بسم الله الرحمن الرحيم

DATA PROCESSING AND STATISTICAL GEOLOGY

EHG 201

REFERNCES

1) Andrew S. C. Ehrenberg, Andrew S. Ehrenberg. A,. (1996). Primer in Data Reduction: An Introductory Statistics, John Wiley & Sons Incorporated, 324pp.

2) Thomas H. Wonnacott, Ronald J. Wonnacott (1990). Introductory Statistics- John Wiley & Sons, Incorporated , 1014 pages

3) Friend, D., (1987). Quantitative Methods. Longman , London, 384pp.

Scheme of Assessment

- Exams: 45%
- Homework: 15%
- Final Exam: 40%

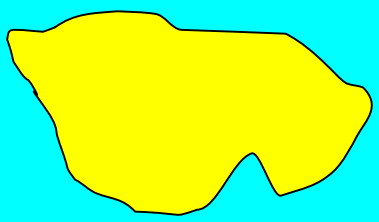
Table of Contents

Definition: statistics is concerned with scientific methods for

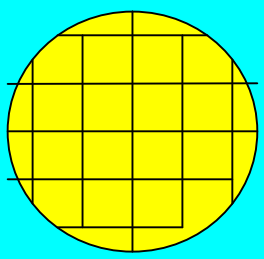
- Collecting;
- Organizing;
- Summarizing,
- Presenting; and
Analyzing data, as well as drawing valid conclusions and making reasonable decisions on the basis of such analysis.

NOTE

There are two important questions for the number and;
the place of collected samples

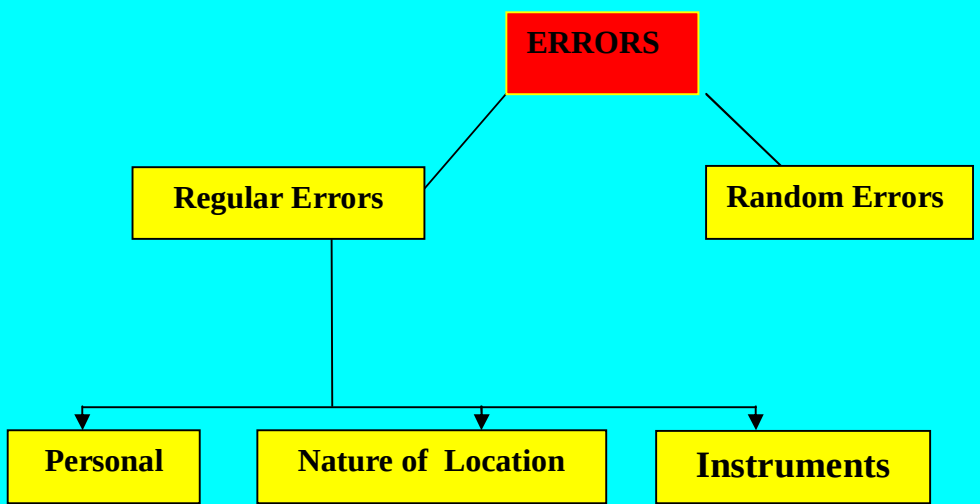


RANDOM



GRIDS

Errors Associated in Data Collection



GENERAL DIFFINITIONS

- **POPULATION:** Which is all subjects possessing a common characteristic that is being studied. and its classified into
  (1) Finite
  (2) Infinite
  (3) Real
  (4) Hypothetical

- **SAMPLE:** A subgroup or subset of the population.

- **DATA :** It's the observations, which classified into (1) Quantitative Data
  (2) Qualitative Data

- **RANDOM VARIABLE:** A variable whose values are determined by chance. Which classified into
  (1) Discrete R. Variables, which assume a finite or countable number of possible values. Usually obtained by counting.
  (2) Continuous R.V, which assume an infinite number of possible values.

Frequency Distribution

Frequency Tables

Example

The following data represent 50 students grades

B C A C D F C B A F D C B C F C D C C B C C
F C A D B D C C B D C F B C D C B A C C E
C D B A F C D

Construct the frequency table using Tally Marks method?

Grade	Marks
A	𝍸
B	𝍸 𝍸
C	𝍸 𝍸 𝍸 𝍸
D	𝍸
F	𝍸 /

NOTE

In case of numerical and large amount of data the above technique not suitable so we will select the following approach (using classes)

(1) Determined the number of classes using the following formula

$$\text{Number of classes} = 1 + 3.3 \log n$$

Where n is a number of data use

EXAMPLE

The concentration of magnesium of 45 groundwater samples were determined and shown in the following table. Put the data in the form of frequency table

15.0	18.1	14.4	14.6	10.9	18.1	18.2	18.3	15.0
16.0	12.6	16.6	**20.7**	19.8	11.6	12.8	15.6	11.0
15.3	9.4	19.5	18.3	14.5	16.6	11.5	16.4	12.5
14.6	11.9	12.5	18.6	13.1	12.1	10.7	17.3	12.4
17.0	**6.3**	16.8	12.5	16.3	14.7	12.7	16.3	11.5

Answer

1- Find the number of classes

$$=1 + 3.3 \log 45$$
$$=1 + 3.3 \log (1.653)$$
$$=6.46$$
$$\approx 7$$

2- Find the range (R)

The range of a set of numbers is the largest value in the set minus the smallest value in the set. Note that the range is a single number, not many numbers

$$\textbf{Range (R) = } X_{\text{maximum}} - X_{\text{minimum}}$$

$$R = 20.7 - 6.3 = 14.4$$

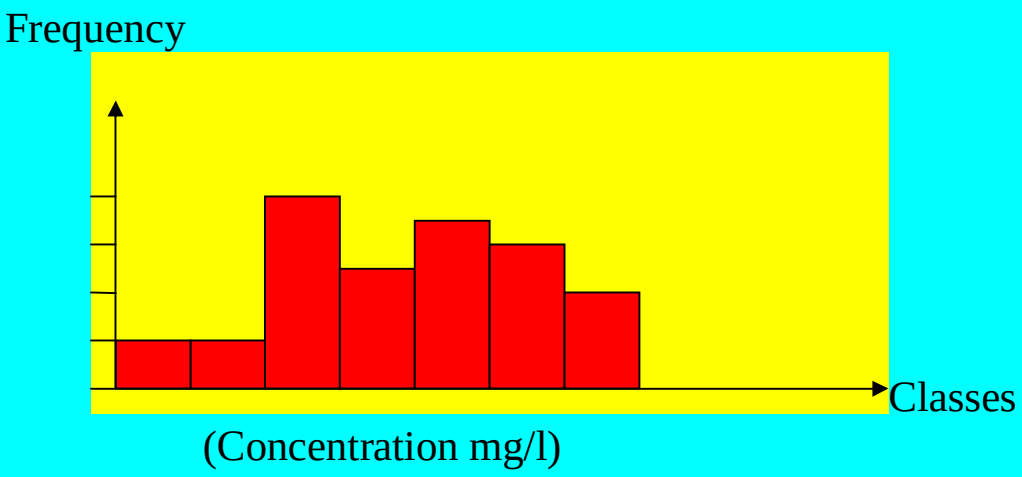
3- Find the length of class

$$\textbf{Length of class} = \frac{\textbf{Range}}{\textbf{No. of Classes}}$$

$$= 14.4/ 7= 2.06 \approx 2.1$$

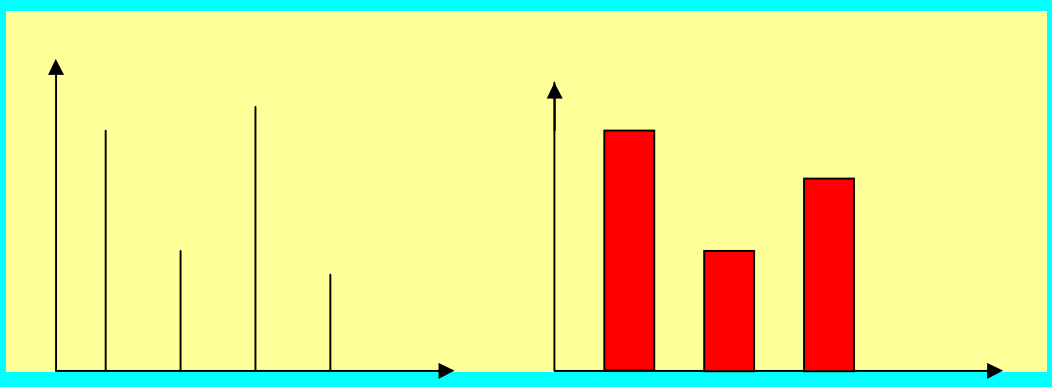
6.25 – 8.35
8.35 – 10.45
10.45 – 12.55
12.55 – 14.65
14 65 - 16.75
16.75 – 18.85
15.85 – 20.95

Classes	center of classes	Frequency	Relative Frequency
6.25 – 8.35	7.30	1	0.022
8.35 – 10.45	9.40	1	0.022
10.45 – 12.55	11.50	12	0.267
12.55 – 14.65	13.60	8	0.178
14 65 - 16.75	15.70	11	0.244
16.75 – 18.85	17.80	9	0.200
18.85 – 20.95	19.90	3	0.067
Total		45	1.00

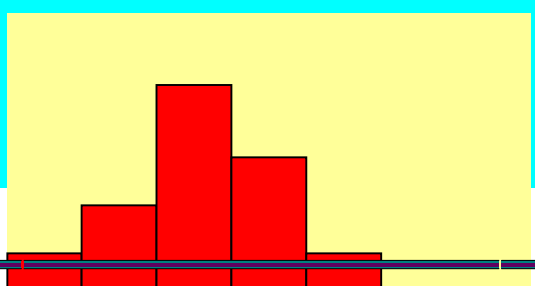


Graphical Representation of Frequency Distribution

1 - Bar Graphs

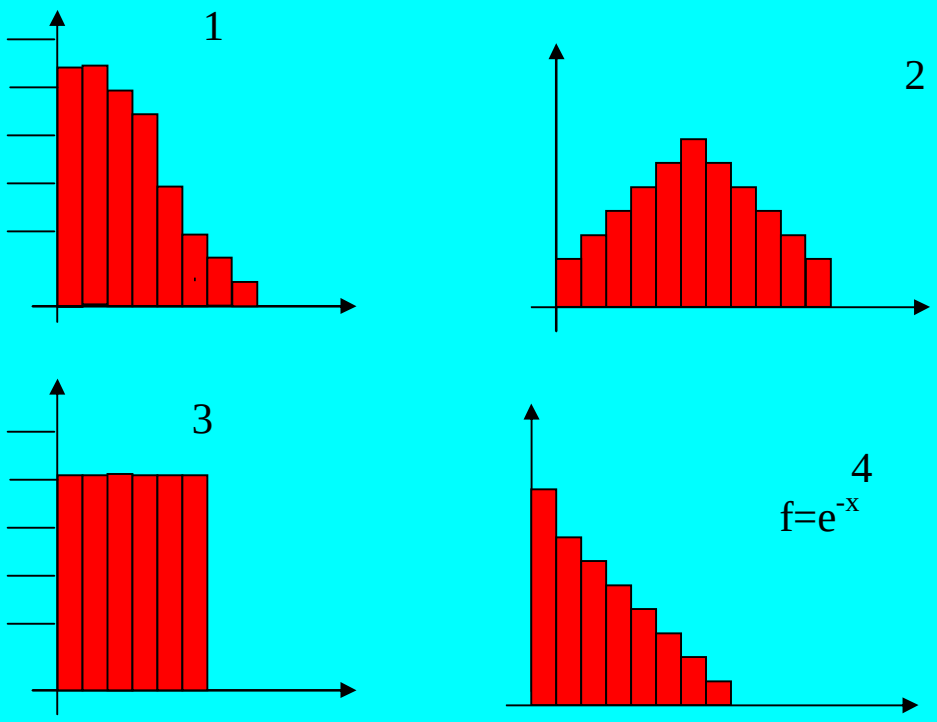


2-Frequency Histogram



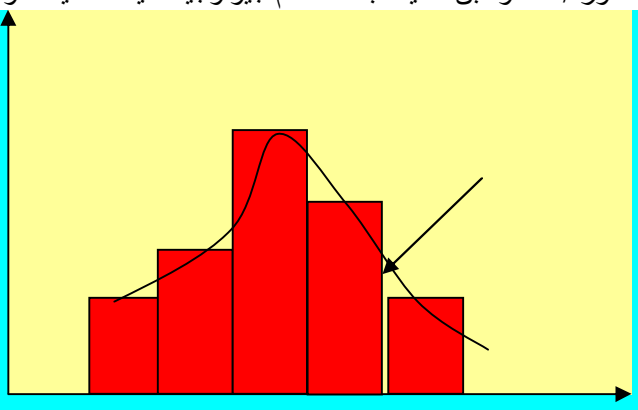
Four types of frequency histogram

1- Log normal distribution
2- Normal distribution
3- Uniform distribution
4- Exponential distribution

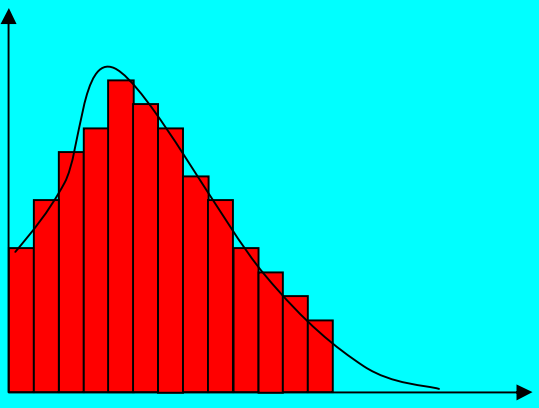


In Figure (1) is asymmetrical and the phenomena in x-direction
In Figure (2) is symmetrical
In Figure (3) is symmetrical and the phenomena is homogenous like grain-size
In Figure (4) is asymmetrical and negative relationship

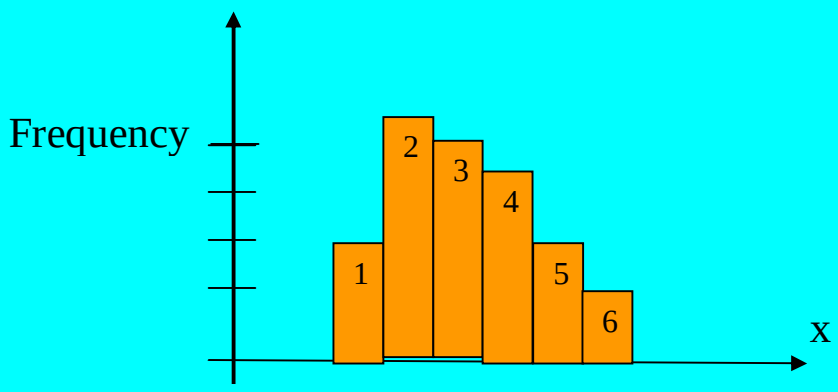
3- Frequency Polygon

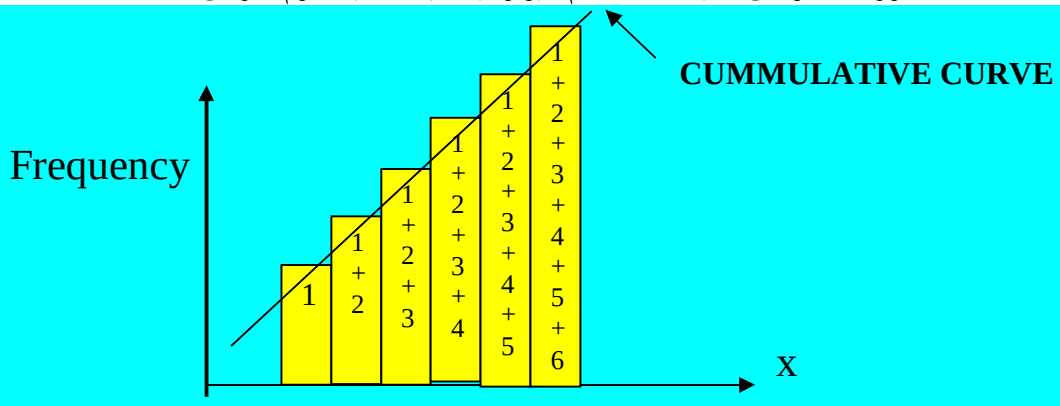


4-Frequency Curve



5- Cumulative Curve





Measures of Central Tendency

Introduction

- Arithmetic Mean ($\bar{X}$)
- Mode (M)
- Median (m)
- Weighted Mean (X_w)
- Geometrical Mean (G_m)
- Harmonic Mean (H_m)
- Moving Average

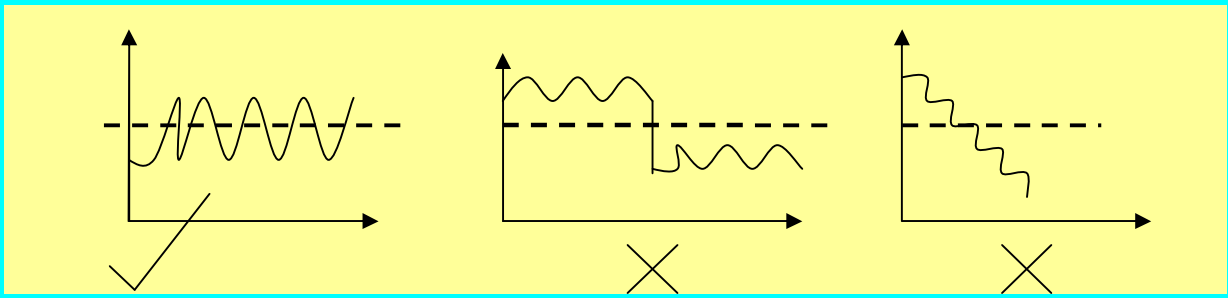
Arithmatic Mean (X)

- **Arithmetic Mean** : is obtained by summing all elements of the data set and dividing by the number of elements

x_1 , x_2 , x_3 , …….x_n

$$\bar{x} = \frac{\sum x_i}{n}$$

Which one can be used an arithmetic mean safely?



Mode (M)

Mode is the data element which occurs most frequently.

Note

Mode may not exist or it may not unique .

Example: The set 2, 2, 5, 7, 9, 9, 9, 10, 10, 11, 12

Answer:

Example: The set 3, 5, 8, 10, 12, 15, 16

Answer:

Example: The set 2, 3, 4, 4, 4, 5, 5, 7, 7, 7, 9

Answer:

Median (m)

The Median is the middle element when the data set is arranged in order of magnitude.

Example (1) The set of numbers 3, 4, 4, 5, 6, 8, 8, 8, 10

Answer:

Example (2) The set of numbers 5, 5, 7, 9, 11, 12, 15, 18
Answer: It has median

What is the relation between the Arithmetic mean, Mode and Median ?

Mean, Median, Mode

- **Positive Skew moves mean to right**
- **Negative Skew moves mean to left**
- **Normal Dist'n has mean = median = mode**
- **Median has highest prob. of occurrence**

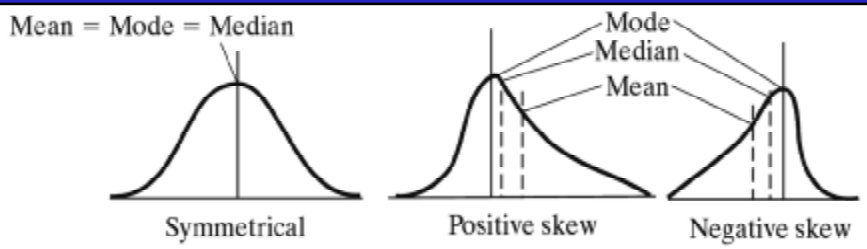


Figure 3.9
Effect of skewness on PDF and relative locations of mean, median, and mode. (From Haan, 1977, Fig. 3.3.).

Example
What is the average, Mean and Median of: 1, 1, 2, 4, 7?

Weigth Mean (X_w)

Sometimes we associate with numbers x_1 , x_2 , x_3 , x_n certain weighting factors or weight w_1 , w_2 , w_3 ,..... w_n depending on the significant or importance attached to the numbers. In this case

$$\bar{X}_w = \frac{w_1.x_1 + w_2.x_2 + \ldots\ldots. w_n.x_n}{w_1 + w_2 + \ldots\ldots w_n} = \frac{\sum wx}{\sum w}$$

Example:

A student's final marks in math, physics, English, and Arabic are 82, 86, 90 and 70 respectively. If the respective credits received for these courses are 3, 5, 3 and 1 : determine an approximate average mark.

Geometric Mean (G_m)

Garithms

t is estimated by the following equation

$$G_m = \sqrt[n]{x_1 * x_2 \ldots x_n}$$

n practice , G_m is computed by logarithms and its equal the arithmetic mean of the ogarithms of full data.

NOTE : The data should not contain negative or zero values

Harmonic Mean (H_m)

The Harmonic mean of a set of N numbers $x_1, x_2, x_3, \ldots x_N$ is the reciprocal of the arithmetic mean of the reciprocals of the numbers

$$H_m = \frac{1}{1/n \sum 1/x} = N/\sum 1/x$$

Note:

Harmonic mean cannot be calculated when the data contain zero value why?

Example

The harmonic mean of the numbers 2, 4, 8 is

Example

What is the harmonic mean of the following

2, 1, 4, 5, 8

What is the relation between the Arithmatic mean, Harmonic mean and Geometric mean?

Arithmetic mean ≥ Geometric mean ≥ Harmonic mean

Skewnss Estimation

It's a degree of symmetrical or far from symmetrical distribution of the data

$$\text{Skewnss} = \frac{\text{Mean - Mode}}{\text{Standard Deviation}}$$

PERSON LOW (1)

$$\text{Skewnss} = \frac{3\,(\text{Mean - Median})}{}$$

PERSON LOW (2)

Later we study the standard deviation, But it can simplify the two above equations and it gives the following

(Arithmetic mean – Mode) = 3(Arithmetic mean –Median)

This is the relation between the Arithmetic mean, Mode and Median

Moving Average

We define moving average of order N to be given the sequence of arithmetic means. It usually applied in the time series.

$x_1, x_2, x_3, \ldots\ldots\ldots\ldots x_n$

Degree of movements $= n - m + 1$

Given a set of numbers $x_1, x_2, x_3, \ldots\ldots\ldots\ldots x_n$

EXAMPLE

Given the numbers 2, 6 , 1, 5, 3, 7, 2 a moving average of order 3 is given by the sequence

It is customary to locate each number in the moving average at its appropriate position relative to the original data . In this example we would write

Original data	2, 6 , 1, 5, 3, 7, 2
Moving average of order 3	3, 4, 3, 5, 4

DEVIATION MEASURES

The degree to which numerical data tend to spread about an average value is called the *variation* or *dispersion* of the data.

Various measures of dispersion or variation are available, the common being the

- Range
- Mean Deviation
- Standard Deviation

(1) Range (**R**) = X_{max} - X_{min}

(2) Mean Deviation (M .D) = $$\frac{\sum |x - \bar{x}|}{N}$$

Where $\bar{x}$ *is the arithmetic mean of the numbers and* $|x - \bar{x}|$ *is the absolute value of the deviation of x from* $\bar{x}$. *The absolute value of a number without the associated sign and is indicated by two vertical lines placed around the number. Thus* $|-6| = 6$, $|+3|$

EXAMPLE

Find the mean deviation of the set of numbers 2, 3, 6, 8, 11

Answer Arithmetic mean = $\bar{X} = \frac{2 + 3 + 6 + 8 + 11}{5} = 6$

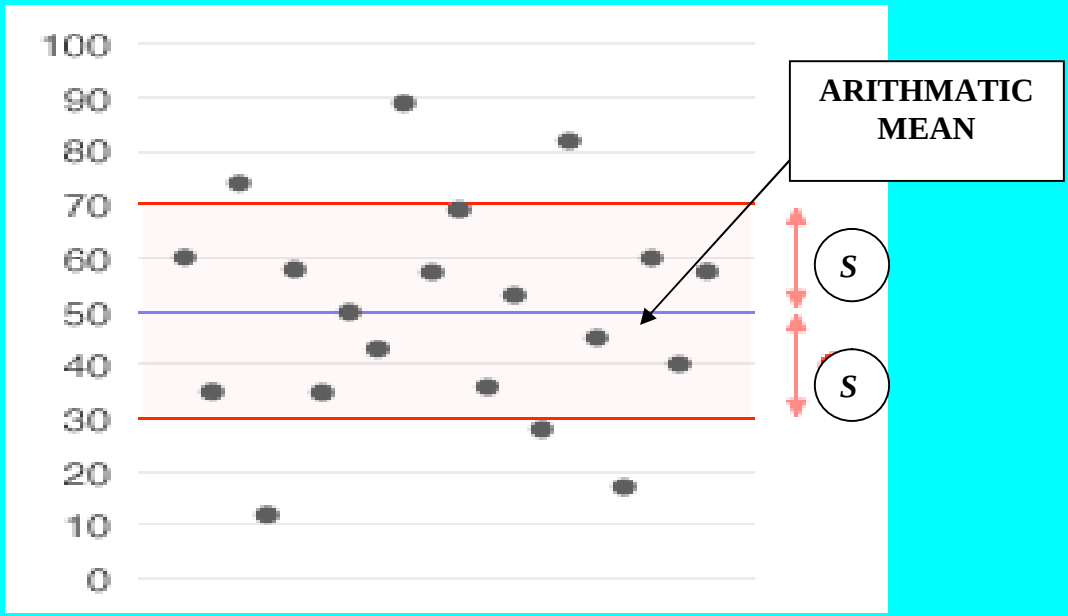
Mean Deviation (M.D) = $$\frac{|2-6| + |3-6| + |6-6| + |8-6| + |11-6|}{5}$$

$$= \frac{|-4| + |-3| + |0| + |2| + |5|}{5} = 4+3+0+2+5 = 2.8$$

(3) Standard Deviation

The standard deviation of a set of N numbers x_1 , x_2, x_3 ,..... x_n is denoted by (s) and is defined by :

$$S = \sqrt{\frac{\sum_{k=1}^{n} (x_k - \bar{x})^2}{n-1}}$$



VARIANCE

It is defined as the square of the standard deviation and is thus given by S^2

$$S^2 = \frac{\sum_{k=1}^{n} (x_k - \bar{x})^2}{n-1}$$

EXAMPLE

Given the numbers, 6, 2, 5, 3 fine the standard deviation and variance.

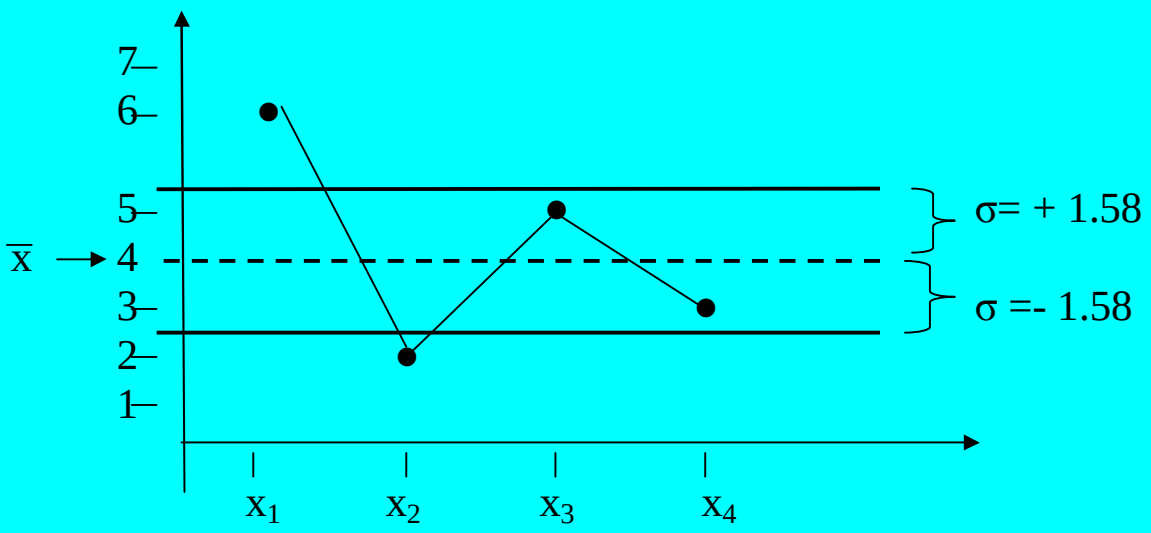
Answer:

- Find the mean $(2 + 6 + 5 + 3)/4 = 4$ and then

$\Sigma (x_i - x)^2 = 4, 4, 1, 1 = 10$

$S = \pm 1.58$ (standard deviation)

$S^2 = 2.5$ (variance)



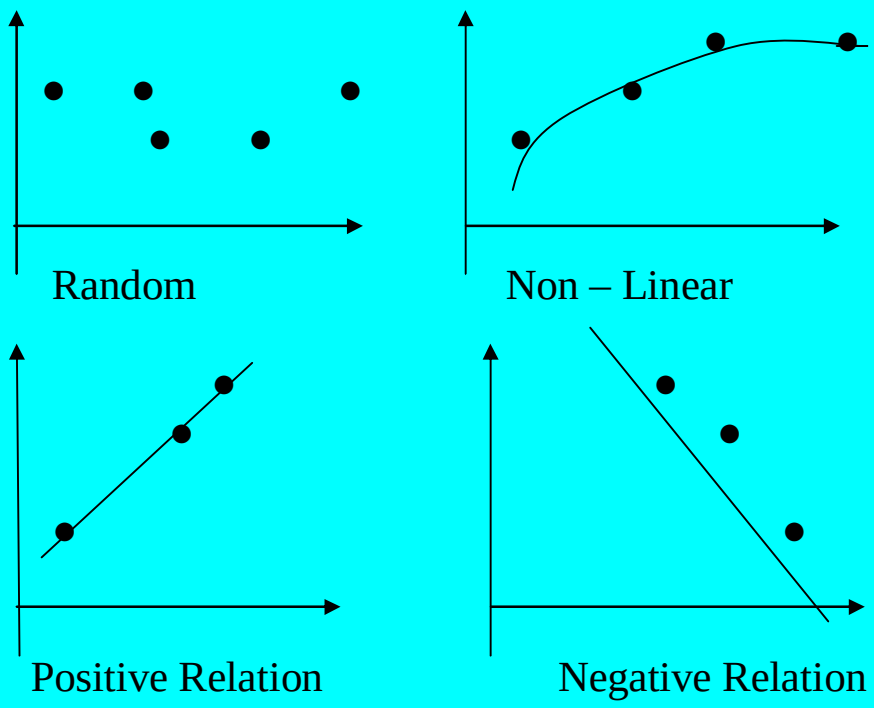
CORRELATION & REGRESSION

- **Simple Linear Regression**

- X (Independent Variable)
- Y (Dependant Variable)

$(x_1, y_1), (x_2, y_2)\ldots\ldots\ldots\ldots(x_n, y_n)$

Scatter Types



Straight line equation

$$y = a + b\,x$$

(a) is the interception where the straight line cross Y
(b) is the Straight line slope

To determine the trend of the line and the (a and b) it is usually used

- ***Least Squares Method***
- ***Normal Equations***

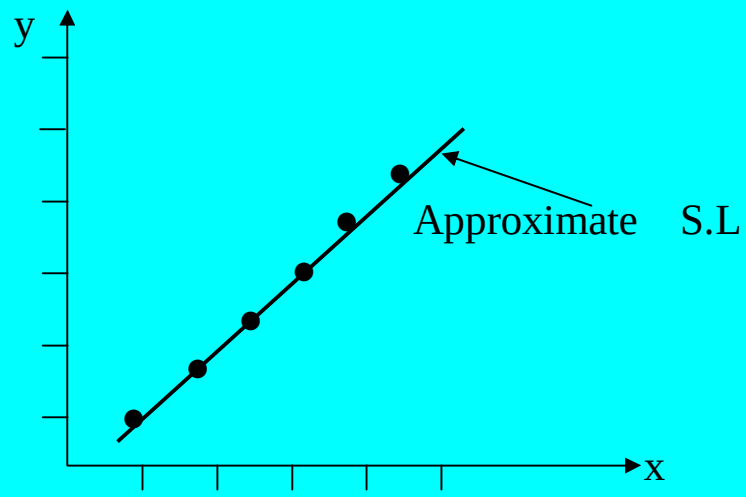
Least Squares Method

Example

Find the straight line equation for the following data

Answer

Y	X
1	2
3	3
7	5
11	7
15	9
17	10



$$Y = a + bx$$

1 = a + b (2) -----(1)

3 = a + b (3) - -----(2)) Complete ……..

(a) =-3

(b) = 2

Straight line equation= y =-3 + 2x

Normal Equations

To find (a) and (b) use the following equation

$$\Sigma y = a N + b \Sigma x$$

$$\Sigma x.y = a \Sigma x + b \Sigma x^2$$ **Where (N) number of data used**

Example

The values of *X* and *Y* were tabulated below

X	Y	X^2	Y^2	$X \cdot Y$
65	68			
63	66			
67	68			
64	65			
68	69			
62	66			
70	68			
66	65			
68	71			
67	67			
69	68			
71	70			
$\Sigma x =$	$\Sigma y =$	$\Sigma x^2 =$	$\Sigma y^2 =$	$\Sigma x.y =$

$$\Sigma y = a N + b \Sigma x$$

$$\Sigma x.y = a \Sigma x + b \Sigma x^2$$

$a = ?$

$b = ?$

Correlation Coefficient (r)

To find the correlation coefficient:

- calculate the variance for Y of the following example

x	y	x^2	x.y
1 2	1 2		

3 4 5 6 7	6 7 10 16 21		
Σ x =	**Σ y =** **Average ($\bar{y}$)=**	**Σ x^2 =**	**Σ x. y =**

$$S^2 = \frac{\Sigma (y - \bar{y})^2}{n} = 45.71$$

Find the S.L equation using the least squares method to calculate (a) and (b) -

a = 4.16 , b = 3.29

$$Y = 4.16 + 3.29\,X$$

X	Y	$\check{Y} = 4.16 + 3.29\,X$	$Y - \check{Y}$	$(Y - \check{Y})^2$
1	1	- 0.87		
2	2	2.42		
3	6	5.71		
4	7	9.0		
5	10	12.29		
6	16	15.58		
7	21	18.87		
				Σ 17.72

- Find the new variance

$$S^2_{\check{y}} = 17.72 / 7 = 2.53$$

- Calculate the correlation coefficient

$$r = \frac{S^2 - S^2_{\check{y}}}{S^2}$$

$$r = 1 - \frac{S^2_{\check{y}}}{S^2}$$

$$r = \pm \sqrt{1 - \frac{S^2_{\check{y}}}{S^2}}$$

$$-1 \geq r \leq +1$$

Time Series Analysis

WHAT IS A TIME SERIES

A time series is a collection of observations of well-defined data items obtained through repeated measurements over time. For example, measuring the value of retail sales each month of the year would comprise a time series

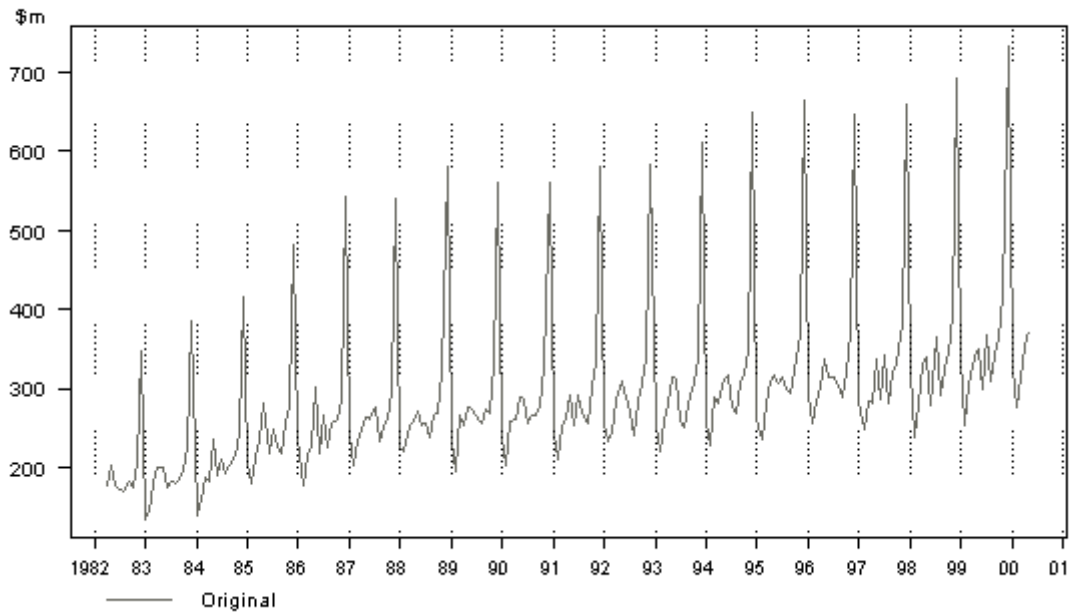


Figure: Monthly Retail Sales in New South Wales (NSW) Retail Department Stores

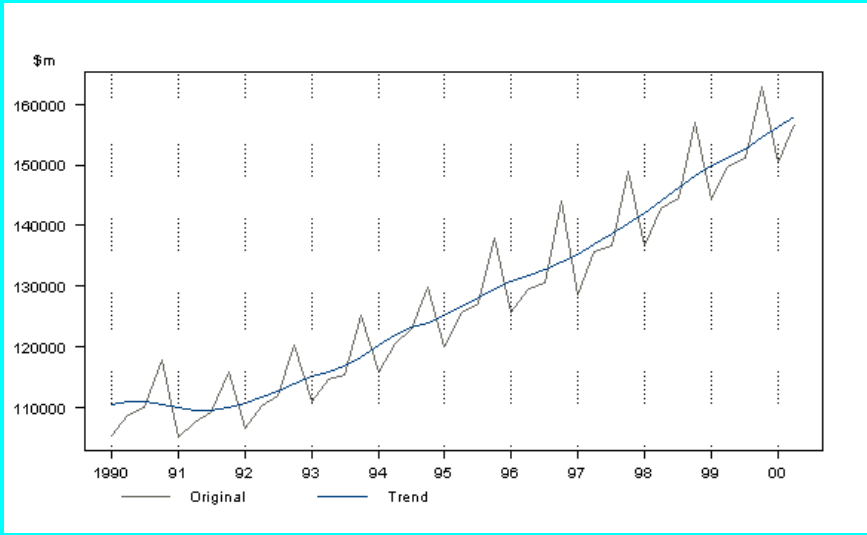


Figure: Quarterly Gross Domestic Product

CHARACTERISTICS MOVEMENTS OF TIME SERIES التحركات المميزة في السلاسل الزمنية

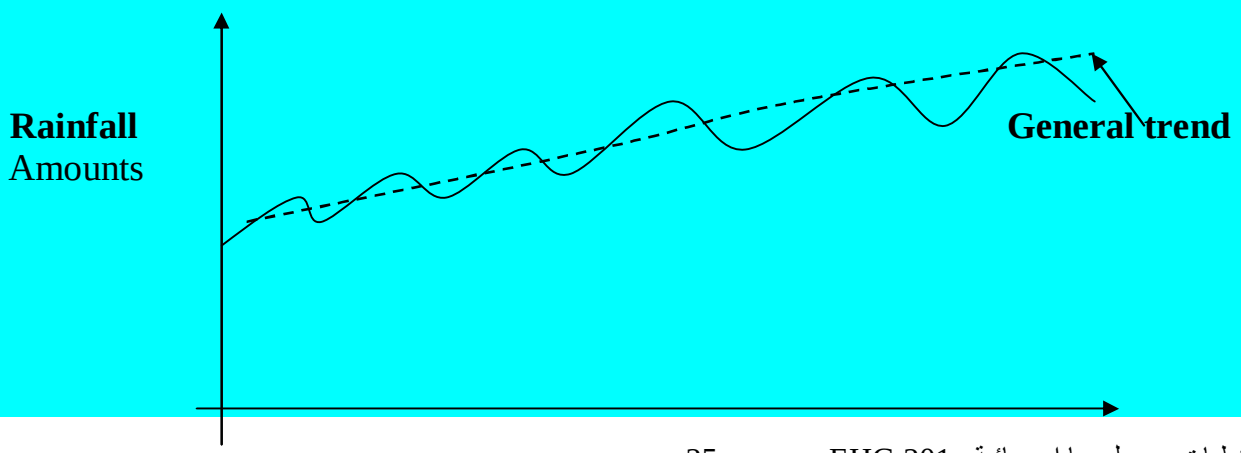
تكشف السلاسل الزمنية عن وجود تحركات او اختلافات مميزة في بعضها او كلها بدرجات مختلفة. وتحليل مثل هذه التحركات لها اهمية كبرى في كثير من الاستخدامات ومنها التنبؤ بالتحركات المستقبلية.

CLASSFICATION OF TIME SERIES MOVEZMENTS تصنيف التحركات في السلاسل الزمنية

يمكن تصنيف التحركات في السلاسل الزمنية الى **اربعة** انماط وتسمى غالبا **مكونات السلسلة الزمنية**

١) التحركات طويلة المدى (الاتجاه العام) General Trend

وتشير الى الاتجاه العام الذي يظهر به الشكل البياني للسلسلة الزمنية على مدى فترة طويلة من الزمن



Years

(٢) تحركات دورية او تغيرات دورية Cyclical Variations

تشير الى التذبذوب طويل المدى حول منحنى خط الاتجاه العام وهي تحركات اذا تكررت بعد فترات زمنية تزيد عن السنة

(٣) التحركات الموسمية Seasonal Variations

وهي تشير الى النمط المتماثل لحركة السلسلة الزمنية في الاشهر المتقابلة خلال سنوات متتالية

(٤) تحركات منتظمة او عشوائية Regular or Random Variations

تشير الى الحركة المنتظمة في السلسلة مثل الفيضانات او الحركة العشوائية مثل الحمم البركانية والهزات الزلزالية

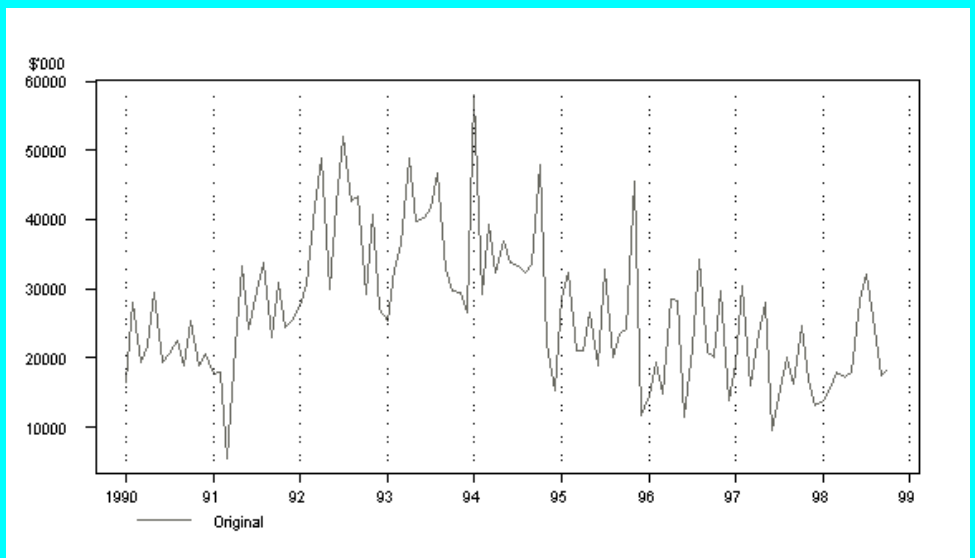
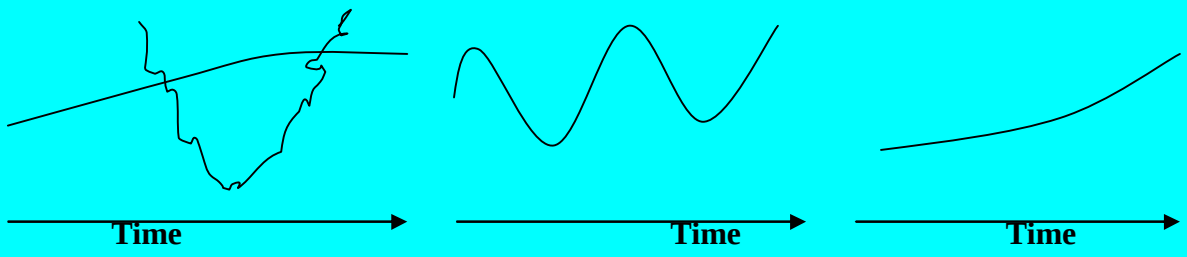


Figure Monthly Value of Building Approvals, Australian Capital Territory (ACT)



Long term trend and cyclical and seasonal movements

Long term trend and cyclical movements

Long term trend

تحليل السلاسل الزمنية

نفترض ان المتغير Y هو حاصل ضرب T, P, S, I

حيث ان

$$Y = T \times P \times S \times I$$

T الاتجاه العام

P التحركات الدورية

S التحركات الموسمية

I التحركات العشوائية او الغير منتظمة

وهذا يشير الى تفكيك السلسلة الزمنية الى المكونات الاساسية لتحركاتها ويمكن تقدير هذه المتغيرات على النحو التالي

١ - الاتجاه العام

يمكن تقديره بعدة طرق

- طريقة المربعات الصغرى
- التمهيد باليد
- طريقة المتوسط المتحرك

٢ - التغيرات الموسمية (الدليل الموسمي)

يحدد المعامل الموسمي بتقدير التغيرات في البيانات في السلاسل الزمنية من شهر الى شهرخلال سنة نموذجية والقيم النسبية التي توضح التغيرات خلال اشهر السنة تسمى

(الدليل الموسمي) او الوسط الحسابي للمتغير خلال كل شهر من اشهر السنة

٣ - المتغيرات الدورية

يمكن تعديل البيانات من الاثر الموسمي والاتجاه العام عن طريق قسمة البيانات ببساطة على القيم الاتجاهية المقابلة و التغيرات الموسمية والتي منها التغيرات الدورية والغير منتظمة

$$Y = T \times S \times C \times I$$

$$\frac{Y}{T \times S} = C \times I$$

٤- التغيرات العشوائية

يمكن تقديرها باستبعاد المتغيرات السابقة وذلك بقسمة البيانات الاصلية على T, S, C

$$\frac{Y}{T \times S \times C} = I$$

مثال توضيحي

الجدول التالي يوضح كميات شراء القمح شهريا بالاطنان في السنوات مابين ٢٠٠٠ وحتى عام ٢٠٠٧م .

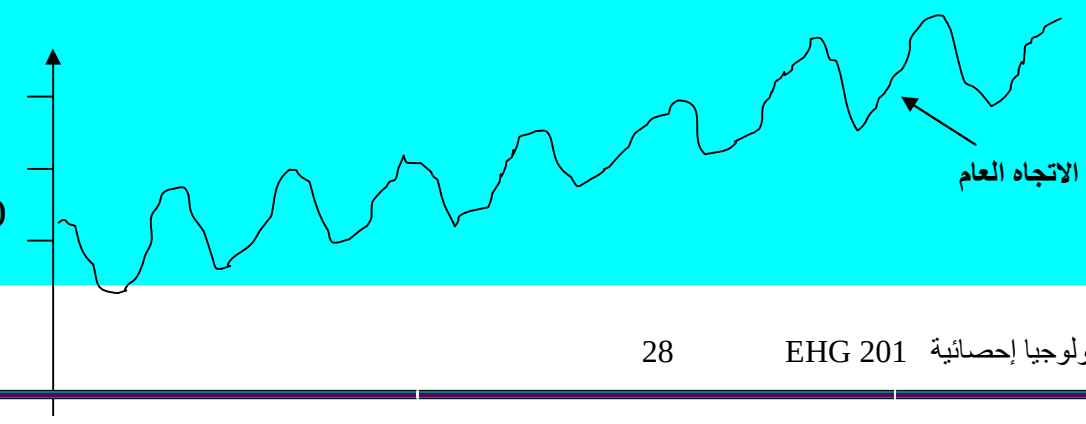
ا- كون الشكل البياني لهذه البيانات

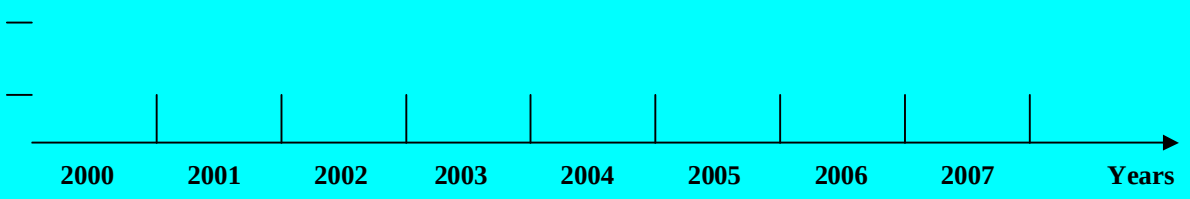
ب- اوجد الدليل الموسمي مستخدما طريقة متوسط النسب المئوية

YEARS	Jan	Feb	Mar	Apri	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec	Total	Mean
2000	318	281	278	250	231	216	223	245	269	302	325	347	3285	274
2001	342	309	299	268	249	236	242	262	288	321	342	364	3522	294
2002	367	328	320	287	269	251	259	284	309	345	367	394	3780	315
2003	392	349	342	311	290	273	282	305	328	364	389	417	4042	337
2004	420	378	370	334	314	296	305	330	356	396	422	452	4373	364
2005	453	412	398	362	341	322	335	359	392	427	454	483	4738	395
2006	487	440	429	393	370	347	357	388	415	457	491	516	5090	424
2007	529	477	463	423	398	380	389	419	448	493	526	560	5505	459

يوضح الجدول المجاميع والمتوسطات الشهرية للسنوات 2007 – 2000

أ) الشكل البياني





السنوات

YEARS	2000	2001	2002	2003	2004	2005	2006	2007
TOTAL	3285	3522	3780	4042	4373	4738	5090	5505
MEAN	273.7	293.5	315.0	336.8	364.4	394.8	424.2	458

ب) بقسمة البينات الشهرية على متوسطات الشهرية لكل سنة مع التعبير لكل نتيجة كنسبة مئوية وكمثال 318/274= 116.2 %
ينتج الجدول التالي :

YEARS	Jan	Feb	Mar	Apri	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec
2000	116.2	102.7	101.6	91.3	84.4	78.9	81.5	89.5	98.3	110.3	118.7	126.8
2001	116.5	105.3	101.9	91.3	84.8	80.4	82.5	89.3	98.1	109.4	116.5	124.0
2002	116.5	104.1	101.6	91.1	85.4	79.7	82.2	90.2	98.1	109.5	116.5	125.1
2003	116.4	103.6	101.5	92.3	86.1	81.1	83.7	90.6	97.4	108.1	116.5	123.8
2004	115.3	103.7	101.5	91.7	86.2	81.2	83.7	90.6	97.7	108.7	115.8	124.0
2005	114.7	104.4	100.8	91.7	86.4	81.6	84.9	90.9	99.3	108.2	115.0	122.3
2006	114.8	103.7	101.1	92.6	87.2	81.8	84.2	91.5	97.8	107.7	115.7	121.6
2007	115.3	104.0	100.9	92.2	86.8	82.8	84.8	91.3	97.7	107.5	114.7	122.1
Total	925.7	831.5	810.9	734.2	687.3	647.5	667.5	723.9	784.4	809.4	928.4	989.7
MEAN	115.7	103.9	101.4	91.8	85.9	80.9	83.4	90.5	9.1	108.7	116.1	123.7

متوسط النسبة المئوية لكل شهر معطى في السطر الاخير . مجموع هذه النسب المئوية هي % 1200.1
وهي فريبة من المجموع % 1200 ولهذا فأن المتوسط في الجدول تعبر عن الدليل الموسمي.