Chapter 2
Graphs, Charts and Tables describing Data

Chapter Goals:

After completing this chapter, you should be able to:

- Construct a frequency distribution both manually and with a computer.
- Construct and interpret a histogram.
- Create and interpret bar chart, pie chart and stem – and – leaf diagrams.
- Present and interpret data in line chart and scatter diagrams.

2.1: Frequency Distribution and Histogram

A *frequency diatribution* is a list or table and it is a way to summarize a large set of data. It is often useful to classify the data into classes or categories and to determine the number of individuals belonging to each class, called the *class frequency*. A tabular arrangement of data by classes together with the corresponding frequencies is called a *frequency distribution* or simply a frequency table. Consider the following problems:

What is a frequency distribution?

1. It is a list or a *table*.
2. containing the *values* of a variable (or a set of ranges within which the data falls)
3. and the corresponding *frequencies* with which each value occure (or frequencies with which data falls within each range).
4. It is used for both types of data discrete and continuous.

Why to use frequency distributions?

1. A frequency distribution is a way to summarize data (*present*).
2. The distribution *condenses* the raw data in a more useful form.
3. and allows for a quick visual interpretation of the data.

Consider the following definitions:

Class Width: The difference between the upper and lower class limits of a given class.

Frequency: The number of observations in a class.

Relative Frequency: The ratio of the frequency of a class to the total number of observations in the data set. Or the proportion of each category.

It is a way to compare between two tables have the same variables with different total number of observation.

$$\text{rf} = \frac{\text{Freq. of i}^{\text{th}}\text{ class}}{\text{Total number of observations}} = \frac{f_i}{\sum_{i=1}^{k} f_i} = \frac{f_i}{n}$$

Cumulative Frequency: The total frequency of all values less than or equal the upper class limit. Or it is the total frequency that less than or equal to the upper class limit.

***Relative Cumulative Frequency*:** The cumulative frequency divided by the total frequency.

Frequency Distribution for Discrete Data:

Example: Assume that an advertiser asks 200 customers how many days per week they read the daily newspaper, then we can summarize the data as follows:

Number of days read	**Freq. f_i**	**Relative freq. *rf***
0	44	0.22
1	24	0.12
2	18	0.09
3	16	0.08
4	20	0.10
5	22	0.11
6	26	0.13
7	30	0.15
Total	200	1.00

The meaning of *rf*:

0.22: 22% of the people in the sample report that they do not read the newspaper.

0.08: 8% of the people in the sample report that they read the newspaper 3 days per week.

0.15: 15% of the people in the sample report that they read the newspaper daily.

Frequency Distribution for Continuous Data:

Example: A manufacturer of insulations randomly selects 20 winter days and records the daily high temperature

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

The steps needed to prepare a frequency distribution for the data set are described below:

Step 1: Write the data in an array (the data have been sorted in ascending or descending order)

Step 2: Find Range = Largest observation – Smallest observation

$R = 58 - 12 = 46$.

Step 3: Divide the range between into classes of equal width. A rule of thumb for the number of classes is $k = \sqrt{n}$.

$$\text{Class width} \approx \frac{\text{Range}}{\text{Number of classes}} = \frac{R}{k}$$

Since we have a sample of size 20, the number of classes in the frequency distribution should be around $\sqrt{20} = 4.47 \approx 5$. In this case, the class width would be approximately $46/5 = 9.2 \approx 10$. The smallest observation is 12. The first class boundary may well start at 12 or little below it, say at 10 (just to avoid the smallest observation, in general, falling on the class boundary). Thus the first class is given by (10, 20]. The second class is given by (20, 30]. Complete the class boundaries for all classes.

Step 4 For each class, compute the midpoint (*class mark*), denoted by (x_i)

$$\text{Midpoint}(x_i) = \frac{\text{Upper endpoint + Lower endpoint}}{2}$$

Step 5 For each class, count the number of observations that fall in that class. This number is called the class frequency.

Step 6: Find the relative frequency of a class is calculated by *f/n*. The cumulative Relative Frequency of a class, denoted by *c.r.f*, The resulting quantity Relative Cumulative Frequency (*r.f/n*) is just the same as Cumulative Relative Frequency and is desirable in a frequency table.

Notes:

1. The classes **MUST** be *mutually exclusive* (the classes contains different values, so each value must be in *only* one class).
2. Also, it **MUST** be *exhaustive* (the classes contain all possible data values).
3. In some statistical packages, the lower boundary of the first class is called the starting point while the class width is called the step size.

For the data, we have the following frequency distribution:

Class	Count	*f*	*midpoint* x_i	*c.f*	*r. f*	*r.c.f*
(10, 20]	///	3	15	3	0.15	0.15
(20, 30]	///// /	7	25	10	0.35	0.50
(30, 40]	/////	4	35	14	0.20	0.70
(40, 50]	////	4	45	18	0.20	0.90
(50, 60]	//	2	55	20	0.10	1.00
Total		20			1.00	

Or you can construct closed intervals from the left:

Class	Count	f	*midpoint* x_i	cf	$r.f$	$r.c.f$
$[10, 20)$	///	3	15	3	0.15	0.15
$[20, 30)$	///// /	6	25	9	0.30	0.45
$[30, 40)$	/////	5	35	14	0.25	0.70
$[40, 50)$	////	4	45	18	0.20	0.90
$[50, 60)$	//	2	55	20	0.10	1.00
Total		20			1.00	

Example: Consider the following data

216	202	208	208	212	202	193	208	206	206
206	213	204	204	204	218	204	198	207	218
204	212	212	205	203	196	216	200	215	202

Construct a frequecny distribution

<u>Histograms:</u> (For GFT's ONLY)

A frequency histogram is a bar diagram where a bar against a class represents the frequency of the class. The classes (intervals) are shown on the horizontal axis, and the frequencies are measured on the vertical axis.

From the histogram, we can get the following informations:

1. It gives an indication where is the approximate *center* of the data.
2. It gives an indication of the degree of spread or *variation*.
3. It gives an idea about the *shape* of the data.

Example: A manufacturer of insulation randomly selects 20 winter days and records the daily high temperature

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

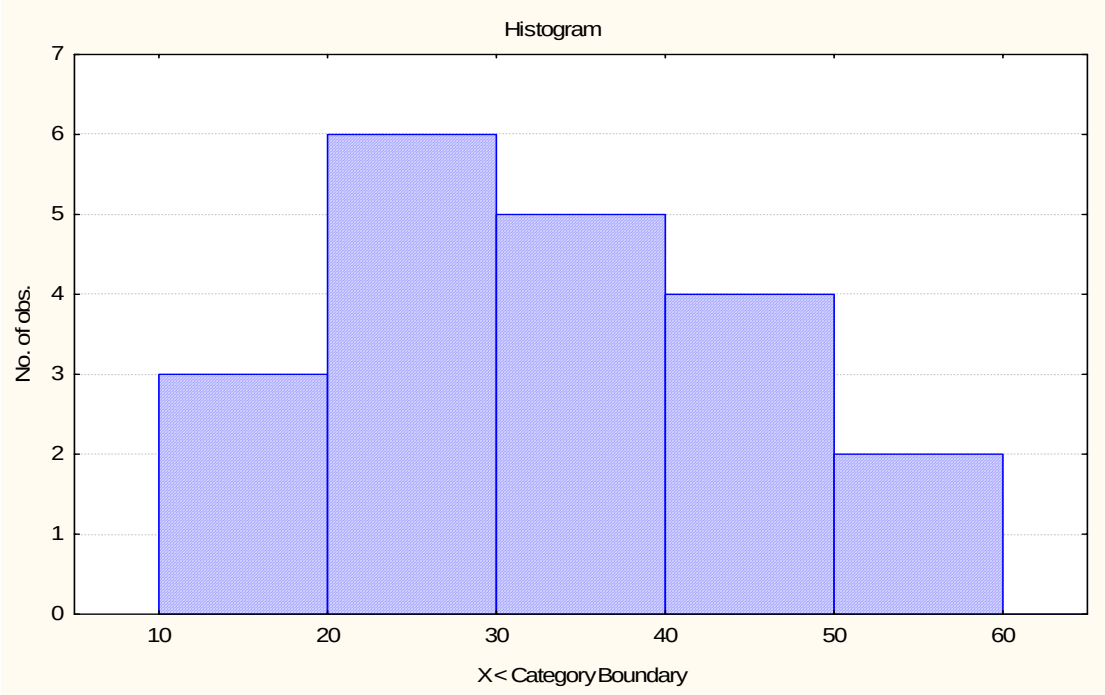
The procedure to construct a frequency histogram for the data set follows:

Step 1: Construct a Grouped Frequency Distribution. (from the example above).

Step 2: Draw the x – y axes.

Step 3: Put the end points of each class on the x – axis and the frequencies on the y – axis.

Step 4: Draw bars (without gaps) between each two endpoints with heights equal to frequency for each class.



Notes:

1. Replacing the frequencies by the *r.f*, we get a *relative freq.* histogram.
2. Replacing the frequencies by the *c.r.f*, we get an *Ogive*.
3. An ogive is construted by connecting the points, that correspond to the class limits, by line segments to form some *polygon*.

Graph Description

The graph can be described in terms of *modality* and *skewness*:

1. Modality: The number of modes or *peaks* in the graph.
2. Skewness: The direction to which the graph has a *tail*.

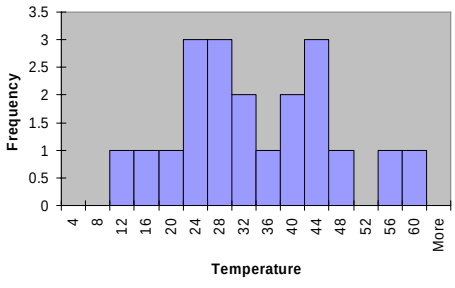
Example: Consider the following data

216	202	208	208	212	202	193	208	206	206
206	213	204	204	204	218	204	198	207	218
204	212	212	205	203	196	216	200	215	202

Construct a frequecny histogram, a relative frequency histogram and an Ogive.

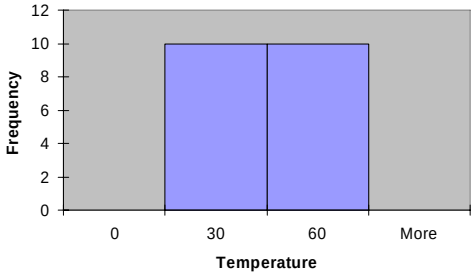
Question: How wide should each interval be? (How many classes should be used?)

1. Many (Narrow class intervals)
   a. May yield a very jagged distribution with gaps of empty classes.
   b. Can give a poor indication of how



frequency varies across classes.

2. Few (wide class intervals)
   a. May cpmpress variation too much and yield a blocky distribution
   b. Can obscure important patterns of variation.



General Guidelines:

1. Class widths can typically be reduced as the number of observations increases.
2. Distributions with numerous observations are more likely to be smooth and have gaps filled since data are plentiful.

Number of data points	Number of classes
Under 50	5 – 7
50 – 100	6 – 10
100 – 250	7 – 12
Over 250	10 – 20

Question: How should the endpoints of each class be determined?

1. Often answered by trial and error, subject to user judgment.
2. The goal is to create a distribution that is neither too "jagged" nor too "blocky".
3. The goal is to appropriately show the pattern of the variation in the data.

2.2: Bar Chart, Pie Chart and Stem and Leaf Diagram

Part 1: Stem and Leaf Diagram:

It is a way to doing preliminary analysis for quantitive data. It is, like the histogram, a way to see distribution details in a data set.

The Method:

Separate the sorted data into leading digits (*stems*) and the trailing digits (*leaves*).

Example: Given the following data

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

Construct a stem and leaf diagram.

Solution:

To construct a stem and leaf diagram for quantitive data use the following steps:

1. **Sort** the data in order array (from min to max).
2. Determine **how you wish to split** the values into stems and leaves.
    - stem = first digit (tens place), leaf = last digit (units place)
3. List all possible **stems in a single column** from the lowest to the highest.
4. For each stem, list all leaves associated with the stem **horizontally**.

Stem	leaves					
1	2	3	7			
2	1	4	4	6	7	7
3	0	2	5	7	8	
4	1	3	4	6		
5	3	8				

Example: Given the following data. Construct a stem and leaf diagram

216	202	208	208	212	202	193	208	206	206
206	213	204	204	204	218	204	198	207	218
204	212	212	205	203	196	216	200	215	202

Example: A sample of n = 25 CPU Times (in seconds) is selected from 1000 CPU times.

1.17	1.61	1.16	1.38	3.53	1.23	3.76	1.94	0.96
4.75	0.15	2.41	0.71	0.02	1.59	0.19	0.82	0.47
2.16	2.01	0.92	0.75	2.59	3.07	1.40		

Construct a Stem and Leaf Plot of the data.

Do Questions 24 and 25 page 55

Part 2: Graphing Categorical Data:

1. **Bar chart:**

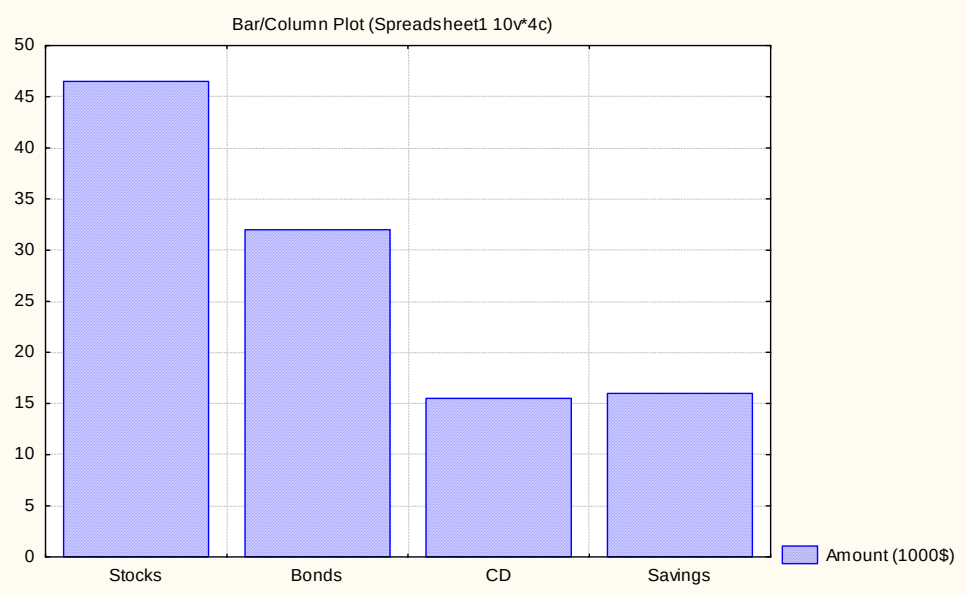
   It is used, mainly, for qualitative data (categorical data), where the height of each bar is the frequency or the percentage for each category, it may be horizontal or vertical.

Example: Consider the following table (*Investment type is a qualitative variable*)

Investment type	*Amount (1000$)*	*Percentage %*	*Cum. Perc. (Pareto)*
Stocks	46.5	*42.27*	**42.27**
Bonds	32.0	*29.09*	**71.36**
CD	15.5	*14.09*	**85.45**
Savings	16.0	*14.55*	**100**
Total	**110**	***100***	

Construct a bar chart

Solution:



Do Questions 20, 28 and 30 page 55 and 56

2. **Pie chart:**

   It is used, mainly, for qualitative data (categorical data), where the size of pie slice is the *percentage* for each category.

Example: Use the data in the previous example to construct a pie chart

Solution:

1. Find the *proportion* (percentage) for each category

$$\%\ \text{of stocks} = \frac{46.5}{110} \times 100\% = 42.27\%$$

$$\%\ \text{of bonds} = \frac{32.0}{110} \times 100\% = 29.09\%$$

$$\%\ \text{of CD} = \frac{15.5}{110} \times 100\% = 14.09\%$$

$$\%\ \text{of savings} = \frac{16.0}{110} \times 100\% = 14.55\%$$

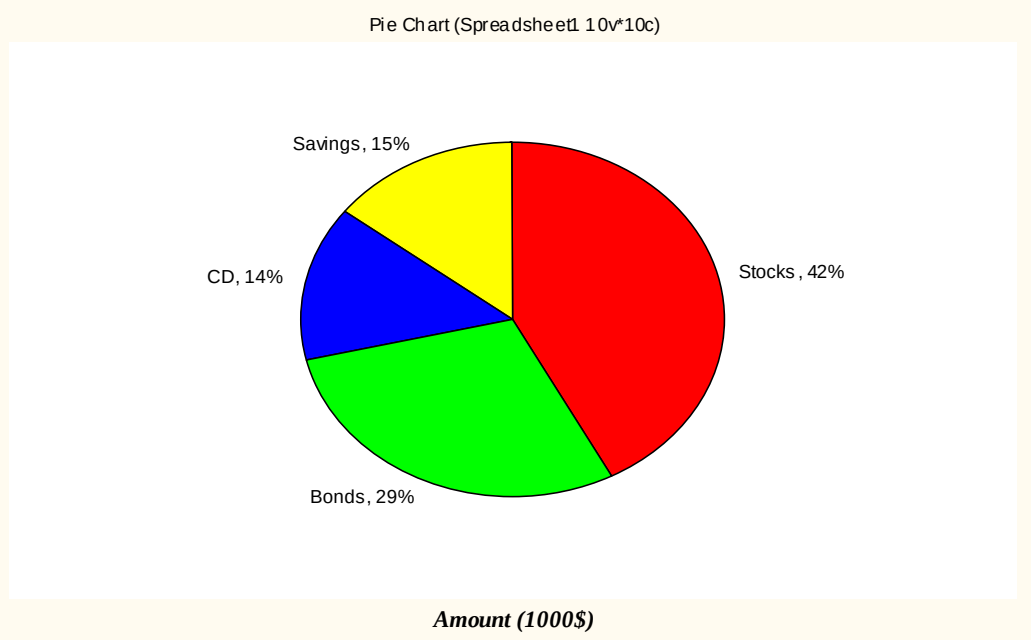
2. Find the angle for each category = the proportion $\times 360^0$

$$\measuredangle \text{of stocks} = \frac{46.5}{110} \times 360^\circ = 152.172^\circ$$

$$\measuredangle \text{of bonds} = \frac{32.0}{110} \times 360^\circ = 104.724^\circ$$

$$\measuredangle \text{of CD} = \frac{15.5}{110} \times 360^\circ = 50.724^\circ$$

$$\measuredangle \text{ of savings} = \frac{16.0}{110} \times 360^\circ = 52.38^\circ$$

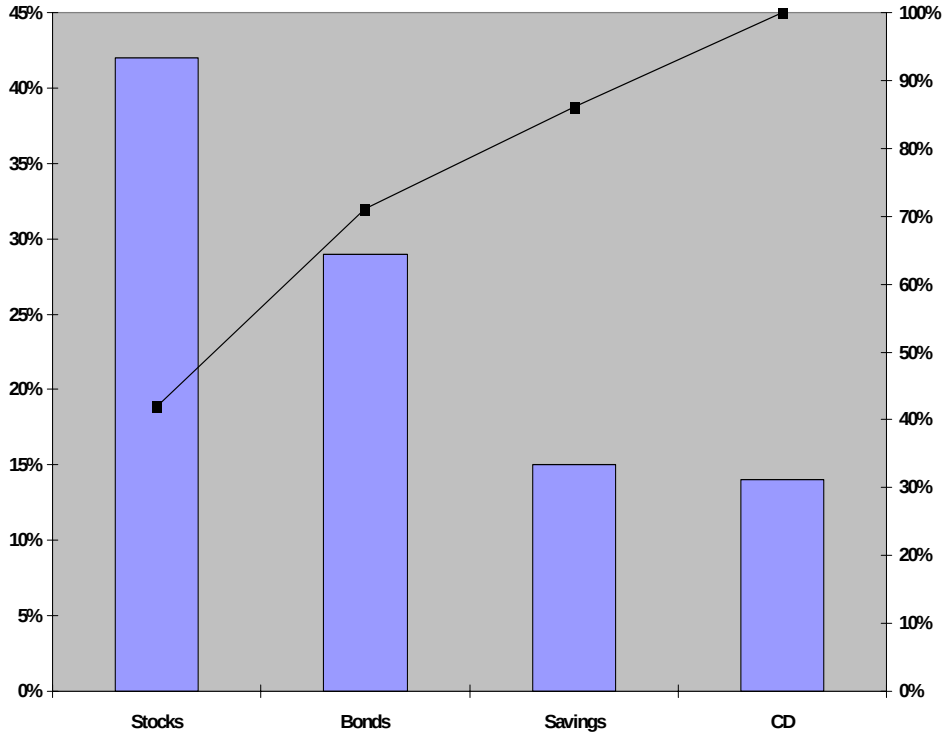


Do Questions 20, 28 and 30 page 55 and 56

3. **Pareto diagram:**

It is very similar to the ogive, but instead of the cum. relative frequencies we use the *percentage* in ascending or descending orders.

Example: Use the data in the previous example to construct a pareto diagram



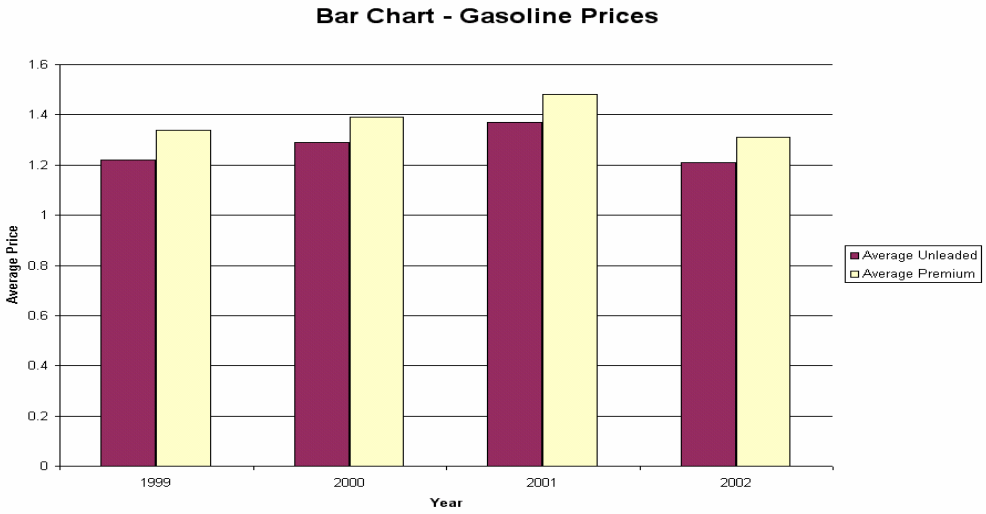
Do Questions 20, 28 and 30 page 55 and 56

Part 3: ***Graphing Multivariate Categorical Data:***

Some time, there are two or more variables that need to be graphed on the same bar chart.

Example: Q 2.21/p.55:

There are variables that need to be graphed on the same bar chart. Categories will be the years. The following is a possible bar chart.



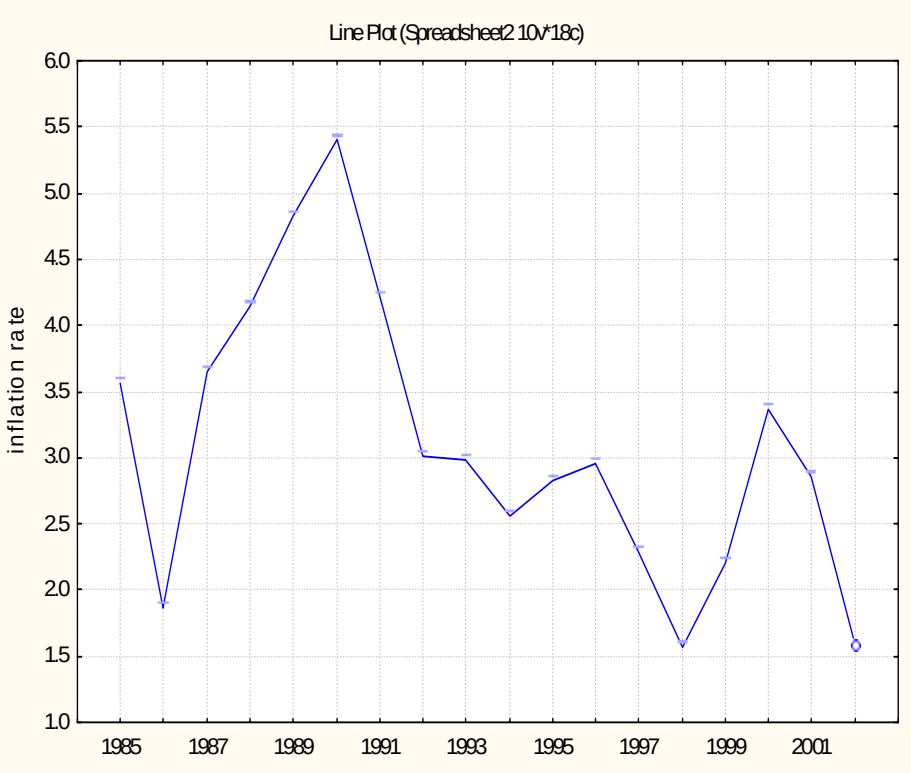
2.3: Line Charts and Scatter Diagram

1. The line charts:

It shows values of one variable (*on the vertical axis*) vs. the time (*on the horixontal axis*), and it is used for time series data to see the trend of the variable across the time

Example: Consider the following table, construct a line chart

Year	*Inflation rate (%)*
1985	3.56
1986	1.86
1987	3.65
1988	4.14
1989	4.82
1990	5.4
1991	4.21
1992	3.01
1993	2.99
1994	2.56
1995	2.83
1996	2.95
1997	2.29
1998	1.56
1999	2.21
2000	3.36
2001	2.85
2002	1.58



Do Q39 page 64

2. The scatter diagram

The scatter diagram: it shows points for bivariate data to see if there is a relationship between the two variables or not. One variable is measured on the vertical axis and the other variable is measured on the horizontal axis.

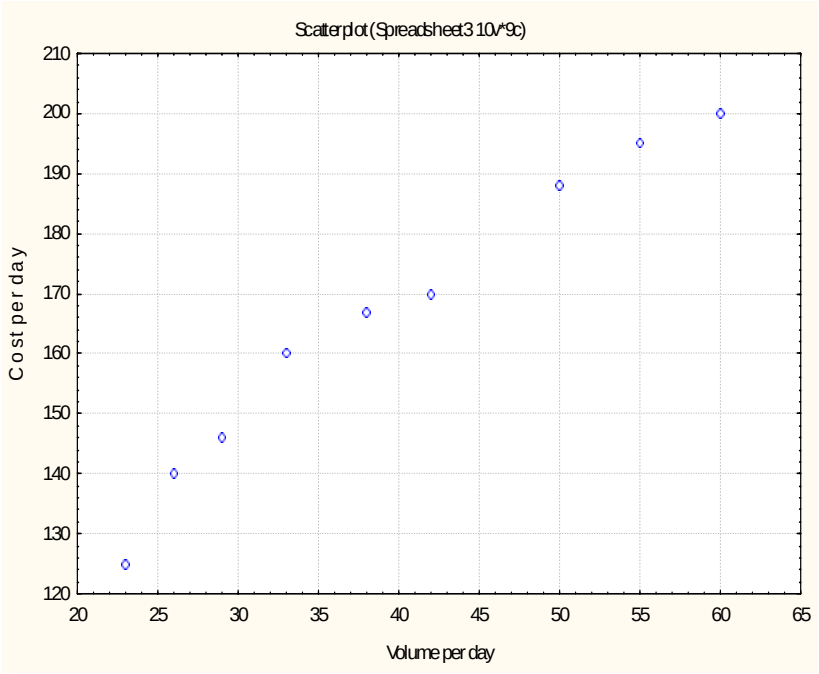
The *dependent* variable **(y)** (*response*): is a set values depend on the *x* variable.
The independent variable **(x)** (*explanatory,* or *predictor*): it is a free variable (fully controlled) and it impacts (affects) the values of the dependent variable.

Dependent (y): is the variable whose variation we wish to explain.

Independent (x): is the variable used to explain the variation in the dependent variable.

Example: Consider the following table, construct a line chart

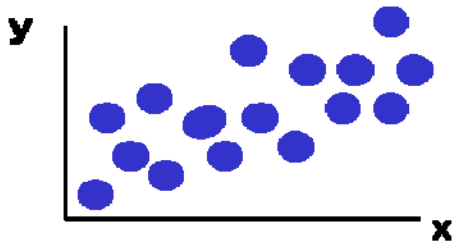
Volume per day (x)	***Cost per day (y)***
23	125
26	140
29	146
33	160
38	167
42	170
50	188
55	195
60	200



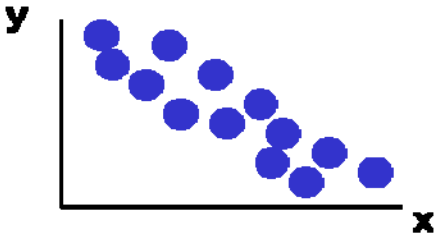
Do Q41 page 64

Types of Relationships:
Linear relationships:

a)



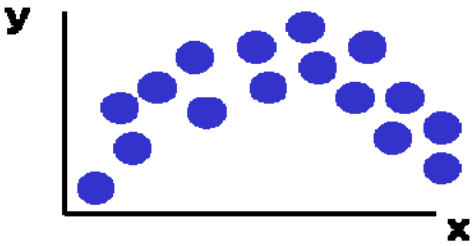
b)



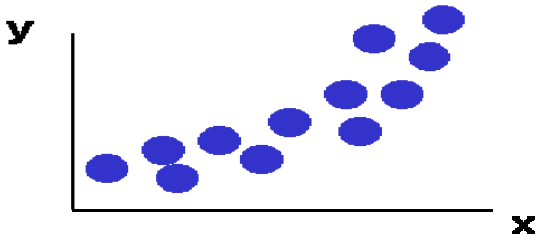
(a) and (b) show a linear relationship between x and y.
In a) as x increase, y also increase (positive or direct relationship)
In b) as x increase, y decrease (negative or inverse relationship)

Curvilinear relationships

c)



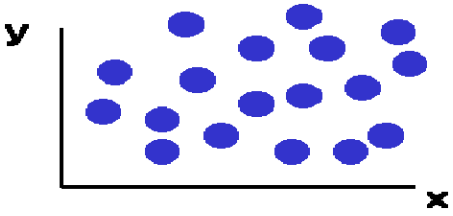
d)



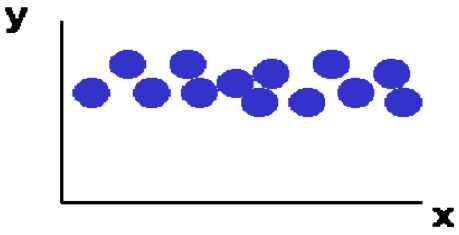
(c) and (d) show that there is no linear relationship between x and y. it may be nonlinear relation between the two variables y and x.

No relationship

e)

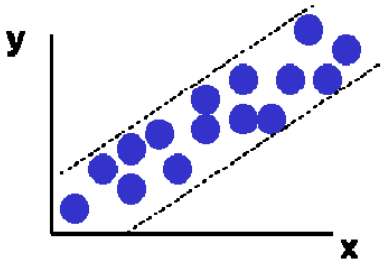


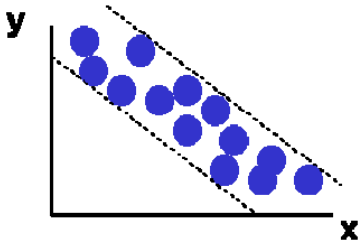
f)



In e) and f): No identifiable relationship between the two variables y and x. This means that as x increases, y sometime increase and some time decrease but with no particular pattern.

Strong relationships





Weak relationships