Displaying Data

Foundations of medical statistics, 05-09 may 2007, University Diabetes Center, King Abdulaziz Uinversity Hospital, King Saud University, Riyadh, Saudi Arabia, 5-9 May, 2007

By
Dr. Khalaf S. Sultan

Department of Statistics and Operations Research
Faculty of Science
King Saud Ubiversity
P.O.Box 2455, Riyadh 11451

ksultan@ksu.edu.sa

Outline

- Basic Concepts
- Statistical data
- Types of Data
- Displaying Continuous Data
- Displaying Categorical Data

Basic Concepts and Definitions

1- **Statistics:**

Statistics is that area of the sciences that teach us how to collect, organize and summarize information about the underlying population. In addition, how we can draw inference about the parameters of the population.

2- **Population:**

A population is the largest group of people or things in which we are interested at a particular time and about which we want to make some statement or conclusions.

3- **Sampling:**

Sampling is a statistical technique enables us to collect information from the underlying population. This information is called a sample. The

number of elements (or individuals) in the sample is called sample size. The collected sample by using the proper sampling techniques is called random sample.

4- **Variables:**

The characteristics to be measured on the population or in the random sample are called variables. The random samples are possible values for the variables.

5- **Descriptive Statistics** [Displaying data – Basic measures]

6- **Statistical Inference** [Point Estimation – Interval Estimation – Hypotheses Testing]

7- **Nonparametric Statistics** [nonparametric methods without any assumption about the distribution of the underlying populations]

8- **Applications** [Regression - Survival Analysis – Reliability – Goodness of fit tests - Time Series]

Statistical Data

Let us start with the following scenarios:

- **Scenario 1:** Suppose that the Dean of the Faculty of Medicine phoned King Abd AlAziz hospital and wanted to know, within an hour, the average time for the cancer patients operated at the hospital. How can you give him the required information?

- **Scenario 2:** Suppose a researcher wants to know the average of weights of a full grown Saudi male who suffer from diabetes? How you can give him reasonable answer to his question?

Statistical study of any problem should depend on some representative data concerning the problem under consideration.

How can we get such data?

Data may be obtained in several ways. It may obtained from

- Available historical records, books, published research papers, etc.
- Designed experiments which are run in the laboratory by the researcher himself or his staff.

- Public opinion surveys through interviews, or questionnaires.

- Computer through simulation of an experiment or a given distribution.

Types of Data

Data may be classified into the following classes:

- Quantitative data: It is the data that takes numerical values and may be classified into two classes:

  - Continuous: Its is the data which can take all possible numerical values in a given interval, e.g. height, temperature, …

  - Discrete data: Its is the data which may take only a countable numbers of distinct numerical values in a

given rage, e.g. number of patients in a certain clinic, ….

- Qualitative data (Categorical data): It is the data which takes non-numerical categories or classes e.g. sex, with categories male and female.

- Interval data: It is the data in which the difference has a meaning but the ratio is meaningless, e.g. the cumulative average of a student.

- Ratio data: It is the data in which the difference has a meaningful order, difference and ratio, e.g. the number of years the student has been in the university.

- Univariate data: It is the data which gives only one measurement on each individual, e.g. age.

- Multivariate data: It is the data which gives two or more measurements on each individual, e.g. age, sex, weight, height, …

Displaying Continuous Data

Given a set of observation will not help to draw conclusion from it. How can we present this set of data by table or a graph to visualize what is going on and to see the main characteristics of the data. When there is a large amount of data, it is not easy to get conclusion from the raw data. So there is a need to summaries the data. This may be done through tables, graphs and some statistics.

Tables

Frequency Table of continuous data

To construct a frequency table of equal class of continuous data recorded to the nearest natural number, apply the following steps:

- Select a suitable number of classes k . It may be the ceiling of $k = \sqrt{n}$, where n is the sample size. Sometimes, you may be able to guess the number of classes from the underlying application. In this cases there is no need to use $k = \sqrt{n}$.
- Find the min. and the max. of the data.

- Find the range of the data $R = \max - \min$.
- Find the class length (assuming equal length) $L = \left\lceil \frac{R}{K} \right\rceil$, the ceiling of $\frac{R}{K}$.
- Find the actual cut points of the classes $C_j = \min + L(j-1), \quad j = 1, 2, ..., k+1.$
- Count the number of observations in each class $(C_{j-1}, C_j),\ j = 1, 2,\ ...,\ k+1.$
- Print each class and its frequency.
- Check that the sum of frequencies is equal to the number of observations in the data.

Example 1:

The following are weights in pounds of 27 children whom visited pediatric clinic in a certain day: 34, 56, 45, 34, 23, 12, 23, 34, 55, 56, 77, 88, 99, 90, 45, 56, 65, 78, 87, 98, 89, 23, 12, 21, 32, 35, 48.

For this data, if we want to construct 4 classes, then we have

$$n = 27$$
$$k = 4$$
$$\min = 12$$
$$\max = 99$$
$$R = 87$$
$$L = \left[\frac{87}{4}\right] = 22$$

The cut points of the classes are: 12, 34, 56, 78 and hence the frequency table is:

Table 1: Frequency Table

Class	Frequency
12 – 33	7
34 – 55	8
56 – 77	5
78 – 99	7
Sum	27

From Table 1, we can see that the jump from any class to the next one $a = 1$. So, the Actual class limits are: $The\ lower\ Limit\ -a/2,\quad The\ upper\ Limit\ +a/2$ as follows

Table 2: Frequency Table

Class	Actual Classes	Frequency f_i
12 – 33	11.5 – 33.5	7
34 – 55	33.5 – 55.5	8
56 – 77	55.5 – 77.5	5
78 – 99	77.5 – 99.5	7

Sum		27

Some other measures can be added to Table 2 as:

- Class mid-point:
  (Lower Limit + Upper Limit)/2
- Relative Frequency = $\frac{f_i}{n}$
- Cumulative Frequency

Table 3: Frequency Table, Relative Frequency and mid-point

Class	Mind-point	Frequency f_i	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency
12 – 33	22.5	7	7	0.259259	0.259259
34 – 55	44.5	8	15	0.296296	0.555556
56 – 77	66.5	5	20	0.185185	0.740741
78 – 99	88.5	7	27	0.259259	1
Sum		27		1.00	

Remarks:

1- Some frequency tables can be constructed with unequal length.

2- Some frequency tables can be constructed with open lower and upper classes.

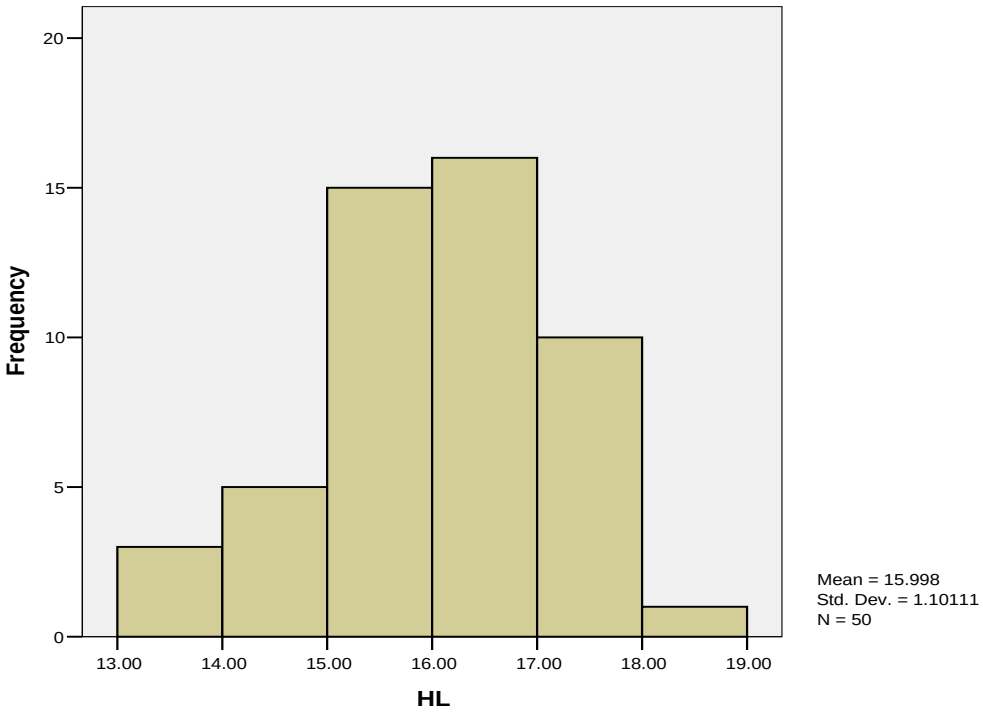
Graphs and Charts for the continuous data

1- Histogram: The histogram is a graphical tool for presenting the frequency table of the continuous data. To constructs he histogram we first place the class boundaries on the horizontal axis and class frequencies on the vertical axis. We then erect rectangles (or bars) in such a way that the height of each bar corresponds to the class frequency and the width of the bar corresponds to the class width. This can be done easily using SPSS.

Example 2:

The following table gives the hemoglobin level (g/dl) of a sample of 50 men.

17.00	15.90	16.20	15.70	13.50
14.60	15.30	13.70	16.40	17.00
14.00	16.40	17.80	15.50	15.80
15.90	13.90	15.90	17.40	17.50
14.20	15.70	17.40	14.40	17.30
17.70	15.20	17.10	17.30	16.30
15.80	16.40	16.20	16.10	15.90
16.20	14.90	16.10	18.30	16.70
15.30	16.80	16.30	15.00	16.10
16.10	15.10	16.50	16.30	15.80



2- Box Plot (Five – Figure Summary): In a set of data, we can calculate five figure-summary, they are: The maximum values, minimum values, median, lower quartile and upper quartile.

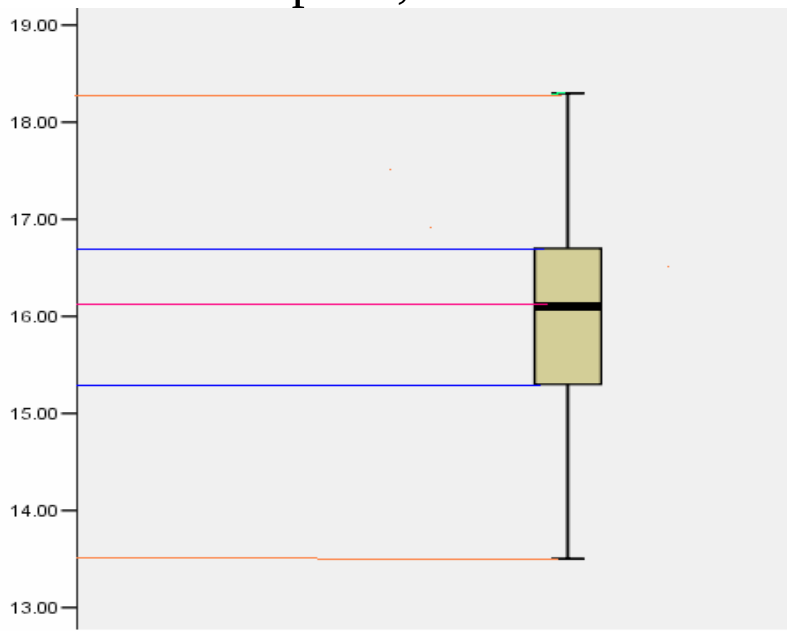
These measures can be calculated from the data as: The box-plot is a graphical display of the five-figure summary. The follows

- Arrange the data in increasing order
- The maximum and minimum values can be easily specified.
- The lower quartile (LQ) is defined as to the $(\frac{n+1}{4})-th$ observation, similarly the upper

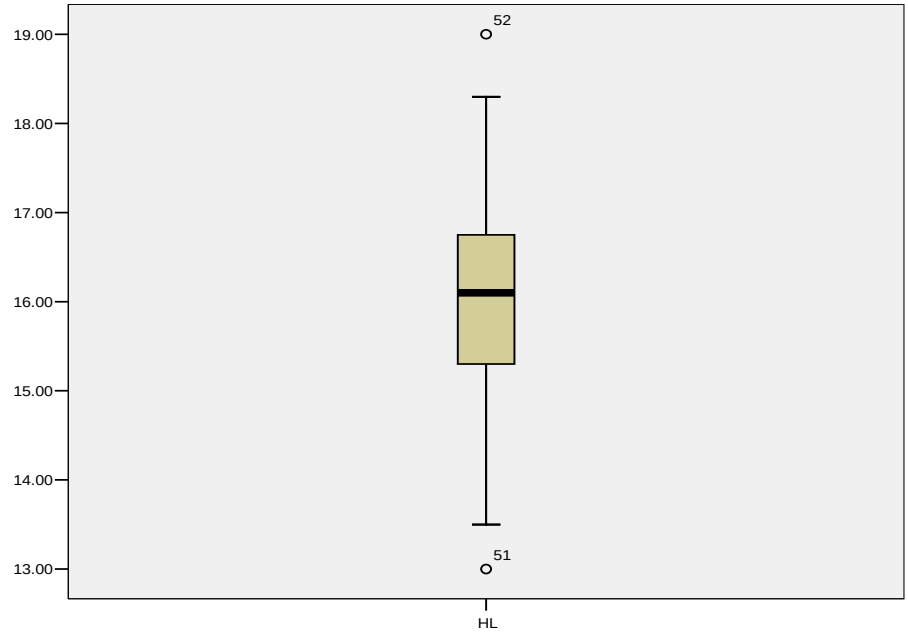
quartile (UQ) is defined to the $3(\frac{n+1}{4})-th$ observation.

- The Inter-Quartile Range $IQR = UQ - LQ$.
- The upper tolerance values is $UQ + 1.5\,IQR$.
- The lower tolerance values is $LQ - 1.5\,IQR$.
- Any sample observation that either as large as the upper tolerance or as small as the lower tolerance values will be declared an outlier.

For the data in Example 2, we have

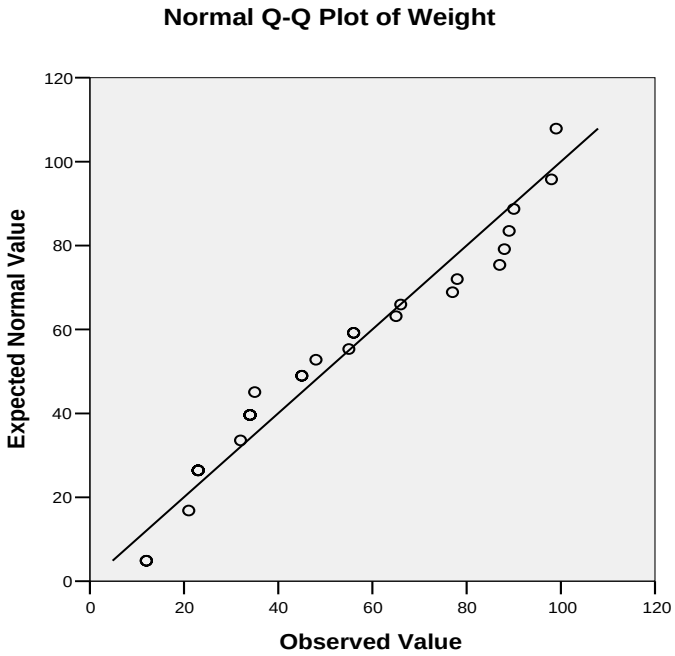


Outliers: If we add the hemoglobin level for another two men to the data given in Example 2, such as 13 and 19. These two values are detected as outlier



3- Q-Q Plot:

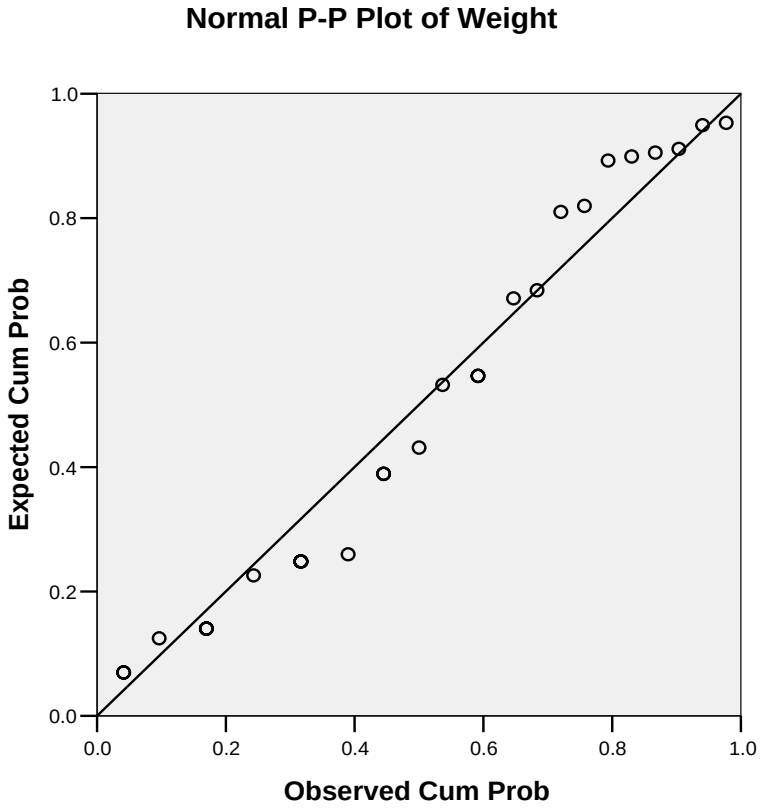
The Q-Q plot is a quick tool to show whether the data fits the considered distribution by using the expected value of each observation against the expected value of the corresponding observation using the distribution assumption. For example the following plot shows the Q-Q plot of the normal distribution to the data in Example 1.



4- P-P Plot:

The P-P plot is like Q-Q plot but using the cumulative probability and the empirical cumulative probability of the observations.

The following plot shows the P-P plot of the normal distribution to the data in Example 1.



Remark:

In the Q-Q plot and P-P plot some other distributions may be used such as: Weibull, Exponential, Pareto, Gamma and Logistic distributions.

5- Scatter Plot:

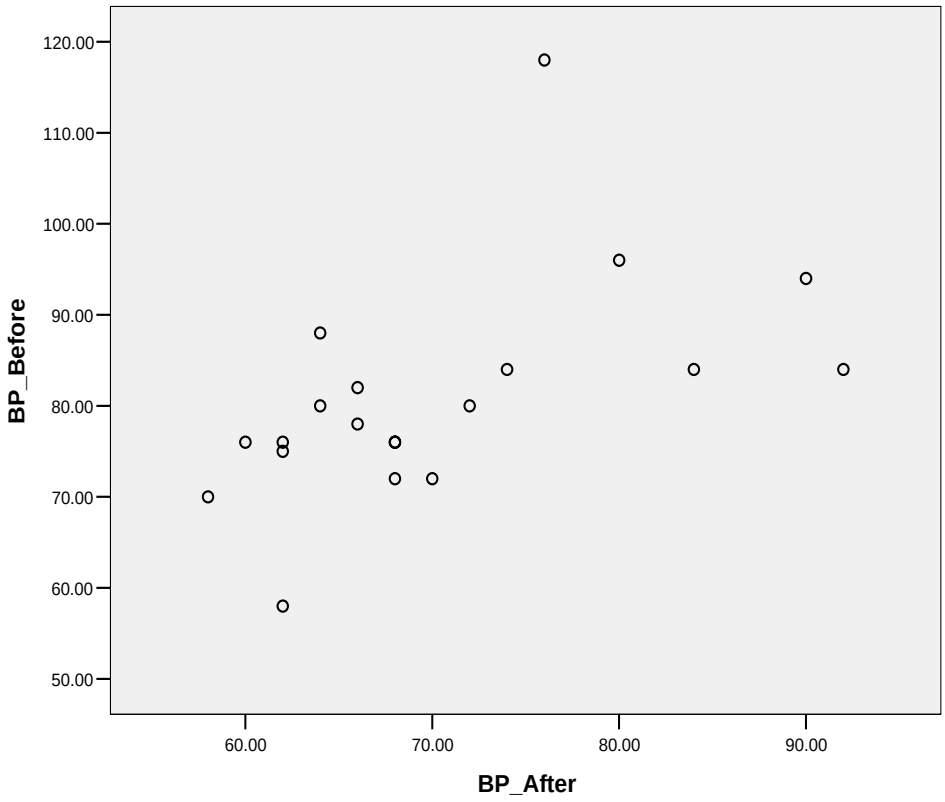
The scatter plot can be used to shoe the graphical presentation of bivariate data. This diagram illustrate the possible relationship between the pair of variable under consideration.

Example 3:
A medical study was performed to determine the effectiveness of two medications in reducing

blood pressures. Twenty patients are taken randomly and their blood pressures before and after medication are given below:

Before Medications		After Medications	
88	94	64	90
70	96	58	80
76	84	62	92
78	76	66	68
80	76	64	60
84	58	74	62
84	82	84	66
72	72	68	70
75	76	62	68
118	80	76	72

The scatter Plot for this bivariate data as:



Displaying Categorical Data

The medical and health decisions are frequently based on some categorical data, such as, discrete data, proportions, ratios, or rates.

Proportion: Many outcomes can be classified as belonging to one of two possible categories: presence and absence, nonwhite and white, male and female, improved and non-improved. Of course, one of these two categories is usually identified as of primary interest: for example, presence in the presence and absence classification, nonwhite in the white and nonwhite classification. We can, in general, relabel the two outcome categories as positive (+) and negative (-). An outcome is positive if the primary category is observed and is negative if the other category is observed.

Tables

1- Frequency Tables:

To construct a frequency table of a given discrete data (categorical data), we apply the following steps:

- Find the different categories of the data
- Partition the data set by categories
- Find the frequency of each category

Example 4

The following data represent the blood group of 24 patients:
A, B, AB, A, B, O, A, B, A, B, AB, AB, O, A, B,AB, A, O, B, B, A, A, A, O.

The frequency table for this data is

Blood Group (Class)	Frequency
A	**9**
B	**7**
AB	**4**
O	**4**
Sum	**24**

2- Cross-Tabulations:

This tool can be use to display the observation collected from a population based on two different variables. Each variable consists of different levels

- **Contingency Tables (2 x 2):** The contingency table can be formed as

Variable A / **Variable B**	**Level 1**	**Level 2**
Level 1	O_{11}	O_{12}

Level 2	O_{21}	O_{22}

Example:

The following data represents the Gender (Male M and Female F) treated by two different medications (M1 and M2)

Gender	Medication	Gender	Medication
M	M1	M	M1
M	M1	M	M1
M	M2	F	M1
F	M1	M	M2
F	M2	F	M2
M	M2	F	M1
F	M2	F	M2
F	M2	M	M1
M	M1	F	M1
F	M2	M	M1

This data can be summarized in the following contingency table

Gender * Medication Crosstabulation

Count

	Medication	Total

		M1	M2	
Gender	F	4	6	10
	M	7	3	10
	Total	11	9	20

- **Two Ways Tables (k1 x k2)**

Variable A / **Variable B**	**Level 1**	**Level 2**	**…**	**Level k1**
Level 1	O_{11}			$O_{1x\,k1}$
Level 2	O_{21}			$O_{2x\,k1}$
. .				
Level k2	$O_{k\,2x1}$	$O_{k\,2x\,2}$		$O_{k\,2x\,k1}$

Example:

The following data represent the blood group (A,B,AB,O) and Gender (Male M and Female F)

Gender	BP	Gender	BP
M	A	M	B
M	B	M	A
M	AB	F	AB
F	A	M	AB
F	AB	F	A
M	O	F	O
F	B	F	A
F	O	M	B
M	O	F	AB
F	B	M	O

Gender * BG Crosstabulation

Count

		BG				Total
		A	AB	B	O	
Gender	F	3	3	2	2	10
	M	2	2	3	3	10
Total		5	5	5	5	20

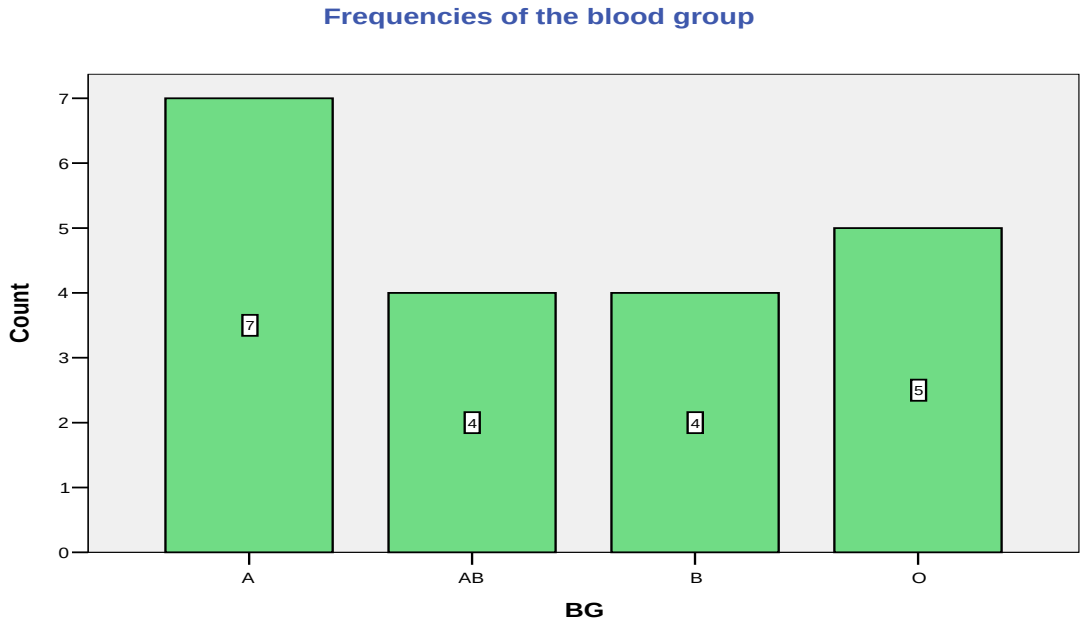
Graphs and Charts

1- Bar Chat

To construct a bar chart of a given data set, we apply the following steps

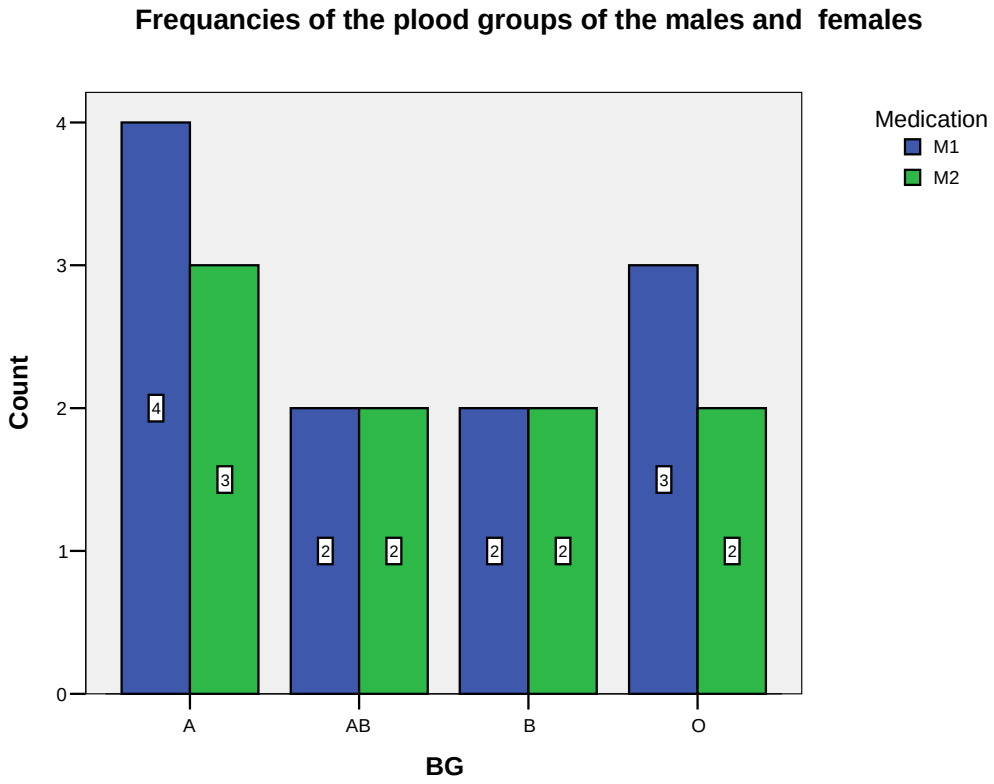
- Find the number of different classes k
- Find the frequency of each category
- For the first class, plot a rectangle such that based equal to a unit length and the height equal to the frequency of the class.
- Leave a space equal (e.g. 1/10 of a unit
- Repeat the last two steps for the other classes.

For the data given in Example 4, we have



Example:

The following data represent the blood group (A, B, AB, O) and Gender (Male M and Female F), we have

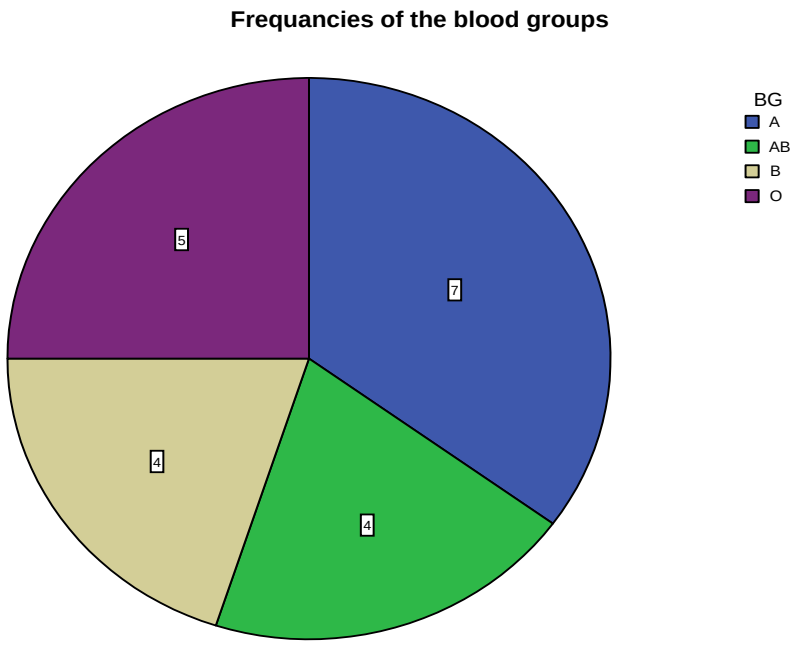


2- Pie Chart: To construct a pie chart of a given data set we apply the following steps**:**

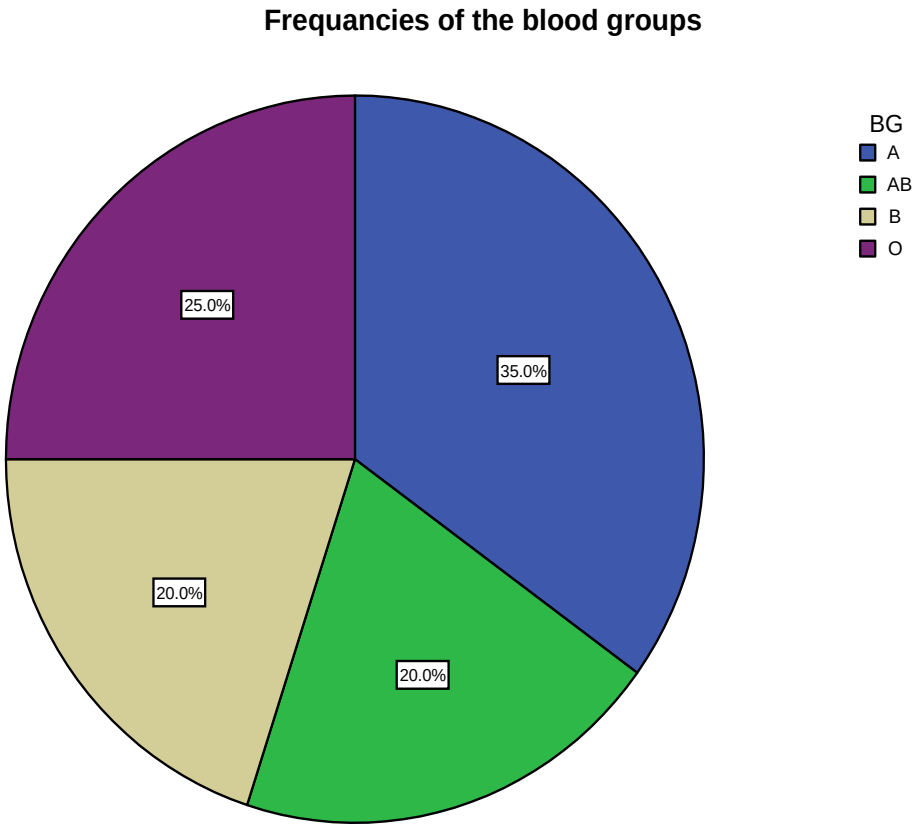
- Find different categories and their frequencies f_i
- Find the relative frequency of each category f_i / n .
- For the j-th class find the angle $\theta_j = \frac{f_j}{n}(360^\circ)$ which represents it, and the cumulative angles of the categories

- Draw a circle of unit radius and define a ray starting from (0, 0) and ending at the unit circle.
- Draw all the required rays representing the cumulative angles.

For the data given in Example 4, we have



The labels inside the figure represent the frequencies and may be changed to the ratios as follows:


Frequancies of the blood groups
BG
A
AB
B
O
35.0%
20.0%
20.0%
25.0%