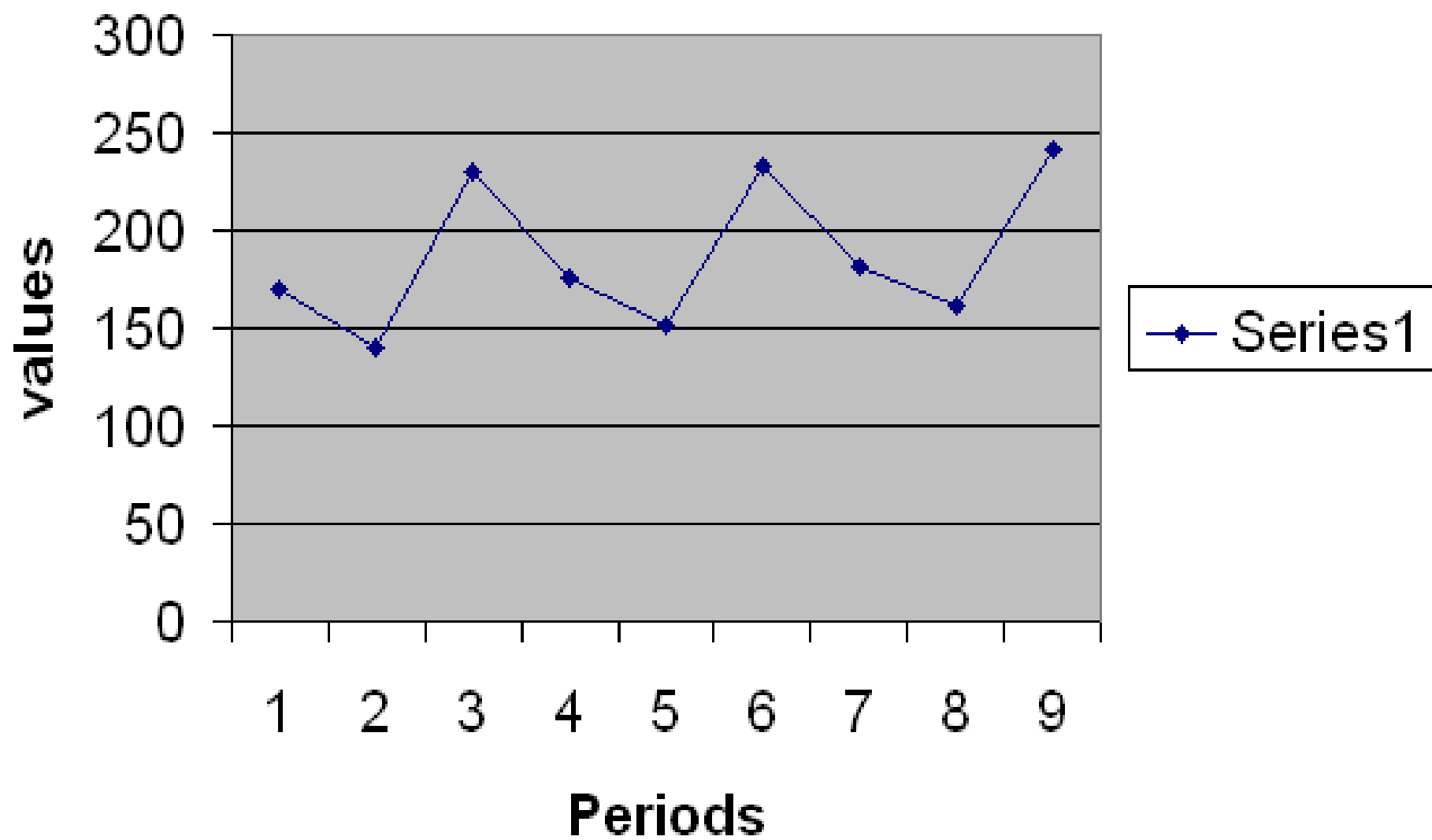
Vertical as period 2

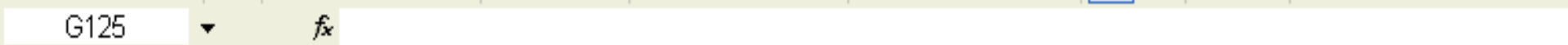
Conclusion

A marked pattern that repeats itself

Well established method to extract trend with strong repeating pattern

DATA



[illegible]

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

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

Another method

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

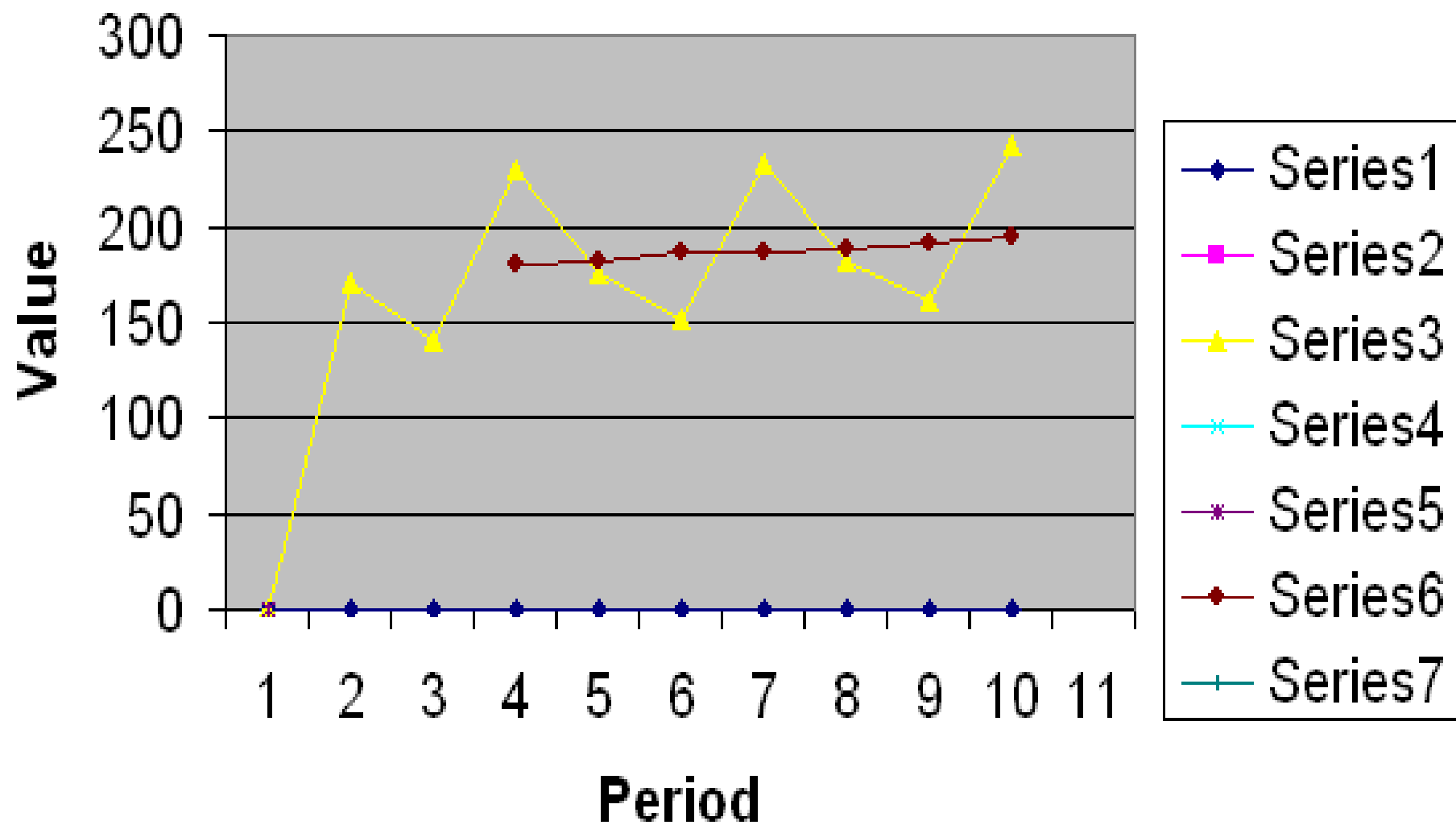
$$\text{Last average} + 2 = 180 + 2 = 182$$

Caution

You may make a mistake

[illegible]

MOVING AVERAGES



ANALYSING SEASONAL VARIATIONS

Find out how much each period differs from trend

Calculate Actual – trend for each period

Day 1, Afternoon

Actual = 180, Trend = 140

Actual – Trend = 180 – 140

=40

BUSINESS MATHEMATICS & STATISTICS

BUSINESS MATHEMATICS & STATISTICS

EXCEL REGRESSION TOOL

P-Value

If we test the hypothesis that the intercept is zero (i.e. Line passes through the origin)

We see that P-value for intercept = 0.07

Adopting 5% significant criterion

This indicates that the population intercept could well be zero

BUSINESS MATHEMATICS & STATISTICS

LECTURE 33

Review Lecture 32
Time Series and
Exponential Smoothing
Part 2

[illegible]

[illegible]

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

[illegible]

FORECAST

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

Regarded as constant amount added to or subtracted from the trends

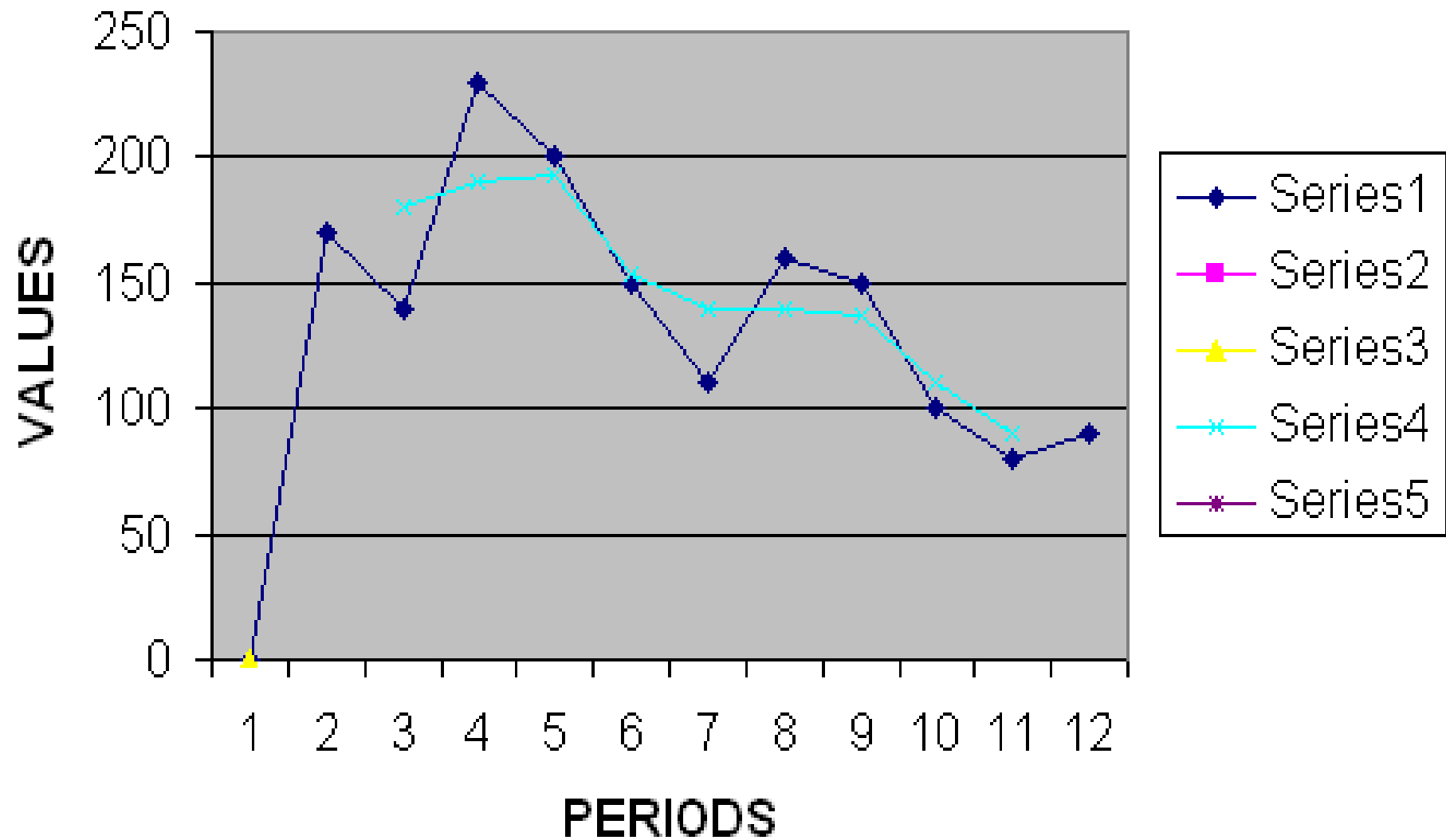
Reasonable Assumption

As seasonal peaks and troughs are roughly of constant size

In Practice

Seasonal variations will not be constant
Will themselves vary as trend increases or decreases
Peaks and troughs can become less pronounced

SEASONAL VARIATIONS

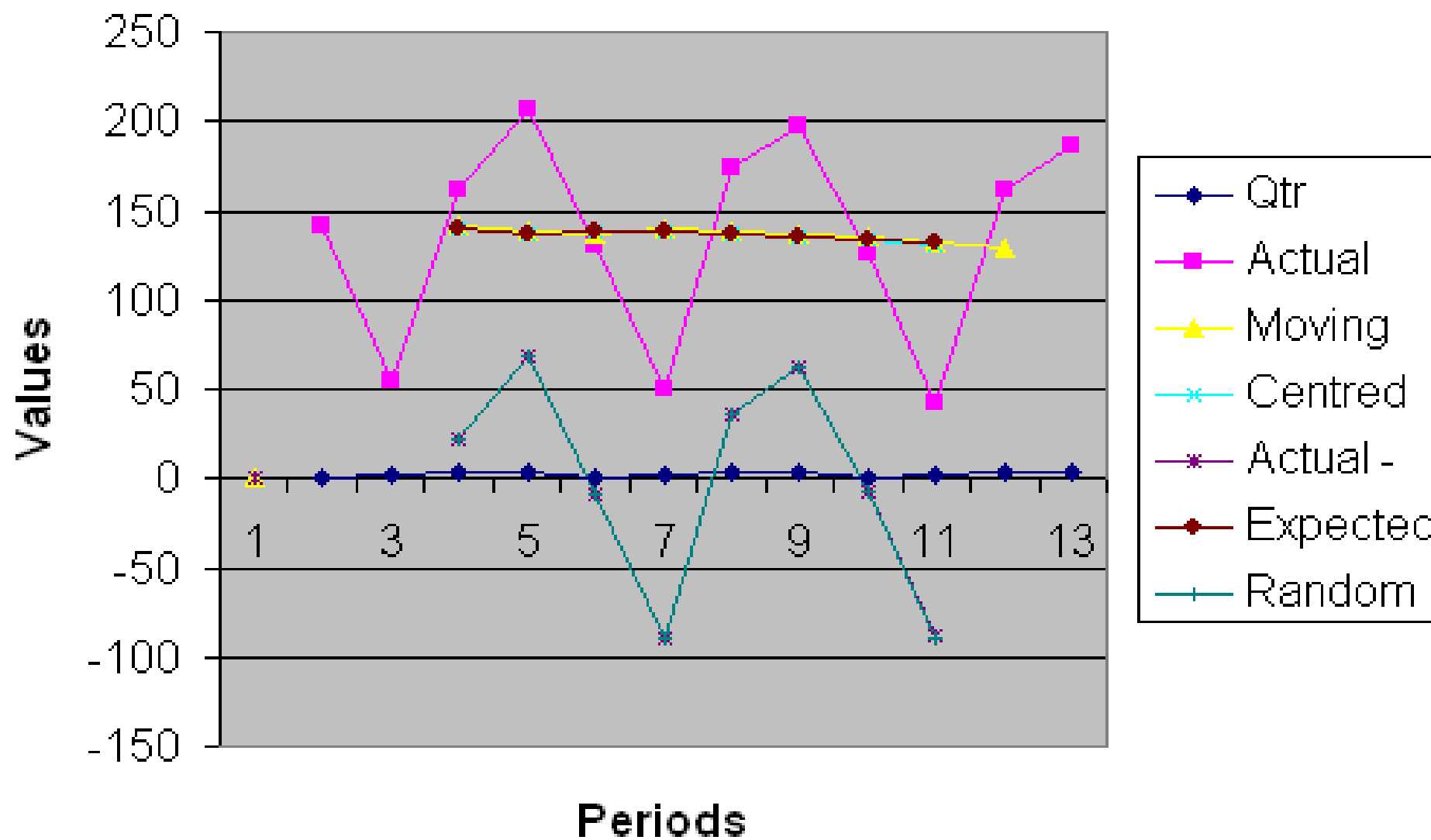


Microsoft Excel - Lecture_33												
Type a question for help												
M65fx												
	B	C	D	E	F	G	H	I	J	K	L	
58												
59			TREND AND SEASONAL VARIATIONS									
60	Quarter		Actual		Moving		Centred		Actual -			
61					Average		M:Average		trend			
62	1		142									
63	2		54									
64	3		162		141		139.5			22.5		
65	4		206		138		137.5			68.5		
66	1		130		137		138.5			-8.5		
67	2		50		140		139.0			-89.0		
68	3		174		138		137.5			36.5		
69	4		198		137		136.0			62.0		
70	1		126		135		133.5			-7.5		
71	2		42		132		130.5			-88.5		
72	3		162		129							

Microsoft Excel - Lecture_33									
File Edit View Insert Format Tools Data Window Help									
Type a question for help									
D104 fx									
	B	C	D	E	F	G	H	I	
87	COMPLETE TABLE								
88	Qtr	Actual	Moving	Centred	Actual -	Expected	Random		
89			Average	M:Average	trend				
90	1	142							
91	2	54				=E92+H85	=C92-G92		
92	3	162	141	139.5	22.5	169.5	-7.5	=C92-G9	
93	4	206	138	137.5	68.5	203.5	2.5		
94	1	130	137	138.5	-8.5	130.5	-0.5		
95	2	50	140	139.0	-89.0	50.0	0.0		
96	3	174	138	137.5	36.5	167.5	6.5		
97	4	198	137	136.0	62.0	202.0	-4.0		
98	1	126	135	133.5	-7.5	125.5	0.5		
99	2	42	132	130.5	-88.5	43	-1.0		
100	3	162	129						
101	4	186							

Microsoft Excel - Lecture_33												
FileEditViewInsertFormatToolsDataWindowHelp Σ 10 € + 0.00 + 0.00												
	B	C	D	E	F	G	H	I	J	K	L	M
75												
76					Quarterly Variations							
77			Spring		Summer		Autumn		Winter			
78	1993		0		0		22.5		68.5			
79	1994		-8.5		-89		36.5		62			
80	1995		-7.5		-88.5		0		0			
81												
82	Total		-16.0		-177.5		59.0		130.5			
83	Average		-8.0		-88.8		29.5		65.3			
84												
85	Rounded		-8		-89		30		65			

COMPLETE TABLE FOR VARIATIONS



FORECAST

Sale steadily declined from 139.0 to 130.5

Over 4 quarters decline = $139.0 - 130.5 = 8.5$

Trend in Spring 1995 = 133.5

Annual decrease = 8.5

Trend in 1996 = $133.5 - 8.5 = 125$

Seasonal variation = -8

Final forecast = $125 - 8 = 117$

FORECASTING IN UNPREDICTABLE SITUATIONS

Two methods studied indicate certain features

- Steady increase in data

- Repeated seasonal variations

Many cases do not conform to these patterns

- There may not be a trend

- No short term pattern

Figures may hover around an average mark

How to forecast under these conditions?

Chart 6

SALES

Week No. Sales

1 4500

2 4000

3 3800

4 4600

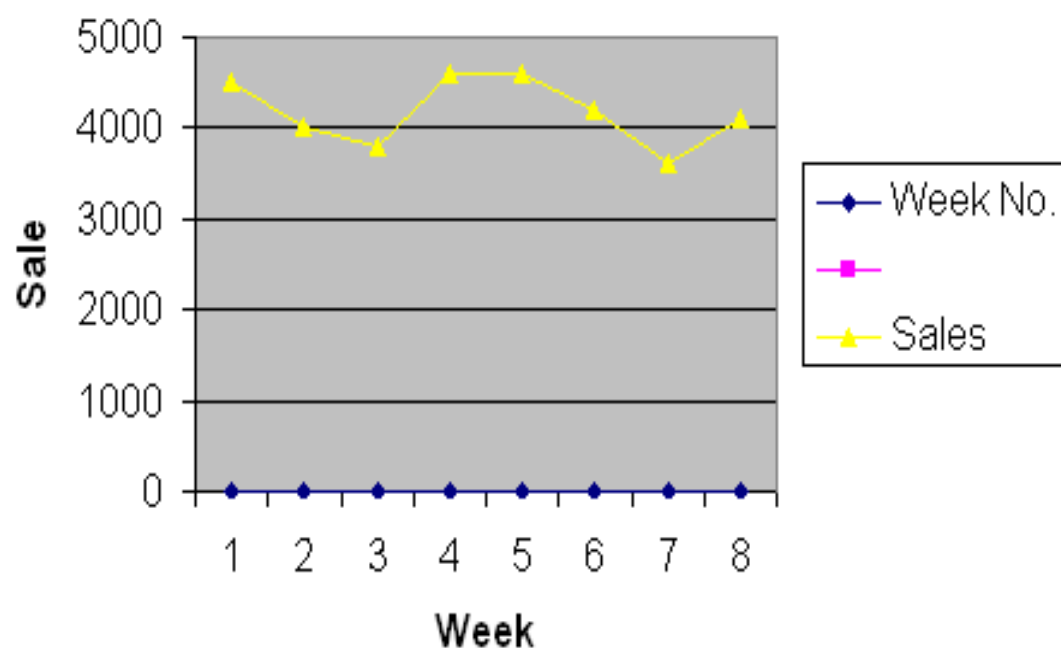
5 4600

6 4200

7 3600

8 4100

SALES



FORECAST

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

Forecast in week 2 is the same as week 1

Actual sale = 4000

Forecast 500 too high

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

new forecast = old forecast + proportion of error α

Or

new forecast = old forecast + $\alpha \times (\text{old actual} - \text{old forecast})$

BUSINESS MATHEMATICS & STATISTICS

BUSINESS MATHEMATICS & STATISTICS

Module 7

Factorials

Permutations and Combinations

(Lecture 34)

Elementary Probability

(Lectures 35-36)

Chi-Square

(Lectures 37)

Binomial Distribution

(Lectures 38)

LECTURE 34

Factorials

Permutations and Combinations

FORECAST

Let $\alpha = 0.3$

Then

$$\begin{aligned}\text{Forecast week 3} &= \text{week 2 forecast} + \alpha \times (\text{week 2 actual} \\ &\quad \text{sale} - \text{week 2 forecast}) \\ &= 4500 - 0.3 \times 500 = 4350\end{aligned}$$

Conclusion

Overestimate is reduced by 30% of the error margin 500

	B	C	D	E	F	G	H	I	
162	Week No.	Sales	Forecast			Error	alpha x Error		
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	Forecast =E164+H164			Error=D164-E164					
	Error Alpha = 0.3								

FORECAST

Method is called **Exponential Smoothing**

α is Smoothing Constant A= Actual F= Forecast

Rule for obtaining a forecast

$$\begin{aligned} F(t) &= F(t-1) + \alpha(A(t-1) - F(t-1)) \\ &= \alpha A(t-1) + (1 - \alpha)F(t-1) \end{aligned}$$

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

Substituting

$$\begin{aligned} F3 &= \alpha A(t-1) + (1 - \alpha) [\alpha A(t-2) + (1 - \alpha)F(t-2)] \\ &= \alpha [A(t-1) + (1 - \alpha)A(t-2)] + (1 - \alpha)^2 F(t-2) \end{aligned}$$

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

$$\begin{aligned} F(t) &= \alpha [A(t-1) + (1 - \alpha)A(t-2) + (1 - \alpha)^2 A(t-3)] + \\ &\quad (1 - \alpha)F(t-3) \end{aligned}$$

WHERE TO APPLY EXPONENTIAL SMOOTHING

What kind of situations?

What are good values of α ?

Criterion = Mean Square Error (MSE)

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 stabilises

Generally alpha between 0.1 and 0.3 performs best

E186							
	B	C	D	E	F	G	H
172	Forecast =E164+H164				Error=D164-E164		
173	ErrorAlpha = 0.3						
174							
175	Week	Actual	Forecast	Error	0.3xError	Error^2	MSE
176	22	2200	2200	0	0	0	0
177	23	2400	2200	200	60	40000	20000
178	24	2600	2260	340	102	115600	51867
179	25	2800	2362	438	131.4	191844	86861
180	26	3000	2493.4	506.6	152.0	256644	120818
181		Forecast =D179+F179				MSE =SUM(H176:H180)/5	
182							

About the Exponential Smoothing dialog box

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.

About the Exponential Smoothing dialog box

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

About the Exponential Smoothing dialog box

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.

About the Exponential Smoothing dialog box

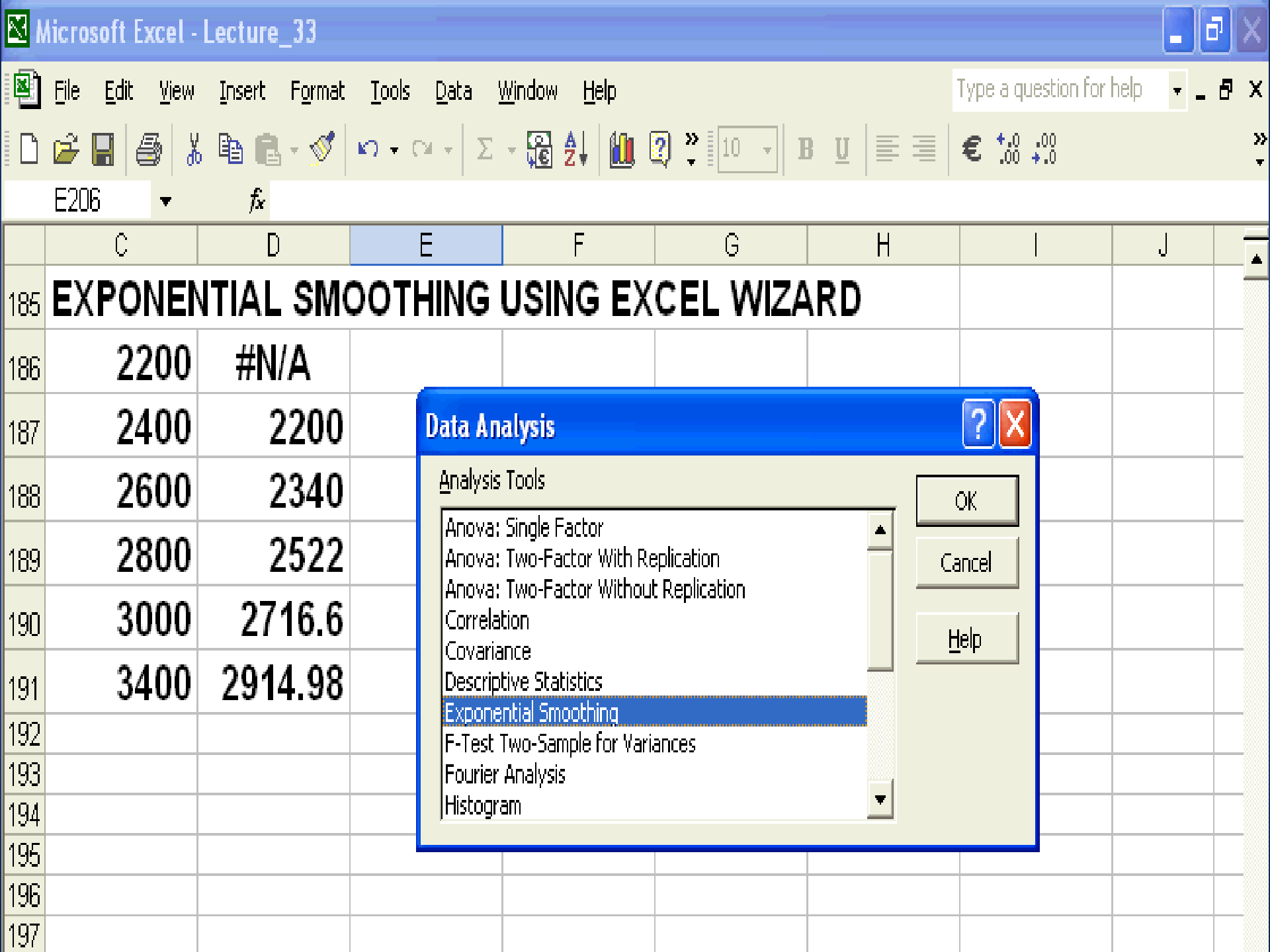
Note The output range must be on the same worksheet as the data used in the input range. For this reason, the New Worksheet Ply and New Workbook options are unavailable.

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.



	C	D	E	F	G	H
185	EXPONENTIAL SMOOTHING USING EXCEL WIZARD					
186	2200					
187	2400					
188	2600					
189	2800					
190	3000					
191	3400					
192						
193						
194						
195						
196						
197						
198						
199						

?

X

Exponential Smoothing

Input

Input Range:

\$C\$186:\$C\$191

Damping factor:

☒ Labels

Output options

Output Range:

\$D\$186:\$D\$191

New Worksheet Ply:

New Workbook

☒ Chart Output

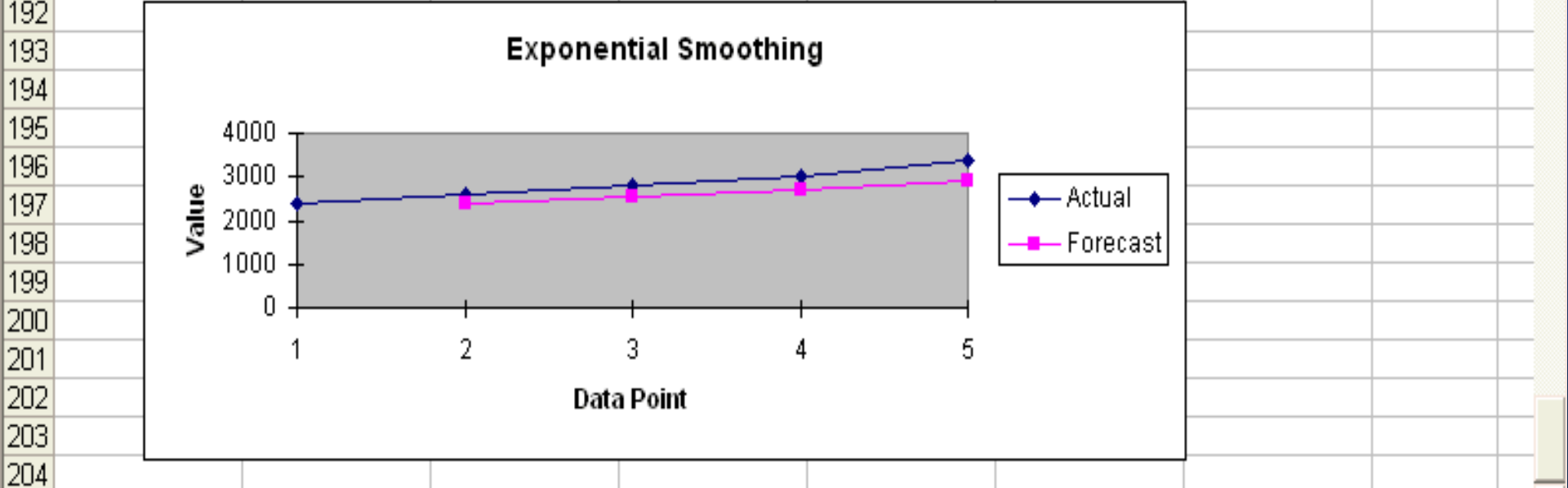
☒ Standard Errors

OK

Cancel

Help

	C205								
	C	D	E	F	G	H	I	J	
185	EXPONENTIAL SMOOTHING USING EXCEL WIZARD								
186	2200	#N/A	#N/A						
187	2400	2400	#N/A						
188	2600	2540	#N/A						
189	2800	2722	#N/A						
190	3000	2916.6	248.2498741						
191	3400								



FACTORIAL

Natural Numbers

1, 2, 3,...

Five Factorial

$$5! = 5.4.3.2.1 \text{ or } 1.2.3.4.5$$

Ten Factorial

$$= 1.2.3.4.5.6.7.8.9.10 = 10.9.8.7.6.5.4.3.2.1$$

In general

$$n! = n(n-1)(n-2)..3.2.1 \text{ or}$$

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

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

FACTORIAL

EXAMPLES

$$10! = 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 3,628,800$$

$$8!/5! = 8 \cdot 7 \cdot 6 \cdot 5! / 5! = 8 \cdot 7 \cdot 6 = 336$$

$$12!/9! = 12 \cdot 11 \cdot 10 \cdot 9! / 9! = 12 \cdot 11 \cdot 10 = 1320$$

$$10!8!/9!5! = 10 \cdot 9!8 \cdot 7 \cdot 6 \cdot 5! / 9!5! = \\ 10 \cdot 8 \cdot 7 \cdot 6 = 3360$$

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

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

PERMUTATIONS

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

$$= 3!/(3-2)! = 3 \cdot 2 = 6$$

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

Number of permutations of n objects taken r at a time =
 ${}^nP_r = n!/(n-r)!$

PERMUTATIONS

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

$${}_nP_n = n(n-1)(n-2)\dots 3.2.1$$

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..

$$N!/n_1!n_2!\dots n_k!$$

PERMUTATIONS

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

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

Formula

$$\begin{aligned} nPr &= n!/n1!n2!..nk! \\ &= 10!/3!1!3!2!1! = 10.9.8.7.6.5.4.3!/3!3!2! \\ &= 50400 \end{aligned}$$

PERMUT

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.

PERMUT

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.

PERMUT

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

The equation for the number of permutations is:

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

B20

fx

	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							

BUSINESS MATHEMATICS & STATISTICS

BUSINESS MATHEMATICS & STATISTICS

LECTURE 35

Review Lecture 34
Elementary Probability
Part 1

COMBINATIONS

Arrangement 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
(r)

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

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, BC.

COMBINATIONS

Examples

$$4C2 = 4!/2!(4-2)! = 4.3.2/2.2 = 6$$

$$5C2 = 5!/2!(5-2)! = 5.4.3!/2.3! = 10$$

Example

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

$$n = 15, r = 11$$

$$15C11 = 15!/11!(15-11)! = 15.14.13.12.11!/11!4! = 15.7.13 = 1365 \text{ ways}$$

Example

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

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

COMBINATIONS

Results of some combinations

$$nC_0 = nC_n = 1$$

$$\text{e.g., } 4C_0 = 4C_4 = 1$$

$$nC_1 = nC_{n-1} = n$$

$$\text{e.g., } 4C_1 = 4C_3 = 4$$

$$nC_r = nC_{n-r}$$

$$\text{e.g., } 5C_2 = 5C_3$$

BINOMIAL EXPANSION

An expression consisting of two terms joined by + or – sign is called a binomial expression.

Expressions such as $(a+b)$, $(a-b)$, $(x+y)^2$ 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.

BINOMIAL EXPANSION

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

$$(x+y)^5 = {}^5C_0.x^5 + {}^5C_1x^4y + {}^5C_2x^3y^2 + {}^5C_3x^2y^3 + {}^5C_4xy^4 + {}^5C_5y^5$$

Solving:

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

BINOMIAL EXPANSION

COEFFICIENTS IN EXPRESSION $(x + y)^n$

Coefficient of first and last term is always 1

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

First term = x^8

Second coefficient =

PROJECT DEVELOPMENT MANAGER'S PROBLEM

A toys manufacturer

Development of new product lines

New toy to be developed

Tied with a new TV series with the same name

40% chance of TV series

1st year estimate = 12,000 units

Profit of Rs. 2 per toy

Without TV series-sale may be 2,000 units

Already 500,000 Rs. Invested

PROJECT DEVELOPMENT MANAGER'S PROBLEM

A rival may sell similar toy

If so sale may be 8000 units

Chance of rival = 50%

Choices:

1. Abandon new product
2. Risk new development

How to tie it all to financial results?

PROBABILITY

Making assessment of chances

A worker out of 600 gets a prize by lottery

Chance of any one individual say Rashid being selected =
 $\frac{1}{600}$

The probability of the event "Rashid is selected" is the
probability of an event occurring

$$P(\text{Rashid}) = \frac{1}{600}$$

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

PROBABILITY

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

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

If out of 600 persons 250 are women,
then the chance of a women being
selected =

$$p(\text{woman}) = 250/600$$

PROBABILITY

In many situations there is no prior knowledge to calculate probabilities

What is the probability of a machine being defective?

Method: Monitor the machine over a period of time

Find out how many times it becomes defective

This experimental or empirical approach

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

PROBABILITY

There are two events A and B

What is the probability of either A or B happening?

What is the probability of A and B happening?
Number of possibilities

Probability of A or B hapenning = Number of
ways A or B can happen/ Total number of
possibilities

= Number of ways A can happen + number of
ways B can happen/ Total number of
possibilities

PROBABILITY

= Number of ways A can happen/ Total number of possibilities + Number of ways B can happen/ Total number of possibilities

= Probability of A happening + Probability of B happening

Condition

A and B must be mutually exclusive

When A and B are mutually exclusive

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

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

40% workforce are women

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

25% females = management grade

30% of males = management grade

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

EXAMPLE

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

$$\text{Total workforce} = 100$$

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

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

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

SET OF MUTUALLY EXCLUSIVE EVENTS

Between them cover all possibilities

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

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

Example

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

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

EXAMPLE

A production line uses 3 machines

Chance that 1st machine

breaks down in any week = $1/10$

Chance for 2nd machine = $1/20$

Chance of 3rd machine = $1/40$

What is the chance that at least one machine breaks down in any week?

EXAMPLE

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

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

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

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

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

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

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

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

EXAMPLE

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?

EXAMPLE

$1C(1/4) > 2C(1/4)$ (Dismissed 1)($1/16 = 4/64$)

$1C(1/4) > 2NC(3/4) > 3C(1/4)$ (Dismissed 2)($3/64$)

$1C(1/4) > 2NC(3/4) > 3NC(3/4)$ (Not dismissed 1)($9/64$)

$1NC(3/4) > 2C(1/4) > 3C(1/4)$ (Dismissed 3)($3/64$)

$1NC(3/4) > 2C(1/4) > 3NC(3/4)$ (Not dismissed 2)($9/64$)

$1NC(3/4) > 2NC(3/4) > 3C(1/4)$ (Not dismissed 3)($9/64$)

$1NC(3/4) > 2NC(3/4) > 3NC(3/4)$ (Not dismissed 4)($27/64$)

EXAMPLE

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

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

$p(\text{late three times but not dismissed}) =$
 $p(\text{not dismissed 1}) + p(\text{not dismissed 2})$
 $+ p(\text{not dismissed 3}) + p(\text{not dismissed 4}) = 9/64 + 9/64 + 9/64 + 27/64 = 54/64$

EXAMPLE

$$\begin{aligned} p(\text{caught}) &= \\ p(\text{dismissed 1}) + p(\text{dismissed 2}) + \\ p(\text{dismissed 3}) &= 4/64 + 3/64 + 3/64 \\ &= 10/64 \end{aligned}$$

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

BUSINESS MATHEMATICS & STATISTICS

BUSINESS MATHEMATICS & STATISTICS

LECTURE 35

Review Lecture 35
Elementary Probability
Part 2

PROBABILITY

Making assessment of chances

A worker out of 600 gets a prize by lottery

Chance of any one individual say Rashid being selected =
 $\frac{1}{600}$

The probability of the event "Rashid is selected" is the
probability of an event occurring

$$P(\text{Rashid}) = \frac{1}{600}$$

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

PROBABILITY

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

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

If out of 600 persons 250 are women, then the chance of a women being selected =

$$p(\text{woman}) = 250/600$$

PROBABILITY

In many situations there is no prior knowledge to calculate probabilities

What is the probability of a machine being defective?

Method: Monitor the machine over a period of time

Find out how many times it becomes defective

This experimental or empirical approach

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

PROBABILITY

There are two events A and B

What is the probability of either A or B happening?

What is the probability of A and B happening? Number of possibilities

Probability of A or B hapenning = Number of ways A or B can happen/ Total number of possibilities

= Number of ways A can happen + number of ways B can happen/ Total number of possibilities

PROBABILITY

= Number of ways A can happen/ Total number of possibilities + Number of ways B can happen/ Total number of possibilities

= Probability of A happening + Probability of B happening

Condition

A and B must be mutually exclusive

When A and B are mutually exclusive

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

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

40% workforce are women

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

25% females = management grade

30% of males = management grade

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

EXAMPLE

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

$$\text{Total workforce} = 100$$

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

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

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

SET OF MUTUALLY EXCLUSIVE EVENTS

Between them cover all possibilities

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

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

Example

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

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

EXAMPLE

A production line uses 3 machines

Chance that 1st machine

breaks down in any week = $1/10$

Chance for 2nd machine = $1/20$

Chance of 3rd machine = $1/40$

What is the chance that at least one machine breaks down
in any week?

EXAMPLE

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

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

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

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

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

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

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

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

BUSINESS MATHEMATICS & STATISTICS

BUSINESS MATHEMATICS & STATISTICS

Module 7

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

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

LECTURE 37

Revision Lecture 36

Patterns of Probability: Binomial, Poisson and Normal
Distributions

Part 1

EXAMPLE

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?

EXAMPLE

$1C(1/4) > 2C(1/4)$ (Dismissed 1) ($1/16 = 4/64$)

$1C(1/4) > 2NC(3/4) > 3C(1/4)$ (Dismissed 2) ($3/64$)

$1C(1/4) > 2NC(3/4) > 3NC(3/4)$ (Not dismissed 1) ($9/64$)

$1NC(3/4) > 2C(1/4) > 3C(1/4)$ (Dismissed 3) ($3/64$)

$1NC(3/4) > 2C(1/4) > 3NC(3/4)$ (Not dismissed 2) ($9/64$)

$1NC(3/4) > 2NC(3/4) > 3C(1/4)$ (Not dismissed 3) ($9/64$)

$1NC(3/4) > 2NC(3/4) > 3NC(3/4)$ (Not dismissed 4) ($27/64$)

EXAMPLE

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

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

$$p(\text{late three times but not dismissed}) = p(\text{not dismissed 1}) + p(\text{not dismissed 2}) + p(\text{not dismissed 3}) + p(\text{not dismissed 4}) = \frac{9}{64} + \frac{9}{64} + \frac{9}{64} + \frac{27}{64} = \frac{54}{64}$$

EXAMPLE

$$\begin{aligned} p(\text{caught}) &= \\ p(\text{dismissed 1}) + p(\text{dismissed 2}) + p(\text{dismissed 3}) &= 4/64 + \\ & 3/64 + 3/64 \\ &= 10/64 \end{aligned}$$

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

EXAMPLE

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?

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

Wrong! Probability greater than 1!

EXAMPLE

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

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

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

EXAMPLE

ALTERNATIVE METHOD

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

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

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

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

40% workforce female

25% females management cadre

30% males 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	

EXAMPLE

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
Management	18	10	28	
Non-Management	42	30	72	
Total	60	40	100	

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

EXAMPLE

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 = $203000/100 = 2030$			

EXPECTED VALUE

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

EXAMPLE

80%: no claim

15%: Claim 5000 Rs.

Remaining 5%: Claim 50000 Rs

Expected value of claim per policy?

$$\begin{aligned} EMV &= 0.8 \times 0 + 0.15 \times 5000 + 0.05 \times 50000 \\ &= 0 + 750 + 2500 \\ &= 3250 \text{ Rs.} \end{aligned}$$

TYPICAL PRODUCTION PROBLEM

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?

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

$n = 1$ to 10

$r = 1$ to 10

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

BUSINESS MATHEMATICS & STATISTICS

BUSINESS MATHEMATICS & STATISTICS

LECTURE 38

Patterns of probability: Binomial, Poisson and Normal
Distributions
Part 2

CUMULATIVE BINOMIAL PROBABILITIES

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

1. Look in column for n
2. Look in column for r
3. Look at column for value of p (0.05 to 0.5)

Example

$$\begin{aligned} n &= 5; r = 4; p = 0.5 \\ p(4 \text{ or more successes in } 5 \text{ trials}) \\ &= 0.1874 = 18.74 \% \end{aligned}$$

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.

BINOMDIST

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.



SUM ✗ ✓ fx =BINOMDIST(A3;A4;A5;FALSE)

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

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

EXAMPLE USING TABLES

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

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

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

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

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

= Chance of 5 or more wet days next week



	D49		
	A	B	C
35	EXAMPLE		
36	Probability of wet days in current month = 60%		
37	Probability of 5 or more wet days next week?		
38	p = 0.6. The tables are for p upto 0.5.		
39	Turn the question. $P(\text{dry}) = 1 - 0.6 = 0.4$		
40	p(5 or more wet days) = (7 - 5 = 2 or less dry days)		
41	p(2 or less dry days) = 1 -p(3 or more dry days)		
42	Data	Description	
43	2	Number of successes in trials	
44	7	Number of independent trials	
45	0.4	Probability of success on each trial	
46	0.4199	=BINOMDIST(A43;A44;A45;TRUE)	
47	At most 2 successes: 1, 2 (0.4199)		

EXAMPLE

8 bit message is transmitted electronically

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

What is the chance that entire message is transmitted correctly)?

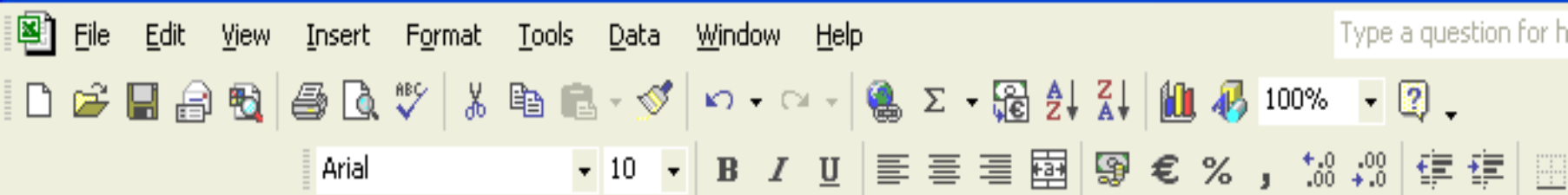
$n = 8; r = 8, p = 0.1; p(\text{exactly 0 errors})?$

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

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

TABLES

$$= 1 - 0.5695 = 0.4305$$



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

EXAMPLE

A surgery is successful for 75% patients
 $p(\text{success in at least 7 cases in randomly selected 9 patients})?$

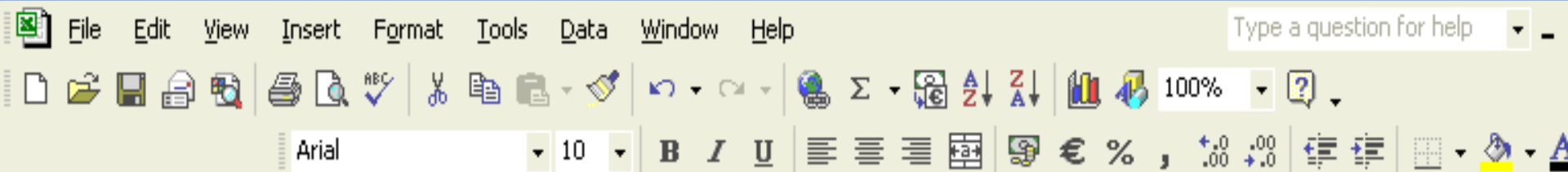
$n = 9$; $p(\text{success}) = 0.75$; $p(\text{at least 7 cases})?$

$p = 0.75$ is outside the table

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

Success at least 7 = Failure 2 or less

$P(\text{failure 2 or less}) = 1 - p(\text{failure 3 or more})$
 $= 1 - 0.3993 = 0.6007 = 60\%$



	C84		
	A	B	C
76	EXAMPLE		
77	Probability of success = 0.75. Failure = 0.25		
78	Probability of 7 successes out of 9? Or 2 or less failures		
79			
80	Data	Description	
81	2	Number of successes in trials	
82	9	Number of independent trials	
83	0.25	Probability of success on each trial	
84	0.6007	=BINOMDIST(B81;B82;B83;TRUE)	
85	At most 2 successes: (0.6007)		
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.

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.

NEGBINOMDIST

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.

B25 ▼ f_x

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

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.

BUSINESS MATHEMATICS & STATISTICS

BUSINESS MATHEMATICS & STATISTICS

BUSINESS MATHEMATICS & STATISTICS

LECTURE 39

Patterns of probability: Binomial, Poisson and Normal
Distributions

Part 3



SUM =CRITBINOM(A68;A69;A70)

	A	B	C	D
65	CRITBINOM(trials,probability_s,alpha)			
66				
67	Data	Description		
68	6	Number of Bernoulli trials		
69	0.5	Probability of a success on each trial		
70	0.75	Criterion value		
71	=	Smallest value for which the cumulative		
72	CRITBINO	binomial distribution is greater than		
73	M(A68;	or equal to a criterion value (4)		
74	A69;A70)			
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 = $p(\text{winning}) \times \text{gain per win} +$
 $p(\text{losing}) \times \text{loss if you loose}$

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

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

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

So on an average you stand to loose 5 Rs.

DECISION TABLES

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?

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
50	0	100	200	300	400	500	240

Expected profit 30 pies

$$\begin{aligned}
 &= 0.1 \times 200 + 0.2 \times 300 + 0.25 \times 300 + 0.2 \times 300 + 0.15 \times 300 + 0.1 \\
 &\quad \times 300 \\
 &= 20 + 60 + 75 + 60 + 45 + 30 \\
 &= 290 \text{ Rs.}
 \end{aligned}$$

Best Profit

35 Pies = Rs. 310

DECISION TREE

TOY MANUFACTURING CASE

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 = $8000 \times 200 - 500000 = 1600000 - 500000 = \mathbf{1100000}$ Rs.

Profit if **no rivals** = $12000 \times 200 - 500000 = 2400000 - 500000 = \mathbf{1900000}$ Rs.

Profit/Loss if no series = $2000 \times 200 - 500000 = 400000 - 500000 = -100000$ Rs. (**No series**)

EMV = Rival markets and no rivals = $0.5 \times \mathbf{1100000} + 0.5 \times \mathbf{1900000} = 1500000$ (**Series**)

EMV = $0.6 \times 1500000 + 0.4 \times -100000 = 900000 - 40000 = \mathbf{860000}$ Rs.

Conclusion

Go ahead

THE POISSON DISTRIBUTION

Either or situation

No data on trials

No data on successes

Average or mean value of successes or failures

Typical Poisson Situation

Characteristics

1. Either/or situation
2. Mean number of successes per unit, m , known and fixed
3. p , chance, unknown but small, (event is unusual)

THE POISSON TABLES OF PROBABILITIES

Give cumulative probability of r or more successes

Knowledge of m 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

$m = 7; r = 9;$

$P(r \text{ or more successes}) = 0.2709$

Values given in 4 decimals

EXAMPLE

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

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

$$m = 8/2 = 4$$

$$r = 0 \text{ (no breakdown)}$$

$$\begin{aligned} p(0 \text{ breakdown}) &= 1 - p(1 \text{ or more breakdowns}) \\ &= 1 - 0.9817 = 0.0183 = 1.83\% \end{aligned}$$

EXAMPLE

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

$$m = 1 \times 500/100 = 5$$

$$\begin{aligned} p(r = \text{less than three}) &= 1 - p(r = 3 \text{ or more}) \\ &= 1 - 0.8753 \\ &= 0.1247 \\ &= 12.47\% \end{aligned}$$

EXAMPLE

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

$$p = 1/100 = 0.1$$

$$m = 6$$

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

BUSINESS MATHEMATICS & STATISTICS

BUSINESS MATHEMATICS & STATISTICS

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



	A	B	C	D	E
1	POISSON(x, mean, cumulative)				
2	Data	Description			
3	2	Number of events			
4	5	Expected mean			
5	=				
6	POIS	Cumulative Poisson probability with the terms above (0.124652)			
7	SON(				
8	A3;				
9	A4;				
10	TRUE				
11	)				
12					
13					
14					
15					
16					

File Edit View Insert Format Tools Data Window Help

Type a question for

100%

Arial 20 B I U

SUM X ✓ fx =POISSON(A17;A18;FALSE)

	A	B	C	D	E
15	POISSON(x, mean, cumulative)				
16	Data	Description			
17	2	Number of events			
18	5	Expected mean			
19	=				
20	POISSON(	Cumulative Poisson probability with			
21	A17;A18;	the terms above (0.084224337)			
22	FALSE)				
23					
24					

THE PATTERN

Binomial and Poisson situations: either/or

Number of times could be counted

In the Candy problem with underweight boxes, there is
measurement of weight

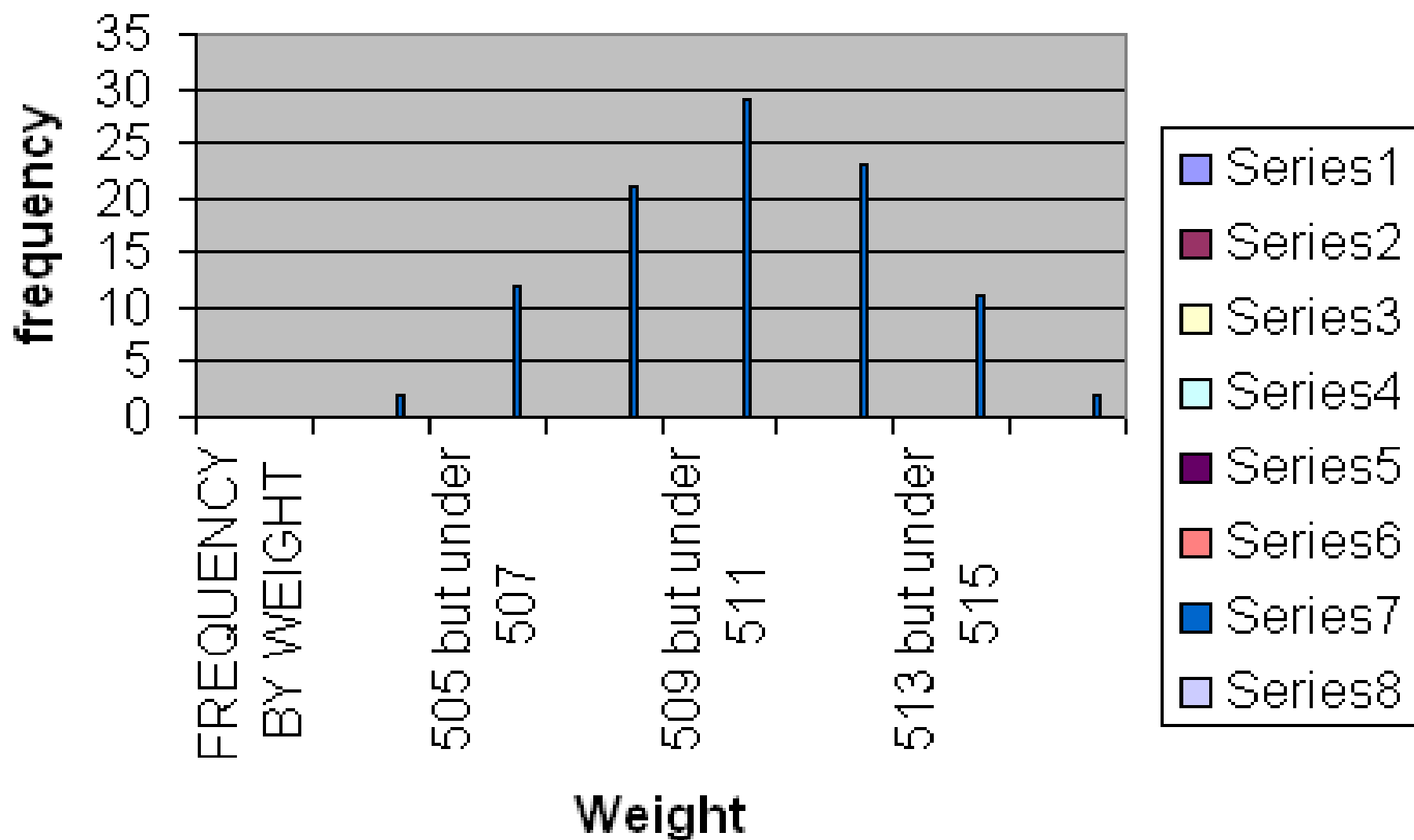
Binomial and Poisson : Discrete probability distributions

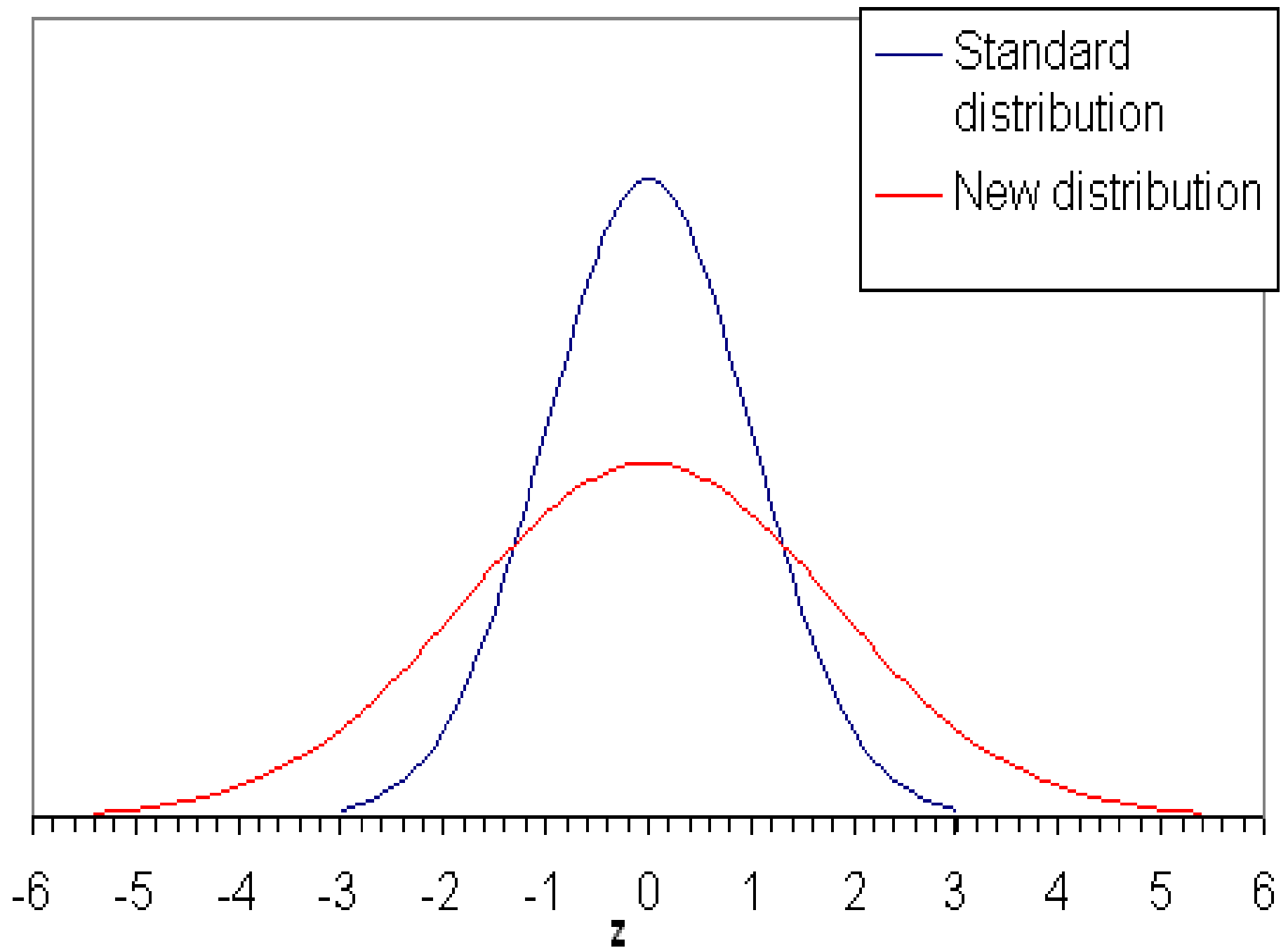
Candy problem is a Continuous probability distribution

F37

 f_x [illegible]

FREQUENCY DISTRIBUTION





NORMAL DISTRIBUTION

Blue Curve

A standard normal distribution (mean = 0 and standard deviation = 1)

Y-axis probability values

X-axis 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.)

NORMAL DISTRIBUTION

Probability 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.

NORMAL DISTRIBUTION

Mean gives the peak of the curve
Standard deviation gives the spread

Weight distribution case

Mean = 510 g

StDev = 2.5 gr

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 table gives the area under one tail
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

1. Look in column 1. Find 2.0
2. Look in column 0.05 and go to intersection of 2.0 and 0.05
3. 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

Find proportion of bags which have weight in excess of 515 g.

Mean = 510. StDev = 2.5 g

$$z = (515 - 510)/2.5 = 2$$

From tables: Area under tail = 0.02275 or 2.28%

EXAMPLE

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 $z = +1$

$$\text{Area} = 0.158$$

Hence

15.8% bags weigh less than 507.5 g

EXAMPLE

What is the probability that a bag filled by the machine weighs less than 512 g?

$$z = (512 - 510)/2.5 = 0.8$$

$$\text{Area under right tail} = 0.2119$$

$$= p(\text{weighs more than 512})$$

$$P(\text{weighs less than 512}) = 1 - p(\text{weighs more than 512})$$

$$= 1 - 0.2119$$

$$= 0.7881$$

EXAMPLE

What percentage of bags weigh between 512 and 515?

$$z1 = (512 - 510)/2.5 = 0.8$$

$$\text{Area 1} = 0.2119$$

$$z2 = (515 - 510)/2.5 = 2$$

$$\text{Area 2} = 0.02275$$

$$p(\text{bags weighs between 512 and 515}) =$$

$$\text{Area 1} - \text{Area 2}$$

$$= 0.2119 - 0.02275$$

$$= 0.18915 = 18.9\%$$

BUSINESS MATHEMATICS & STATISTICS

BUSINESS MATHEMATICS & STATISTICS

LECTURE 41

Estimating from Samples: Inference Part 1

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.

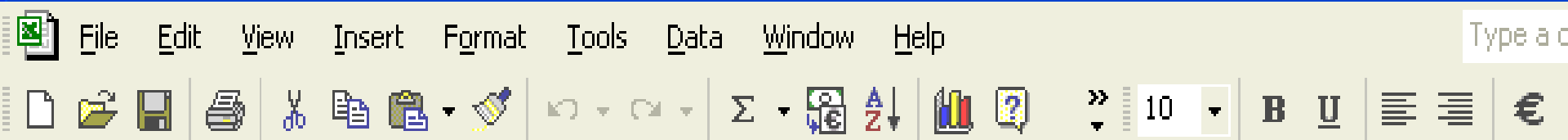
NORMDIST

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



	A	B	C
1	NORMDIST(x, mean, standard_dev, cumulative)		
2	Data Description		
3	42	Value for which you want the distribution	
4	40	Arithmetic mean of the distribution	
5	1.5	Standard deviation of the distribution	
6	0.9		
7		Cumulative distribution function for	
8		the terms above (0.908789)	
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.

SUM    =NORMSDIST(1.333333)

	A	B	
11	NORMSDIST(z)		
12	A	B	
13	Formula	Description (Result)	
14	=NORMSDIST(1.333333)		
15	Normal cumulative distribution function		
16	at 1.333333 (0.908789)		
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.

SUM =NORMINV(A23;A24;A25)

	A	B	
20	NORMINV(probability, mean, standard_dev)		
21	A	B	
22	Data	Description	
23	0.90879	Probability corresponding to the normal distribution	
24	40	Arithmetic mean of the distribution	
25	1.5	Standard deviation of the distribution	
26	=NORMINV(A23;A24;A25)		
27	Inverse of the normal cumulative		
28	distribution for the terms above (42)		

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

SUM    =NORMSDIST(1.333333)

	A	B	
11	NORMSDIST(z)		
12	A	B	
13	Formula	Description (Result)	
14	=NORMSDIST(1.333333)		
15	Normal cumulative distribution function		
16	at 1.333333 (0.908789)		
17			

SAMPLING VARIATIONS

Electronic components are despatched by a manufacturer
in boxes of 500

A small number of faulty components is unavoidable

Customers agree to a rate of 2%

One customer recently found 25 faulty components (5%)

Was this box representative of production as a whole?

SAMPLING VARIATIONS

The box represents a sample from the whole output
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?

EXAMPLE

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

Houselds E and F: Do not

It was decided to take random samples of 3 households

List all possible samples and find % of each sample using
corn flakes

EXAMPLE

Sample	% cf users	Sample	% cf users
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% cf users
12 contain 67% cf users
4 contain 33% cf 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% cf 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

$$\text{Mean} = 100\% \times 0.2 + 67\% \times 0.6 + 33\% \times 0.2 = 67\%$$

Mean of the sampling distribution is the true percentage for the population as a whole

Must make allowance for variability in samples

CONDITIONS FOR SAMPLE SELECTION

1. Number of items in the sample, n , is fixed and known in advance
2. Each item either has or has not the desired characteristic
3. 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

BUSINESS MATHEMATICS & STATISTICS

BUSINESS MATHEMATICS & STATISTICS

LECTURE 42

Estimating from Samples: Inference

Part 2

EXAMPLE

In a factory 25% workforce is women
How likely is it that a random sample of 80 workers
contains 25 or more women?

Mean = 25%

$$\begin{aligned}\text{STEP} &= [25(100 - 25)/80]^{1/2} \\ &= [(25 \times 75)/80]^{1/2} \\ &= 4.84\%\end{aligned}$$

$$\% \text{ women in sample} = (25/80) \times 100 = 31.25\%$$

$$\begin{aligned}z &= (31.25 - 25)/4.84 \\ &= 1.29\end{aligned}$$

$$\begin{aligned}P(\text{sample contains 25 women}) &: (\text{Look for p against } z = 1.29) \\ &= 0.985 \text{ or about } 10\%\end{aligned}$$

APPLICATIONS OF STEP

1. What is the probability that such a sample will arise?
2. Estimate the percentage P from information obtained from a single sample
3. How large a sample will be required in order to estimate a population percentage with a given degree of accuracy?

CONFIDENCE LIMITS

A market researcher wishes to conduct a survey to determine % consumers buying company's products

He selects a sample of 400 consumers at random

He finds that 280 of these are purchasers of the product

What can he conclude about % of all consumers buying the product?

CONFIDENCE LIMITS

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

$$\text{STEP} = [(70(100 - 70)/400)]^{1/2} = 2.29\%$$

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
= 85% 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

A sample of 60 students contains 12 who are left handed or
20%

Find the range with 95% confidence in which the entire left
handed students fall

$$\begin{aligned}\text{Range} &= 20 \pm 2 \times \text{STEP} \\ &= 20 \pm 2 \times [(20 \times 80)/60]^{1/2} \\ &= 9.67\% \text{ and } 30.33\%\end{aligned}$$

ESTIMATING PROCESS SUMMARY

1. Identify n and P (the sample size and percentag in the sample)
2. Calculate STEP using these values
3. The 95% confidence interval is approximately $P \pm 2$ STEP

99% confidence

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

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?

$s = 15$; $n = 64$; population mean = 100

$$\text{STEM} = 15/(64)^{1/2} = 15/(64)^{1/2} = 15/8 = 1.875$$

$$\begin{aligned} z &= 100 - 95 / \text{STEM} \\ &= 5/1.875 = 2.67 \end{aligned}$$

This gives a probability of 0.0038

So the chance that the **average IQ of the sample** is below 95 is very small

BUSINESS MATHEMATICS & STATISTICS

BUSINESS MATHEMATICS & STATISTICS

LECTURE 43

Hypothesis testing : Chi-Square Distribution Part 1

EXAMPLE

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

$$\text{STEM} = \text{s.d}/(n)^{1/2}$$

s.d is not known

Use s.d of sample as an approximation

$$\text{STEM} = 5/(100)^{1/2} = 0.5$$

95% confidence interval = $225 \pm 2 \times 0.5$ or

From 224 to 226 g

PROBLEM 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 **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
Training Manager would like to assess the effect of retraining
if any

Particular questions

1. Is quality of product better than produced before retraining?
2. Has the speed of machines increased?
3. 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

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

$$\begin{aligned} &= [P(100 - P)/n]^{1/2} \\ &= [4(100 - 4)/400]^{1/2} = 0.98\% \end{aligned}$$

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)

$$P = 5$$

$$\begin{aligned}\text{STEP} &= [5(100 - 5)/400]^{1/2} \\ &= [5 \times 95/400]^{1/2} \\ &= 1.09\end{aligned}$$

Range at 95 % Confidence Limits

$$\begin{aligned}&= 5 \pm 2 \times 1.09 \\ &= 2.82 \% \text{ to } 7.18 \%\end{aligned}$$

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**
4. (A) If the sample falls **outside the interval**, **reject the null hypothesis** (sample differs significantly from the population %)
(B) 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 that population percentages will be applicable

1. 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
2. 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

1. 99% interval requires $2.58 \times \text{STEP}$. Interval wider. Less likely to conclude that something is significant
2. (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)

BUSINESS MATHEMATICS & STATISTICS

BUSINESS MATHEMATICS & STATISTICS

LECTURE 44

Hypothesis testing : Chi-Square Distribution
Part 2

FURTHER POINTS ABOUT HYPOTHESIS TESTING

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

$$4 - 1.64 \times 0.98 = 2.39\%$$

New figure = 3.5%

No reason to conclude that things have improved

FURTHER POINTS ABOUT HYPOTHESIS TESTING

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

$4 \pm 2 \times 1.96 \times 0.98 = 2.08\% \text{ to } 5.92\%$

New figure = 3.5%

No reason to conclude that things have improved

HYPOTHESES ABOUT MEANS

Before the course

Worker X took 2.5 minutes to produce 1 item

StDev = 0.5 min

After the course

64 items; mean = 2.58 min

Null hypothesis

No change after the course

STEM

$$= s.d./n^{1/2} = 0.5/(64)^{1/2} = 0.0625$$

Range

$$= 2.5 \pm 2 \times 0.0625 = 2.375 \text{ to } 2.625 \text{ min}$$

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 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
 - (A) If sample z is smaller, do not reject the Null Hypothesis
 - (B) 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 StDev of sample and have to approximate population StDev then we use **Student's t-distribution**

STUDENT'S t-DISTRIBUTION

Very much like normal distribution

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 StDev by that of the sample

EXAMPLE

Mean training time for population = 10 days

Sample mean for 8 women = 9 days

Sample StDev = 2 days

To approximate population StDev by a sample divide the sum of squares by $n - 1$

$$STEM = 2/(8)^{1/2} = 0.71$$

Null Hypothesis : There is no difference in overall training time between men and women

$$t\text{-value} = (\text{sample mean} - \text{population mean})/STEM$$

$$= (9 - 10)/0.71 = -1.41$$

$$\text{For } n = 8, v = 8 - 1 = 7;$$

For 5%(.05) significance level looking at 0.025 (2-tailed)

$$t = 2.365 \text{ (Calculated table value)}$$

Decision: Do not reject the Null Hypothesis

SUMMARY - I

If underlying population is normal &

We know the StDev

Then

Distribution of sample means is normal

With StDev = 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 $\text{StDev} = \text{STEM} = \text{population s.d.}/(n)^{1/2}$

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 $\text{StDev} = \text{STEM} = \text{sample s.d}/(n)^{1/2}$
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 StDev = Rs. 10

50 Workers from Maintenance had a mean of Rs. 130 with
StDev = 12

Is there a difference in wages between workers?

Difference of two sample means = $s[(1/n_1) + (1/n_2)]^{1/2}$

$$s = [(n_1 \cdot s_1^2 + n_2 \cdot s_2^2) / (n_1 + n_2)]^{1/2}$$

$$N_1 = 30; n_2 = 50; s_1 = 10; s_2 = 12$$

$$s = [(30 \times 100 + 50 \times 144) / (30 + 50)]^{1/2} = 11.29$$

Standard Error of Difference in Sample Means (STEDM)

$$= 11.29(1/30 + 1/50)^{1/2} = 2.60$$

$$Z = (\text{difference in sample means} - 0) / \text{STEDM}$$

$$= 120 - 130 / 2.60 = -3.85$$

Well outside the critical z for 5% significance

**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

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 (χ^2)

Degrees of freedom = $(r-1) \times (c-1) = (3-1)(2-1) = 2$

Critical value of chi-squared at 5% (and $v = 2$) = 5.991

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 χ^2
4. Calculate degrees of freedom (rows minus 1) x (columns minus 1); look up the critical χ^2 under the selected significance level
5. Compare the calculated value of χ^2 from the sample with value from the table; if the sample χ^2 is smaller (within the interval) don't reject the null hypothesis; if it is bigger (outside) reject the null hypothesis

	A	B	C	D	E	F	G
1	STATUS	SEX	AGE				
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				

Microsoft Excel - Emp					
File Edit View Insert Format Tools Data Window Help					
Type a question for help					
D15 fx					
	A	B	C	D	E
1	t-Test: Two-Sample Assuming Unequal Variances				
2					
3		38	42		
4	Mean	37.913	37.571		
5	Variance	37.356	75.252		
6	Observations	23	35		
7	Hypothesized Mean Di	0			
8	df	56			
9	t Stat	0.1758			
10	P(T<=t) one-tail	0.4305			
11	t Critical one-tail	1.6725			
12	P(T<=t) two-tail	0.8611			
13	t Critical two-tail	2.0032			
14					

CHITEST

Returns the test for independence. **CHITEST** returns the value from the **chi-squared (χ^2) distribution** for the statistic and the appropriate degrees of freedom. You can use χ^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.

SUM =CHITEST(A3:B5;A7:B9)

	A	B	C
	Men (Actual)	Women (Actual)	Description
2			
3	58	35	Agree
4	11	25	Neutral
5	10	23	Disagree
6	Men (Expected)	Women (Expected)	Description
7	45.35	47.65	Agree
8	17.56	18.44	Neutral
9	16.09	16.91	Disagree
10	A3:B5;A7: B9)	The χ^2 statistic for the data above is 16.16957 with 2 degrees of freedom (0.000308)	
11			

BUSINESS MATHEMATICS & STATISTICS

BUSINESS MATHEMATICS & STATISTICS

LECTURE 45

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:

A set of decision variables, x_j .

An objective function, $\sum c_j x_j$.

A set of constraints, $\sum a_{ij} x_j \leq b_i$.

THE FORMAT FOR AN LP MODEL

Max or min $\sum c_j x_j = c_1 x_1 + c_2 x_2 + \dots + c_n x_n$

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

Define decision variables

Hand-write objective

Formulate math model of objective function

Hand-write each constraint

Formulate math model for each constraint

Add non-negativity conditions

Introduction to Linear Programming

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

Introduction to Linear Programming

The Importance of Linear Programming

There are efficient solution techniques that solve linear programming models

The output generated from linear programming packages provides useful “what if” analysis

Introduction to Linear Programming

Assumptions of the linear programming model

The parameter values are known with certainty

The objective function and constraints exhibit constant returns to scale

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

A Production Problem – A Prototype Example

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

Technological input

Model A requires 2 kg of plastic and
3 minutes of labour per dozen.

Model B requires 1 kg of plastic and
4 minutes of labour per dozen

A Production Problem – A Prototype Example

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.

The current production plan consists of:

Model A = 450 dozen

Model B = 100 dozen

Profit = Rs. 410,000 per week

$$800(450) + 500(100)$$


A Production Problem – A Prototype Example

Management is seeking
a production schedule
that will increase
the company's profit

A Production Problem – A Prototype Example

A linear programming model
can provide
an insight
and
an intelligent solution
to this problem

A Production Problem – A Prototype Example

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

A Production Problem – A Prototype Example

$$\begin{aligned} \text{Max } & 800X_1 + 500X_2 && \text{(Weekly profit)} \\ \text{subject to } & && \\ & 2X_1 + 1X_2 \leq 1000 && \text{(Plastic)} \\ & 3X_1 + 4X_2 \leq 2400 && \text{(Production Time)} \\ & X_1 + X_2 \leq 700 && \text{(Total production)} \\ & X_1 - X_2 \leq 350 && \text{(Mix)} \\ & X_j \geq 0, \quad j = 1, 2 && \text{(Nonnegativity)} \end{aligned}$$

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 PROBLEM

The dentist employs three assistants and uses two operators
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

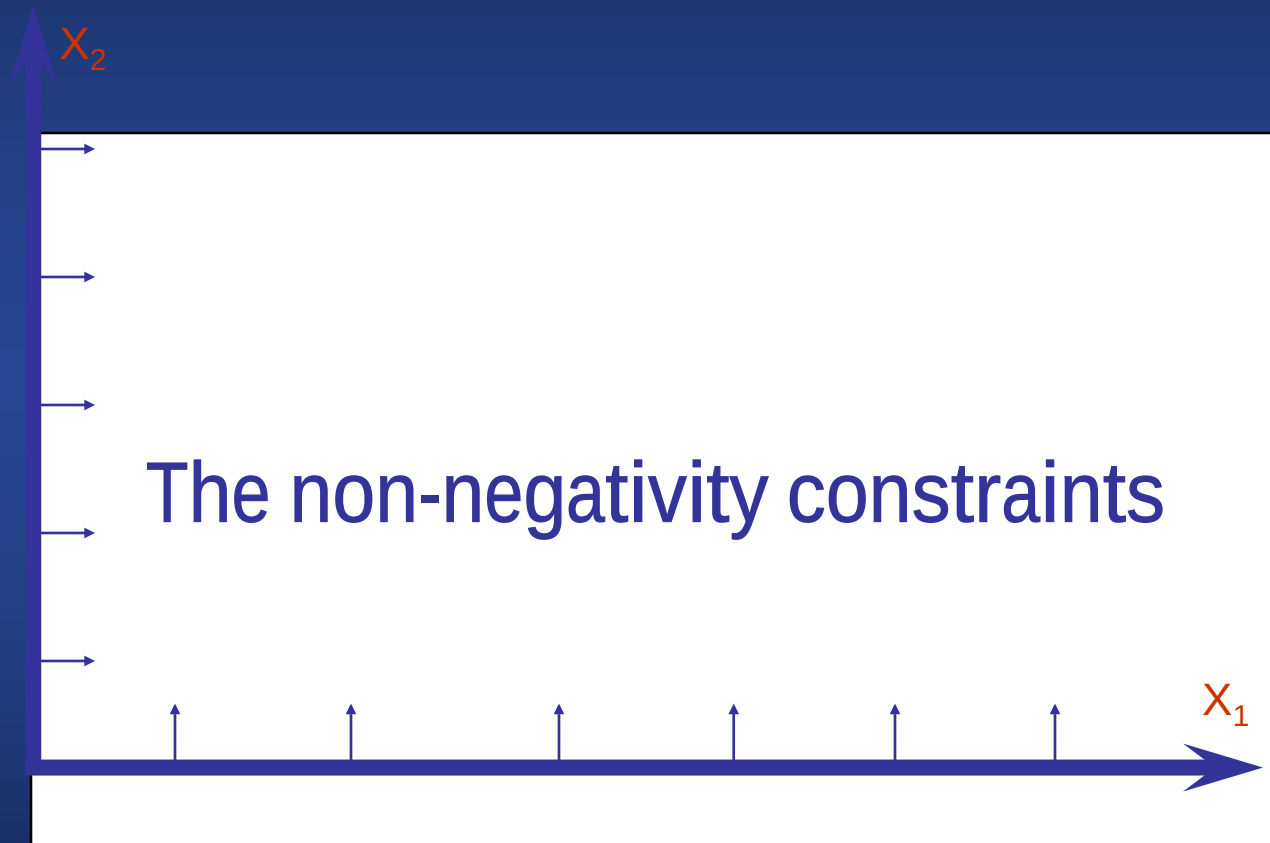
The set of all points
that satisfy
all the constraints
of the model
is called a

FEASIBLE REGION

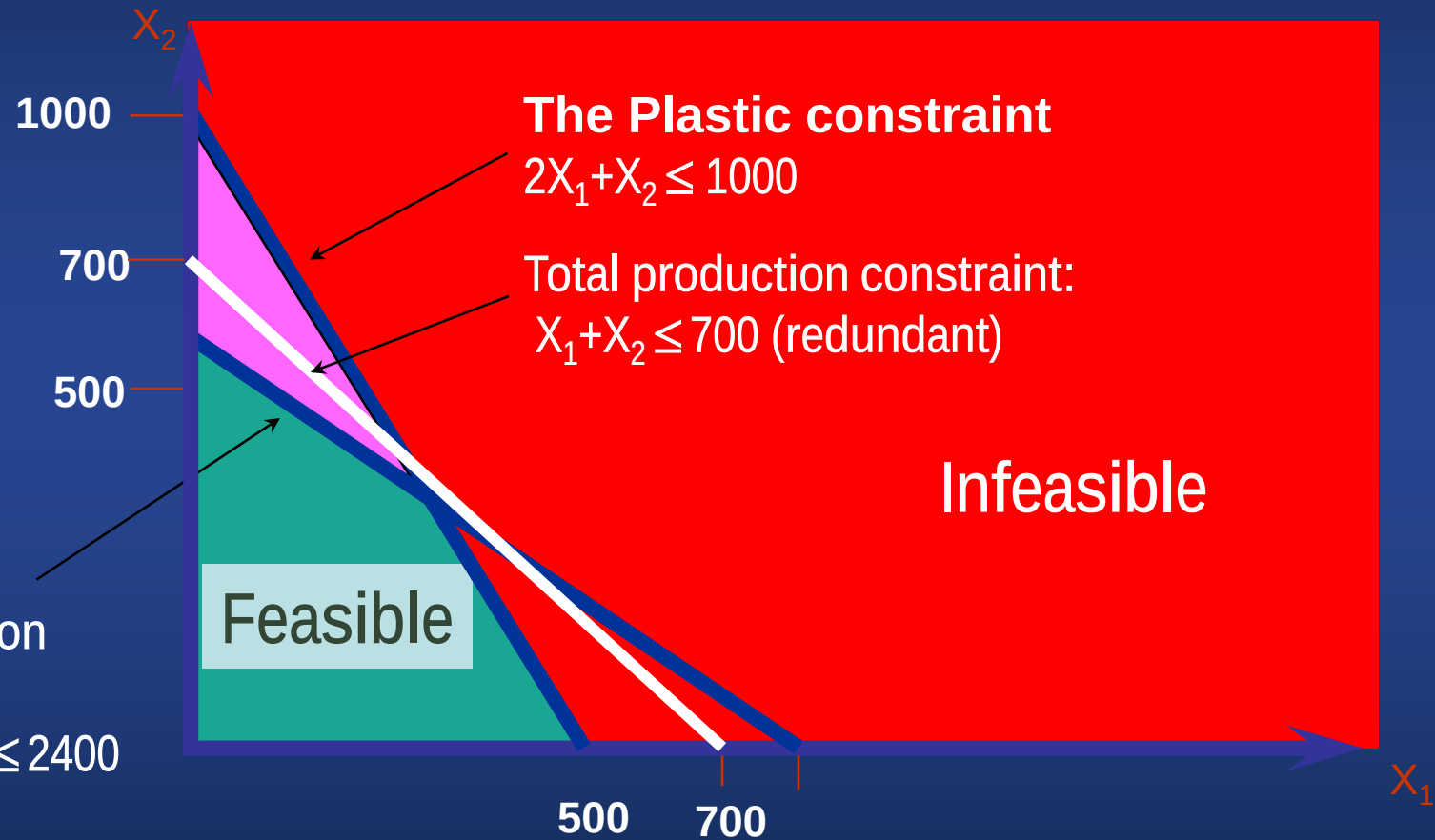
THE GRAPHICAL ANALYSIS OF LINEAR PROGRAMMING

Using a graphical presentation
we can represent
all the constraints,
the objective function,
and
the three types of
feasible points

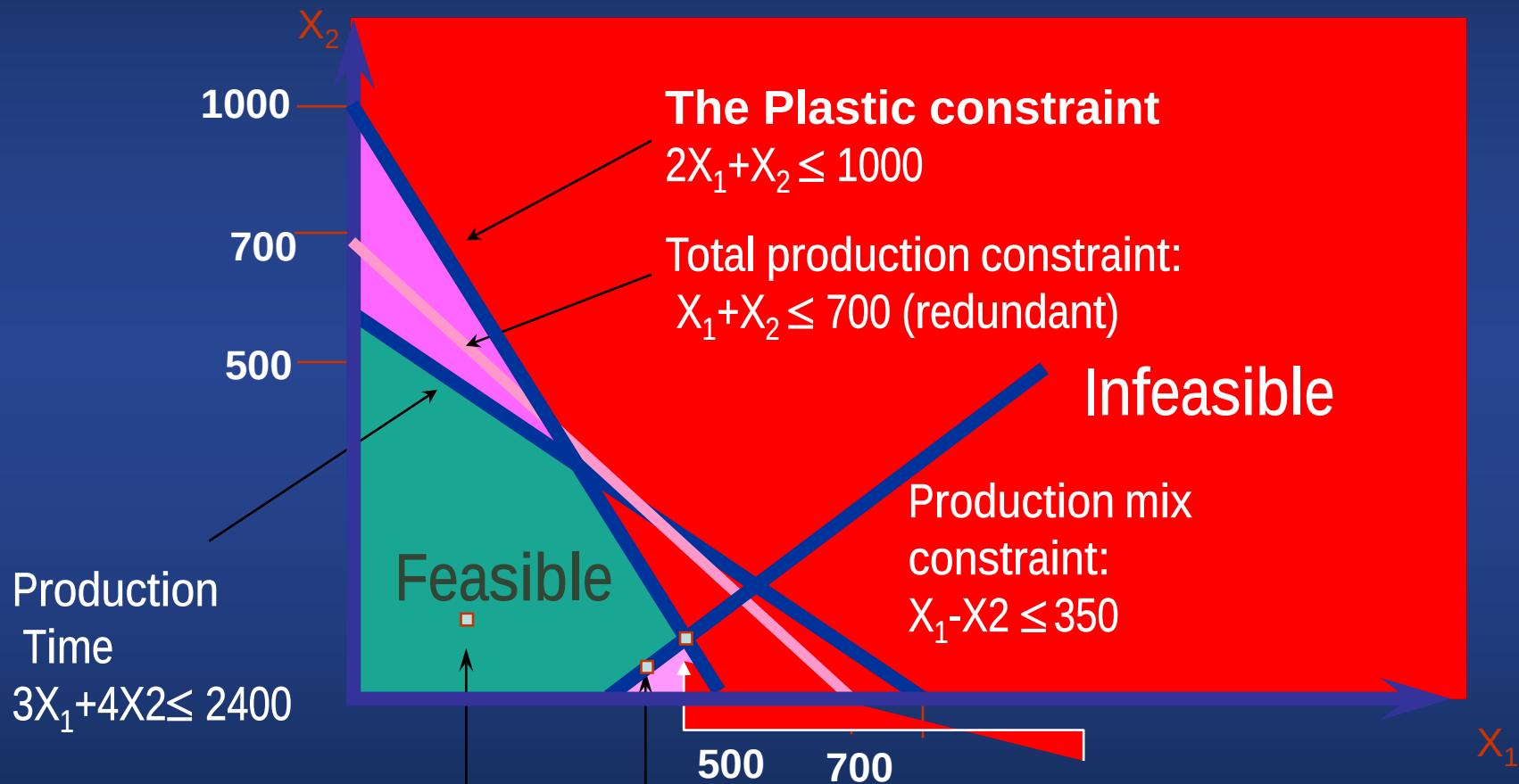
GRAPHICAL ANALYSIS – THE FEASIBLE REGION



GRAPHICAL ANALYSIS – THE FEASIBLE REGION



GRAPHICAL ANALYSIS – THE FEASIBLE REGION

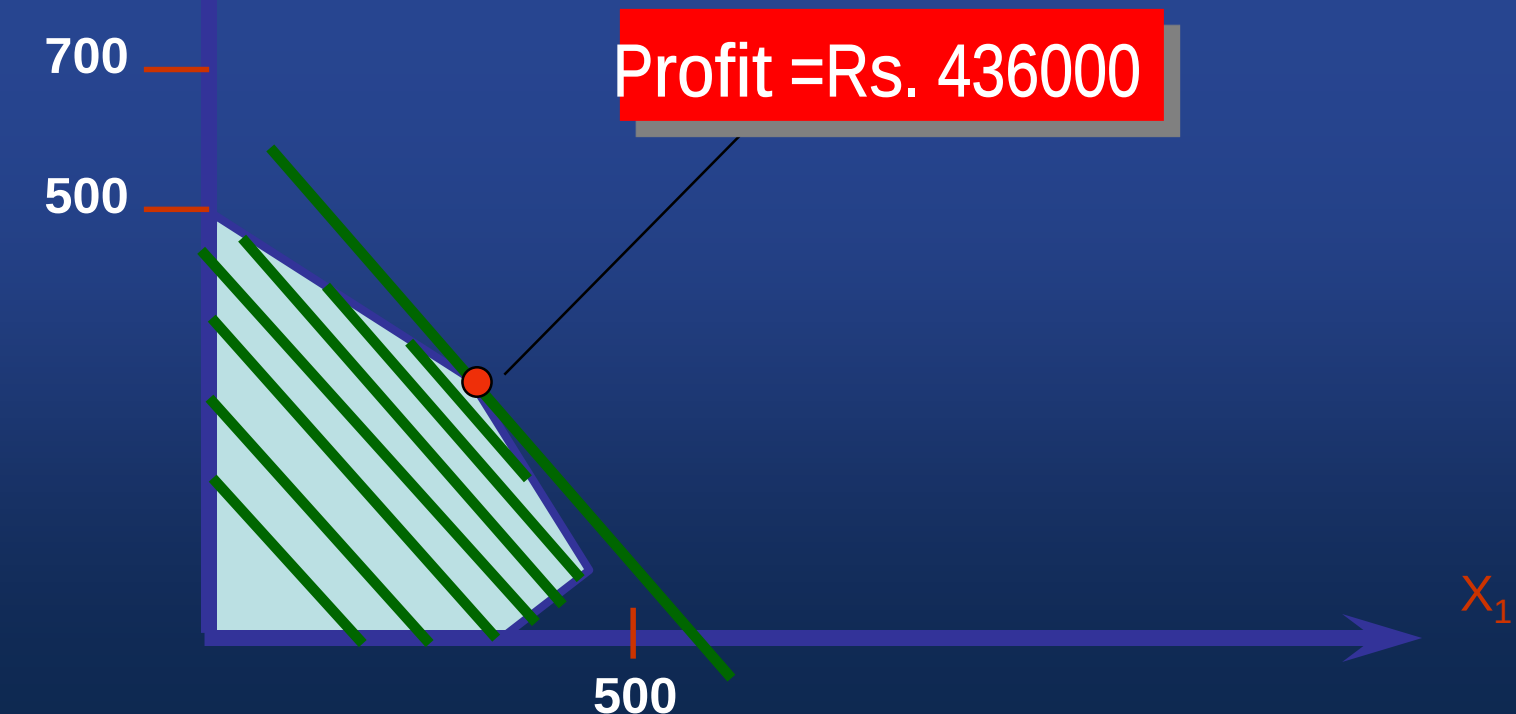


Interior points. Boundary points. Extreme points.

There are three types of feasible points

THE SEARCH FOR AN OPTIMAL SOLUTION

x_2 Start at some arbitrary profit, say profit = Rs.200,000...
Then increase the profit, if possible...
...and continue until it becomes infeasible



SUMMARY OF THE OPTIMAL SOLUTION

Model A = 320 dozen

Model B = 360 dozen

Profit = Rs. 436000

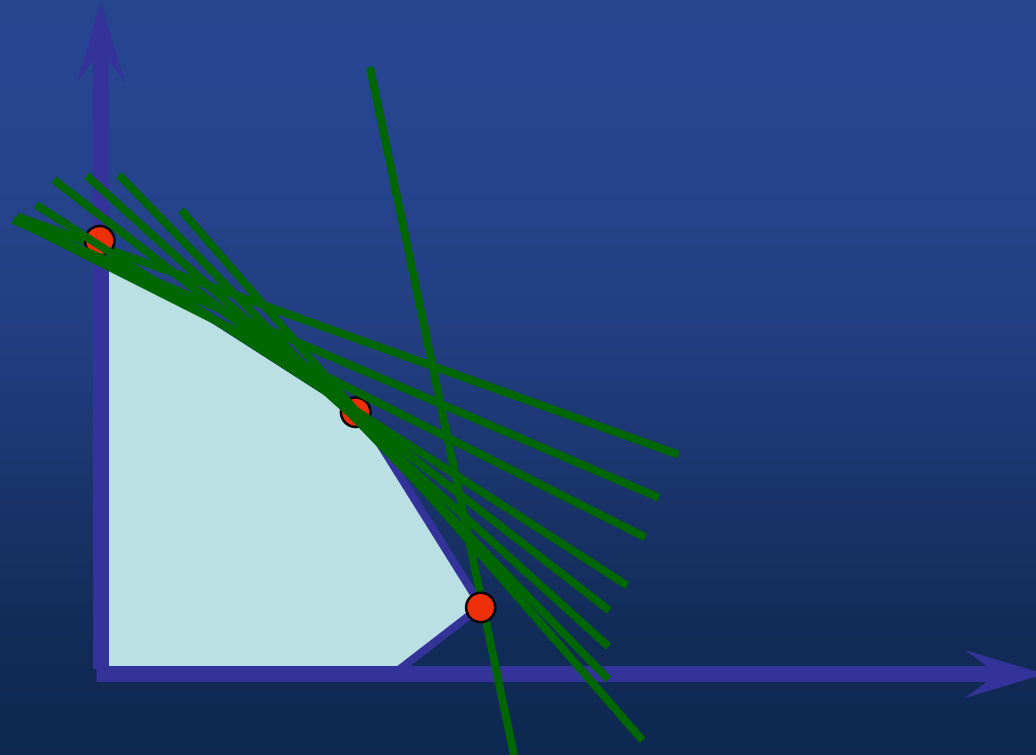
This solution utilizes all the plastic and all the production hours

Total production is only 680 (not 700)

Model a production does not exceed Model B production at all

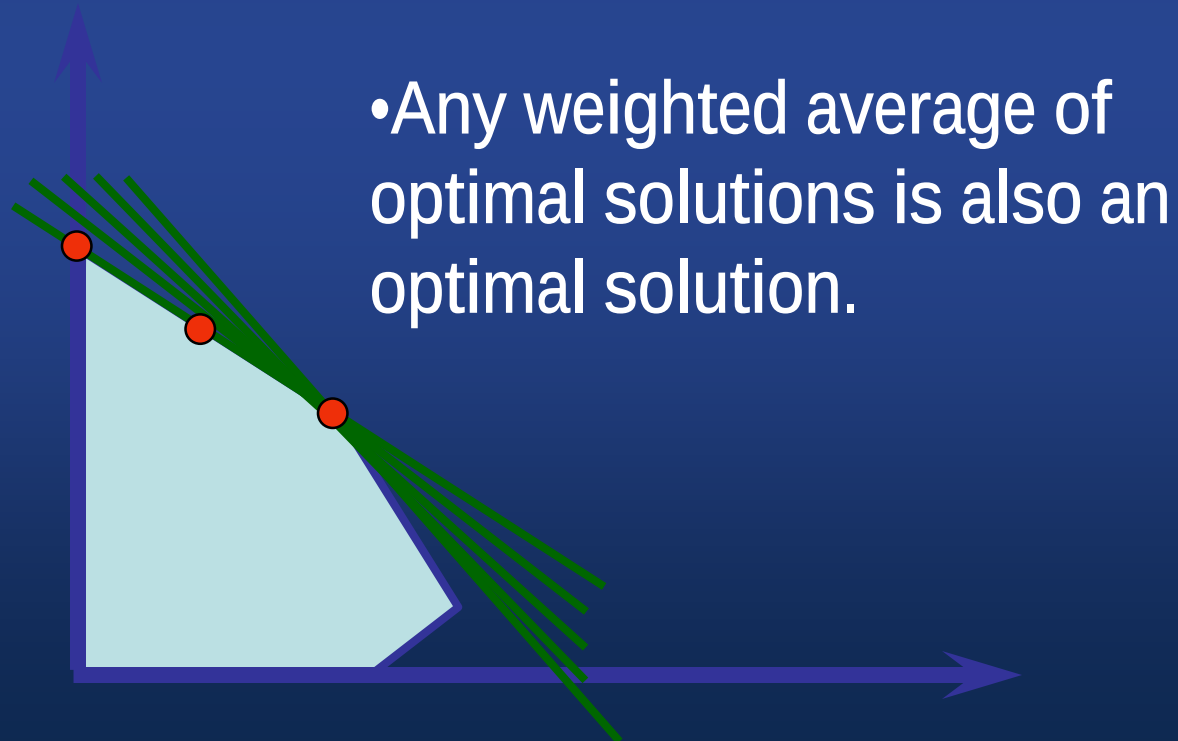
EXTREME POINTS AND OPTIMAL SOLUTIONS

- If a linear programming problem has an optimal solution, an extreme point is optimal.



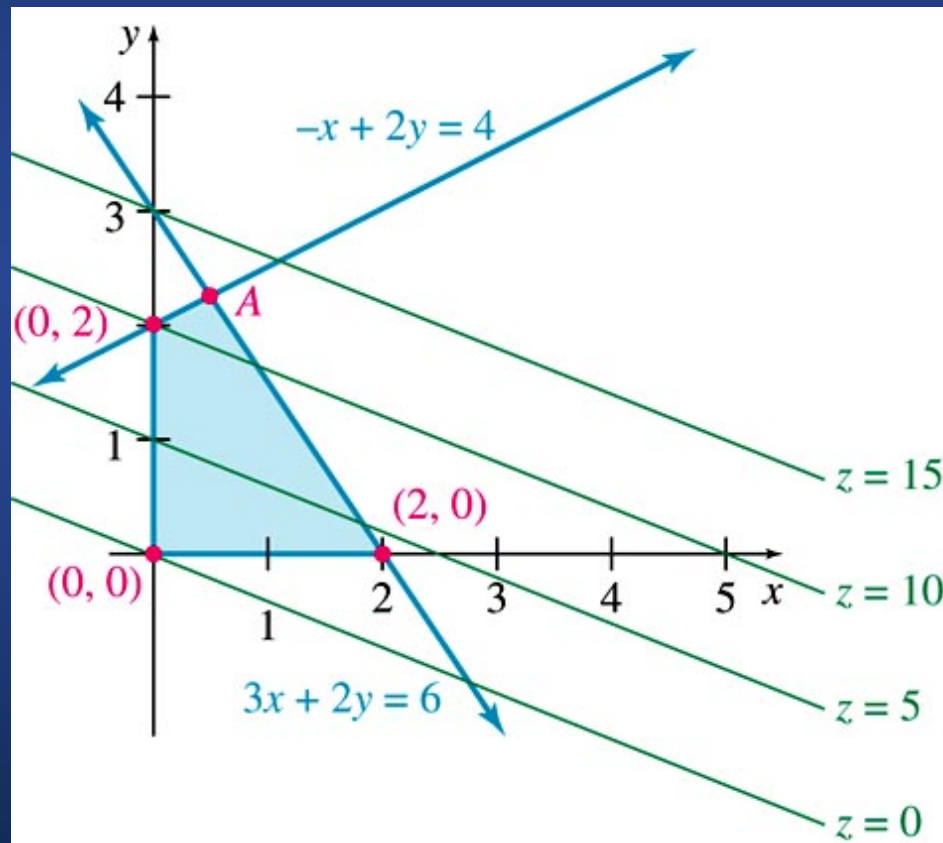
MULTIPLE OPTIMAL SOLUTIONS

- For multiple optimal solutions to exist, the objective function must be parallel to one of the constraints

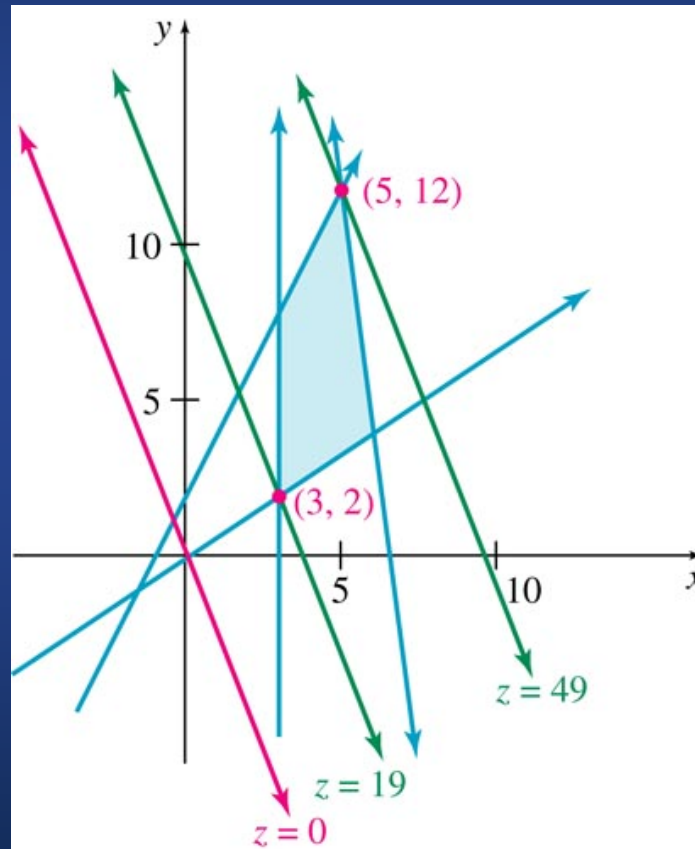


BUSINESS MATHEMATICS & STATISTICS

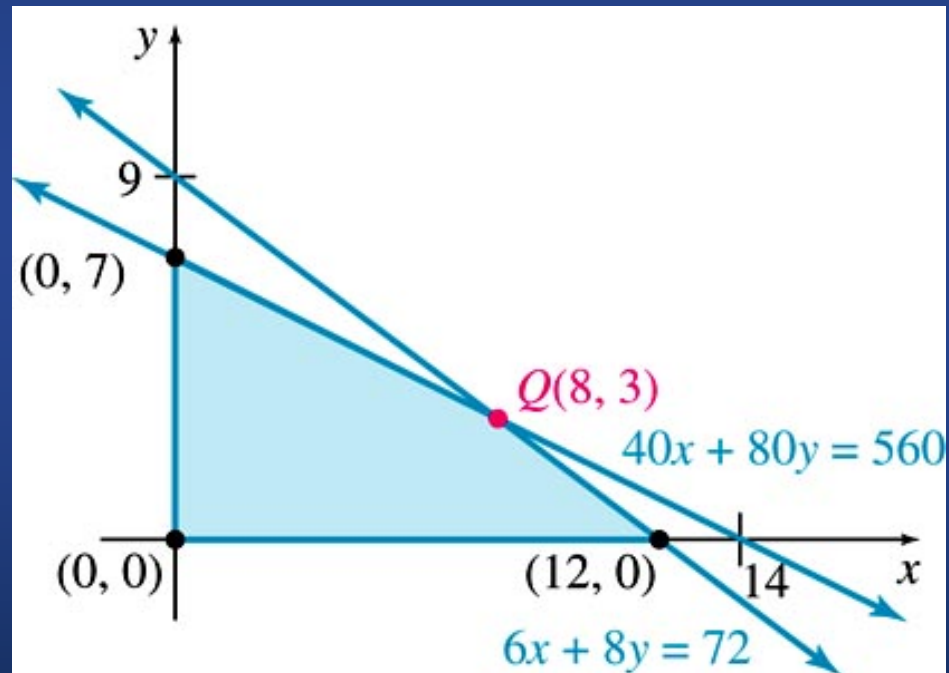
Example



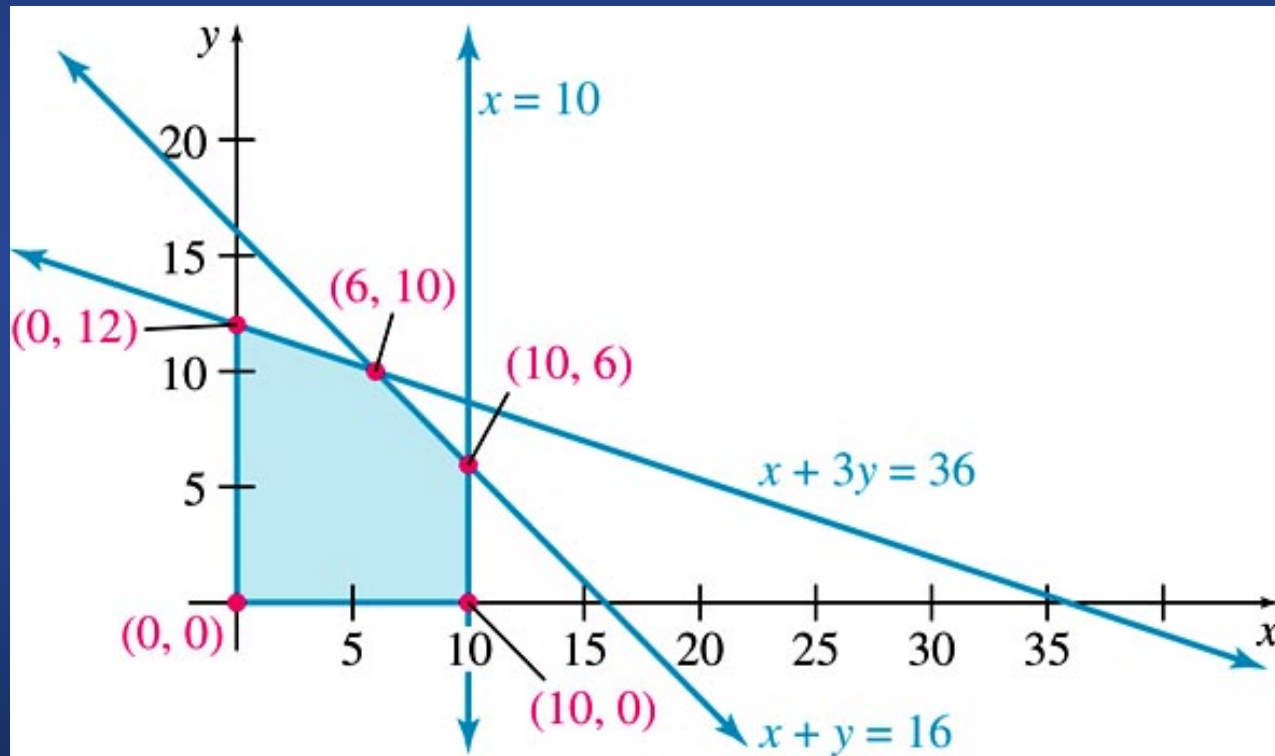
Example



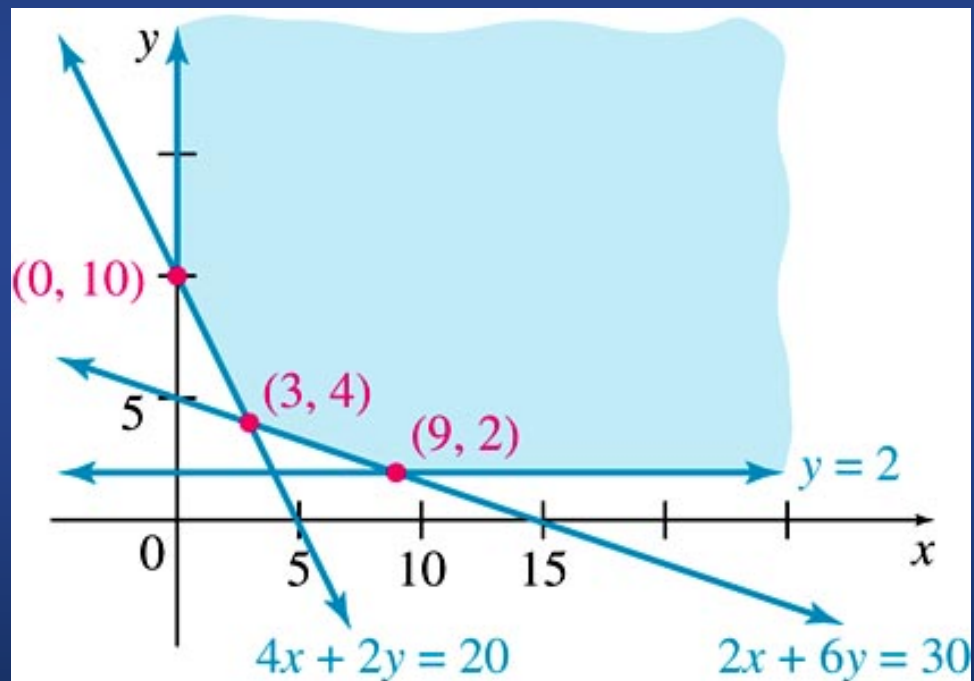
Example



Example



Example



Example

