STAT 2361
STATISTICS FOR BUSINESS AND ECCONOMICS
STAT 2311
STATISTICS 1
LECTURE NOTES

PREPEARED BY MOHAMMAD MADIAH

CHAPTER 2
Descriptive Statistics
Tabular and Graphical Methods

Organization of Data

Always studies result in data represented by a large collection of numbers (**raw data**). So we need to **organize** data for the purpose of analyzing and interpreting. Usually, we organize data using a table called a **frequency distribution.**

A frequency distribution for qualitative data lists all categories (names labels) and the number of elements in each category (frequency). The following example shows a frequency distribution for a qualitative data.

Example 1

Twenty five persons were asked to taste two types of soft drinks, A and B, and indicate the taste of A was superior (S), the same (M), or inferior (I) to that of B. The responses are listed below.

S	I	I	M	M	S	M	M	S	M	S	I	I
M	M	S	I	I	M	S	M	S	M	S	M	

Solution

The variable in the problem has three **categories**: superior (S), the same (M) and inferior (I). We record these categories in the first column of the following table and frequencies in the second column.

Response	Frequency	Relative Frequency	Percentage
S	8	$\frac{8}{25}=0.32$	32%
M	11	$\frac{11}{25}=0.44$	44%
I	6	$\frac{6}{25}=0.24$	24%
Total	25	$\frac{25}{25}=1$	100%

The **relative frequency** of a category is defined by,

$$\text{Relative frequency} = \frac{\text{category frequency}}{\text{total frequencies}}.$$

Graphical Presentation of Qualitative data

1. Bar Graph

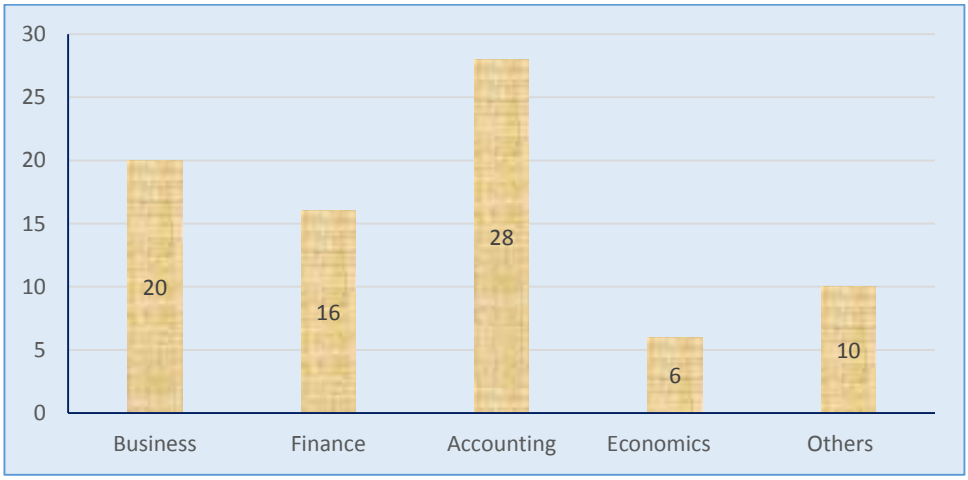
A bar graph is a graph made of **bars** whose heights represent the frequencies (or relative frequencies) of respective categories.

Example 2

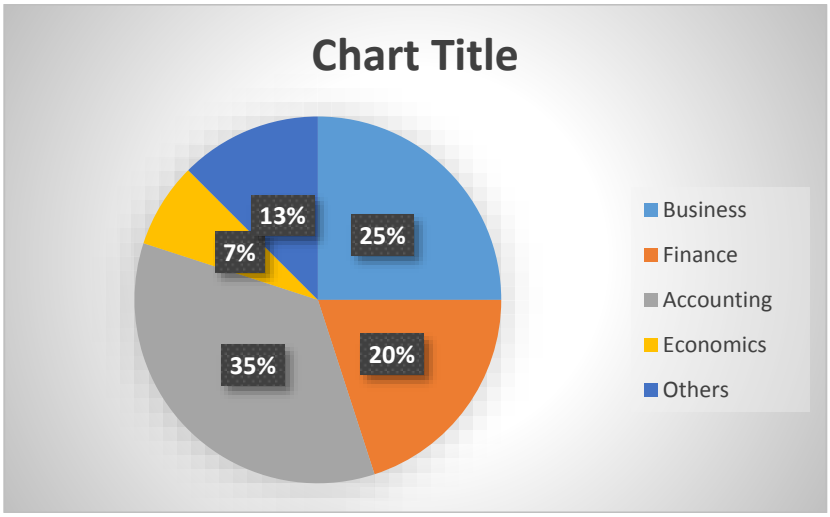
Consider the following frequency distribution for a data obtained from a sample of BZU students.

Major	Number of students
Business	20
Finance	16
Accounting	28
Economics	6
Others	10
Total	80

To construct the bar graph, mark the various categories on the horizontal axis, and frequencies on the vertical axis. Then draw one bar for each category such that the height of the bar represents the frequency of the corresponding category.



2. **Pie Chart:** A circle divided into sectors that represent the relative frequencies or percentages of the categories.



Organizing and Graphing Quantitative Data

Numerical data sets containing a large collection of numbers often arranged into **groups** or **classes**. A **frequency distribution** is grouping the data into disjoint classes showing the number of observations in each (the frequency). The frequency distribution has two columns, class's column and frequencies column.

To construct a frequency distribution:

1. Determine the class interval or **width**. The approximate class width **W** (for a given number of classes) is given by:

$$W = \frac{\text{largest data value} - \text{smallest data value}}{\text{number of classes}}$$

Usually this approximate class width is rounded to a convenient number (next integer), which is then used as the class width. Notice that rounding this number may slightly change the number of classes initially intended.

2. Determine the **limits** of each class. Each class in a frequency distribution has two limits: the **lower** class limit **L** and the **upper** class limit **U**. If W is the class width, then;

U = L + W – 1

3. Count the number of elements in each class.

When constructing a frequency distribution make sure that:

1. Each item of the data is fit into one and only one class (The first class must contain the smallest data value, and the last one must contain the largest data value).
2. All classes are with the same width (this is an optional choice).
3. The classes are mutually exclusive. That is there is no intersection between any two classes.
4. Use from 5 to 12 classes (this is an optional choice).

Example 3

The following data represents the results of a statistics test

32	55	95	89	76	78	81	45	67	91
81	86	73	74	83	60	71	74	90	51
43	31	95	89	76	62	94	65	45	75

Construct a frequency distribution with 7 classes

Solution

Since the largest data value is 95 and the smallest data value is 31, and the number of classes is 7, the class width is;

$$w = \frac{95 - 31}{7} = 9.14$$

Round this number to the next integer 10, that is, the class width is 10.
The first class must contain 31. So, we should start with the data value 31 or any value less 31. Let's choose the number 30 to be the lower limit for the first class. So, the upper limit is 30 +10 – 1 = 39.
The first class is 30 – 39; the second is 40 – 49, and so on. Now, count the number of scores in each class. See the following table:

Class	Frequency
30 – 39	2
40 – 49	3
50 – 59	2
60 – 69	4
70 – 79	8
80 – 89	6
90 – 99	5
Total	30

Representing Quantitative data by Graphs

Histogram

A histogram is a graph in which either class limits or class boundaries (true limits) are marked on a horizontal axis, and frequencies, relative frequencies, or percentages are marked on the vertical axis. To construct a histogram:

- Find class boundaries of each class (Subtract 0.5 from the lower limit and add 0.5 to the upper limit of each class).
- Construct a set of adjoining rectangles (bars) having as a base the width of class and as height the frequency of each class.

Cumulative Distributions

The cumulative frequency distribution uses the number of classes, class widths, and class limits developed for the frequency distribution. However, rather than showing the frequency of each class, the cumulative frequency distribution shows the number of data items with values less than or equal to the upper class limit of each class.

A curve that represents the cumulative frequency distribution of grouped data on a graph is called a Cumulative Frequency Curve or an Ogive.
An ogive graph plots **cumulative frequency** on the y-axis and **class upper limit** along the x-axis

Example 4

The following frequency distribution shows the daily expenditure of a sample of BZU students.

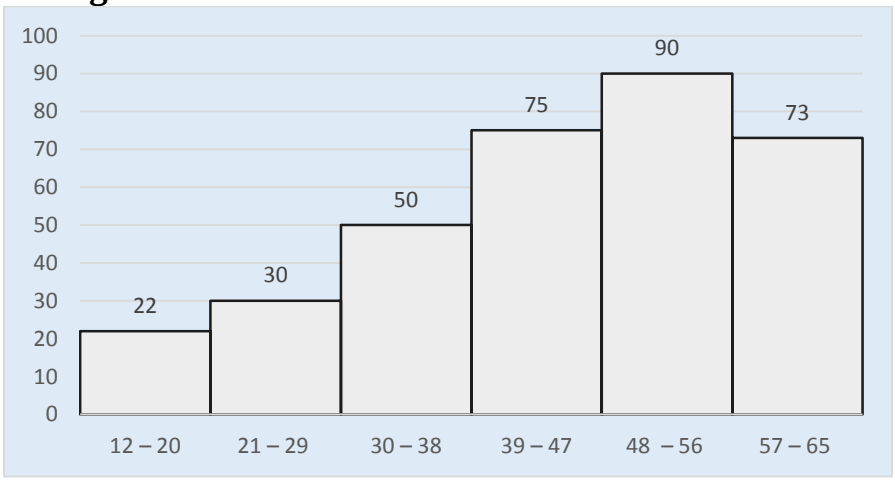
Class	Frequency
12 – 20	22
21 – 29	30
30 – 38	50
39 – 47	75
48 – 56	90
57 – 65	73
Total	

Answer the following questions

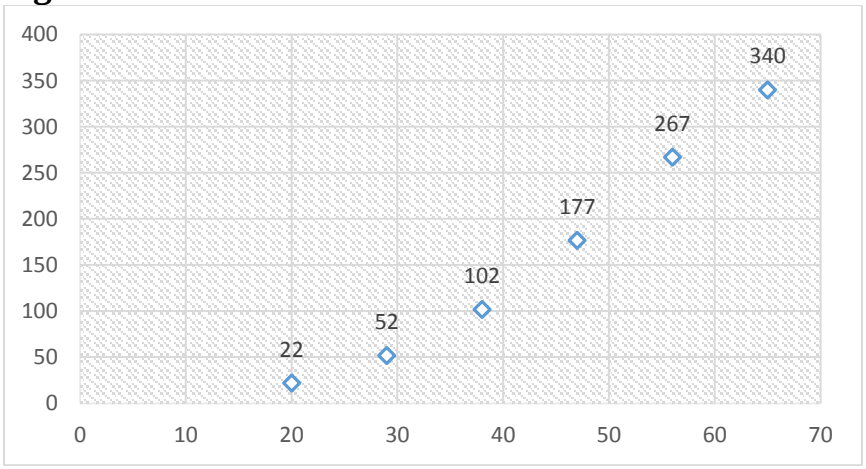
1. What is the sample size?
   Total number of elements = 340
2. What is the class width?
   w = U – L + 1 = 22 – 12 +1 =
3. Add a column that shows the frequency distribution/percentage of each class.
   Relative frequency = (class frequency)/ Total frequencies
4. Add a column that shows the midpoint of each class.
   Midpoint = (upper limit + lower limit)/ 2
5. Add a column that shows the true limits (boundaries) of each class.
   True limits of a class: add 0.5 to the upper limit and subtract 0.5 from the lower limit (of each class).
6. Construct a histogram for the above data. Comment on the shape of the distribution.
7. Construct a cumulative frequency distribution and an Ogive for the above data.

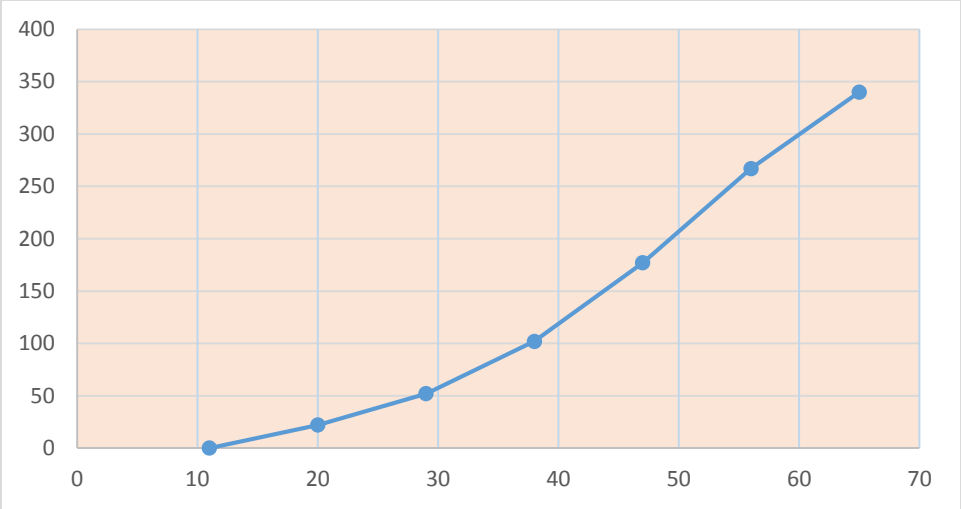
Class	Frequency	R.F / %	True limits	Midpoint	Cumulative frequency Relative C. F
12 – 20	22	22/340 6.5%	11.5 – 20.5	(12 + 20) /2 = 16	**22** 22/340 = 6. 5%
21 – 29	30	30/340 8.8%	20.5 – 29.5	25	22 + 30 = **52** 52/340 = 15.3%
30 – 38	50			34	52 + 50 = **102**
39 – 47	75			43	102 + 75 = **177**
48 – 56	90			52	177 + 90 = **267**
57 – 65	73	73/340 21.47%		61	267 + 73 = **340** 340/340 = 100%
Total	**340**	**100**			

Histogram



Ogive





One of the most important uses of a histogram is to provide information about the **shape**, or form, of a distribution.

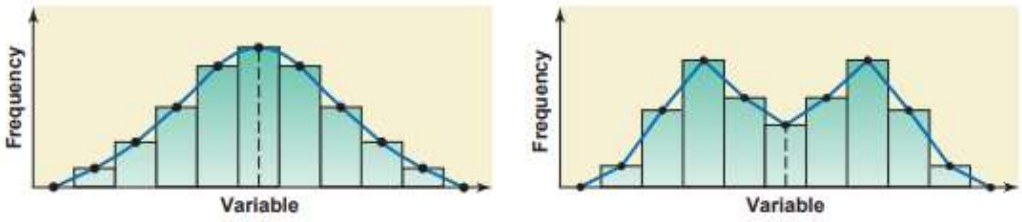


Figure 2.8 Symmetric histograms.

A **skewed histogram** is nonsymmetric. For a skewed histogram, the tail on one side is longer than the tail on the other side. A **skewed-to-the-right histogram** has a longer tail on the right side (see Figure 2.9*a*). A **skewed-to-the-left histogram** has a longer tail on the left side (see Figure 2.9*b*).

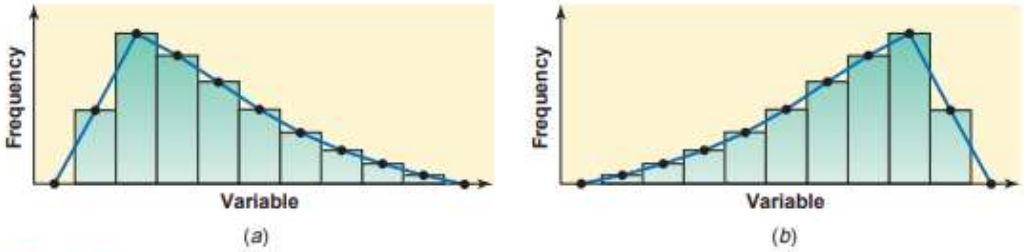


Figure 2.9 (*a*) A histogram skewed to the right. (*b*) A histogram skewed to the left.

2.4 Crosstabulations and Scatter Diagrams

A cross-tabulation (or crosstab) is a **tabular summary** for two variables (either qualitative or quantitative). The classes for one variable are represented by the rows, the classes of the other variable are represented by the columns. Cross-tabulation tables provide a wealth of information about the **relationship** between the variables.

Example 1

a researcher is interewsted in whether there is a relationship between gender and salary at her company. She randomly sampled some men and women on the company and asked them about their salaries in ILS. Her results appear in the table below.

Salary (ILS)	**Gender**		
	Men	Women	Total
1450 – 5000	120	80	200
5001 – 7500	140	100	240
7501 – 10000	30	15	45
More than 10000	10	5	15
Total	300	200	**500**

- Variables:
  - Gender, qualitative, ordinal
  - Salary, quantitative, ratio
- Number of elements (sample size): 500

1. What is the sample size?
   Total number of elements = 500 employees.
2. What is the number of employees with salaries more than 7500 ILS?
   Number = 45 + 15 = 60
3. What is the percentage of women employees?
   Percentage of women = (200/500) (100%) = 40%
4. What is the relative frequency of employees with salaries more than 10000 ILS?
   Relative frequency = 15/500 = 0.03.
5. Among the men employees, what is the percentage of those with salaries between 7501 – 10000?
   Percentage = 30 / 300 = 0.1*100%= 10%.

6. Among the employees with salaries less than or equal to 7500 ILS, what is the percentage of women employees?
   Percentage = (100 + 80)/ (240 +200) = 180 / 440 = 40.91%

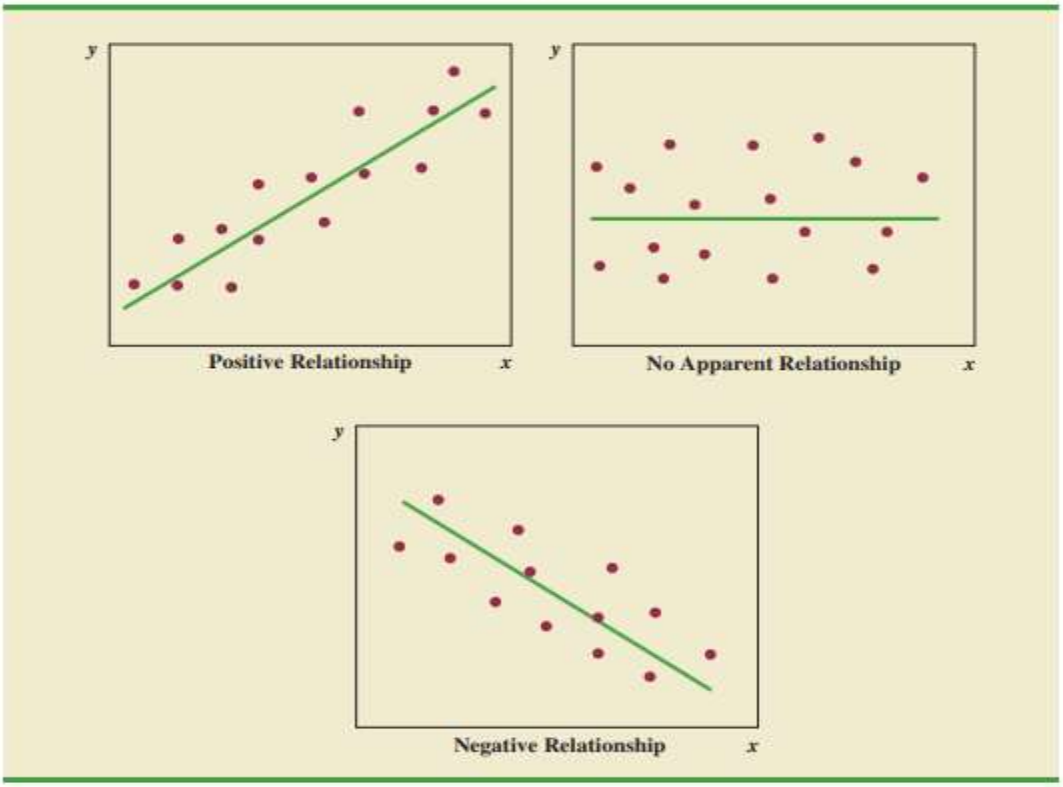
Scatter Diagrams and trend lines.

Crosstabulation is a method that can be used to summarize the data for two variables and help reveal the relationship between the variables. In most cases, a graphical display is more useful for recognizing **patterns** and **trends** in the data.

A scatter diagram (plot) is a graphical display of the relationship between two quantitative (numerical variables x (**independent**), y (**dependent**) variables, and a trend line is a line that provides an approximation of the relationship.
Let x and y be two numerical variables. Select a sample of size n, then we have the following observations for the two variables:
$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$

The first step to get some idea about possible relationship between two variables is constructing a scatter diagram. It is **an indicator** "about" the relation between the two variables.
To construct a scatter diagram we graph pairs of numerical data, with one variable on each axis. (Independent == Horizontal. Dependent == Vertical).

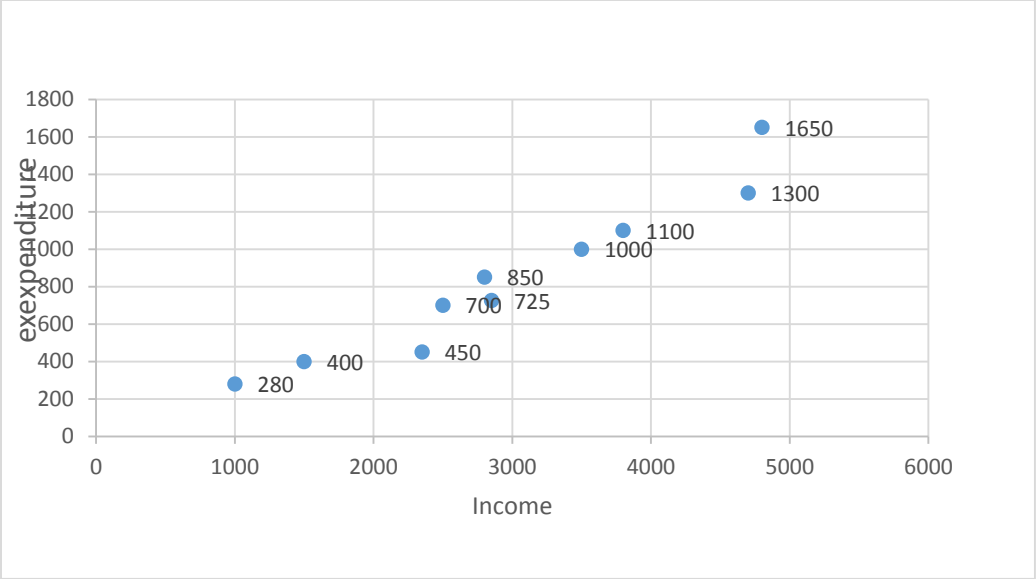
FIGURE 2.8 TYPES OF RELATIONSHIPS DEPICTED BY SCATTER DIAGRAMS



Example 2

The income and food expenditures of ten households is given in the following table. Construct a scatter diagram for this distribution.

Household	Income X	Food expenditure Y
1.	3500	1000
2.	4700	1300
3.	2500	700
4.	3800	1100
5.	1500	400
6.	2800	850
7.	2350	450
8.	4800	1650
9.	1000	280
10.	2850	725



There is a **positive linear relationship** between the income and the daily expenditure