محاضرات

١٠٩ احص

احصاء حيوي

١٤٣٨ -١٤٣٩هـ

محـــاضـــرات

١٠٩ احص

احصاء حيوي

قسم الإحصاء وبحوث العمليات ـكلية العلوم
جامعة الملك سعود
الفصل الدراسي الأول ١٤٣٨/١٤٣٩
١٠٩ـاحص

Stat ١٠٩

OfficRoom No. 2A03
email: Sabunasrah@ksu.edu.sa
building (25)

Week	Title
1	Introduction to Bio-Statistics, (1.1-1.4)
2	types of data and graphical representation, (1.1-1.4)
3	Descriptive statistics: Measures of Central tendency- Mean , median, mode (2.1-2.6 Excluding stem plot percentiles)
3+4	Measures of dispersion-Range, Standard deviation, coefficient of Variation. (2.1-2.6 Excluding stem plot percentiles)
4+5	Calculating Measures from an Ungrouped Frequency Table (2.1-2.6 Excluding stem plot percentiles)
5+6	Basic probability. Conditional probability, concept of independence, sensitivity, specificity….. (3.1-3.6)
6	Bayes Theorem for predictive probabilities. (3.1-3.6)
(TEST No. 1))	
7	Some discrete probability distributions: cumulative probability (4.1-4.4)
8	Binomial, and Poisson -their mean and variance (4.1-4.4Excluding the use of binomial and Poisson tables).
9	Continuous probability distributions: Normal distribution-Z-table (4.5-4.8)
10	Sampling with and without replacement, sampling distribution of one and two sample means and one and two proportions. (5.1-5.7 Excluding sampling without Replacement)
(TEST No. 2)	
11	Sampling with and without replacement, sampling distribution of one and two sample means and one and two proportions. (5.1-5.7 Excluding sampling without Replacement)
12	Statistical inference: Point and interval estimation, Type of errors, Concept of P-value (6.2-6.6. 7.1-7.6 Excluding Variances not equal page 181-182)
13	Testing hypothesis about one and two samples means and proportions including paired data – different cases under normality. (6.2-6.6. 7.1-7.6 Excluding Variances not equal page 181-182)
14	Testing hypothesis about one and two samples means and proportions including paired data – different cases under normality. (6.2-6.6. 7.1-7.6 Excluding Variances not equal page 181-182)
Text Book	**Biostatistics: Basic Concepts and Methodology for the Heath Sciences by Wayne W. Daniel. [9th ed.] Books available from university book store below SAMBA bank. The book costs 70 Riyals for students.**

- **مواعيد الامتحانات تحدد لاحقا بالاتفاق مع الكلية وتعلن على موقع السنة التحضيرية ومواقع المدرسين**

CHAPTER 1: Getting Acquainted with Biostatistics

1.1 Introduction:

The course "Biostatistics" (STAT-145) is about information; how it is obtained, how it is analyzed, and how it is interpreted.

The objective of the course is to learn:

(1) How to organize, summarize, and describe data.
(Descriptive Statistics)

(2) How to reach decisions about a large body of data by examine only a small part of the data.
(Inferential Statistics)

1.2 Some Basic Concepts:

Data:

Data is the raw material of statistics. There are two types of data:

(1) Quantitative data
(numbers: weights, ages, …).

(2) Qualitative data
(words or attributes: nationalities, occupations, …).

Statistics:

Statistics is the field of study concerned with:

(1) The collection, organization, summarization, and analysis of data. (Descriptive Statistics)

(2) The drawing of inferences and conclusions about a body of data (population) when only a part of the data (sample) is observed. (Inferential Statistics)

Biostatistics:

When the data is obtained from the biological sciences and medicine, we use the term "biostatistics".

Sources of Data:

1. Routinely kept records.
2. Surveys.
3. Experiments.
4. External sources. (published reports, data bank, ...)

Population:

- A population is the largest collection of entities (elements or individuals) in which we are interested at a particular time and about which we want to draw some conclusions.
- When we take a measurement of some variable on each of the entities in a population, we generate a population of values of that variable.
- Example: If we are interested in the weights of students enrolled in the college of engineering at KSU, then our population consists of the weights of all of these students, and our variable of interest is the weight.

Population Size (N):

The number of elements in the population is called the population size and is denoted by N.

Sample:

- A sample is a part of a population.
- From the population, we select various elements on which we collect our data. This part of the population on which we collect data is called the sample.
- Example: Suppose that we are interested in studying the characteristics of the weights of the students enrolled in the college of engineering at KSU. If we randomly select 50 students among the students of the college of engineering at KSU and measure their weights, then the weights of these 50 students form our sample.

Sample Size (n):

The number of elements in the sample is called the sample

size and is denoted by n.

Variables:

The characteristic to be measured on the elements is called variable. The value of the variable varies from element to element.

Example of Variables:

(1) No. of patients (2) Height
(3) Sex (4) Educational Level

Types of Variables:

(1) Quantitative Variables:

A quantitative variable is a characteristic that can be measured. The values of a quantitative variable are numbers indicating how much or how many of something.

Examples:

(i) Family Size (ii) No. of patients
(iii) Weight (iv) height

Types of Quantitative Variables:

(a) Discrete Variables:

There are jumps or gaps between the values.

Examples: - Family size ($x = 1, 2, 3, \ldots$)
- Number of patients ($x = 0, 1, 2, 3, \ldots$)

(b) Continuous Variables:

There are no gaps between the values.

A continuous variable can have any value within a certain interval of values.

Examples: - Height ($140 < x < 190$)
- Blood sugar level ($10 < x < 15$)

(2) Qualitative Variables:

The values of a qualitative variable are words or attributes indicating to which category an element belong.

Examples:

- Blood type - Nationality
- Students Grades - Educational level

Types of Qualitative Variables:

(a) Nominal Qualitative Variables:

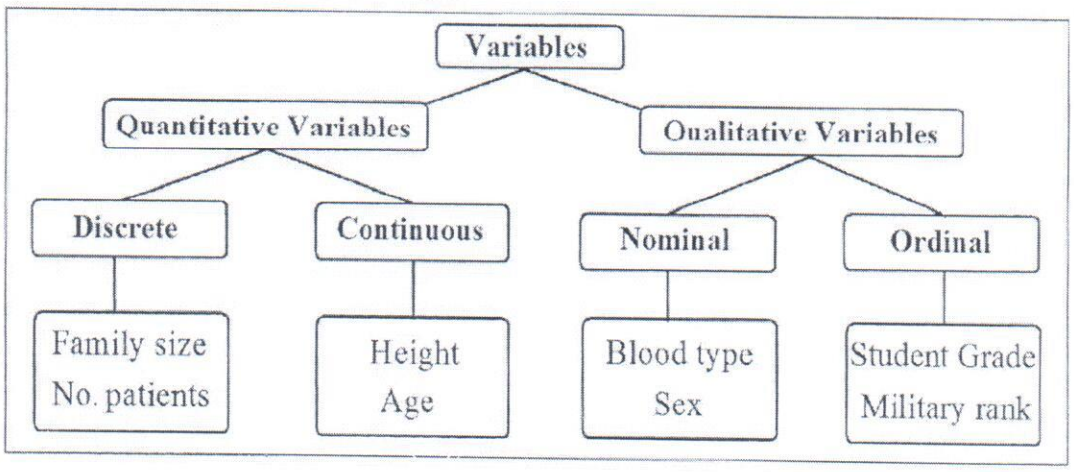
A nominal variable classifies the observations into various mutually exclusive and collectively non-ranked categories. The values of a nominal variable are names or attributes that can not be ordered or sorted or ranked.

Examples: - Blood type (O, AB, A, B)
- Nationality (Saudi, Egyptian, British, …)
- Sex (male, female)

(b) Ordinal Qualitative Variables:

An ordinal variable classifies the observations into various mutually exclusive and collectively ranked categories. The values of an ordinal variable are categories that can be ordered, sorted, or ranked by some criterion.

Examples: - Educational level (elementary, intermediate, …)
- Students grade (A, B, C, D, F)
- Military rank



1.4 Sampling and Statistical Inference:

There are several types of sampling techniques, some of which are:

(1) Simple Random Sampling:

If a sample of size (n) is selected from a population of size (N) in such a way that each element in the population has the same chance to be selected, the sample is called a simple random sample.

(2) Stratified Random Sampling:

In this type of sampling, the elements of the population are classified into several homogenous groups (strata). From each group, an independent simple random sample is drawn. The sample resulting from combining these samples is called a stratified random Sample.

Exercises:

Q6: For each of the following variables indicate whether it is quantitative or qualitative variable:

(a)The blood type of some patient in the hospital.(Qualitative nominal)

(b) Blood pressure level of a patient.

.(Qualitative ordinal)

(c) Weights of babies born in a hospital during a year. Quantitative continues

(d) Gender of babies born in a hospital during a year. Qualitative nominal

(e) The distance between the hospital to the house Quantitative continues

(f) Under-arm temperature of day-old infants born in a hospital. Quantitative continues

Q7: For each of the following situations, answer questions (a) through (d):

(a) What is the population?

(b) What is the sample in the study?

(c) What is the variable of interest?

(d) What is the type of the variable?

Situation A: A study of 300 households in a small southern town revealed that if she has school-age child present.

(a) Population: All households in a small southern town.

(b) Sample: 300 households in a small southern town.

(c) Variable: Does households had school age child present.

(d) Variable is qualitative nominal.

- Situation B: A study of 250 patients admitted to a hospital during the past year revealed that, Distance the patient live away from the hospital

(a) Population: All patients admitted to a hospital during the past year.

(b) Sample: 250 patients admitted to a hospital during the past year.

(c) Variable: Distance the patient live away from the hospital

(d) Type of variable: Variable is Quantitative continuous.

Choose the right answer:

1-The variable is a

a. subset of the population.

b. parameter of the population.

c. relative frequency.

d. characteristic of the population to be measured.

e. class interval.

2-Which of the following is an example of discrete variable

a. the number of students taking statistics in this term at KSU..

b. the time to exercise daily.

c. whether or not someone has a disease

d. height of certain buildings

e. Level of education

3.Which of the following is not an example of discrete variable

a. the number of students at the class of statistics.

b. the number of times a child cry in a certain street.

c. the time to run a certain distance.

d. the number of buildings in a certain street.

e. number of educated persons in a family.

4.Which of the following is an example of qualitative variable

a. the blood pressure.

b. the number of times a child brush his/her teeth.

c. whether or not someone fail in an exam.

d. Weight of babies at birth.

e. the time to run a certain distance.

5.The continuous variable is a

a. variable with a specific number of values.

b. variable which cant be measured.

c. variable takes on values within intervals.

d. variable with no mode.

e. qualitative variable.

6. which of the following is an example of continuous variable

a. The number of visitors of the clinic yesterday.

b. The time to finish the exam.

c. The number of patients suffering from certain disease.

d. Whether or not the answer is true.

7. The discrete variable is

a-qualitative variable.

b-variable takes on values within interval.

C-variable with a specific number of values.

d-variable with no mode.

8-Which of the following is an example of nominal variable :

a-age of visitors of a clinic.

b-The time to finish the exam.

c-Whether or not a person is infected by influenza.

d-Weight for a sample of girls .

9-The nominal variable is a

a-A variable with a specific number of values

b-Qualitative variable that can't be ordered.

c-variable takes on values within interval.

d-Quantitative variable .

10-Which of the following is an example of nominal variable :

a-The number of persons who are injured in accident.

b-The time to finish the exam.

c-Whether or not the medicine is effective.

d-Socio-economic level.

11-The ordinal variable is :

a-variable with a specific number of values.

b-variable takes on values within interval.

c-Qualitative variable that can be ordered.

d-Variable that has more than mode.

CHAPTER 2: Strategies for Understanding the Meaning of Data:

2.1 Introduction:

In this chapter, we learn several techniques for organizing and summarizing data so that we may more easily determine what information they contain. Summarization techniques involve:

- frequency distributions
- descriptive measures

2.2 The Ordered Array:

A first step in organizing data is the preparation of an ordered array.

An ordered array is a listing of the values in order of magnitude from the smallest to the largest value.

Example:

The following values represent a list of ages of subjects who participate in a study on smoking cessation:

55 46 58 54 52 69 40 65 53 58

The ordered array is:

40 46 52 53 54 55 58 58 65 69

2.3 Grouped Data: The Frequency Distribution:

To group a set of observations, we select a suitable set of contiguous, non-overlapping intervals such that each value in the set of observations can be placed in one, and only one, of the intervals. These intervals are called "class intervals".

Example:

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

17.0	17.7	15.9	15.2	16.2	17.1	15.7	17.3	**13.5**	16.3
14.6	15.8	15.3	16.4	13.7	16.2	16.4	16.1	17.0	15.9
14.0	16.2	16.4	14.9	17.8	16.1	15.5	**18.3**	15.8	16.7
15.9	15.3	13.9	16.8	15.9	16.3	17.4	15.0	17.5	16.1
14.2	16.1	15.7	15.1	17.4	16.5	14.4	16.3	17.3	15.8

We wish to summarize these data using the following class

intervals:

13.0 – 13.9 , 14.0 – 14.9 , 15.0 – 15.9 ,
16.0 – 16.9 , 17.0 – 17.9 , 18.0 – 18.9

Solution:

Variable = X = hemoglobin level (continuous, quantitative)
Sample size = n = 50
Max= 18.3
Min= 13.5

Class Interval	Tally	Frequency
13.0 – 13.9	|||	3
14.0 – 14.9	𝍸	5
15.0 – 15.9	𝍸 𝍸 𝍸	15
16.0 – 16.9	𝍸 𝍸 𝍸 |	16
17.0 – 17.9	𝍸 𝍸	10
18.0 – 18.9	|	1

The grouped frequency distribution for the hemoglobin level of the 50 men is:

Class Interval (Hemoglobin level)	Frequency (no. of men)
13.0 – 13.9	3
14.0 – 14.9	5
15.0 – 15.9	15
16.0 – 16.9	16
17.0 – 17.9	10
18.0 – 18.9	1
Total	n=50

Notes:

1. Minimum value ∈ first interval.
2. Maximum value ∈ last interval.
3. The intervals are not overlapped.
4. Each value belongs to one, and only one, interval.
5. Total of the frequencies = the sample size = n

Mid-Points of Class Intervals:

- Mid-point $= \dfrac{\text{upper limit} + \text{lower limit}}{2}$

True Class Intervals:

- d = gap between class intervals
- d = lower limit – upper limit of the preceding class interval
- true upper limit = upper limit +d/2
- true lower limit = lower limit – d/2

Class Interval	True Class Interval	Mid-point	Frequency
13.0 – **13.9**	12.95 - 13.95	13.45	3
14.0 – 14.9	13.95 - 14.95	14.45	5
15.0 – 15.9	14.95 - 15.95	15.45	15
16.0 – 16.9	15.95 - 16.95	16.45	16
17.0 – 17.9	16.95 - 17.95	17.45	10
18.0 – 18.9	17.95 – 18.95	18.45	1

For example:

Mid-point of the 1st interval = (13.0+13.9)/2 = 13.45

:

Mid-point of the last interval = (18.0+18.9)/2 = 18.45

Note:

(1) Mid-point of a class interval is considered as a typical (approximated) value for all values in that class interval.

For example: approximately we may say that:

there are 3 observations with the value of 13.45

there are 5 observations with the value of 14.45

:

there are 1 observation with the value of 18.45

(2) There are no gaps between true class intervals. The end-point (true upper limit) of each true class interval equals to the start-point (true lower limit) of the following true class interval.

Cumulative frequency:

Cumulative frequency of the 1st class interval = frequency.

Cumulative frequency of a class interval

= frequency + cumulative frequency of the preceding class interval

Relative frequency and Percentage frequency:

Relative frequency = frequency/n

Percentage frequency = Relative frequency × 100%

Class Interval	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency	Percentage Frequency	Cumulative Percentage Frequency
13.0 – 13.9	3	3	0.06	0.06	6%	6%
14.0 – 14.9	5	8	0.10	0.16	10%	16%
15.0 – 15.9	15	23	0.30	0.46	30%	46%
16.0 – 16.9	16	39	0.32	0.78	32%	78%
17.0 – 17.9	10	49	0.20	0.98	20%	98%
18.0 – 18.9	1	50	0.02	1.00	2%	100%

From frequencies:

The number of people whose hemoglobin levels are between 17.0 and 17.9 = 10

From cumulative frequencies:

The number of people whose hemoglobin levels are less than or equal to 15.9 = 23

The number of people whose hemoglobin levels are less than or equal to 17.9 = 49

From percentage frequencies:

The percentage of people whose hemoglobin levels are between 17.0 and 17.9 = 20%

From cumulative percentage frequencies:

The percentage of people whose hemoglobin levels are less than or equal to 14.9 = 16%

The percentage of people whose hemoglobin levels are less than or equal to 16.9 = 78%

Displaying Grouped Frequency Distributions:

For representing frequency (or relative frequency or percentage frequency) distributions, we may use one of the following graphs:

- The Histogram
- The Frequency Polygon

Example:

Consider the following frequency distribution of the ages of 100 women.

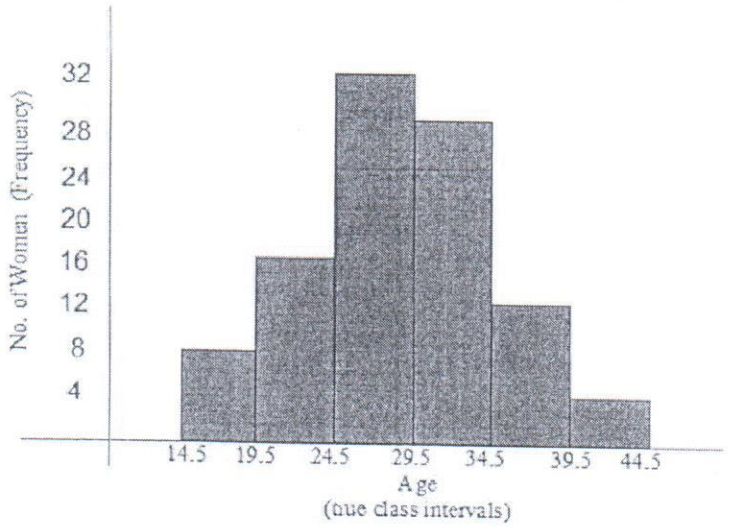
True Class Interval (age)	Frequency (No. of women)	Cumulative Frequency	Mid-points
14.5 - 19.5	8	8	17
19.5 - 24.5	16	24	22
24.5 - 29.5	32	56	27
29.5 - 34.5	28	84	32
34.5 - 39.5	12	96	37
39.5 - 44.5	4	100	42
Total	n=100		

Width of the interval:

W =true upper limit – true lower limit = 19.5 – 14.5 = 5

(1) Histogram:

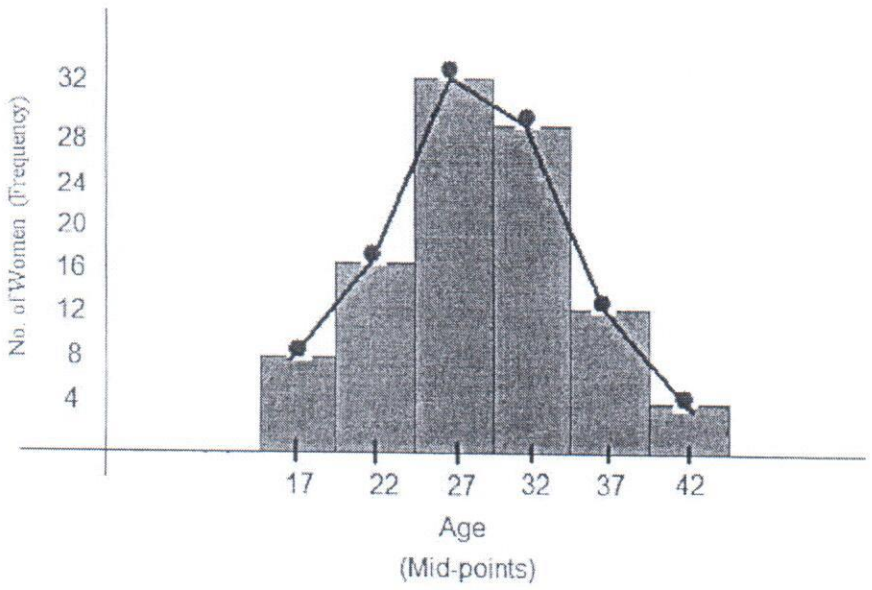
Organizing and Displaying Data using Histogram:



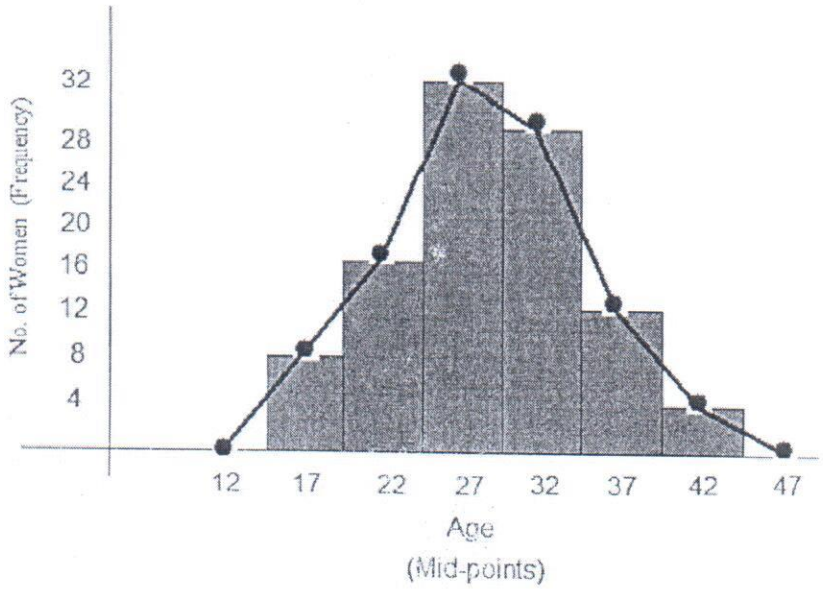
(2) The Frequency Polygon:

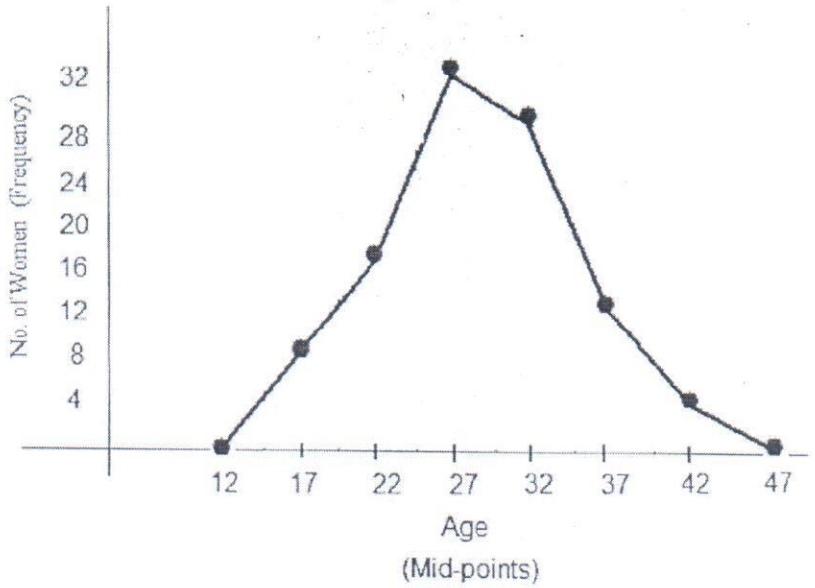
Organizing and Displaying Data using Polygon:

Polygon (Open)



Polygon (Closed)





Exercises on chapter 2

Complete the table ,then answer the following questions

Text Book : Basic Concepts and Methodology for the Health Sciences

Class interval (Age)	Mid – interval	Frequency Freq (f)	Cumulative Frequency	Relative Frequency R.f	Cumulative Relative Frequency
30 – 39	34.5	11	11	0.0582	0.0582
40 – 49	44.5	46	57	0.2434	-
50 – 59	54.5	-	127	-	0.6720
60 – 69	-	45	-	0.2381	0.9101
70 – 79	74.5	16	188	0.0847	0.9948
80 – 89	84.5	1	189	0.0053	1
Total		189		1	

Text Book : Basic Concepts and Methodology for the Health Sciences

Example :

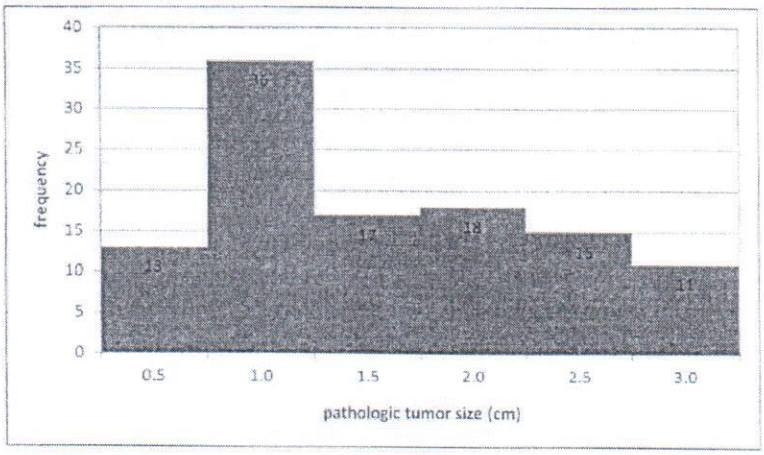
- From the above frequency table, complete the table then answer the following questions:
- 1-The number of objects with age less than 50 years ?
- 2-The number of objects with age between 40-69 years ?
- 3-Relative frequency of objects with age between 70-79 years ?
- 4-Relative frequency of objects with age more than 69 years ?
- 5-The percentage of objects with age between 40-49 years ?

- 6- The percentage of objects with age less than 60 years ?
- 7- Variable is .....
- Type of variable is .....

8. True class for the third interval (50-59) is .....

- 9- The width of the interval (W) ?
- 10- Sample size is .....

the following histogram show the frequency distribution of pathologic tumor size (in cm) for a sample of cancer patients:



Answer the following

1- The variable is .....

2. The type of the variable is .....

3- The sample size is .....

4- The name of the chart.....

5- The number of cancer patient with lowest Pathologic tumor size is .....

6-The percent of cancer patients with approximate level of Pathologic tumor size = 2.....

2.4 Descriptive Statistics: Measures of Central Tendency:

(Measures of location)

In the last section we summarize the data using frequency distributions (tables and figures). In this section, we will introduce the concept of summarization of the data by means of a single number called "a descriptive measure".

A descriptive measure computed from the values of a sample is called a "statistic".

A descriptive measure computed from the values of a population is called a "parameter".

For the variable of interest there are:

(1) "N" population values.
(2) "n" sample of values.

- Let $X_1, X_2, \ldots, X_N$ be the population values (in general, they are unknown) of the variable of interest.
  The population size = N

- Let $x_1, x_2, \ldots, x_n$ be the sample values (these values are known).
  The sample size = n.

(i) A **parameter** is a measure (or number) obtained from the population values: $X_1, X_2, \ldots, X_N$.
- Values of the parameters are unknown in general.
- We are interested to know true values of the parameters.

(ii) A **statistic** is a measure (or number) obtained from the sample values: $x_1, x_2, \ldots, x_n$.
- Values of statistics are known in general.
- Since parameters are unknown, statistics are used to approximate (estimate) parameters.

Measures of Central Tendency:
(or measures of location):

The most commonly used measures of central tendency are: the mean – the median – the mode.

- The values of a variable often tend to be concentrated around the center of the data.
- The center of the data can be determined by the measures of central tendency.
- A measure of central tendency is considered to be a typical (or a representative) value of the set of data as a whole.

Mean:

(1) The Population mean (μ):

If $X_1, X_2, \ldots, X_N$ are the population values, then the population mean is:

$$\mu = \frac{X_1 + X_2 + \cdots + X_N}{N} = \frac{\sum_{i=1}^{N} X_i}{N} \qquad \text{(unit)}$$

- The population mean μ is a parameter (it is usually unknown, and we are interested to know its value)

(2) The Sample mean ($\bar{x}$):

If $x_1, x_2, \ldots, x_n$ are the sample values, then the sample mean is:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \text{(unit)}$$

- The sample mean $\bar{x}$ is a statistic (it is known – we can calculate it from the sample).
- The sample mean $\bar{x}$ is used to approximate (estimate) the population mean μ.

Example:

Suppose that we have a population of 5 population values:

$$X_1 = 41,\ X_2 = 30,\ X_3 = 35,\ X_4 = 22,\ X_5 = 27.\ (N=5)$$

Suppose that we randomly select a sample of size 3, and the sample values we obtained are:

$$x_1 = 30,\ x_2 = 35,\ x_3 = 27.\ (n=3)$$

Then:

The population mean is:

$$\mu = \frac{41+30+35+22+27}{5} = \frac{155}{5} = 31 \qquad \text{(unit)}$$

The sample mean is:

$$\bar{x} = \frac{30+35+27}{3} = \frac{92}{3} = 30.67 \qquad \text{(unit)}$$

Notice that $\bar{x} = 30.67$ is approximately equals to $\mu = 31$.

Note: The unit of the mean is the same as the unit of the data.

Advantages and disadvantages of the mean:

Advantages:

- Simplicity: The mean is easily understood and easy to compute.
- Uniqueness: There is one and only one mean for a given set of data.
- The mean takes into account all values of the data.

Disadvantages:

- Extreme values have an influence on the mean. Therefore, the mean may be distorted by extreme values.

For example:

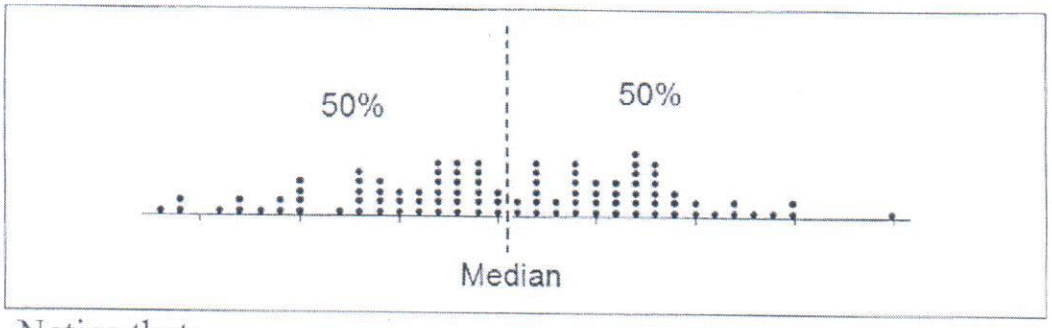
Sample	Data	mean
A	2 4 5 7 7 10	5.83
B	2 4 5 7 7 100	20.83

- The mean can only be found for quantitative variables.

Median:

The median of a finite set of numbers is that value which divides the **ordered array** into two equal parts. The numbers in the first part are less than or equal to the median and the numbers in the second part are greater than or equal to the

median.



Notice that:

50% (or less) of the data is $\leq$ Median

50% (or less) of the data is $\geq$ Median

Calculating the Median:

Let $x_1, x_2, \ldots, x_n$ be the sample values. The sample size (n) can be odd or even.

- First we order the sample to obtain the ordered array.
- Suppose that the ordered array is:

$$y_1, y_2, \ldots, y_n$$

- We compute the rank of the middle value (s):

$$rank = \frac{n+1}{2}$$

- If the sample size (n) is an odd number, there is only one value in the middle, and the rank will be an integer:

$$rank = \frac{n+1}{2} = m \qquad \text{(m is integer)}$$

The median is the middle value of the **ordered** observations, which is:

$$\text{Median} = y_m .$$

Ordered set → (smallest to largest)	y_1	y_2	…	y_m middle value	…	y_n
Rank (or order) →	1	2	…	m	…	n

- If the sample size (n) is an even number, there are two values in the middle, and the rank will be an integer plus 0.5:

$$rank = \frac{n+1}{2} = m + 0.5$$

Therefore, the ranks of the middle values are (m) and (m+1). The median is the mean (average) of the two middle values of the **ordered** observations:

$$\text{Median} = \frac{y_m + y_{m+1}}{2}.$$

Ordered set →	y_1	y_2	…	y_m middle value	y_{m+1} middle value	…	y_n
Rank (or order) →	1	2	…	**m**	**m+1**	…	*n*

Example (odd number):
Find the median for the sample values: 10, 54, 21, 38, 53.
Solution:
.$n = 5$ (odd number)
There is only one value in the middle.
The rank of the middle value is:

$$rank = \frac{n+1}{2} = \frac{5+1}{2} = 3. \quad (m=3)$$

Ordered set →	10	21	**38** (middle value)	53	54
Rank (or order) →	1	2	**3** **(m)**	4	5

The median =38 (unit)

Example (even number):
Find the median for the sample values: 10, 35, 41, 16, 20, 32
Solution:
.$n = 6$ (even number)
There are two values in the middle.
The rank is:

$$rank = \frac{n+1}{2} = \frac{6+1}{2} = 3.5 = 3 + 0.5 = m+0.5 \quad (m=3)$$

Therefore, the ranks of the middle values are:

.m = 3 and m+1 = 4

Ordered set →	10	16	**20**	**32**	35	41
Rank (or order) →	1	2	**3 (m)**	**4 (m+1)**	5	6

The middle values are 20 and 32.

The median = $= \frac{20+32}{2} = \frac{52}{2} = 26$ (unit)

Note: The unit of the median is the same as the unit of the data.

Advantages and disadvantages of the median:

Advantages:

- Simplicity: The median is easily understood and easy to compute.
- Uniqueness: There is only one median for a given set of data.
- The median is not as drastically affected by extreme values as is the mean. (i.e., the median is not affected too much by extreme values).

For example:

Sample	Data	median
A	9 4 5 9 2 10	7
B	9 4 5 9 2 100	7

Disadvantages:

- The median does not take into account all values of the sample.
- In general, the median can only be found for quantitative variables. However, in some cases, the median can be found for ordinal qualitative variables.

Mode:

The mode of a set of values is that value which occurs most frequently (i.e., with the highest frequency).

- If all values are different or have the same frequencies, there will be no mode.
- A set of data may have more than one mode.

Example:

Data set	Type	Mode(s)
26, 25, 25, 34	Quantitative	25
3, 7, 12, 6, 19	Quantitative	No mode
3, 3, 7, 7, 12, 12, 6, 6, 19, 19	Quantitative	No mode
3, 3, 12, 6, 8, 8	Quantitative	3 and 8
B C A B B B C B B	Qualitative	B
B C A B A B C A C	Qualitative	No mode
B C A B B C B C C	Qualitative	B and C

Note: The unit of the mode is the same as the unit of the data.

Advantages and disadvantages of the mode:

Advantages:

- Simplicity: the mode is easily understood and easy to compute..
- The mode is not as drastically affected by extreme values as is the mean. (i.e., the mode is not affected too much by extreme values).

For example:

Sample	Data	Mode
A	7 4 5 7 2 10	7
B	7 4 5 7 2 100	7

- The mode may be found for both quantitative and qualitative variables.

Disadvantages:

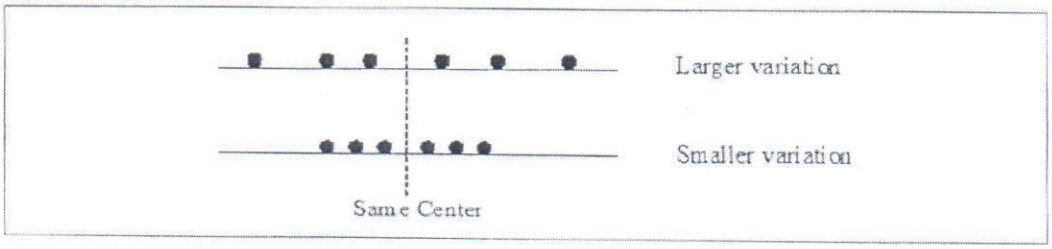
- The mode is not a "good" measure of location, because it depends on a few values of the data.
- The mode does not take into account all values of the sample.
- There might be no mode for a data set.
- There might be more than one mode for a data set.

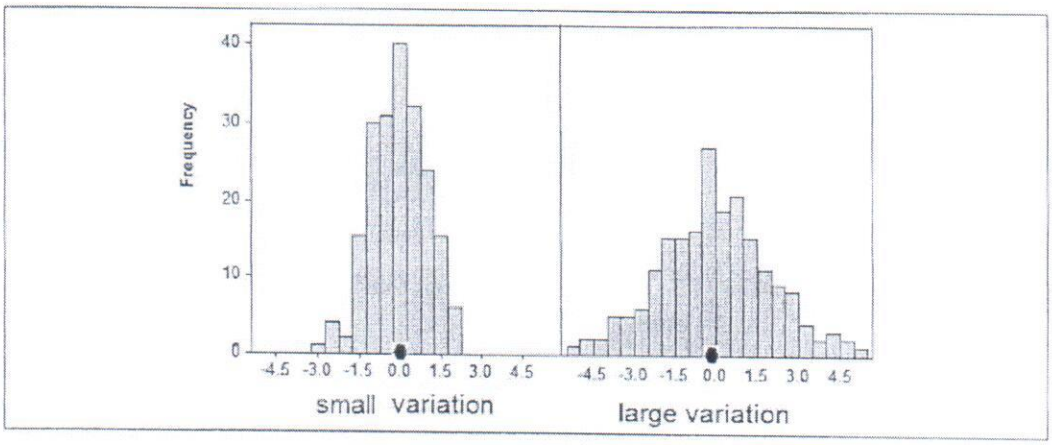
2.6 Descriptive Statistics: Measures of Dispersion (Measures of Variation):

The dispersion (variation) of a set of observations refers to the variety that they exhibit. A measure of dispersion conveys information regarding the amount of variability present in a set of data. There are several measures of dispersion, some of which are: Range, Variance, Standard Deviation, and Coefficient of Variation.

The variation or dispersion in a set of values refers to how spread out the values is from each other.

- The dispersion (variation) is small when the values are close together.
- There is no dispersion (no variation) if the values are the same.





The Range:

The Range is the difference between the largest value (Max) and the smallest value (Min).

$$\text{Range } (R) = \text{Max} - \text{Min}$$

Example:

Find the range for the sample values: 26, 25, 35, 27, 29, 29.

Solution:

.max = 35

.min = 25

Range (R) = 35 – 25 = 10 (unit)

Notes:

1. The unit of the range is the same as the unit of the data.
2. The usefulness of the range is limited. The range is a poor measure of the dispersion because it only takes into account two of the values; however, it plays a significant role in many applications.

The Variance:

The variance is one of the most important measures of dispersion.

The variance is a measure that uses the mean as a point of reference.

- The variance of the data is small when the observations are close to the mean.
- The variance of the data is large when the observations are spread out from the mean.
- The variance of the data is zero (no variation) when all observations have the same value (concentrated at the mean).

Deviations of sample values from the sample mean:

Let $x_1, x_2, \ldots, x_n$ be the sample values, and $\bar{x}$ be the sample mean.

The deviation of the value x_i from the sample mean $\bar{x}$ is:

$$x_i - \bar{x}$$

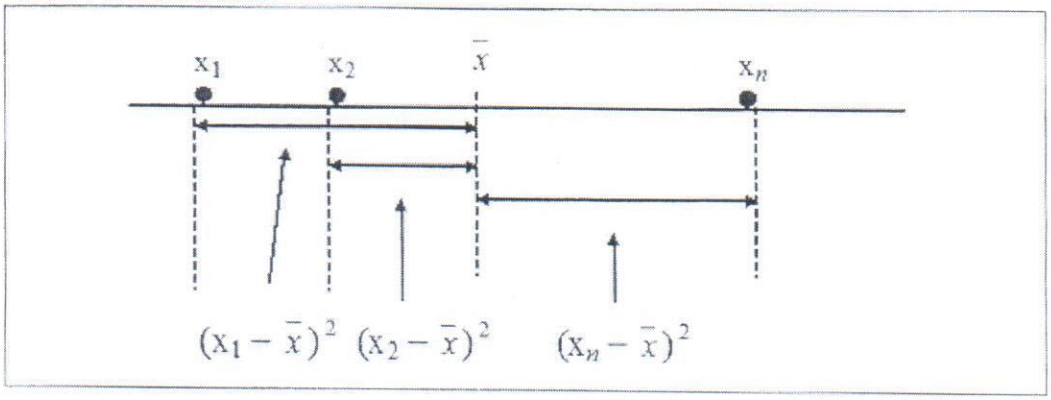
The squared deviation is:

$$(x_i - \bar{x})^2$$

The sum of squared deviations is:

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

The following graph shows the squared deviations of the values from their mean:



(1) The Population Variance σ^2 :

(Variance computed from the population)

Let $X_1, X_2, \ldots, X_N$ be the population values. The population variance (σ^2) is defined by:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}$$

$$= \frac{(X_1-\mu)^2+(X_2-\mu)^2+\cdots+(X_N-\mu)^2}{N} \quad (unit)^2$$

where, $\mu = \frac{\sum_{i=1}^{N} X_i}{N}$ is the population mean, and (N) is the population size.

Notes:

- σ^2 is a parameter because it is obtained from the population values (it is unknown in general).
- $\sigma^2 \geq 0$

(2) The Sample Variance S^2:

(Variance computed from the sample)

Let $x_1, x_2, \ldots, x_n$ be the sample values. The sample variance (S^2) is defined by:

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

$$= \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} \qquad (unit)^2$$

where $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ is the sample mean, and (n) is the sample size.

Notes:

- S^2 is a statistic because it is obtained from the sample values (it is known).
- S^2 is used to approximate (estimate) σ^2.
- $S^2 \geq 0$
- $S^2 = 0 \quad \Leftrightarrow$ all observation have the same value
  $\Leftrightarrow$ there is no dispersion (no variation)

Example:

We want to compute the sample variance of the following sample values: 10, 21, 33, 53, 54.

Solution:

n=5

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{5} x_i}{5} = \frac{10+21+33+53+54}{5} = \frac{171}{5} = 34.2$$

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{5}(x_i - 34.2)^2}{5-1}$$

$$S^2 = \frac{(10-34.2)^2 + (21-34.2)^2 + (33-34.2)^2 + (53-34.2)^2 + (54-34.2)^2}{4}$$

$$= \frac{1506.8}{4} = 376.7 \quad (\text{unit})^2$$

Another Method for calculating sample variance:

x_i	$(x_i - \bar{x}) = (x_i - 34.2)$	$(x_i - \bar{x})^2 = (x_i - 34.2)^2$
10	-24.2	585.64
21	-13.2	174.24

x_i	$(x_i - \bar{x}) = (x_i - 34.2)$	$(x_i - \bar{x})^2 = (x_i - 34.2)^2$
33	-1.2	1.44
53	18.8	353.44
54	19.8	392.04
$\sum_{i=1}^{5} x_i = 171$	$\sum_{i=1}^{5} (x_i - \bar{x}) = 0$	$\sum (x_i - \bar{x})^2 = 1506.8$

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i}{5} = = \frac{171}{5} = 34.2 \quad \text{and} \quad S^2 = \frac{1506.8}{4} = = 376.7$$

Standard Deviation:

The variance represents squared units, therefore, is not appropriate measure of dispersion when we wish to express the concept of dispersion in terms of the original unit.

- The standard deviation is another measure of dispersion.
- The standard deviation is the square root of the variance.
- The standard deviation is expressed in the original unit of the data.

(1) Population standard deviation is: $\sigma = \sqrt{\sigma^2}$ (unit)

(2) Sample standard deviation is: $S = \sqrt{S^2}$ (unit)

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

Example:

For the previous example, the sample standard deviation is

$$S = \sqrt{S^2} = \sqrt{376.7} = 19.41 \quad \text{(unit)}$$

Coefficient of Variation (C.V.):

- The variance and the standard deviation are useful as measures of variation of the values of a single variable for a single population.
- If we want to compare the variation of two variables we cannot use the variance or the standard deviation because:

1. The variables might have different units.
2. The variables might have different means.

- We need a measure of the relative variation that will not depend on either the units or on how large the values are. This measure is the coefficient of variation (C.V.).
- The coefficient of variation is defined by:

$$\text{C.V.} = \frac{S}{\bar{x}} \times 100\%$$

- The C.V. is free of unit (unit-less).
- To compare the variability of two sets of data (i.e., to determine which set is more variable), we need to calculate the following quantities:

	Mean	Standard deviation	C.V.
1st data set	$\bar{x}_1$	S_1	$C.V_1 = \frac{S_1}{\bar{x}_1} 100\%$
2nd data set	$\bar{x}_2$	S_2	$C.V_2 = \frac{S_2}{\bar{x}_2} 100\%$

- The data set with the larger value of CV has larger variation.
- The relative variability of the 1st data set is larger than the relative variability of the 2nd data set if $C.V_1 > C.V_2$ (and vice versa).

Example:

Suppose we have two data sets:

1st data set: $\bar{x}_1 = 66$ kg, $S_1 = 4.5$ kg

$$\Rightarrow C.V_1 = \frac{4.5}{66} * 100\% = 6.8\%$$

2nd data set: $\bar{x}_2 = 36$ kg, $S_2 = 4.5$ kg

$$\Rightarrow C.V_2 = \frac{4.5}{36} * 100\% = 12.5\%$$

Since $C.V_2 > C.V_1$, the relative variability of the 2nd data set is larger than the relative variability of the 1st data set.

If we use the standard deviation to compare the variability of the two data sets, we will wrongly conclude that the two data sets have the same variability because the standard deviation of both sets is 4.5 kg.

How to use calculator

Mode (3:stat) ___ Then ___(1: 1- VAR)

يظهر جدول لإدخال البيانات

Example: 2, 4,6, 9

2 = , 4= , 6 = , 9=

لإيجاد الوسط والتباين والانحراف المعياري

نضغط

Shift (1)

وتختار الرقم 5:Var

ويظهر في الشاشه

1:n	2: $\overline{X}$
3: xσn	4: xσn-1

For mean : Push 2

For Sample standard deviation :Push 4

For Population standard deviation :Push 3

*To find sample variance = square (Sample standard deviation)

* To find Population variance = square (Population standard deviation)

Example:

For a random sample wrest size of 12 women (in cm)will be as:

16 10 49 15 6 15 8 19 11 22 13 17

(a) The mean Ans: 16.75

(b) The median Ans: 15

(c) The mode Ans: 15

(d) The range Ans: 43

(e) The variance Ans: 124.0227

(f) The standard deviation Ans: 11.1365

(g) The coefficient of variation Ans: 66.49%

Example 2.5.3 Page 46:

- Suppose two samples of human males yield the following data:(which one is more variable.)

	Sampe1	Sample2
Age	25-year-olds	11year-olds
Mean weight	135 pound	60 pound
Standard deviation	10 pound	10 pound

Sample 2 has more variation than sample 1

$$C.V(sample1) = \frac{S}{\overline{X}} x100 = \frac{10}{135} x100 = 7.41\%$$

$$C.V(sample2) = \frac{S}{\overline{X}} x100 = \frac{10}{60} x100 = 16.67\%$$