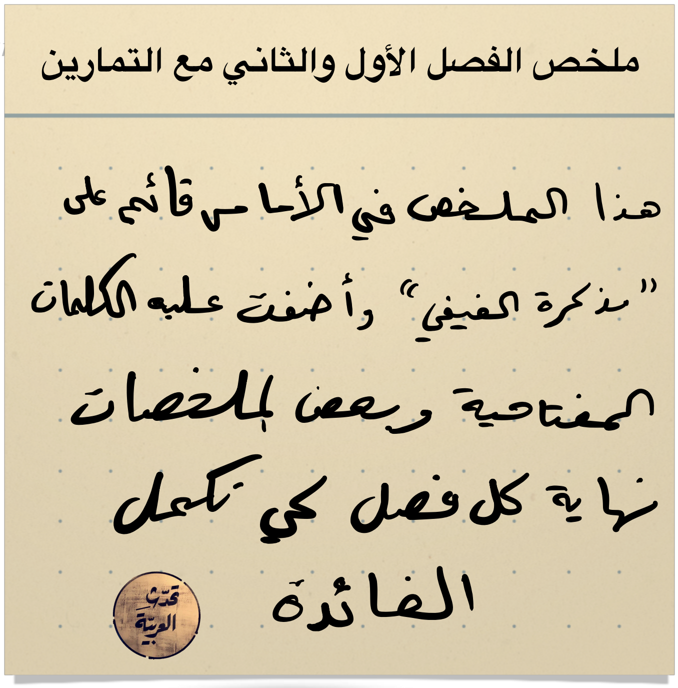


ملخص الفصل الأول والثاني مع التمارين

هذا الملخص في الأساس قائم على
"مذكرة الفيفي" وأضفت عليه الكلمات
المفتاحية وبعض الملخصات
نهاية كل فصل كي تكمل
الفائدة

STAT 101

Prepared by

Abdulrahman Alfaifi

King Saud University

Department of Statistics and Operation Research

alfaifi.stat.ksu@gmail.com

@AlfaifiStat

A Alfaifi

Chapter 1:
Descriptive Statistics

DEFINITION 1.0.1 (Data)

Data is a collection of information collected by means of experiments, observations or real life events and stored in a proper format (the word data is derived from a Latin word 'datum').

DEFINITION 1.0.2 (Statistics)

Statistics is a branch of science deals with collection, organization, presentation, analysis, interpretation of data and take the appropriate decisions.

DEFINITION 1.1.1 (Descriptive Statistics)

Descriptive statistics consist of methods and techniques which are used for presenting and summarizing data in tables or graph forms and provide some numerical measures for it.

DEFINITION 1.1.2 (Population)

Population is a set of all things (which have at least one common characteristic (or feature)) that will be subjected to a study to obtain inferences for a specific problem. The elements of population are called individuals.

DEFINITION 1.1.3 (Sample)

A sample is a subset of population, which is used to collect information and to make inferences about the entire population.

DEFINITION 1.1.4 (Inferential Statistics)

Inferential statistics is some methods and techniques that can be used for drawing conclusions about the entire population using the observations from the samples taken from that population.

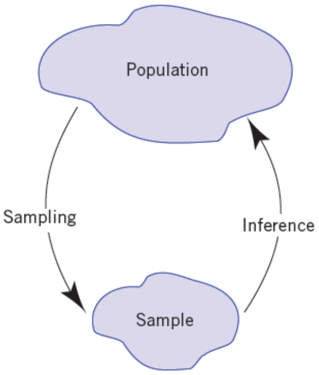


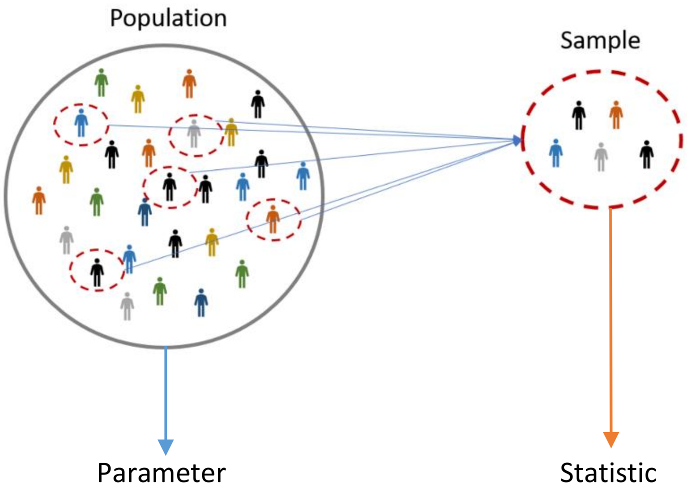
Figure 1.1.1 (Relationship between population and sample)

DEFINITION 1.1.5 (Parameter)

Parameter is a certain quantity or quality for describing a characteristic or phenomenon in a given population that summarizes the data for the entire population.

DEFINITION 1.1.6 (Statistic)

Statistic is a certain quantity or quality for describing a characteristic or phenomenon of a sample that summarizes the data for the entire sample.



DEFINITION 1.1.7 (Variables)

A variable is a map (or a function) X defined on the population (or sample) and takes values in an arbitrary set M. That means:

$$X : Population\ (or\ Sample) \longrightarrow \mathrm{M}$$

This variable measures: A characteristic, feature or factor (that varies from one individual to another) in the population.

TYPES OF VARIABLES

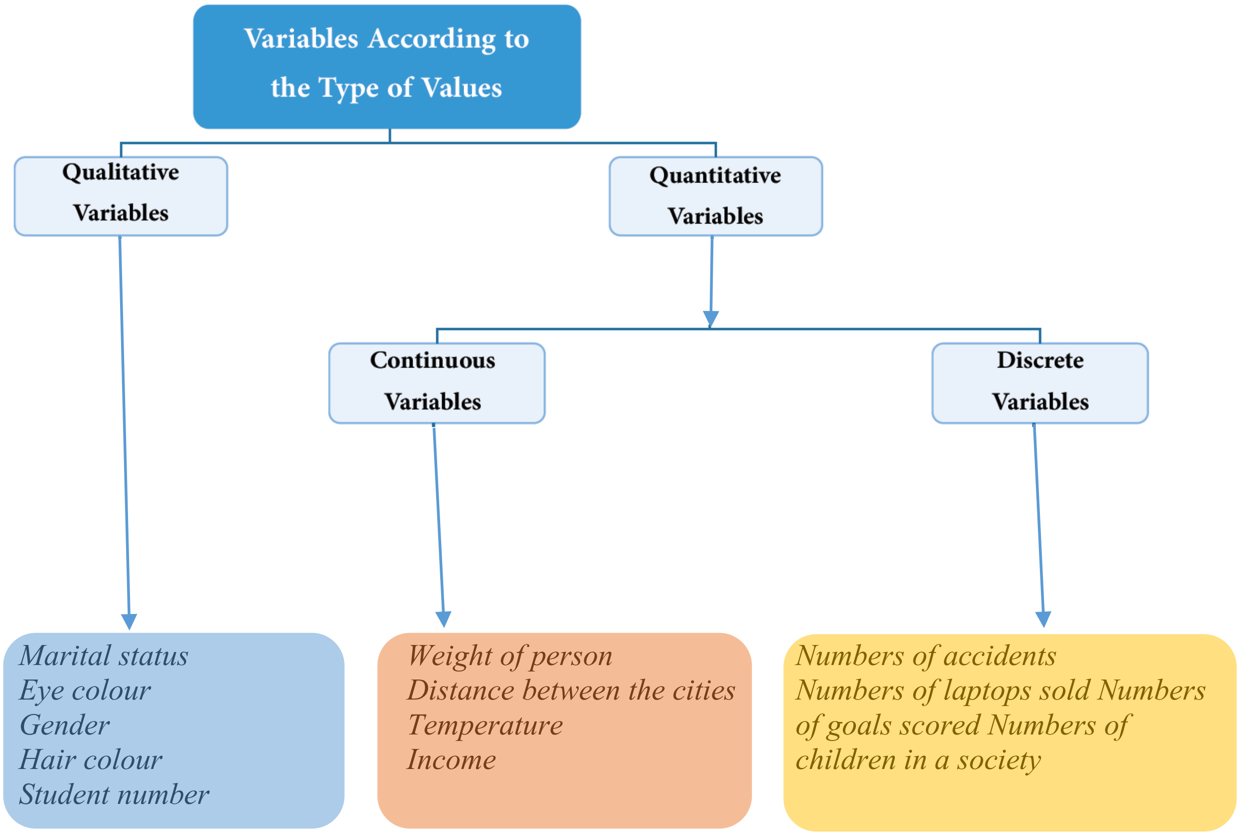
DEFINITION 1.1.8 (Qualitative or Categorical Variable)
A qualitative variable is a variable that takes non-numeric values or numeric values which indicate an attribute or property.

DEFINITION 1.1.9 (Quantitative Variable)
A quantitative variable is a variable which takes numerical values, and these numerical values can be undergoing mathematical operations (or calculation operations), or it has a measurement unit.

لها وحدة قياس وتقدر رتبتها تصاعدياً وتنازلياً.

DEFINITION 1.1.10 (Discrete Variable)
A discrete variable is a variable which takes finite or infinite countable number of values.

DEFINITION 1.1.11 (Continuous Variable)
A continuous variable is a variable which takes uncountable number of values.



Question 1: Classify each variable as *Qualitative* or *Quantitative*.	**The answer**
The variable that records lengths of rivers in Asia.	
The variable that records specializations of teachers.	
The variable that records heights of people in Riyadh.	
The variable that records colors of flowers in gardens.	

Question 1: Classify each variable as *Qualitative* or *Quantitative*.	**The answer**
The variable that records types of taxi cars in KSA.	
The variable that records ages of boys in a kindergarten.	
The variable that records ID of students in CFY (Common First Year).	
The variable that records weights of fruit boxes in a store.	

Question 1: Classify each variable as *Qualitative* or *Quantitative*.	**The answer**
The variable that records colors of cars.	
The variable that records ID of students.	
The variable that records sizes of shoes.	
The variable that records numbers of cows in cow farms.	

Question 1: Classify each variable as *Qualitative* or *Quantitative*.	**The answer**
The variable that records the time needed to finish manufacturing cars in a factory.	
The variable that records colors of flowers in a garden.	
The variable that records types of trees.	
The variable that records lives of machines.	

Question 1: Classify each variable as *Qualitative* or *Quantitative*.	**The answer**
The variable that records types of cars.	*Qualitative*
The variable that records heights of people in a country.	*Quantitative*
The variable that records the temperature in classrooms.	*Quantitative*
The variable that records colors of flowers in a forest.	*Qualitative*

Question 1: Classify each variable as *Qualitative* or *Quantitative*.	**The answer**
The variable that records skin color of people.	***Qualitative***
The variable that records numbers of animals in forestry.	***Quantitative***
The variable that records clothing sizes of men.	***Qualitative***
The variable that records lengths of revers in the continents.	***Quantitative***

Question 2: Classify each variable as *Continuous* or *Discrete*.	**The answer**
The variable that records heights of children in a school.	
The variable that records numbers of schools in KSA cities.	
The variable that records colors of pepper in Riyadh markets.	
The variable that records the exam completion time for students in KSU.	

Question 2: Classify each variable as *Continuous* or *Discrete*.	**The answer**
The variable that records ages of students in CFY.	
The variable that records temperature of patients in Riyadh hospitals.	
The variable that records production time of radios in a factory.	
The variable that records types of trees in cities of KSA.	

Question 4: Put the right word or symbol in its proper position:
Statistics, Descriptive statistics, Population, Sample, range, Interquartile range, Inferential statistics, Data, Statistic, Parameter, Proportion, Population.

A .... is a subset of population which is used to collect information and to make inferences about the entire population.

A .... is a numerical characteristics of a population that summarizes the data for the entire population.

.... is a collection of information collected by means of experiments, observations or real life events and stored in a proper format.

The .... of given raw data is defined as the difference between the greatest and smallest values.

Question 4: Put the right word or symbol in its proper position:
discrete variables, continuous variables, statistics, descriptive statistics, inferential statistics, statistic, parameter, proportion, population .

.... is a branch of science deals with collection, organization, presentation, analysis, interpretation of data and take the appropriate decisions.

A pie chart is a simple way of representing the .... of each class or category of data on a circular disk, so that each category is allocated a circular sector representing it.

.... is a certain quantity or quality for describing a characteristic or phenomenon of a sample that summarizes the data for the entire sample.

The variables, which take finite or infinite countable number of values, are called ....

Frequency table for qualitative data

▶ **EXAMPLE 1.2.1 (Qualitative data):** Consider the month of the birth of 25 members of a community, where we find the following raw data:

June	July	January	December	March
March	April	September	August	May
May	February	July	February	June
June	April	February	November	August
January	July	April	June	December

$$\text{The relative frequency of a class} = \frac{\text{The frequency of class}}{\text{The sum of frequencies}}$$

$$R\cdot f = \frac{f_i}{\sum f_i}$$

$$\text{The percent frequency of a class} = (\text{The relative frequency}) \times 100\%$$

$$P\cdot f = R\cdot f \times 100$$

Month	Frequency	Relative Frequency	Percent Frequency
January	2	$2/25 = \mathbf{0.08}$	$0.08\times 100 = \mathbf{8}\ \%$
February	3	$3/25 = \mathbf{0.12}$	$0.12\times 100 = \mathbf{12}\ \%$
March	2	$2/25 = \mathbf{0.08}$	$0.08\times 100 = \mathbf{8}\ \%$
April	3	$3/25 = \mathbf{0.12}$	$0.12\times 100 = \mathbf{12}\ \%$
May	2	$2/25 = \mathbf{0.08}$	$0.08\times 100 = \mathbf{8}\ \%$
June	4	$4/25 = \mathbf{0.16}$	$0.16\times 100 = \mathbf{16}\ \%$
July	3	$3/25 = \mathbf{0.12}$	$0.12\times 100 = \mathbf{12}\ \%$
August	2	$2/25 = \mathbf{0.08}$	$0.08\times 100 = \mathbf{8}\ \%$
September	1	$1/25 = \mathbf{0.04}$	$0.04\times 100 = \mathbf{4}\ \%$
October	0	$0/25 = \mathbf{0}$	$0\times 100 = \mathbf{0}\ \%$
November	1	$1/25 = \mathbf{0.04}$	$0.04\times 100 = \mathbf{4}\ \%$
December	2	$2/25 = \mathbf{0.08}$	$0.08\times 100 = \mathbf{8}\ \%$
Total	$n = \sum_i f_i = 25$	1	100%

▶ **EXAMPLE 1.2.5:** Consider the blood groups of the 40 persons below.

O	O	A	B	A	O	A	A	A	O
B	O	B	O	O	A	O	O	A	A
A	A	AB	A	B	A	A	O	O	A
O	O	A	A	A	O	A	O	O	AB

Construct the frequency table for the above data.

The Answer: The frequency table for the above example is given as follows:

Table 1.2.3 (Frequency Table for blood group of 40 persons)

Blood group	Frequency	Relative Frequency	Percent Frequency
O	16	0.40	40 %
A	18	0.45	45 %
B	4	0.10	10 %
AB	2	0.05	5 %
Total	40	1	100%

DEFINITION 1.3.1 (Pie chart)

A pie chart is a simple way of representing the proportion of each class or category of data on a circular disk, so that each category is allocated a circular sector representing it.

$$(\textbf{The relative frequency of the class } (i)) \times (360) = \ldots \textbf{ degree}$$

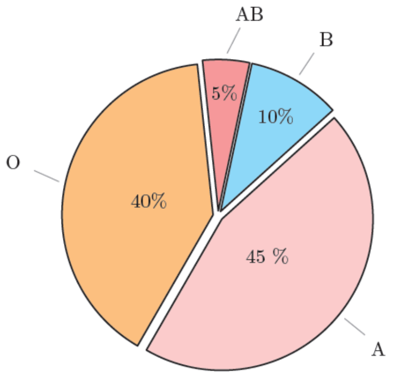
For category (AB) the measure angle is $0.05 \times 360 = 18$ **deg**

For category (B) the measure angle is $0.10 \times 360 = 36$ **deg**

For category (A) the measure angle is $0.45 \times 360 = 162$ **deg**

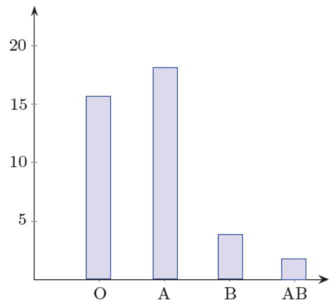
For category (O) the measure angle is $0.40 \times 360 = 144$ **deg**

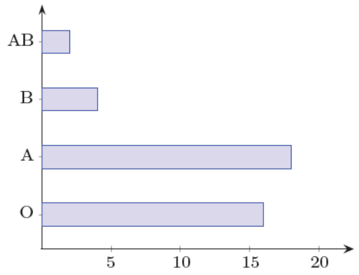
Therefore, the pie chart for example 1.2.5 is given as follows:



DEFINITION 1.3.2 (Bar chart)

A bar chart is a representation of data of discrete variable with finite values (qualitative or quantitative). This is done through vertical or horizontal bars; so that it draws over each statement a bar with height (or length) equals to the frequency of that statement.





DEFINITION 1.3.5 (Component, or Stacked bar chart)

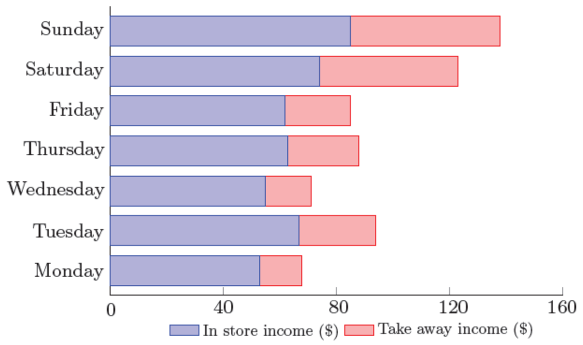
Component bar chart is a bar chart, where we can represent each component by a section in the bar, whose size is proportional to its contribution in the class.

▶ **EXAMPLE 1.3.4:** Let us consider the income at a café in a particular week.

Table 1.3.3

Day	In store Income (in $)	Take away Income (in $)	Total income (in $)
Monday	53	15	68
Tuesday	67	27	94
Wednesday	55	16	71
Thursday	63	25	88
Friday	62	23	85
Saturday	74	49	123
Sunday	85	53	138

The following is the component (Stacked) bar chart for the above data:



DEFINITION 1.3.4 (Multiple bar chart)

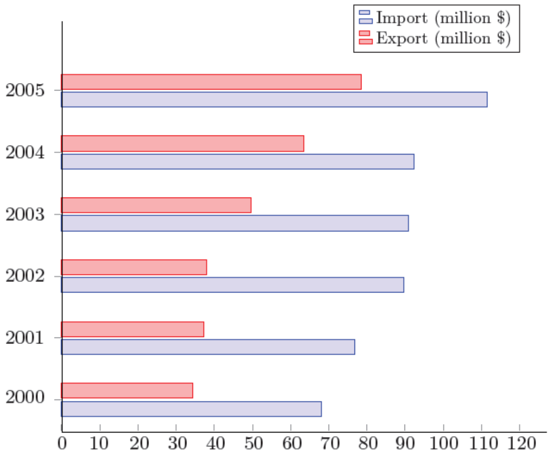
A multiple bar chart is a bar chart, where we can use it to represent multiple inter related variables by clustering bars side by side.

▶ **EXAMPLE 1.3.3:** To demonstrate a multiple bar chart, consider the following import and export data in a country.

Table 1.3.2

Year	Imports $ (in billions)	Exports $ (in billions)
2000	68.15	34.44
2001	76.71	37.33
2002	89.78	37.98
2003	90.95	49.59
2004	92.43	63.35
2005	111.39	78.44

The following is the multiple bar chart for the above data:

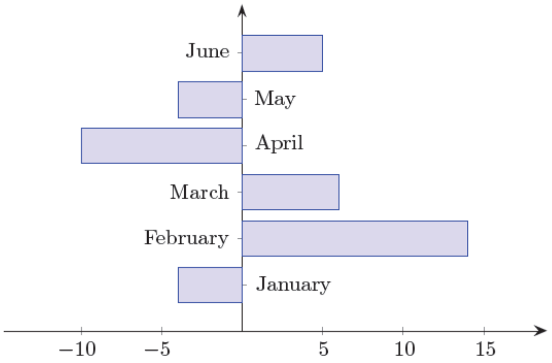


▶ **EXAMPLE 1.3.2:** In the following table, we consider the changes in income of a company from January to June.

Table 1.3.1

Month	Change in Income
January	–4 %
February	14 %
March	6 %
April	–10 %
May	–4 %
June	5 %

The following is the two-way bar chart for the above data:



FREQUENCY TABLE (DISCRETE QUANTITATIVE DATA)

▶ **EXAMPLE 1.2.2 (Discrete quantitative data):** Consider the number of children in 40 families of a society, where we find the following raw data:

4	1	2	0	2	0	1	2
0	3	0	4	0	1	1	2
3	1	2	4	0	1	0	2
4	0	1	1	2	3	0	4
0	2	0	5	2	3	1	0

Number of Children	Frequency	Relative Frequency	Percent Frequency
0	12	0.300	30 %
1	9	0.225	22.5 %
2	9	0.225	22.5 %
3	4	0.100	10 %
4	5	0.125	12.5 %
5	1	0.025	2.5 %
Total	**40**	**1**	**100%**

▶ **EXAMPLE 1.2.7:** We want to see in how many subjects each student failed in the 5th standard. Therefore, we consider a sample of 40 students. So we find the following data.

0	1	2	3	1	2	2	1
1	2	0	2	1	0	1	0
1	1	2	1	2	1	3	1
2	1	1	0	0	2	1	1
0	1	2	2	2	1	0	1

Then we construct the frequency table for this data as follow:

Table 1.2.4 (Frequency table for 40 students who failed)

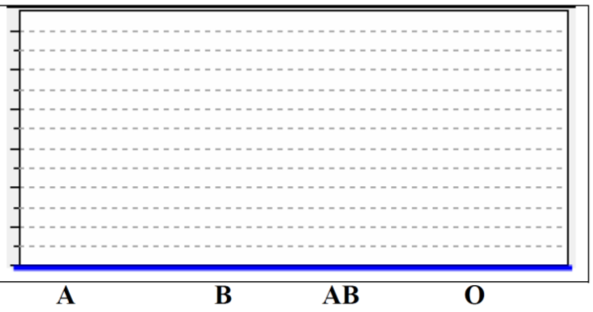
No. of subjects in which student failed	Frequency	Relative Frequency	Percent Frequency
0	8	0.20	20 %
1	18	0.45	45 %
2	12	0.30	30 %
3	2	0.05	5 %
Total	40	1	100%

Question 6: The following two data sets represent blood groups of **30** boys and **30** girls in kindergarten:

Boys	B	A	A	B	O	O	O	A	AB	A	B	O	AB	A	O
	A	O	A	A	O	A	B	A	O	B	AB	O	O	B	AB
Girls	O	B	A	O	O	B	B	O	A	AB	A	O	A	O	A
	B	AB	B	B	O	AB	B	B	A	A	AB	O	O	B	A

a) Complete the following frequency table for the above data, and draw the **multiple bar graph** for them.

Blood group	Frequency for Boys	Frequency for Girls
A		
B		
AB		
O		
Total		

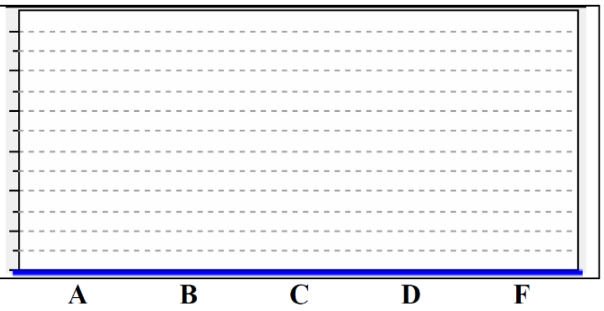


Question 6: The following data sets represent grades of **32** students in CFY:

Grades	D	A	C	B	C	C	B	A	D	C	B	C	D	D	C	B
	C	B	A	C	F	C	B	D	B	F	D	B	C	B	D	D

Complete the following frequency table for the above data, and draw the **bar graph** for them.

Grade	Frequency	Relative Frequency
A		
B		
C		
D		
F		
Total		



FREQUENCY DISTRIBUTION TABLE (CONTINUOUS QUANTITATIVE DATA)

a. Calculate the range (we denote it by $\mathbf{R}$) of the data which is given by the difference of the greatest x_ℓ and the smallest x_s value of the data. This means:

$$\mathbf{R} = x_\ell - x_s$$

b. Determine the number of classes (or categories) k to be formed. Generally, 5-20 categories are good for the analysis. Then the length (or height) of each class (we denote it by **C**) is determined by the following relation:

$$\text{Length} = \frac{\text{Range+one measuring unite}}{\text{No. of classes}} \quad \Leftrightarrow \quad \mathbf{C} = \frac{\mathbf{R}\text{+ one measuring unite}}{k}$$

c. If the length of a class is calculated, using the above formula comes out to be a fraction value t. Then (for easier viewing of classes by data dump) we can take a value u greater than t, or take (if the length of the class is greater than 1) the smallest integer greater than the fraction as the class length.

d. The lower limit of the first class limit is the minimum value in the data. The upper limit of the first class limit is calculated by adding the number $(\mathbf{C}-1)$ to the lower limit of the current class. The lower limit of the next class limit is calculated by adding 1 to the upper limit of the previous class limit, and we become the upper limit of this class by adding the number $(\mathbf{C}-1)$ to the lower limit of this class. As such, the rest of the class limits are built.

e. To make the class boundaries subtract 0.5 unit from the lower limit of each class limit and add 0.5 unit to the upper limit of each class limit. This means that the length class $\mathbf{C}$, which is previously calculated is for the class boundaries.

f. It is common practice to represent the classes in a frequency distribution table of a continuous quantitative variable by the class midpoint. Class midpoint is the center

$$\text{Class midpoint} = \frac{\text{Upper limit of the class + Lower limit of the class}}{2}$$

▶ **EXAMPLE 1.2.8:** In the shopping center recorded sales of traditional accessories for girls, whose prices are between 1 and 25 SR, we had the following data estimated at SR.

4	1	7	9	12	16	17	7	12	19
22	24	3	2	8	6	13	24	14	11
18	16	23	20	1	2	6	25	15	7
11	12	16	17	21	22	15	17	14	5
7	8	12	13	20	23	13	19	18	12

We will construct the frequency distribution table for this data by using 5 classes.

First, we calculate the range of the given data. Note that we have in that example, the greatest value is 25 and the smallest value is 1. Therefore, we have:

$$\mathbf{R} = 25 - 1 = 24$$

Therefore, the length of class boundary is given by:

$$\mathbf{C} = (24+1)/5 = 5$$

Therefore, the length of class limit equals to $\mathbf{C} - 1 = 5 - 1 = 4$.

Class Limit	Class Boundaries	Class Midpoint	Frequency	Relative Frequency	Less than	Ascending Cumulative Frequency (ACF) F_i
1 - 5	$0.5 \to 5.5$	3	7	0.14	5.5	**7**
6 - 10	$5.5 \to 10.5$	8	9	0.18	10.5	7+9 = **16**
11 - 15	$10.5 \to 15.5$	13	14	0.28	15.5	7+9 +14 = **30**
16 - 20	$15.5 \to 20.5$	18	12	0.24	20.5	7+9 +14+12 = **42**
21 - 25	$20.5 \to 25.5$	23	8	0.16	25.5	7+9 +14+12 8 = **50**
Total	---------------	--------	50	1	-------	---------------------------

Class Boundaries	Frequency	greater than	Descending Cumulative Frequency (DCF) Φ_i
$0.5 \to 5.5$	7	0.5	**50**
$5.5 \to 10.5$	9	5.5	50-7 = **43**
$10.5 \to 15.5$	14	10.5	50-9-7 = **34**
$15.5 \to 20.5$	12	15.5	50-14-9-7 = **20**
$20.5 \to 25.5$	8	20.5	50-12-14-9-7 = **8**
--------------	50	-------	----------------------------

▶ **EXAMPLE 1.2.9:** Consider the mileage of 40 cars per liter of fuel in a particular city, so we get the following results:

12	16	15	12	19	17	18	16	14	13
12	20	12	15	16	20	16	15	12	18
16	17	19	15	16	17	15	16	15	14
12	13	14	15	16	17	18	19	20	20

We construct the frequency distribution table for this data in the following manner:

We have the range for the given data equal to $\mathbf{R} = 20 - 12 = 8$.

Now we determine the number of classes using the following relation:

$$k = \left\lfloor 3.322 \log n \right\rfloor = \left\lfloor 3.322 \log 40 \right\rfloor = \left\lfloor 5.322 \right\rfloor = 5$$

So the class boundary length equals to $\mathbf{C} = (8+1)/5 = 1.8$. We will take $\mathbf{C} = 2$, Therefore, the length of class limit equal to $\mathbf{C} - 1 = 2 - 1 = 1$.

Based on the above, we can present the frequency distribution table for the above data in the following table:

Class Limit	Class Boundaries	Class Midpoint	Frequency	Relative Frequency	Ascending Cumulative Frequency F_i	Descending Cumulative Frequency Φ_i
12-13	11.5→13.5	12.5	8	0.20	8	40
14-15	13.5→15.5	14.5	10	0.25	18	32
16-17	15.5→17.5	16.5	12	0.30	30	22
18-19	17.5→19.5	18.5	6	0.15	36	10
20-21	19.5→21.5	20.5	4	0.10	40	4
Total	----------	--------	40	1	----------	----------

▶ **EXAMPLE 1.2.10:** We will construct the cumulative relative and the cumulative percentages frequencies by using the data presented in the previous example 1.2.9.

We have:

Table 1.2.7

Class Boundaries	Frequency	Ascending Cumulative Frequency F_i	Ascending Cumulative Relative Frequencies	Ascending Cumulative Percentages Frequencies
11.5→13.5	8	8	$8/40 = \mathbf{0.20}$	$0.20\times100 = \mathbf{20}$
13.5→15.5	10	18	$18/40 = \mathbf{0.45}{=}0.20+0.25$	$0.45\times100 = \mathbf{45}$
15.5→17.5	12	30	$30/40 = \mathbf{0.75}{=}0.45+0.30$	$0.75\times100 = \mathbf{75}$
17.5→19.5	6	36	$36/40 = \mathbf{0.90}{=}0.75+0.15$	$0.90\times100 = \mathbf{90}$
19.5→21.5	4	40	$40/40 = \mathbf{1}.00{=}0.90+0.10$	$1.00\times100 = \mathbf{100}$
Total	40	----------	---------	---------

DEFINITION 1.3.6 (Histogram)

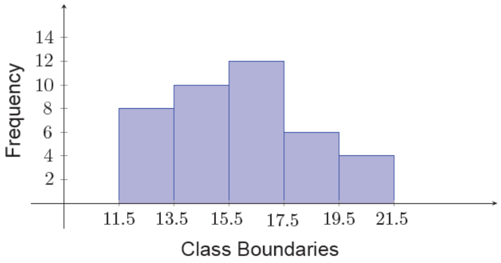
A histogram is a graphical display used for data generated by continuous variables. It is a graph in which class boundaries are marked on the horizontal axis and the frequencies are marked on a vertical axis, and is constructed by drawing a rectangular column above each actual category so that its height equals the frequency of that category.

▶ **EXAMPLE 1.3.5:**

Consider the data from Example 1.2.9, where we have:

Class Boundaries	11.5 → 13.5	13.5 → 15.5	15.5 → 17.5	17.5 → 19.5	19.5 → 21.5	Total
Frequency	8	10	12	6	4	40

Then the histogram for the given data is as follow:



I- Symmetric Histogram:

Unimodal: Histogram with one peak.

Bimodal: Histogram with two peaks.

Multimodal: Histogram with more than two peaks.

Uniform: Histogram with no peaks (all classes have the same frequency).

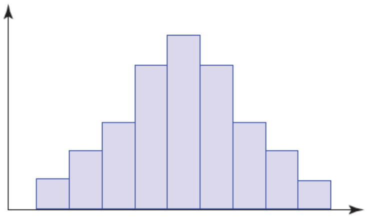


Figure 1.3.7-a (Unimodal Histogram)

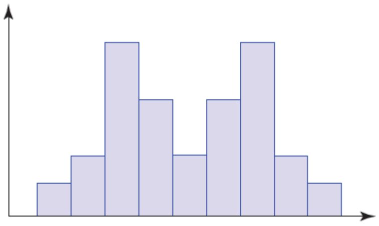


Figure 1.3.7-b (Bimodal Histogram)

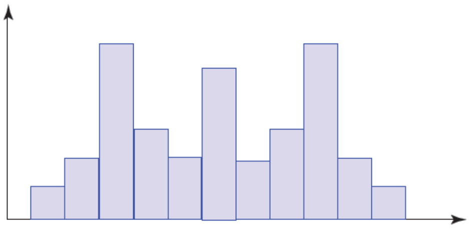


Figure 1.3.7-c (Multimodal Histogram)

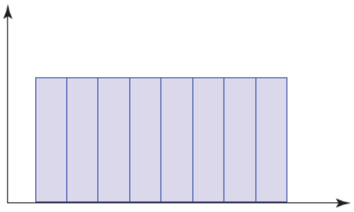


Figure 1.3.7-d (Uniform Histogram)

II- Skewed Histogram:

DEFINITION 1.3.7 (Skewedness)
Histograms are called as skewed if they are non-symmetric. In such histograms, bins on one side have high frequency which decreases as we move to the other side. The side with lower frequency is said to have a longer tail.

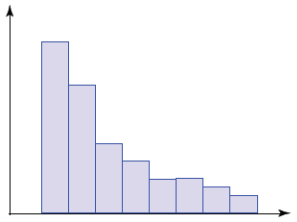


Figure 1.3.8-a (Right Skewed Histogram)

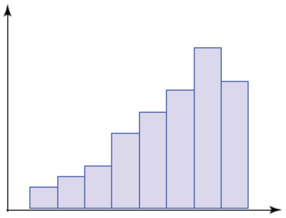


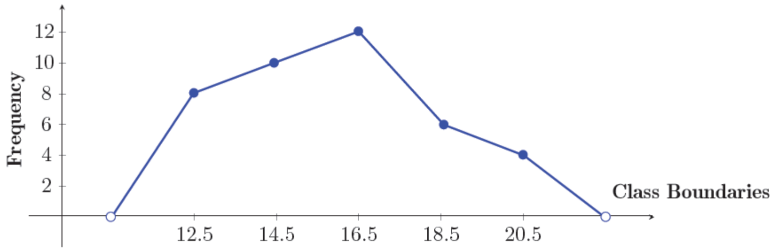
Figure 1.3.8-b (Left Skewed Histogram)

DEFINITION 1.3.8 (Polygon)
The frequency polygon is a polygon which connects with a straight line the points (x_i, f_i), whereas $x_i$and f_i are the midpoint and the frequency of class boundary i respectively, and it closes from the left to the middle of a default class boundary located before the first class boundary, and from the right to the middle of a default class boundary located after the last class boundary.

▶ **EXAMPLE 1.3.6:** Consider the frequency table of car mileage data from example 1.2.9.

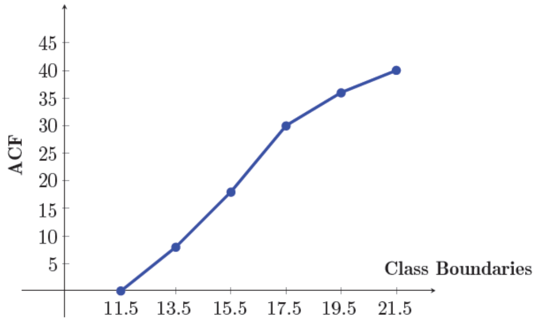
Class Boundaries	11.5 → 13.5	13.5 → 15.5	15.5 → 17.5	17.5 → 19.5	19.5 → 21.5	Total
Frequency	8	10	12	6	4	40

As we can see the class midpoints are 12.5, 14.5, 16.5, 18.5 and 20.5 respectively. We now construct the frequency polygon for this data.



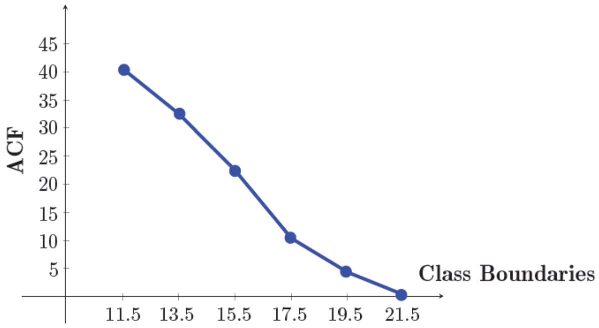
DEFINITION 1.3.9 (Ascending Cumulative Frequency Polygon (ACFP))
The ascending cumulative frequency polygon is a polygon which connects with a straight line the points (b_i, F_i), whereas b_i and F_i are the upper bound and the ascending cumulative frequency of class boundary i respectively, and closes from the left to the beginning of the first class boundary.

The ascending cumulative frequency polygon for the data from Example 1.2.9 is presented in the following graph.



DEFINITION 1.3.10 (Descending Cumulative Frequency Polygon (DCFP)):
The descending cumulative frequency polygon (DCFP) is a polygon which connects with a straight line the points (b_i, Φ_i), whereas b_i and Φ_i are the lower bound and the descending cumulative frequency of class boundary i respectively, and closes from the right to the end of the last class boundary.

The descending cumulative frequency polygon for the data from Example 1.2.9 is presented in the following graph.



Section 1.4

MEASURES OF CENTRAL TENDENCY AND POSITION

Measure	*Raw Data*	*Frequency table*
Mean $\bar{x}$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\bar{x} = \frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i x_i$	$\bar{x} = \frac{1}{\sum_{i=1}^{k} f_i}\sum_{i=1}^{k} f_i x_i$
Median $\tilde{x}$	*First: arrange in increasing order* *even* $\frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$ *odd* $x_{\frac{n+1}{2}}$	*First: find ACF* $\tilde{x} = \tilde{L} + \frac{\frac{\sum f_i}{2}(\tilde{F}-\tilde{f})}{\tilde{f}} \times C$
Mode $\hat{x}$	*The observation with the highest frequency*	$\hat{x} = \hat{L} + \frac{d_1}{d_1 - d_2} \times C$

$\tilde{L}$ is the lower limit of the median class boundary,

$\tilde{F}$ is the cumulative frequency of the median class boundary,

$\tilde{f}$ is the frequency of the median class boundary,

C is the class length of the median class boundary.

$\hat{L}$ is the lower limit of the modal class boundary,

d_1 is the difference between the frequency of the modal class and the frequency of the previous class directly,

d_2 is the difference between the frequency of the modal class and the frequency of the next class directly,

C is the class length of the modal class boundary.

• *Mean:*

▶ **EXAMPLE 1.4.1:** Calculate the mean of the following data:

20 18 15 15 14 12 11 9 7 6 4 1

According to the definition of the mean, we have:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{20+18+15+15+14+12+11+9+7+6+4+1}{12} = \frac{132}{12} = 11$$

DEFINITION 1.4.1.b (Mean for Organized Data in a Frequency Table)

We suppose that the data (of a discrete variable) are given by the following Frequency table:

Table 1.4.1

i	Values	Frequency
1	x_1	f_1
2	x_2	f_2
$\vdots$	$\vdots$	$\vdots$
m	x_m	f_m
Total	--------	$\sum f_i = n$

Then, the mean for these data (we denote it by $\overline{x}$) is given by the following relation:

$$\overline{x} = \frac{1}{\sum f_i}\sum_{i=1}^{m} f_i\, x_i = \frac{1}{n}\sum_{i=1}^{m} f_i\, x_i$$

▶ **EXAMPLE 1.4.2:** Let us consider the following data (No. of subjects in which student failed):

Table 1.4.2

i	No. of subjects in which student failed	Frequency
1	0	8
2	1	18
3	2	12
$m = 4$	3	2
Total	----------	40

The mean of the number of subjects in which students failed is calculated as:

$$\overline{x} = \frac{1}{\sum f_i}\sum_{i=1}^{m} f_i\, x_i = \frac{(0\times 8)+(1\times 18)+(2\times 12)+(3\times 2)}{8+18+12+2} = \frac{0+18+24+6}{40} = \frac{48}{40} = 1.2$$

DEFINITION 1.4.1.c (Mean for Organized Data in a Frequency Distribution Table)
We suppose that the data (of a continuous variable) are given by the following frequency distribution table:

Table 1.4.3

i	Class Boundaries	Class Midpoint x_i	Frequency f_i	Ascending Cumulative Frequency (ACF)
1	$b_0 \to b_1$	x_1	f_1	$F_1 = f_1$
2	$b_1 \to b_2$	x_2	f_2	$F_2 = f_1 + f_2$
$\vdots$	$\vdots \ \vdots \ \vdots$	$\vdots$	$\vdots$	$\vdots \ \vdots \ \vdots \ \vdots$
k-1	$b_{k-2} \to b_{k-1}$	x_{k-1}	f_{k-1}	$F_{k-1} = f_1 + f_2 + \dots + f_{k-1}$
k	$b_{k-1} \to b_k$	x_k	f_k	$F_k = f_1 + f_2 + \dots + f_k$
Total	-----------	--------	$\sum f_i$	-----------------

Then, the mean for these data (we denote it by $\overline{x}$) is given by the following relation:

$$\overline{x} = \frac{1}{\sum f_i} \sum_{i=1}^{k} f_i \, x_i$$

EXAMPLE 1.4.3 (For Continuous quantitative data): Let us consider the following data (the mileage of 40 cars per liter of fuel in a particular city):

Table 1.4.4

i	Class Boundaries	Class Midpoint	Frequency
1	11.5→13.5	12.5	8
2	13.5→15.5	14.5	10
3	15.5→17.5	16.5	12
4	17.5→19.5	18.5	6
$k = 5$	19.5→21.5	20.5	4
Total	------------	----------	40

The mean of the mileage of the cars is calculated as:

$$\overline{x} = \frac{12.5 \times 8 + 14.5 \times 10 + 16.5 \times 12 + 18.5 \times 6 + 20.5 \times 4}{40} = \frac{636}{40} = 15.9$$

DEFINITION 1.4.2 (Weighted Mean)

Let $x_1, x_2, \ldots, x_\ell$ be observed values and their weights are $w_1, w_2, \ldots, w_\ell$ respectively. Then the weighted mean (it is a mean, and we denote it by $\overline{x}$) of this data is given by the following relation:

$$\overline{x} = \frac{1}{\sum w_i} \sum_{i=1}^{\ell} w_i x_i$$

▶ **EXAMPLE 1.4.4:** Consider a student in King Saud University received the following grades in the first semester:

Courses	Grades	Credit hours
Mathematics	B	4
Statistics	A	3
English	C	3
Physics	C	4

Where grades $A = 5,\ B = 4,\ C = 3$ and $D = 2$ points.

Now to calculate the Grade Point Average (GPA) for this student in this semester we have

Courses	Grades	Points (x_i)	Credit Hours (w_i)	$w_i\ x_i$
Mathematics	B	4	4	16
Statistics	A	5	3	15
English	C	3	3	9
Physics	C	3	4	12
Total	------------	----------	14	52

Then, the Grade Point Average is calculated as:

$$\text{GPA} = \overline{x} = \frac{1}{\sum w_i} \sum_{i=1}^{\ell} w_i x_i = \frac{52}{14} = 3.71$$

Advantages of The Mean

- It is quick and easy to compute.
- All values are considered by calculating the mean.
- It is one and only one value for a set of data.

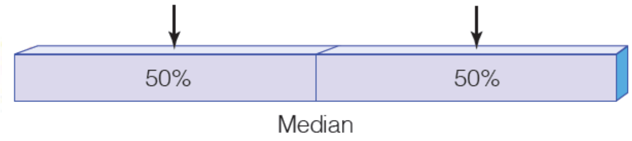
Disadvantages of The Mean

- Mean is not defined for qualitative data.
- Since it considers all the observed values, it is highly affected by the extreme values.
- It becomes not applicable if a data is lost.

• *Median:*

DEFINITION 1.4.3 (Median)

Median (we denote it by $\tilde{x}$) is that value which divides the data in two halves after ordering them, in ascending or descending order.



▶ **EXAMPLE 1.4.6:** Calculate the median of the finishing times of 7 bike racers who had finishing times as:

28 22 26 29 21 23 24

We first arrange them in increasing order:

$$21,\ 22\ ,\ 23\ ,\ 24\ ,\ 26\ ,\ 28\ ,\ 29$$

$$x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5 \quad x_6 \quad x_7$$

$$\tilde{x} = x_{\frac{7+1}{2}} = x_4 = 24$$

▶ **EXAMPLE 1.4.7:** Calculate the median of the finishing times of 8 bike racers who had finishing times as:

28 22 26 29 21 23 24 35

In the similar manner as the above example, we first arrange the data as:

$$21,\ 22\ ,\ 23\ ,\ 24\ ,\ 26\ ,\ 28\ ,\ 29\ ,\ 35$$

$$x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5 \quad x_6 \quad x_7 \quad x_8$$

$$\tilde{x} = \frac{x_{\frac{8}{2}} + x_{\frac{8}{2}+1}}{2} = \frac{x_4 + x_5}{2} = \frac{24+26}{2} = 25$$

▶ **EXAMPLE 1.4.8:** Consider the following data:

Table 1.4.5-a

Subjects	No. of subjects in which student failed	Frequency (Number of students whose failed in the subject)
English	2	12
Mathematics	1	18
Statistics	0	8
Chemistry	3	2

We arrange the representative values of the data in the given table, so we get the following frequency table:

i	Subjects	No. of subjects in which student failed	Frequency	ACF
1	Statistics	0	8	**8**
2	Mathematics	1	18	8+18 = **26**
3	English	2	12	26+12 = **38**
$m = 4$	Chemistry	3	2	38 +2 = **40**
Total		---------	40	------------

- We use the following formula to calculate the median for continuous quantitative data:

$$\boxed{\tilde{x} := \tilde{L} + \frac{\frac{1}{2}\sum f_i - \left(\tilde{F} - \tilde{f}\right)}{\tilde{f}} \times \mathrm{C}}$$

Where, $\tilde{L}$ is the lower limit of the median class boundary,

$\tilde{F}$ is the cumulative frequency of the median class boundary,

$\tilde{f}$ is the frequency of the median class boundary,

C is the class length of the median class boundary.

▶ **EXAMPLE 1.4.9:** We will calculate the median for the data in example 1.4.3.

Table 1.4.6

i	Class Boundaries	Class Midpoint	Frequency	Relative Frequency	Ascending Cumulative Frequency
1	11.5→13.5	12.5	8	0.20	8
2	13.5→15.5	14.5	10	0.25	18
3	15.5→17.5	16.5	12	0.30	30
4	17.5→19.5	18.5	6	0.15	36
5	19.5→21.5	20.5	4	0.10	40
Total		-------	40	1	-----------

First, we find the median class. In this example $n = 40$, therefore, $\frac{1}{2}\sum f_i = 20$. Then, the median class is $15.5 - 17.5$. So, using the following formula:

$$\tilde{x} := \tilde{L} + \frac{\frac{1}{2}\sum f_i - \left(\tilde{F} - \tilde{f}\right)}{\tilde{f}} \times \mathrm{C}$$

Where we have, $\tilde{F} = 30$, $\tilde{f} = 12$ and $\mathrm{C} = 2$, then we get:

$$\tilde{x} = 15.5 + \frac{20 - \left(30 - 12\right)}{12} \times 2 = 15.5 + 0.33 = 15.83$$

Advantages of The Median

- It is easy to compute and understand.
- It is not affected by outliers or extreme values.
- It can be used even if you loss some data (known argument) that is not in the middle.

Disadvantages of The Median

- It does not take all values into account.
- It is not used in many statistical tests.
- It cannot be identified for qualitative data

• *Mode:*

DEFINITION 1.4.4 (The Mode)
The mode (we denote it by $\hat{x}$) of data is a value or observation, which has the highest frequency.

CALCULATING THE MODE

▶ **EXAMPLE 1.4.10:** We consider the following data that represents grades of 12 students in an exam:

A, A, C, A, D, A, B, B, C, D, A, B

So we find that the mode of this data is A.

▶ **EXAMPLE 1.4.11:** The following data represent the time spent in 10 Km race for ten bicycle runners:

28 22 26 29 21 23 28 28 25 29

If we put this data in a frequency table, then we have:

Table 1.4.7

Finishing Time	21	22	23	25	26	28	29
Frequency	1	1	1	1	1	3	2

For this data, we note that the mode is equal to 28.

▶ **EXAMPLE 1.4.12:** Consider the following data representing the age (in years) of 14 students:

12	11	13	14	13	12	11
12	13	12	12	13	14	13

The ages 12 and 13 in this example have the highest frequency. Hence the variable is said to be bimodal. i.e. having two modes 12 and 13.

▶ **EXAMPLE 1.4.14:** The finishing times of 7 bike racers who had finishing times as:

28 22 26 29 21 23 24

If we put this data in a frequency table, then we have:

Table 1.4.8

Finishing Time	21	22	23	24	26	28	29
Frequency	1	1	1	1	1	1	1

In this example we note that all data have the same frequency 1. Therefore, the given data have no mode.

▶ **EXAMPLE 1.4.15:** The following table represents marks (out of 10) obtained by 20 students:

Table 1.4.10

Marks (out of 10)	Frequency	Marks (out of 10)	Frequency
2	1	6	3
3	2	7	3
4	3	9	4
5	2	10	2

So, we note that the mode of the marks obtained by students is 9.

▶ **EXAMPLE 1.4.16:** We consider the data in example 1.2.5 (the data about the blood group of 40 persons), which we can display in the following table:

Table 1.4.11

Blood group	O	A	B	AB
Frequency	16	18	4	2

Then we find that the mode of the blood group is A.

For Organized Data in a Frequency Distribution Table (*data of a continuous variable*):

$$\boxed{\hat{x} = \hat{L} + \frac{d_1}{d_1 + d_2} C}$$

Where, $\hat{L}$ is the lower limit of the modal class boundary,

d_1 is the difference between the frequency of the modal class and the frequency of the previous class directly,

d_2 is the difference between the frequency of the modal class and the frequency of the next class directly,

C is the class length of the modal class boundary.

▶ **EXAMPLE 1.4.17:** Consider the time taken by 24 persons in a certain run race.

Table 1.4.12

Seconds	Class Boundaries	Frequency
1	50.5→55.5	2
2	55.5→60.5	7
3	60.5→65.5	8
4	65.5→70.5	4
5	70.5→75.5	3

We note that then the class which has frequency greater than the frequency of the previous and subsequent classes directly is the third class, and this class is unique. Therefore, the modal class for the given table is the class 60.5-65.5. So, we have:

$$\hat{x} = \hat{L} + \frac{d_1}{d_1 + d_2} C = 60.5 + \frac{8-7}{(8-7)+(8-4)} \times 5 = 61.5$$

▶ **EXAMPLE 1.4.18:** The minutes spent per week by the teenagers in watching movies are given by the following table:

Table 1.4.13

Number of Minutes per week	Number of Teenagers
$0 \to 90$	26
$90 \to 180$	32
$180 \to 270$	65
$270 \to 360$	75
$360 \to 450$	60
$450 \to 540$	42
Total	300

Then we find the class modal for the above data is 270⟶360 because it has the highest frequency 75. Therefore, we have:

$$\hat{x} = \hat{L} + \frac{d_1}{d_1 + d_2} C = 270 + \frac{10}{10 + 15} \times 90 = 306$$

Advantages of The Mode

- It is quick and easy to compute.
- It can be evaluated for both quantitative and qualitative data.
- It is not affected by extreme values.

Disadvantages of The Mode

- There may be more than one mode for a certain data set.
- Sometimes, there is no mode for a given data set.
- It may not reflect the center of the distribution very well.

THE RELATIONSHIPS AMONG THE MEAN, MEDIAN AND MODE

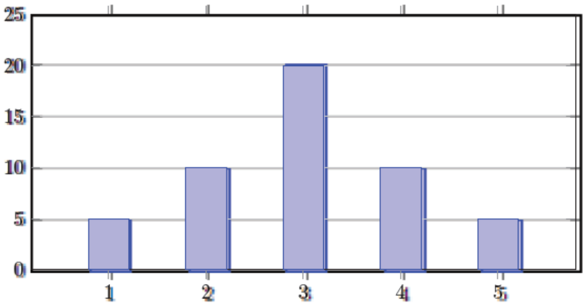


Figure 1.4.2 (A symmetric distribution with mode = mean = median = 3)

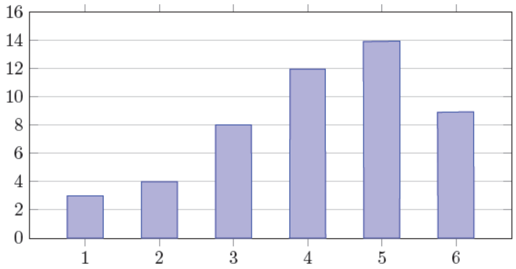


Figure 1.4.3 (A left skewed distribution with mode greater than median greater than mean)

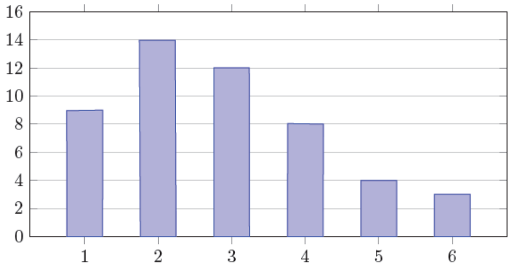


Figure 1.4.4 (A right skewed distribution with mean greater than median greater than mode)

In general, we can explain the relationship among the three measures of central tendency using the following graphs.

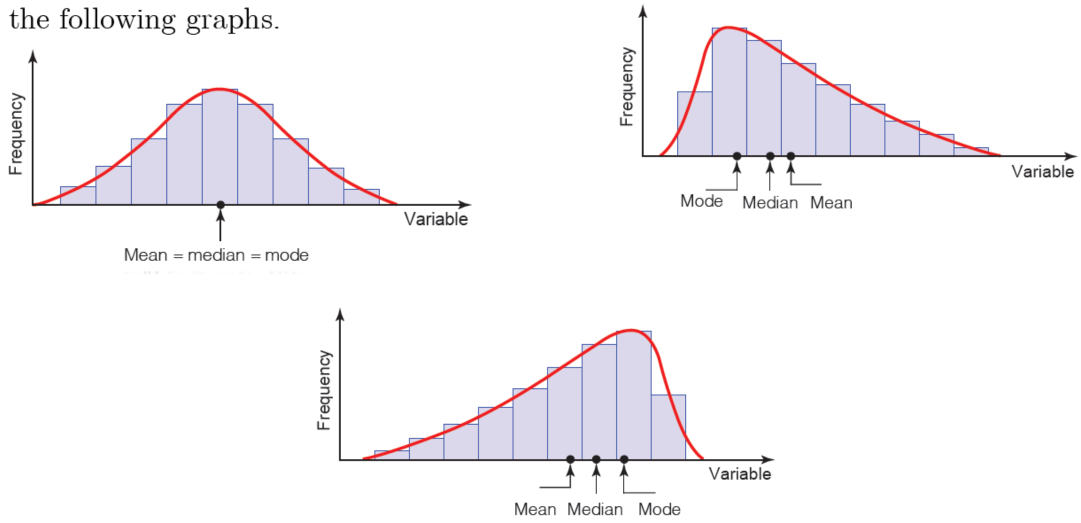


Figure 1.4.5 (The relationship among the three measures of tendency)

DEFINITION 1.4.5 (Percentiles)

The percentiles (denoted by P_1 , P_2 ,…and P_{99}) of a variable divide the ordered observed values into 100 equal parts. Median is the 50th percentile and it divides ordered data into two equal halves. The percentile P_1 divides the ordered observed data into 1% from bottom and 99% from top. Similarly, any jth percentile, P_j divides the ordered observed value into two parts such that j % observed values are below this value and (100 - j) % observed values are above this value. The following graph explains the concept of percentiles.

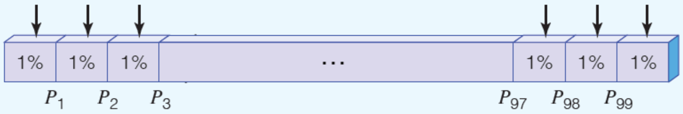


Figure 1.4.6 (The concept of percentiles)

How can we calculate the percentiles?

Let x_1 , x_2 , ..., x_n be arranged data. Then:

- We calculate the **rank** of r^{th} percentile (we denote it by p_r) by the following relation:

$$p_r = \frac{r\ (n+1)}{100} \qquad ;\ r = 1,\ 2, ..., 99$$

- The r^{th} **percentile** P_r can be calculated by the following relation:

$$P_r = x_k + s\,(x_{k+1} - x_k) \qquad ;\ r = 1,\ 2, ..., 99$$

Where k is the integer part of p_r , and the number s is the rest of p_r .

▶ **EXAMPLE 1.4.19:** Calculate the 35th percentile of the data given below.

40	51	92	10	36	60	70	36	36	40
80	39	53	56	60	60	70	72	88	92
50	92	20	70	38	95	56	60	88	70

Solution: We first arrange the data in the increasing (or ascending) order as follows:

10	20	36	36	36	38	39	40	40	50
51	53	56	56	60	60	60	60	70	70
70	70	72	80	88	88	92	92	92	95

Here we have $n = 30$. So $p_{35} = \dfrac{r\ (n+1)}{100} = \dfrac{35\ (30+1)}{100} = 10.85$

Also we have $k = 10$ and $s = 0.85$. Therefore, we become that:

$$\begin{aligned} P_{35} = x_k + s\,(x_{k+1} - x_k) &= x_{10} + 0.85\,(x_{11} - x_{10}) \\ &= 50 + 0.85(51 - 50) = 50.85 \end{aligned}$$

DEFINITION 1.4.6 (Deciles)

The deciles (denoted by D_1 , D_2 ,…and D_9) divide the ordered data into 10 equal parts. D_1 is the 10th percentile; D_2 is 20th percentile and so on till D_9 which is the 90th percentile. The following graph explains the concept of deciles.

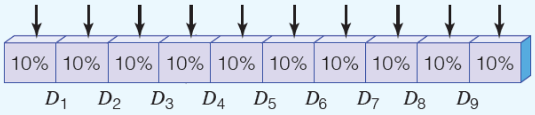


Figure 1.4.7 (The concept of deciles)

How can we calculate the deciles?

Let x_1 , x_2 , ..., x_n be arranged data. Then:

- We calculate the **rank** of r^{th} decile, whose denoted by d_r , and is calculated by the following relation:

$$d_r = \frac{r\ (n+1)}{10} \qquad ; r = 1,\ 2, \dots, 9$$

- The r^{th} **decile** D_r can be calculated by the following relation:

$$D_r = x_k + s\,(x_{k+1} - x_k) \qquad ; r = 1,\ 2, \dots, 9$$

Where k is the integer part of d_r , and the number s is the rest of d_r .

For example, refer to the example 1.4.19 we calculate the sixth decile. We have $n = 30$. So,

$$d_6 = \frac{r\ (n+1)}{10} = \frac{6\ (30+1)}{10} = 18.6$$

Also we have $k = 18$ and $s = 0.6$. Therefore, we become that:

$$D_6 = x_k + s\,(x_{k+1} - x_k) = x_{18} + 0.6\,(x_{19} - x_{18}) = 60 + 0.6(70 - 60) = 66$$

DEFINITION 1.4.7 (Quartiles)

The Quartiles (denoted by Q_1 ,Q_2 and Q_3) divide the ordered data into 4 equal parts. The first quartile Q_1 is 25th percentile, the second quartile Q_2 is 50th percentile which is also the median of the data and the third quartile Q_3 is 75th percentile. The following graph explains the concept of quartiles.

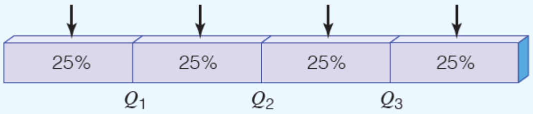


Figure 1.4.8 (The concept of quartiles)

How can we calculate the quartiles?

Let x_1 , x_2 , ..., x_n be arranged data. Then:

- We calculate the **rank** of r^{th} quartile, whose denoted by q_r , and is calculated by the following relation:

$$q_r = \frac{r\ (n+1)}{4} \qquad ; r = 1,\ 2,3$$

- The r^{th} **quartile** Q_r can be calculated by the following relation:

$$Q_r = x_k + s\,(x_{k+1} - x_k) \qquad ; r = 1,\ 2,3$$

Where k is the integer part of q_r , and the number s is the rest of q_r .

For example, refer to the example above we calculate the first quartile. We have $n = 30$. So,

$$q_1 = \frac{r\ (n+1)}{4} = \frac{(30+1)}{4} = 7.75$$

Also we have $k = 7$ and $s = 0.75$. Therefore, we become that:

$$\begin{aligned} Q_1 = x_k + s\,(x_{k+1} - x_k) &= x_7 + 0.75\,(x_8 - x_7) \\ &= 39 + 0.75(40 - 39) = 39.75 \end{aligned}$$

▶ **EXAMPLE 1.4.20:** Find the quartiles for the data given below:

28 22 26 29 21 23 24

We first arrange the data in the increasing order as follows:

21 22 23 24 26 28 29

Then:

For Q_1 we have the rank $q_1 = \dfrac{r\,(n+1)}{4} = \dfrac{(7+1)}{4} = 2$

Also we have $k = 2$ and $s = 0$. Therefore, we become that:

$$Q_1 = x_k + s\,(x_{k+1} - x_k) = x_2 = 22$$

For Q_2 (the median) we have the rank $q_2 = \dfrac{r\,(n+1)}{4} = \dfrac{2(7+1)}{4} = 4$

Also we have $k = 4$ and $s = 0$. Therefore, we become that:

$$Q_2 = x_k + s\,(x_{k+1} - x_k) = x_4 = 24$$

For Q_3 we have the rank $q_3 = \dfrac{r\,(n+1)}{4} = \dfrac{3(7+1)}{4} = 6$

Also we have $k = 6$ and $s = 0$. Therefore, we become that:

$$Q_3 = x_k + s\,(x_{k+1} - x_k) = x_6 = 28$$

▶ **EXAMPLE 1.4.21:** Find the quartiles of the data given below:

39 40 41 56 7 8 15 36

We first arrange the data in the increasing order as follows:

7 8 15 36 39 40 41 56

In a similar way to the above we find:

$$q_1 = 2.25 \quad \Rightarrow \quad Q_1 = 8 + 0.25(15-8) = 9.75$$
$$q_2 = 4.5 \quad \Rightarrow \quad Q_2 = 36 + 0.5(39-36) = 37.5$$
$$q_3 = 6.75 \quad \Rightarrow \quad Q_3 = 40 + 0.75(41-40) = 40.75$$

DEFINITION 1.4.8 (Extreme Value)

We say that a value x of given data is said to be extreme if one of the following relations is realizing:

$$x < LF := Q_1 - 1.5\,(Q_3 - Q_1) \quad [LF \text{ is the abbreviation of “Lower Fence”}]$$

or

$$x > HF := Q_3 + 1.5\,(Q_3 - Q_1) \quad [HF \text{ is the abbreviation of “Higher Fence”}]$$

▶ **EXAMPLE 1.4.22:** Refer to the example 1.4.21, we find:

$$Q_1 - 1.5\,(Q_3 - Q_1) = 9.75 - 1.5(40.75 - 9.75) = -36.75$$

$$Q_3 + 1.5\,(Q_3 - Q_1) = 40.75 + 1.5(40.75 - 9.75) = 87.25$$

So, the given data haven't an extreme value.

DEFINITION 1.4.9 (Five Numbers)

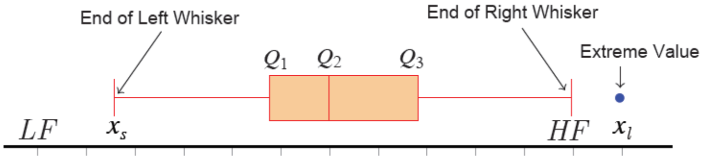
Five Numbers are a summary of the variable data which includes the below mentioned five characteristics:

$$\overset{X_s}{\text{Smallest}}\text{ value, } Q_1,\ Q_2,\ Q_3 \text{ and } \overset{X_l}{\text{Largets}}\text{ value}$$

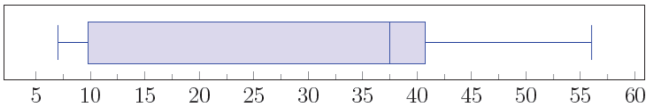
The five numbers summary for example 1.4.21 is given by **7, 9.75, 37.5, 40.75** and **56**.

DEFINITION 1.4.10 (Box Plot)

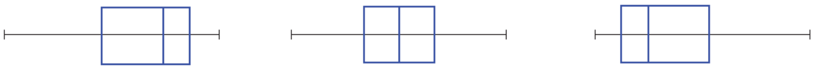
The box plot of given data is the graphical representation of its five numbers summary.



The box plot for the data given in Example 1.4.21 with five numbers summary are 7, 9.75, 37.5, 40.75 and 56 is given below



The following graph represents three different sets of data displaying the symmetric and the skewed distribution of data.



Left - Skewed

Symmetric

Right - Skewed

(11.5 marks)
Question 8: Let **5, 13 , 8 , 6 , 10, 6** be data of a sample. Then:

a) Calculate the **mean** for the given data.

....

....

b) Calculate the **median** for the given data.

....

....

....

....

c) How many **modes** do the given data have? **Determine** them.

....

....

b) Calculate the **standard deviation** for the given data.

....

....

....

....

e) Calculate the **coefficient of variation** for the given data.

....

....

....

e) Calculate Q_1 , D_3 and P_{75} for the given data.

For Q_1 : ....

....

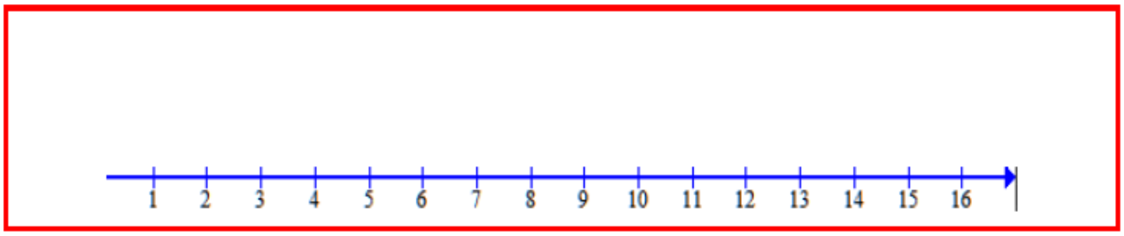
For D_3 : ....

....

For P_{75} : ....

....

h) Draw the **box plot** for the given data and determine the **five numbers**.



(11.5 marks)

Question 8: Let **8, -1 , 7 , 7 , 8, 5 , 15** be data of a sample. Then:

a) Calculate the **mean** for the given data.

....

....

b) Calculate the **median** for the given data.

....

....

....

c) How many **modes** do the given data have? **Determine** them.

....

....

d) Calculate the **standard deviation** for the given data.

....

....

....

....

e) Calculate the **z-score** for the value **8**.

....

....

....

e) Calculate Q_3, D_7 and P_{25} for the given data.

For Q_3 : ....

....

For D_7 : ....

....

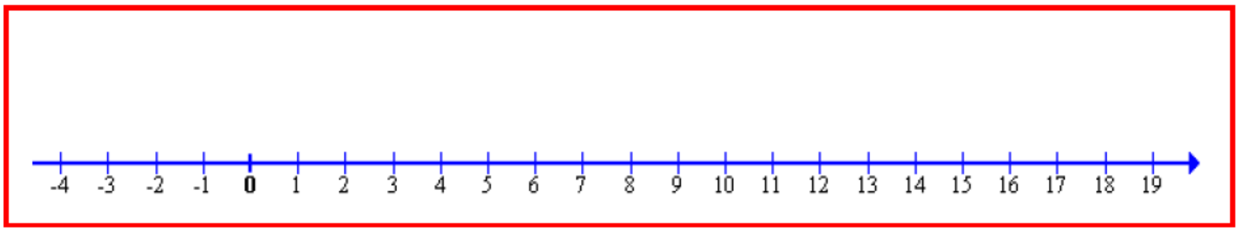
For P_{25} : ....

....

g) Does the given data have extreme values? Why?

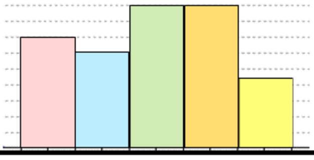
....

h) Draw the **box plot** for the given data and determine the **five numbers** on the graph.



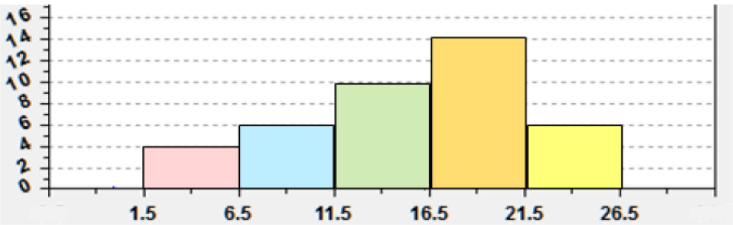
Question 3: Determine whether of the following statements is *True* or *False*.	**The answer**
The range of data is sensitive to extreme values.	
The interquartile range is the best measure for dispersion.	
If the mean of **8, 3, x, x+4, 5, 10** is **8**, then $x = 8$.	
If some data are missing, then we can always use the median.	

Question 3: Determine whether of the following statements is *True* or *False*.	**The answer**
Always the value of mean for raw data equals to the value of mean for grouped data.	
Always we can use the median as measure of tendency.	
If the median of the ordered data **-3, 5, 6, x, x+1, x+3, 11, 15, 16** is **7.5**, then $x = 7$.	
When we have extreme values, then the interquartile range is an appropriate measure for the dispersion.	

Question 5: Put ✓ at the correct answer.	
The data **3, 3, 3, 3, 3, 3, 3, 3**.	Have no standard deviation.
	Have standard deviation equal to 1.
	Have standard deviation equal to 0.
Data with the histogram 	Have two modes.
	Have one mode.
	Have three modes.

Question 5: Put ✓ at the correct answer.	
If we have the ordered data: **-5, 0.5, ?, 2, 4, ?, 4.5, ?, 7, 9, 11, 51**	Then we can calculate the interquartile range.
	Then we cannot calculate the interquartile range.
If we have $Q_2 = Q_3$ for data, then	we cannot draw the box plot of this data.
	we can draw the box plot of this data.
If we have $\overline{x} < \tilde{x} < \hat{x}$ for continuous data, then	the distribution of data is left skewed.
	the distribution of data is right skewed.
If two data sets have the same variance, then	these data sets have the same mean.
	it is possible to have these data sets the same mean.

Question 7: If we have data with the following histogram:



Then:

a) Complete the following frequency distribution table for the given data in the previous figure:

Class Limit	Class Boundaries	Midpoint	Frequency	Relative Frequency	Percentage %	A.C.F.
Total						

b) Calculate the **median** for the given data.

....

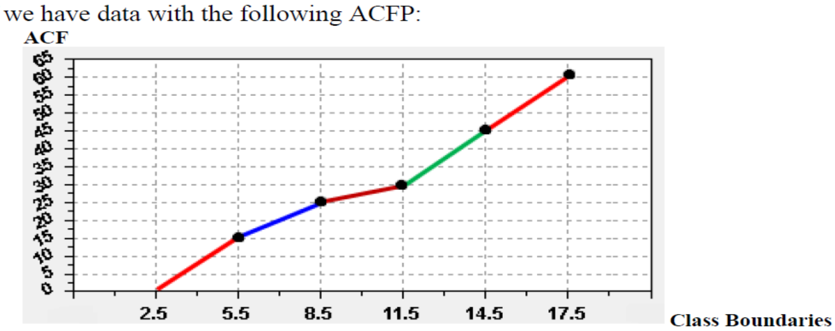
....

....

c) Calculate the **range** for the given data.

....

Question 7: If we have data with the following ACFP:



Then:

a) Complete the following frequency distribution table for the given data in the previous figure:

Class Limit	Class Boundaries	Midpoint	Frequency	Relative Frequency	Percentage %	A.C.F.
Total						

b) Calculate the **mode(s)** for the given data.

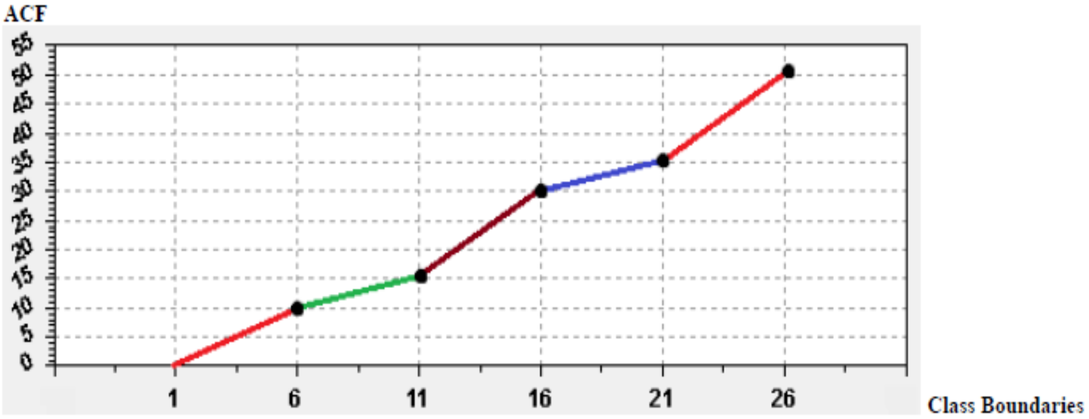
....

....

....

c) Calculate the **range** for the given data.

Question 5: Consider data given by the following ascending cumulative frequency polygon (ACFP):



Then:

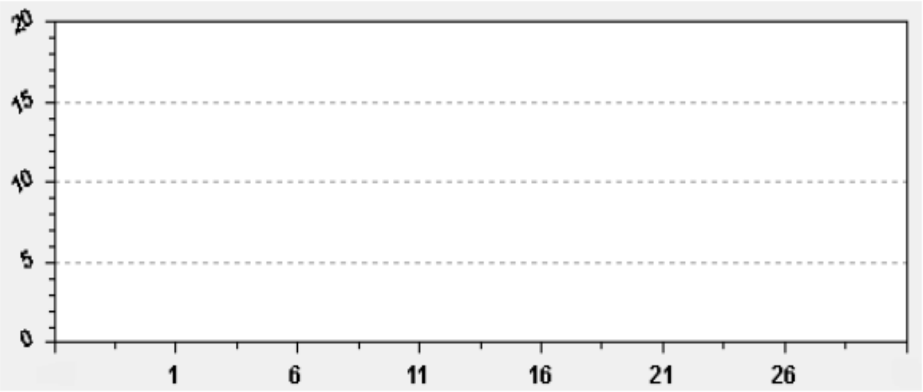
a) Complete the following frequency distribution table for the above ACFP.

Class Limit	Class Boundaries	Midpoint	Frequency	Relative Frequency	Percentage Frequency	ACF
Sum						

b) How many modes have the given data in the above frequency distribution table. Calculate them (or it).

c) Calculate the **range** of data for the above frequency distribution table.

d) Draw the **histogram** of the above frequency distribution table.



Section 1.5

MEASURES OF VARIABILITY

Measure	*Raw Data*	*Frequency table*
Variance S^2 $(unit)^2$	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ $S^2 = \frac{1}{n(n-1)}[n\sum_{i=1}^{n}{x_i}^2 - (\sum_{i=1}^{n}x_i)^2]$	$S^2 = \frac{1}{(\sum_{i=1}^{k}f_i)-1}\sum_{i=1}^{k}f_i(x_i - \bar{x})^2$
Standard deviation S $(unit)$	$S = +\sqrt{S^2}$	$S = +\sqrt{S^2}$
Range R $(unit)$	*Maximum - Minimum* $R = x_l - x_s$	*Midpoint 1st class – Midpoint of kth class* $R = x_k - x_1$
Interquartile range IQR	$IQR = Q_3 - Q_1$	$IQR = Q_3 - Q_1$
Coefficient of variation CV	$CV = \frac{S}{\bar{x}} \times 100\%$	$CV = \frac{S}{\bar{x}} \times 100\%$

DEFINITION 1.5.10 (*z*-scores)

Let $x_1, x_2, \ldots., x_n$ be raw data with mean $\overline{x}$ and standard deviation $S > 0$. Then the standard score of a value x_i for some i (z-scores and one denotes it by z_{x_i}) of data converts the data in such manner that the resultant data have a mean 0 and a standard deviation 1. The following formula is used to calculate the standard score of a data:

$$z_{x_i} = \frac{x_i - \overline{x}}{S}$$

DEFINITION 1.5.1 (Variance for raw data)

Let $x_1, x_2, ..., x_n$ be raw data with mean $\overline{x}$ and $n \geq 2$. Then the variance of this data (one denote it by S^2) is given by the following relation:

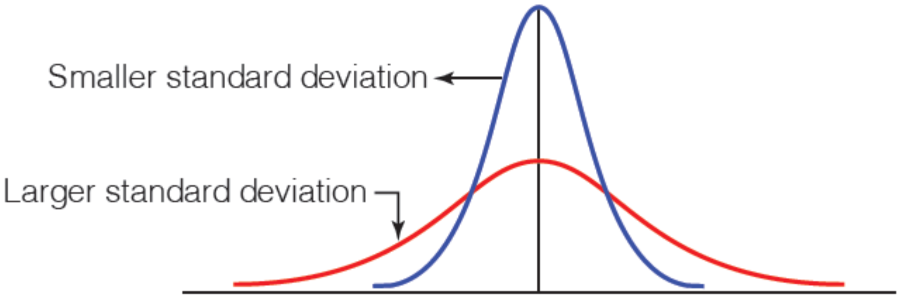
$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

DEFINITION 1.5.2 (Standard Deviation)

The standard deviation (one denotes it by S) is the positive square root of the variance and calculated by the following relation:

$$S = +\sqrt{S^2}$$

$$\boxed{S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]}$$



▶ **EXAMPLE 1.5.1:** Let 2, 3, 6, 8, 10, 13 and 14 be given data. Then the mean of this data is:

$$\overline{x} = \frac{2+3+6+8+10+13+14}{7} = \frac{56}{7} = 8$$

Variable values	Squared Variable values	Deviation from mean $\left(x_i - \overline{x}\right)$	Squared Deviation $\left(x_i - \overline{x}\right)^2$
2	4	$2-8=-6$	36
3	9	-5	25
6	36	-2	4
8	64	0	0
10	100	2	4
13	169	5	25
14	196	6	36
$\sum_{i=1}^{7} x_i = 56$	$\sum_{i=1}^{7} x_i^2 = 578$	0	$\sum_{i=1}^{7} (x_i - \overline{x})^2 = 130$

Therefore, we have:

$$S = +\sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n-1}} = +\sqrt{\frac{130}{6}} = +\sqrt{21.667} = 4.65$$

$$S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right] = \frac{1}{6}\left(578 - \frac{(56)^2}{7}\right) = \frac{1}{6}\left(578 - 448\right) = \frac{1}{6}(130) = 21.667$$

And then we have $S = +\sqrt{21.667} = 4.65$.

DEFINITION 1.5.3 (Variance for grouped data in a frequency table)

We suppose that the data are given by the Frequency table 1.4.1. Then, the variance for these data is given by the following relation:

$$S^2 = \frac{1}{\left(\sum_{i=1}^{m} f_i\right) - 1} \sum_{i=1}^{m} f_i \left(x_i - \overline{x}\right)^2$$

▶ **EXAMPLE 1.5.2:** Let us consider the following data:

Table 1.5.2

i	Subjects	Frequency
1	2	6
2	5	10
3	7	16
4	12	8
Total	----------	40

The mean of the data for the above table is calculated as:

$$\overline{x} = \frac{1}{\sum f_i} \sum_{i=1}^{m} f_i\, x_i = \frac{(6\times 2) + (10\times 5) + (16\times 7) + (8\times 12)}{40}$$

$$= \frac{12 + 50 + 112 + 96}{40} = \frac{270}{40} = 6.75$$

We calculate the variance by the following relation:

$$S^2 = \frac{1}{\left(\sum_{i=1}^{m} f_i\right) - 1} \sum_{i=1}^{m} f_i \left(x_i - \overline{x}\right)^2$$

$$= \frac{6\,(2-6.75)^2 + 10\,(5-6.75)^2 + 16\,(7-6.75)^2 + 8\,(12-6.75)^2}{39} = \frac{387.5}{39} = 9.94$$

Therefore, the standard deviation becomes that:

$$S = +\sqrt{S^2} = +\sqrt{9.94} = 3.15$$

DEFINITION 1.5.4 (Variance for grouped data in a frequency distribution table) We suppose that the data are given by the Frequency table 1.4.3. Then, the variance for these data is given by the following relation:

$$S^2 = \frac{1}{\left(\sum_{i=1}^{m} f_i\right) - 1} \sum_{i=1}^{m} f_i \left(x_i - \overline{x}\right)^2$$

▶ **EXAMPLE 1.5.3:** Let us consider the following frequency distribution table:

Table 1.5.3

Class boundary	Midpoint	Frequency
4 → 10	7	8
10 → 16	13	10
16 → 22	19	18
22 → 28	25	12
28 → 34	31	2
Total		**50**

The mean of the data for the above table is calculated as:

$$\overline{x} = \frac{1}{\sum f_i} \sum_{i=1}^{m} f_i\, x_i = \frac{(8\times 7)+(10\times 13)+(18\times 19)+(12\times 25)+(2\times 31)}{50}$$
$$= \frac{56+130+342+300+62}{50} = \frac{890}{50} = 17.8$$

We calculate the variance by the following relation.

$$S^2 = \frac{1}{\left(\sum_{i=1}^{m} f_i\right) - 1} \sum_{i=1}^{m} f_i \left(x_i - \overline{x}\right)^2$$
$$= \frac{8\,(7-17.8)^2 + 10\,(13-17.8)^2 + 18\,(19-17.8)^2 + 12\,(25-17.8)^2 + 2\,(31-17.8)^2}{49}$$
$$= \frac{2160}{49} = 44.08$$

Therefore, the standard deviation becomes that:

$$S = +\sqrt{S^2} = +\sqrt{44.08} = 6.64$$

DEFINITION 1.5.5 (Range for raw data)

We have introduced the definition of the range earlier when building a frequency distribution table, were we had $\mathbf{R} = x_\ell - x_s$. Whereas x_l is the greatest value of data, and x_s is the smallest value of data.

DEFINITION 1.5.6 (Range for organized data in a frequency table)

We suppose that the data are given by the Frequency table 1.4.1. Then, the range for these data is given by the following relation:

$$\mathbf{R} = x_m - x_1$$

Where x_m is the value of the last subject, and x_1 is the first subject in the frequency table.

▶ **EXAMPLE 1.5.4:** Let us consider the following data:

Table 1.5.4

i	Subjects	Frequency
1	2	4
2	5	10
3	7	16
$m = 4$	12	8
Total	----------	38

The range of the number of subjects is calculated as:

$$\mathbf{R} = x_m - x_1 = 12 - 2 = 10$$

DEFINITION 1.5.7 (Range for grouped data)

We conceder grouped data in a frequency distribution table as in Table 1.4.3. Then, the range is defined as follow:

$$\mathrm{R} = x_k - x_1$$

Where x_k is the middle point of last class, and x_1 is the middle point of the first class.

▶ EXAMPLES 1.5.5:

1. We conceder the following sets of data:

$$\mathrm{X}: \; 4,\; 8,\; 7,\; 3,\; 5,\; 10,\; 24,\; 5$$

$$\mathrm{Y}: \; 10, 7, 9, 11, 11, 8, 9, 7$$

Then we find the range:

For data (X) equal to $\mathrm{R}_{\mathrm{X}} = x_\ell - x_s = 24 - 3 = 21$.

For data (Y) equal to $\mathrm{R}_{\mathrm{Y}} = x_\ell - x_s = 11 - 7 = 4$.

DEFINITION 1.5.8 (Interquartile Range)

The Interquartile Range (one denotes it by IQR) of given data is defined as the difference between the third quartile and the first quartile.

$$IQR = Q_3 - Q_1$$

IQR approximately gives us the range of the middle 50% of the observed values and hence it is also sometimes called as mid-spread.

▶ **Example 1.5.6:** Find the Interquartile range of the data given in Example 1.4.20.

Solution: In the example 1.4.20 we calculated the quartiles of the given data which were as $Q_1 = 22$, $Q_2 = 24$ and $Q_3 = 28$. Therefore, we get that:

$$IQR = Q_3 - Q_1 = 28 - 22 = 6$$

DEFINITION 1.5.9 (Coefficient of Variation)

Let $x_1, x_2, \ldots., x_n$ be raw data with mean $\overline{x} \neq 0$ and standard deviation S. Then the coefficient of variation (we denote it by CV) is calculated as:

$$CV = \frac{S}{\overline{x}} \times 100 \;\%$$

The coefficient of variation is a useful measure of variation to compare between sets of data with different units (measures).

DEFINITION 1.5.10 (*z*-scores)

Let $x_1, x_2, \ldots., x_n$ be raw data with mean $\overline{x}$ and standard deviation $S > 0$. Then the standard score of a value x_i for some i (*z*-scores and one denotes it by z_{x_i}) of data converts the data in such manner that the resultant data have a mean 0 and a standard deviation 1. The following formula is used to calculate the standard score of a data:

$$z_{x_i} = \frac{x_i - \overline{x}}{S}$$

▶ **EXAMPLE 1.5.7:** Let 2, 5, 3, 3, 7 be given data. Then to calculate the z - scores for this data we must calculate the mean and the standard deviation of data. We find $\overline{x} = 4$ and $S = 2$. So we get:

$$z_2 = \frac{x_1 - \overline{x}}{S} = \frac{2-4}{2} = \frac{-2}{2} = -1 \quad \& \quad z_5 = \frac{x_2 - \overline{x}}{S} = \frac{5-4}{2} = \frac{1}{2}$$

$$z_3 = \frac{x_3 - \overline{x}}{S} = \frac{3-4}{2} = \frac{-1}{2} \quad \& \quad z_7 = \frac{x_5 - \overline{x}}{S} = \frac{7-4}{2} = \frac{3}{2}$$

Thus we find that the standardized values $-1, \frac{1}{2}, \frac{-1}{2}, \frac{-1}{2}, \frac{3}{2}$ have mean (we denote it by $\overline{z}$):

$$\overline{z} = \frac{-1 + 0.5 - 0.5 - 0.5 + 1.5}{5} = \frac{0}{5} = 0$$

And variance (we denote it by S_z^2):

$$\begin{aligned} S_z^2 &= \frac{1}{n-1} \sum_{i=1}^{n} \left(z_i - \overline{z}\right)^2 \\ &= \frac{(-1-0)^2 + (0.5-0)^2 + (-0.5-0)^2 + (-0.5-0)^2 + (1.5-0)^2}{4} \\ &= \frac{4}{4} = 1 \end{aligned}$$

Therefore, we get:

$$S_z = +\sqrt{S_z^2} = +\sqrt{1} = 1$$

▶ **EXAMPLE 1.5.8:** Let 2, 5, 7, 7, 8, 9, 9, 9, 7 and 0, 0, 1, 5, 1, 5, 7, 9, 3, 4 be degrees of students in two classes A and B respectively. Now let's see which of the students having 5, are the best in terms of level.

For this, we calculate the z-score for both students. To calculate the z - scores we must calculate the mean and the standard deviation for the two sets of data.

We find $\overline{x}_A = 7$, $S_A = 2.29$, $\overline{x}_B = 3.5$ and $S_B = 3.06$. So we get:

$$\begin{aligned} z_{5,A} &= \frac{x - \overline{x}_A}{S_A} \\ &= \frac{5-7}{2.29} = -0.873 \end{aligned}$$

$$\begin{aligned} z_{5,B} &= \frac{x - \overline{x}_B}{S_B} \\ &= \frac{5-3.5}{3.06} = 0.4902 \end{aligned}$$

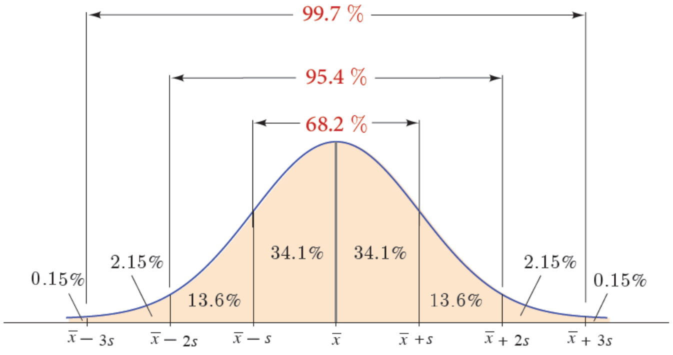
Therefore, we find $z_{5,A} < z_{5,B}$. This means that the student who has 5 in class B has a best level than the student who has 5 in class A.

THE EMPIRICAL RULE

If a data set has an approximately bell-shaped relative frequency histogram,

1. Approximately 68.2% of the data lie within one standard deviation of the mean, that is, in the interval with endpoints $\overline{x} \pm S$.
2. Approximately 95.4% of the data lie within two standard deviations of the mean, that is, in the interval with endpoints $\overline{x} \pm 2S$.
3. Approximately 99.7% of the data lies within three standard deviations of the mean, that is, in the interval with endpoints $\overline{x} \pm 3S$.

The following graph illustrates the concept of the empirical rule.



▶ **EXAMPLE 1.5.9:** Scores of some tests (as IQ test) have a bell-shaped distribution with mean $\mu = 100$ and standard deviation $S = 10$. Discuss what the Empirical Rule implies concerning individuals with scores of 110, 120, and 130.

Solution: The Empirical Rule states that:

1. Approximately 68.2% of the IQ scores in the population lie between 90 and 110,
2. Approximately 95.4% of the IQ scores in the population lie between 80 and 120, and,
3. Approximately 99.7% of the IQ scores in the population lie between 70 and 130.

▶ **EXAMPLE 1.5.10:** We will assume that apartment's prices in a certain Saudi city have a bell-shaped with mean 500000 (in S.R.) and standard deviation 100000. We will determine the price range for which at least 95.4% of the houses will sell.

The empirical rule states that 95.4% of the data values will fall within 2 standard deviations of the mean. Thus,

$$\begin{aligned}\left(\bar{x} - 2S, \bar{x} + 2S\right) &= \left(500000 - 2 \times 100000,\ \ 500000 + 2 \times 1000000\right)\\ &= \left(300000,\ \ 700000\right)\end{aligned}$$

Hence, at least 95.4% of all apartments sold will have a price range from SR 300000 to SR 700000.

DEFINITION 1.0.1 (Data)

Data is a collection of information collected by means of experiments, observations or real life events and stored in a proper format (the word data is derived from a Latin word ‘datum’).

DEFINITION 1.0.2 (Statistics)

Statistics is a branch of science deals with collection, organization, presentation, analysis, interpretation of data and take the appropriate decisions.

DEFINITION 1.1.1 (Descriptive Statistics)

Descriptive statistics consist of methods and techniques which are used for presenting and summarizing data in tables or graph forms and provide some numerical measures for it.

DEFINITION 1.1.2 (Population)

Population is a set of all things (which have at least one common characteristic (or feature)) that will be subjected to a study to obtain inferences for a specific problem. The elements of population are called individuals.

DEFINITION 1.1.3 (Sample)

A sample is a subset of population, which is used to collect information and to make inferences about the entire population.

DEFINITION 1.1.4 (Inferential Statistics)

Inferential statistics is some methods and techniques that can be used for drawing conclusions about the entire population using the observations from the samples taken from that population.

DEFINITION 1.1.5 (Parameter)

Parameter is a certain quantity or quality for describing a characteristic or phenomenon in a given population that summarizes the data for the entire population.

DEFINITION 1.1.6 (Statistic)

Statistic is a certain quantity or quality for describing a characteristic or phenomenon of a sample that summarizes the data for the entire sample.

DEFINITION 1.1.7 (Variables)

A variable is a map (or a function) X defined on the population (or sample) and takes values in an arbitrary set M. That means:

$$X : Population\ (or\ Sample) \longrightarrow \mathrm{M}$$

This variable measures: A characteristic, feature or factor (that varies from one individual to another) in the population.

DEFINITION 1.1.8 (Qualitative or Categorical Variable)

A qualitative variable is a variable that takes non-numeric values or numeric values which indicate an attribute or property.

DEFINITION 1.1.9 (Quantitative Variable)

A quantitative variable is a variable which takes numerical values, and these numerical values can be undergoing mathematical operations (or calculation operations), or it has a measurement unit. لها وحدة قياس وتقدر رتبتها تصاعدياً وتنازلياً.

DEFINITION 1.1.10 (Discrete Variable)

A discrete variable is a variable which takes finite or infinite countable number of values.

DEFINITION 1.1.11 (Continuous Variable)

A continuous variable is a variable which takes uncountable number of values.

DEFINITION 1.3.1 (Pie chart)

A pie chart is a simple way of representing the proportion of each class or category of data on a circular disk, so that each category is allocated a circular sector representing it.

DEFINITION 1.3.2 (Bar chart)

A bar chart is a representation of data of discrete variable with finite values (qualitative or quantitative). This is done through vertical or horizontal bars; so that it draws over each statement a bar with height (or length) equals to the frequency of that statement.

DEFINITION 1.3.4 (Multiple bar chart)

A multiple bar chart is a bar chart, where we can use it to represent multiple inter related variables by clustering bars side by side.

DEFINITION 1.3.5 (Component, or Stacked bar chart)

Component bar chart is a bar chart, where we can represent each component by a section in the bar, whose size is proportional to its contribution in the class.

DEFINITION 1.3.6 (Histogram)

A histogram is a graphical display used for data generated by continuous variables. It is a graph in which class boundaries are marked on the horizontal axis and the frequencies are marked on a vertical axis, and is constructed by drawing a rectangular column above each actual category so that its height equals the frequency of that category.

DEFINITION 1.3.7 (Skewedness)

Histograms are called as skewed if they are non-symmetric. In such histograms, bins on one side have high frequency which decreases as we move to the other side. The side with lower frequency is said to have a longer tail.

DEFINITION 1.3.8 (Polygon)

The frequency polygon is a polygon which connects with a straight line the points (x_i, f_i), whereas x_i and f_i are the midpoint and the frequency of class boundary i respectively, and it closes from the left to the middle of a default class boundary located before the first class boundary, and from the right to the middle of a default class boundary located after the last class boundary.

DEFINITION 1.3.10 (Descending Cumulative Frequency Polygon (DCFP)):

The descending cumulative frequency polygon (DCFP) is a polygon which connects with a straight line the points (b_i, Φ_i), whereas b_i and Φ_i are the lower bound and the descending cumulative frequency of class boundary i respectively, and closes from the right to the end of the last class boundary.

Advantages of The Mean

- It is quick and easy to compute.
- All values are considered by calculating the mean.
- It is one and only one value for a set of data.

Disadvantages of The Mean

- Mean is not defined for qualitative data.
- Since it considers all the observed values, it is highly affected by the extreme values.
- It becomes not applicable if a data is lost.

DEFINITION 1.4.3 (Median)

Median (we denote it by $\tilde{x}$) is that value which divides the data in two halves after ordering them, in ascending or descending order.

Advantages of The Median

- It is easy to compute and understand.
- It is not affected by outliers or extreme values.
- It can be used even if you loss some data (known argument) that is not in the middle.

Disadvantages of The Median

- It does not take all values into account.
- It is not used in many statistical tests.
- It cannot be identified for qualitative data

DEFINITION 1.4.4 (The Mode)

The mode (we denote it by $\hat{x}$) of data is a value or observation, which has the highest frequency.

Advantages of The Mode

- It is quick and easy to compute.
- It can be evaluated for both quantitative and qualitative data.
- It is not affected by extreme values.

Disadvantages of The Mode

- There may be more than one mode for a certain data set.
- Sometimes, there is no mode for a given data set.
- It may not reflect the center of the distribution very well.

DEFINITION 1.4.8 (Extreme Value)

We say that a value x of given data is said to be extreme if one of the following relations is realizing:

$$x < LF := \mathrm{Q}_1 - 1.5\,(\mathrm{Q}_3 - \mathrm{Q}_1) \quad [LF \text{ is the abbreviation of “Lower Fence”}]$$

or

$$x > HF := \mathrm{Q}_3 + 1.5\,(\mathrm{Q}_3 - \mathrm{Q}_1) \quad [HF \text{ is the abbreviation of “Higher Fence”}]$$

DEFINITION 1.4.9 (Five Numbers)

Five Numbers are a summary of the variable data which includes the below mentioned five characteristics:

$\overset{x_s}{\text{Smallest}}$ value, Q_1, Q_2, Q_3 and $\overset{x_l}{\text{Largets}}$ value

DEFINITION 1.4.10 (Box Plot)

The box plot of given data is the graphical representation of its five numbers summary.

1

1.3

organisation for raw data

group data → frequency disterbution table.

For?

continuous quantitative data

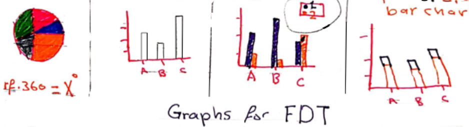
ungrouped data → frequency table

For?

1-qualitative data
2-discrete quantitative data

graphs for FT

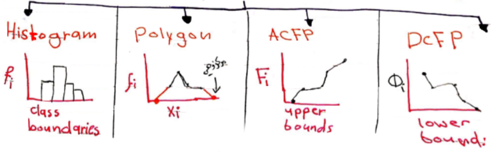
Pie chart | Bar chart | multiblechart | compnent or stacked bar chart



$rf\cdot 360 = X^\circ$

Graphs for FDT

Histogram | Polygon | ACFP | DcFP



2

1.3

Types of histogram.

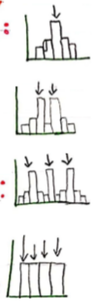
متناظر Symmatric

1) Uni modal:

2) Bimodal:

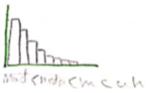
3) Multimodal:

4) uniform:

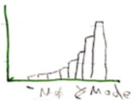


ملتوي skewed

1) Right skewed histogram



2) Left skewed histogram



3

1.4

Measures of central tendency

raw data

mean:

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

median:

n

زوجي

$\tilde{X} = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}$

فردي

$\tilde{X} = X_{\frac{n+1}{2}}$

mode:

highest f_i

frequency T.

Mean:

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

Median

① $n = \sum f_i$

even $\frac{n}{2}$ or odd $\frac{n+1}{2}$

$F_i \geq \square$

② تحديد الميديان كلاس

③ الـ median = قيمة الميديان كلاس

Mode:

highest f_i

Frequenx distribution T.

Mean:

$\bar{X} = \frac{1}{\sum f_i}\cdot\sum_{i=1}^{k} f_i X_i$

midpoint

Median

① $X = \frac{\sum f_i}{2}$

② Median class: ACF $\geq X$

③ $\tilde{X} = \tilde{L} + \frac{\frac{\sum f_i}{2} - (\tilde{F} - \tilde{f})}{\tilde{f}}\cdot C$

$\tilde{L}$ = lower bound
$\tilde{F}$ = ACF for medianclass
$\tilde{f}$ = frequency of median class

mode

① تحديد الـ Modal class تكرارها أكبر

② ونحدد تكراري الفئة التي تسبقها وتليها.

③ $\hat{X} = \hat{L} + \frac{d_1}{d_1 + d_2}\cdot C$

$\hat{L}$: lower bound
d_1: f modal class − f_i pervious
d_2: f modal class − f_i after

4

Measures of position

* اول خطوة فيهم ترتيب الداتا

percentiles	Deciles	Quartiles
1) Rank:	1) Rank:	1) Rank
$p_r = \frac{r(n+1)}{100} = k.s$	$d_r = \frac{r(n+1)}{10} = k.s$	$q_r = \frac{r(n+1)}{4} = k.s$
2) percentile:	2) ~~per~~ decile:	2) Quantile:
$P_r = X_k + s(X_{k+1} - X_k)$	$D_r = X_k + s(X_{k+1} - X_k)$	$Q_r = X_k + s(X_{k+1} - X_k)$

$P_{25} = Q_1$, $P_{75} = Q_3$, $P_{50} = Q_2 = D_5 =$ median

Steps of a BoxPlot:

١) ايجاد الأعداد الخمسة $X_S\ Q_1\ Q_2\ Q_3\ X_L$

٢) حساب HF و LF لإيجاد (الإكستريم فاليوز) or (outliners) نفس المعنى

٣) تشكل خط افقي بتقسيم مناسب

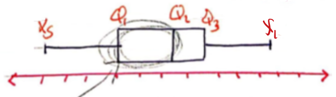
٤) تحديد الأعداد الثلاثة $Q_1\ Q_2\ Q_3$ ← نضع نقطة ونكتب اسم العدد

٥) نرسم صندوق مغلق من Q_1 الى Q_3 ونضع خط عمود على كل Q

٦) نهاية السواعد

توجد قيم متطرفة: نمد السواعد إلى الحاجز (الـ HF أو LF)

لا توجد قيم متطرفة: نمد السواعد الى اصغر واكبر قيمة



left skewed

5

Distribution table steps:

1. find range: $R = X_L - X_S$
2. find length of class boundry $C = \frac{R+1}{K}$ → given / → calculate $[3.322 \log n]$
3. length of class limit = C−1
4. draw the table
5. First class limit = X_S + length of class limit and so ...
6. class boundray: lower limit −0.5 → upper limit +0.5
7. midpoint = $\frac{\text{lower limit} + \text{upper limit}}{2}$ or $\frac{\text{lower bond} + \text{upper bound}}{2}$

$\frac{f_i}{\sum f_i}$ F_i Φ_i

class limit	class boundary	mid point	f_i	rf	P.rf	ACF	DCF

Extrame value

$x < LF := Q_1 - 1.5\,(Q_3 - Q_1)$

$x \to HF := Q_3 + 1.5\,(Q_3 - Q_1)$

$IQR = Q_3 - Q_1$ | $SK = \frac{\overline{X} - \hat{x}}{5}$

The value of SK is discussed as follows:

a. If $SK > 0$, then the distribution of data is right skewed.
b. If $SK = 0$, then the distribution of data is symmetric.
c. If $SK < 0$, then the distribution of data is left skewed.

$S^2 = \frac{\sum(\bar{x} - x_i)^2}{n-1}$ | $S = \sqrt{s^2}$ | $CV = \frac{S}{\overline{X}} \times 100\%$ | $Z_i = \frac{X_i - \overline{X}}{S}$

Chapter 2: Probability

DEFINITION 2.1.2 (Permutations)

Any an arrangement of r distinct objects, from a set of $n \in \mathbb{N}$ different objects called permutation.

$$\boxed{n\,\mathrm{Pr} = \frac{n\,!}{(n-r)\,!} \qquad ;\, 0 \le r \le n}$$

▶ **EXAMPLE 2.1.2** How many ways one can arrange in order any three of the 8 letters of the alphabet l, m, o, p, q, r, s, t .

Solution: The permutation formula gives:

$$_8\mathrm{P}_3 = \frac{8\,!}{(8-3)\,!} = \frac{8\,!}{5\,!} = \frac{8 \times 7 \times 6 \times (5\,!)}{5\,!} = 8 \times 7 \times 6 = 336.$$

DEFINITION 2.1.3 (Combinations)

Any an unordered group of r distinct objects, from a set of n different objects is called combination. We use the notation $n\mathcal{C}r$ to represent the total number of different combinations of size r that can be selected from n distinct objects and read as "n chosen r".

$$\boxed{n\mathcal{C}r = \frac{n\,!}{r\,!\;(n-r)\,!} \qquad ;\, 0 \le r \le n}$$

▶ **EXAMPLE 2.1.4** How many different unordered groups of any three of the six letters l, m, n, o, p and q ?

Solution: With $n = 6$ and $r = 3$, one has:

$$_6\mathcal{C}_3 = \frac{6\,!}{3\,!\;(6-3)\,!} = \frac{6 \times 5 \times 4 \times (3\,!)}{3 \times 2 \times (3\,!)} = 20$$

DEFINITION 2.1.4 (Cardinal Number of a Set)

Let Ω be a given set, then $|\Omega|$ denote the number of all elements in Ω, and this number is called the cardinal number of Ω.

DEFINITION 2.2.1 (Probability Science)

Probability Science is a branch of mathematics that deals with theoretical mathematical models of random experiments.

DEFINITION 2.2.2 (Space of Elementary Events)

Suppose that we have a random experiment. Then the set of all possible results of this random experiment is called the space of elementary events, and denoted by Ω.

DEFINITION 2.2.3 (Discrete Space)

If a space of elementary events Ω is either finite or countable infinite, then it is called a discrete space.

DEFINITION 2.2.4 (Continuous Space)

If the space of elementary events Ω consists uncountable number of outcomes, then it is called a continuous space.

DEFINITION 2.2.5 (Algebra of Events)

Suppose that we have a random experiment with a space of elementary events Ω. Then a collection $\mathscr{A}$ of subset of Ω is said to be an algebra on Ω if and only if the following condition are verified:

1. $\Omega \in \mathscr{A}$.

2. For any two elements A and $B \in \mathscr{A}$. Then $A \cup B \in \mathscr{A}$.

3. For any element $A \in \mathscr{A}$. Then $\overline{A} \in \mathscr{A}$.

DEFINITION 2.2.6 (Simple Event and Compound Event)

An event A is said to be a simple event if contains only one elementary event (one outcome).

An event A is said to be a compound event if contains at least two elementary events (at least two outcomes).

DEFINITION 2.2.7 (Impossible Event)

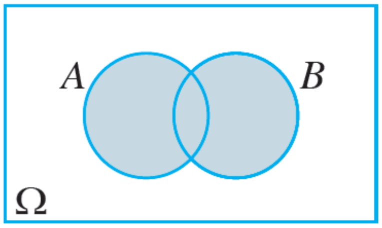
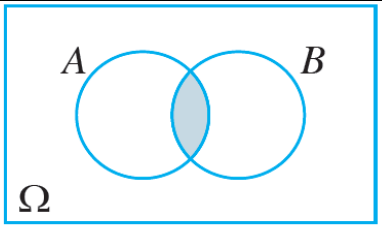
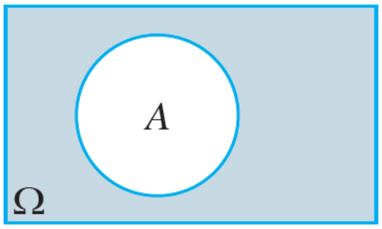
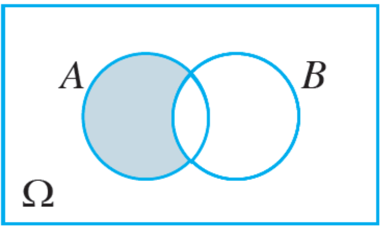
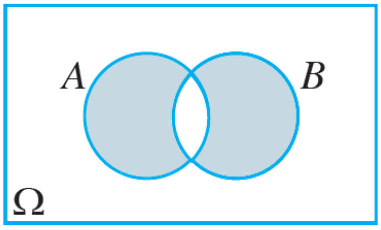
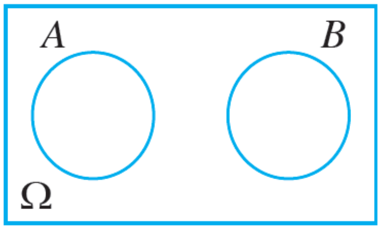
For an event $A \in \mathscr{A}$ we know that it is impossible an outcome of the experiment belong to $A \cap \overline{A}$. Therefore, the event $A \cap \overline{A}$ is called an impossible event, and since in set theory $A \cap \overline{A} = \varnothing$, so one denotes the impossible event by $\varnothing$ also.

DEFINITION 2.2.8 (Certain Event)

For an event $A \in \mathscr{A}$ we know that it is surely an outcome of the experiment belongs to $A \cup \overline{A}$. Therefore, the event $A \cup \overline{A}$ is called a certain event, and since in set theory $A \cup \overline{A} = \Omega$, so one denotes the certain event by Ω also.

DEFINITION 2.2.9 (Mutually Exclusive Events)

Two events A and B of $\mathscr{A}$ are mutually exclusive if they cannot occur at the same time.

$A \cup B$	
$A \cap B$	
$\bar{A}$	
$A \cap \bar{B}$ $A \backslash B$	
$(A \cap \bar{B}) \cup (\bar{A} \cap B)$ $A \Delta B$	
$A \cap B = \emptyset$	

DEFINITION 2.2.10 (Pair Wise Mutually Exclusive Event)

Events $A_1, A_2, A_3, \ldots$ of $\mathscr{A}$ are said to be pair wise mutually exclusive if:

$$A_i \cap A_j = \varnothing \qquad ; \forall\, i \neq j$$

DEFINITION 2.3.1 (Probability)

Probability is a numerical measure of the likelihood that a specific event will occur.

DEFINITION 2.3.2 (Relative Frequency of Event)

If $n(A)$ represents the number of times (trials) that event A occurs among N trials of a given experiment, then $f_A = \dfrac{n(A)}{N}$ represents the relative frequency of occurrence of A on these trials.

DEFINITION 2.3.3 (Probability Measure)

Let Ω be a space of elementary events of a random experiment, and $\mathscr{A} \subseteq 2^{\Omega}$ is a σ-algebra on Ω. Furthermore, we suppose that P a real set function on $\mathscr{A}$ with the following properties:

a. We have $P(\varnothing) = 0$.

b. For any sequence $A_1, A_2, \ldots, A_n, \ldots \in \mathscr{A}$ with $A_i \underset{i \neq j}{\cap} A_j = \varnothing$, then:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

c. We have $P(\Omega) = 1$.

Then one say that P a probability measure (or a probability function).

THEOREM 2.3.1

For a random experiment with probability space $[\Omega, \mathscr{A}, P]$, then:

a. For any event A of $\mathscr{A}$ we have $P(\overline{A}) = 1 - P(A)$.

b. For any two events A and B of $\mathscr{A}$ with $A \subset B$, we get $P(B \setminus A) = P(B) - P(A)$.

c. For any two events A and B of $\mathscr{A}$ with $A \subset B$, then $P(A) \leq P(B)$ (monotonicity property of the measure P).

d. For any event A of $\mathscr{A}$ follows from the monotonicity property that:

$$P(\varnothing) = 0 \leq P(A) \leq 1 = P(\Omega)$$

THEOREM 2.3.2

Let $[\Omega, \mathscr{A}, P]$ is a probability space of a random experiment. Now, if Ω is finite (we suppose $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$), then we can calculate the probability of any event $A \in \mathscr{A}$ by the following relation:

$$P(A) = \sum_{i\,;\omega_i \in A} P(\{\omega_i\})$$

$$\boxed{P(A \mid B) = \frac{P(A \cap B)}{P(B)}}$$

DEFINITION 2.4.3 (Statistical Independency of Events)

Let $[\Omega, \mathscr{A}, P]$ be a probability space of a random experiment, and $A_1, A_2, ..., A_n \in \mathscr{A}$ are given events. Then $A_1, A_2, ..., A_n$ said to be statistically independent if and only if for any permutation $(i_1\ i_2\ \dots\ i_k)$ with $k \in \{2, 3, \dots, n\}$ of the numbers $1, 2, \dots, n$ we have:

$$P\left(A_{i_1} \cap A_{i_2} \cap ... \cap A_{i_k}\right) = P\left(A_{i_1}\right) \cdot P\left(A_{i_2}\right) \cdot ... \cdot P\left(A_{i_k}\right)$$

If the last relation satisfy for the special case $k = 2$, then one say that the events $A_1, A_2, ..., A_n$ are pair wise independent.

	Probabilities	Verbal Description
Mutually Exclusive	$P(A \cap B) = 0$	Both events cannot happen
Independent	$P(A \mid B) = P(A)$ $P(A \cap B) = P(A) \cdot P(B)$	The occurrence of B does not affect the occurrence of A

TOTAL PROBABILITY

DEFINITION 2.4.1

Let $[\Omega, \mathscr{A}, P]$ be a probability space of a random experiment. Then events $Z_1, Z_2, ..., Z_n$ of $\mathscr{A}$ are a partition of Ω, if:

a) $Z_i \neq \varnothing$ for all i

b) $Z_i \cap Z_j = \varnothing$ for all $i \neq j$

c) $\bigcup_{i=1}^{n} Z_i = \Omega$

THEOREM 2.4.1 (Total Probability Formula)

Let $[\Omega, \mathscr{A}, P]$ be a probability space of a random experiment. If $Z_1, Z_2, \ldots, Z_n$ are events of $\mathscr{A}$, constitute a partition of the space of elementary events Ω such that $P(Z_k) \neq 0$ for $k = 1,2,\ldots,n$, then for any $B \in \mathscr{A}$ we have:

$$P(B) = \sum_{k=1}^{n} P(Z_k)\, P(B|Z_k)$$

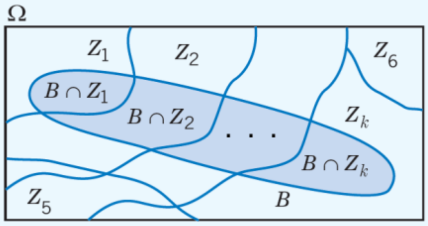


Figure 2.4.2

THEOREM 2.4.2 (Bayes' Theorem)

Let $[\Omega, \mathscr{A}, P]$ be a probability space of a random experiment. If $Z_1, Z_2, \ldots, Z_n \in \mathscr{A}$ are a partition of a space of elementary events Ω such that $P(Z_k) \neq 0$ for $k = 1,2,\ldots,n$, then for any $B \in \mathscr{A}$ such that $P(B) \neq 0$:

$$P(Z_i \mid B) = \frac{P(Z_i)\ P(B \mid Z_i)}{\sum_{k=1}^{n} P(Z_k)\, P(B \mid Z_k)} \qquad ; i = 1,2,\ldots,n$$

$$\Omega = \begin{Bmatrix} (1,1), (1,2), (1,3), (1,4), (1,5), (1,6), \\ (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), \\ (3,1), (3,2), (3,3), (3,4), (3,5), (3,6), \\ (4,1), (4,2), (4,3), (4,4), (4,5), (4,6), \\ (5,1), (5,2), (5,3), (5,4), (5,5), (5,6), \\ (6,1), (6,2), (6,3), (6,4), (6,5), (6,6) \end{Bmatrix}$$

Question 6: We roll two identical fair dice (at the same time) and we take the summation of numbers of outcomes. Then:

a) Determine the probability space of this random experiment.

………

………

………

b) Calculate the probability that the sum is an even number.

………

………

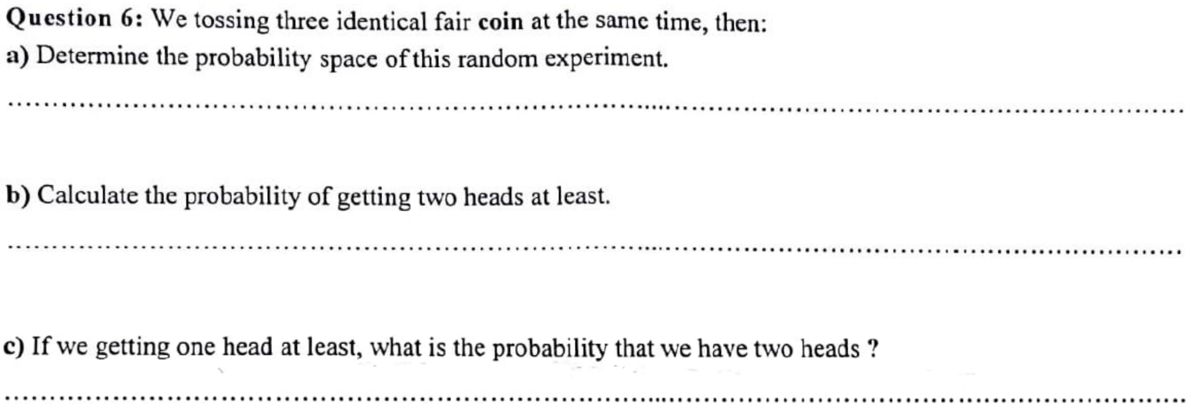
Question 7: We roll a fair die two times and multiply the apparent numbers. Then:

a) Determine the probability space of this random experiment.

………

b) Calculate the probability that the product is a multiple of **5**.

………



Question 6: We tossing three identical fair **coin** at the same time, then:

a) Determine the probability space of this random experiment.

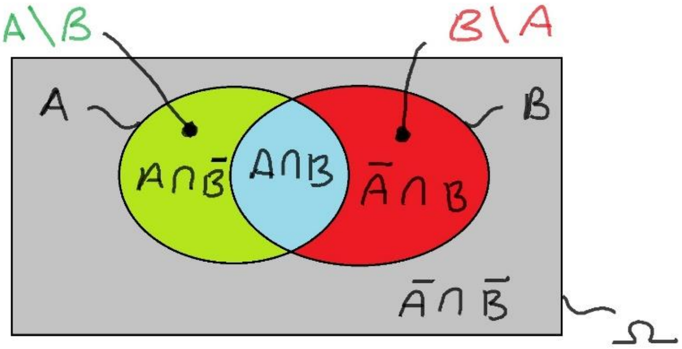
………

b) Calculate the probability of getting two heads at least.

………

c) If we getting one head at least, what is the probability that we have two heads ?

………



1. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
2. $P(A \cap B) = 0$ *if A and B mutually exclusive.*
3. $P(A \cap B) = P(A) \times P(B)$ *if A and B independent*
4. $P(A|B) = \frac{P(A\cap B)}{P(B)}$
5. $P(\bar{A}) = 1 - P(A)$

Question 5: Let $[\Omega, \mathscr{A}, P]$ be a probability space, and we consider A and B two events of $\mathscr{A}$ with $P(A) = \mathbf{0.45}$, $P(A \setminus B) = \mathbf{0.25}$ and $P(A \cup B) = \mathbf{0.55}$. Then calculate the following probabilities:

$P(A \cap B) =$ ……………………………………………………………………………………………

$P(B) =$ ……………………………………………………………………………………………

$P(B \setminus A) =$ ……………………………………………………………………………………………

$P(\bar{A} \cup \bar{B}) =$ ……………………………………………………………………………………………

$P(\Omega | B) =$ ……………………………………………………………………………………………

Question 5: Let $[\Omega, \mathscr{A}, P]$ be a probability space, and A and B are two mutually exclusive events of $\mathscr{A}$ with $P(A \setminus B) = \mathbf{0.30}$ and $P(B \setminus A) = \mathbf{0.25}$. Then:

a) Calculate the following probabilities:

$P(A) =$ ……………………………………………………………………………………………

……………………………………………………………………………………………

$P(B) =$ ……………………………………………………………………………………………

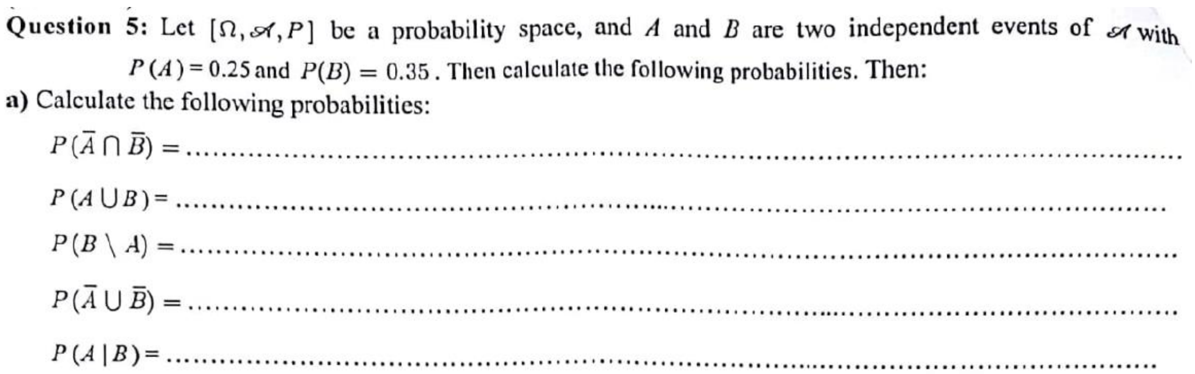
$P(A \cup B) =$ ……………………………………………………………………………………………

$P(\bar{A} \cap \bar{B}) =$ ……………………………………………………………………………………………

$P(B | A) =$ ……………………………………………………………………………………………

b) Are the events A and B independent **?**

……………………………………………………………………………………………



Question 5: Let $[\Omega, \mathscr{A}, P]$ be a probability space, and A and B are two independent events of $\mathscr{A}$ with $P(A) = 0.25$ and $P(B) = 0.35$. Then calculate the following probabilities. Then:

a) Calculate the following probabilities:

$P(\bar{A} \cap \bar{B}) =$ ……………………………………………………………………………………

$P(A \cup B) =$ ……………………………………………………………………………………

$P(B \setminus A) =$ ……………………………………………………………………………………

$P(\bar{A} \cup \bar{B}) =$ ……………………………………………………………………………………

$P(A \mid B) =$ ……………………………………………………………………………………

Question 7: Five publishing houses P_1, P_2, P_3, P_4 and P_5 have to print a dictionary. Each of them produces the same number of dictionaries. The percentage of incomplete dictionaries in these presses are **2%**, **3%**, **5%**, **4%** and **1%** respectively. We withdraw a dictionary randomly from the total production of these presses. Then:

a) What is the probability that the dictionary is complete?

………

b) If we have found that, the drawn dictionary is incomplete, what is the probability that this dictionary will be printed by the publishing house P_5 **?**

………

Question 6: In a factory for the production of switches there are **4** lines that produce **15%, 25%, 20%** and **40%**. The percentage of defected switches in the production of these lines are **3%, 2%, 3%** and **1%** respectively. We draw a switch randomly of the total production of the factory, then:

a) What is the probability that the switch is defected **?**

………

b) If we found that the switch is defected, what is the probability that this switch produced by the second line **?**

………

1

Counting Rule

oR → $n_1 + n_2 + n_3$

and → $n_1 \cdot n_2 \cdot n_3$

$${}_nP_r = \frac{n!}{(n-r)!} \qquad {}_nC_r = \frac{n!}{(n-r)!\,r!}$$

$|\Omega|$ = number of all elements

to be Algebra: $\mathcal{A}$

1. $\Omega \in \mathcal{A}$
2. $A, B \in \mathcal{A}$, then $A \cup B \in \mathcal{A}$
3. $A \in \mathcal{A}$, $\bar{A} \in \mathcal{A}$

to be σ-algebra $\mathcal{A}$:

$A_1, A_2, \ldots A_n \in \mathcal{A}$

$\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$

$2^{|\Omega|}$ = number of subsets

Mutually exclusive A, B → $A \cap B = \phi$, $P(A \cap B) = 0$

Pair wise Mutually Exclusi

$A_i \cap A_j = \phi$

RF = relative Frequency

$$P(A) = \frac{n(A)}{N}$$

$P(\bar{A}) = 1 - P(A)$

Classical

- $P(A) = \frac{|A|}{|\Omega|}$ or $\frac{1}{|\Omega|}$
  - $\frac{|A|}{|\Omega|}$ → Compound
  - $\frac{1}{|\Omega|}$ → Simple

$P(\bar{A} \cap \bar{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B)$

$P(\bar{A} \cup \bar{B}) = P(\overline{A \cap B}) = 1 - P(A \cap B)$

$P(\bar{A} \cap B) = P(B) - P(A \cap B) = P(B \setminus A)$

$P((A \cap \bar{B}) \cup (\bar{A} \cap B)) = P(A \cap \bar{B}) + P(\bar{A} \cap B)$
→ only A or B

$A \cup \bar{A} = \Omega$

$A \cap \bar{A} = \phi$

$A \cup \Omega = \Omega$

$A \cap \Omega = A$

2

∪ Additive Rule

oR
either, one of them : Key words

Not Mutually exclusive الحالة العامة

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ → For two events

$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$
→ For three events

Mutually exclusive

$P(A \cup B) = P(A) + P(B)$

$P(A \cup B \cup C) = P(A) + P(B) + P(C)$ ، A, B, C Pair wise Mutually exclusive

∩ Multiplication Rule
Conditional Rule

Key word: and, given that, both of them, Neither

dependent = Not independent

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad P(B|A) = \frac{P(A \cap B)}{P(A)}$$

$P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$

Independent $P(A|B) = P(A)$ $P(B|A) = P(B)$

$P(A \cap B) = P(A) \cdot P(B)$

3

- $P(A_1 \cup A_2 | B) = P(A_1|B) + P(A_2|B)$
- if $B \subseteq C$ → $P(C|B) = 1$
- $P(\bar{A}|B) = 1 - P(A|B)$
- $P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \ldots$
- total Probability formula:

$$P(B) = P(B|Z_1) \cdot P(Z_1) + P(B|Z_2) \cdot P(Z_2) + \ldots$$
$$= \sum P(B|Z_i) \cdot P(Z_i)$$

Bayes' theorem:
$$P(Z_i|B) = \frac{P(Z_i)\,P(B|Z_i)}{\sum P(Z_k) \cdot P(B|Z_k)}$$

To be Partition:

1. $Z_i \neq 0$
2. $Z_i \cap Z_j = \phi$
3. $\bigcup_{i=1}^{\infty} Z_i = \Omega$

T. Hadeer Ahmed
0557265401
منصة بيمز