Second Lecture

Tabular and Grapgical Presentation

Tabular Presentation : **العرض الجدولي**

There are two main types of statistical tables, namely:

1. The simple tables : It is a table on which the data is distributed according to one trait . Usually consists of two columns, first is represented the trait or phenomenon divisions to classes or groups, and the second are shows the number of fixed observations for each class or group .

Classes	Number of students
60-62	5
63-65	15
66-68	45
69-71	27
72-74	8
total	100

2. Composite table: It is a table which the data is distributed according to the two traits or phenomena or more at the same time usually consists of rows and columns.
The rows represented classes or groups of one trait .
The columns represent are classes or groups of other trait .

Frequency distribution table or frequency table : جدول التوزيع التكراري أو الجدول التكراري

It is a simple table that consists of two columns, the first which contain the variable values are divided into sections or groups called classes and the second shows the observation number of each class which called frequency.

Sum important Definitions : بعض التعاريف المهمة

Ungrouped data (Raw data) : بيانات غير مبوبة

The first data which are collected and can't be useful .

Grouped data : بيانات مبوبة

The data which are presented in table and are ready to be analyzed .

Classes : الفئات

Is a groups which the values of variable were included .

Class Limits : حدود الفئات

Every class have two limits , lower class limit and upper class limit .

True class limits : الحدود الحقيقية للفئات

Every class have two true class limits , lower true class limit and upper true class limit .

To account these limits when the limits are integers number :

Lower true class limit = lower class limit – 0.5

Upper true class limit = upper class limit + 0.5

Class Length : طول الفئة

It is the range between two class limits .

Class Mark (y_i) : مركز الفئة

Is the range mid-point between both class limits .

Class Frequency (f_i) : تكرار الفئة

Is the number of observations which located in the range of that class .

There are many methods to account class length from frequency table as following :

(1) . when the class limits are correct numbers only :

Class length = upper limit – lower limit + 1

62 – 60 +1 = 3

(2) . class length = upper true limit – lower true limit

= 62.5 – 59.5 = 3

(3). class length = The deference between two successive upper or lower class limits :

= 63 – 60 = 3 or 65 – 62 = 3

(4) . class length = The deference between two successive upper true or lower true class limits . = 62.5 - 59.5 = 3 or 65.5 – 62.5 = 3

(5) . class length = The deference between two successive class mark .

= 64 – 61 = 3

We can also account the true limits of each class by any of these following methods :

(1). Lower true limit = class mark – 0.5 (class length)

= 61 – 0.5 (3) = 59.5

While , upper true limit = class mark + 0.5 (class length)

= 61 + 0.5 (3) = 62.5

$$\text{(2) Lower true class limit} = \frac{\text{Lower limit + upper limit of previous class}}{2}$$

Lower true class limit of the 2nd class = $\frac{63+62}{2} = 62.5$

Upper true class limit of the 2nd class = $\dfrac{\text{Upper limit + lower limit of the following class}}{2}$

Upper true class limit of the 2nd class = $\frac{65+66}{2}=65.5$

To account class mark of each class there are two method as following :

(1). Class mark = $\dfrac{lower\,\text{lim}it+upper\,\text{lim}it}{2}$

Class mark of the 1st class = $\frac{60+62}{2}=61$

(2). Class mark = $\dfrac{\text{Lower true limit + upper true limit}}{2}$

Class mark of the 1st class = $\frac{59.5+62.5}{2}=61$

<u>General steps to create a frequency distribution tables are</u>: الخطوات العامة لإنشاء جدول التوزيع التكراري

(1). extracting the variable range. استخراج المدى
(2). Determine the number of classes . تحديد عدد الفئات
(3). Find the class length . إيجاد طول الفئة
(4). writing of class limits . كتابة حدود الفئات
(5). Find the frequency numbers of each class . إيجاد عدد التكرارات لكل فئة

<u>Example</u> : The following values represents the ages of 20 students , established frequency table ?
24 , 19 , 22 , 24 , 23 , 19 , 25 , 21 , 18 , 20 , 21 , 25 , 26 , 23 , 18 , 24 , 21 , 19 , 22 , 25 .

<u>Solution :</u>

(1). Range = upper value – lower value = 26 – 18 = 8

(2). Choice the number classes = 2.5 * $\sqrt[4]{observation}$ = 2.5 * $\sqrt[4]{20}$ = 2.5*2 = 5

(3). Find class length = $\frac{range}{classnumber}=\frac{8}{5}=1.6\approx 2$

(4). Writing class limits :

(5). Find class frequency:

classes	Frequency(f_i)	True class limits	Class mark(y_i)	R.F	R.F %
18-19	5	17.5-19.5	18.5	0.25	25
20-21	4	19.5-21.5	20.5	0.2	20
22-23	4	21.5-23.5	22.5	0.2	20
24-25	6	23.5-25.5	24.5	0.3	30
26-27	1	25.5-27.5	26.5	0.05	5
	$\sum f_i = 20$				

To account true limits : لحساب الحدود الحقيقية للفئات

Lower true limit of 1st class = lower class limit – 0.5 = 18 - 0.5 = 17.5

Upper true limit of 1st class = upper class limit + 0.5 = 19 + 0.5 = 19.5

And so the rest classes . وهكذا بقية الفئات

To account class mark (y_i) : لحساب مراكز الفئات

Class mark of 1st class = $\frac{lower\lim it + upper\lim it}{2} = \frac{18+19}{2} = 18.5$

And so the rest classes . وهكذا بقية الفئات

Relative Frequency Distribution (R.F) : **جدول التوزيع التكراري النسبي**

A table shows the importance of relative per class , and calculated as follows :

R.F = $\frac{f_i}{\sum f_i}$ so , the 1st class R.F of the previous example = $\frac{5}{20} = 0.25$

And so the rest classes . وهكذا بقية الفئات

Usually put relative frequency as a percentage by multiplying R.F *100 as follows :

P.R.F of 1st class = 0.25 * 100 = 25 And so the rest classes . وهكذا بقية الفئات

Cumulative Distribution : **التوزيعات المتجمعة**

In some cases there may be a need to know the number of values or observations of less or more than a certain value and the tables that contain information is called cumulative distribution tables. There are two types of these tables :

(1) . **Less than cumulative distribution table**. **جدول التوزيع التكراري التجمعي التصاعدي**

It gives us a number of observations which their value is less than the lower limit for a certain class and we will symbolize the cumulative frequency for any Class by (F_i) . This table is consists of two columns: the first column that shows the classes limits , and the second column which it showes the less than cumulative frequency, as the following :

frequency of first class = F0 = 0

frequency of second class (F_1) = f_1 .

frequency of third class = $F_2 = f_1+f_2$

frequency of forth class = F_3 = f1+ f2 + f3 .

Thus, so that the cumulative frequency of latter class = $F_n = \Sigma f_i$.

Less than cumulative distribution table

Class limits	Less than cumulative frequency
Less than 18	= F0 = 0
Less than 20	= f_1 = 5
Less than 22	= f_1+f_2 = 9
Less than 24	=f1+ f2 + f3 = 13

Less than 26	=f1+ f2 + f3 +f_4 =19
Less than 28	=f1+ f2 + f3 +f_4+f_5 =20

(2) . **More than cumulative distribution table** :

It gives us a number of observations which their value is more than the lower limit for certain class . this table also consist of two columns :

The first column , the class limits will be written on it .

The second column , the more than cumulative frequency will be written on it as follows :

The first class frequency = $F_1 = \sum f_i$

The second class frequency = $F_2 = \sum f_i - f_1$

The third class frequency = $F_3 = \sum f_i - f_1 - f_2$ or $\sum f_i - (f_1 + f_2)$ Thus , as show down :

Class limits	More than cumulative frequency
More than 18	$= \sum f_i = 20$
More than 20	$= \sum f_i - f_1 = 15$
More than 22	$= \sum f_i - (f_1 + f_2) = 11$
More than 24	$= \sum f_i - (f_1 + f_2 + f_3) = 7$
More than 26	$= \sum f_i - (f_1 + f_2 + f_3 + f_4) = 1$
More than 28	$= \sum f_i - (f_1 + f_2 + f_3 + f_4 + f_5) = 0$

Sometimes , we reflects about less than or more than cumulative frequency as relative cumulative frequency or percentage , so , at this case :

The relative cumulative frequency of any class = $\frac{f_i}{\sum f_i}$

But the percentage cumulative frequency =($\frac{f_i}{\sum f_i}$) * 100

Example (1) .

Complete the following frequency table :

classes	Frequency f_i	True limits	Class mark Y_i	Relative frequency RF	Percentage frequency Rf %
	2		4		
	5		9		
	10		14		
	25		19		
	8		24		
total	$\sum f_i = 50$				

Solution :

Class length = The deference between two successive class mark .

= 9 – 4 = 5

The 1st class lower true limit = class mark of first class – 0.5 (class length)

= 4 – 0.5 (5) = 1.5

The 1st class upper true limit = class mark of first class + 0.5 (class length)

= 4 + 0.5 (5) = 6.5

And so the rest classes . وهكذا بقية الفئات

Or we can additive the class length to the lower true class limit of first class to gain the lower true limit of second class :

5 + 1.5 = 6.5 and thus .

Then we can additive the class length to upper true limit of first class to gain the upper true limit of second class .

5 + 6.5 = 11.5

But the 1st lower class limit = lower true limit of 1st class + 0.5

= 1.5 + 0.5 = 2

And the 1st upper class limit = upper true limit of 1st class – 0.5

= 6.5 – 0.5 = 6 and so the rest classes . وهكذا بقية الفئات

Relative frequency of any class = $\frac{f_i}{\sum f_i}$

R.f of first class = $\frac{2}{50} = 0.04$

Percentage R.f = ($\frac{f_i}{\sum f_i}$) * 100

P.R.f of first class = 0.04 * 100 = 4 وهكذا بقية الفئات كما موضح في الجدول أدناه

Classes	Frequency f_i	True limits	Class mark Y_i	Relative frequency RF	Percentage frequency Rf %
2 _ 6	2	1.5 _ 6.5	4	0.04	4
7 _ 11	5	6.5 _ 11.5	9	0.10	10
12 _ 16	10	11.5 _16.5	14	0.20	20
17 _ 21	25	16.5 _ 21.5	19	0.50	50
22 _ 26	8	21.5 _ 26.5	24	0.16	16
total	$\sum f_i = 50$				

Example (2).

If you know the variable observation number = 50 ($\sum f_i = 50$) , so from following relative frequency table find the frequency , classes mark , true limits and percentage relative frequency for this table :

class	Relative frequency Rf
20 _39	0.12
40 _ 59	0.28
60 _ 79	0.36
80 _ 99	0.20
100 _ 119	0.04
total	

Solution :

The relative frequency of any class $= \frac{f_i}{\sum f_i}$ so , Class frequency = relative frequency * total frequency

First class frequency = 0.12 * 50 = 6

Second class frequency = 0.28 * 50 = 14 thus , وهكذا

But , class mark = lower limit + upper limit / 2 so ,

First class limit = 20 + 39 / 2 = 29.5

Second class limit = 40 + 59 / 2 = 49.5 thus , وهكذا

But , class length = upper limit - lower limit + 1

class length = 39 – 20 + 1 = 20

But , lower true limit of any class = class mark – 0.5 (class length)

First lower true class limit = 29.5 – 0.5 (20) = 19.5

First upper true class limit = 29.5 + 0.5 (20) = 39.5 thus , وهكذا

But , percentage frequency = R. f * 100

Percentage frequency of first class = 0.12 * 100 = 12 thus , as show in the following table :

class	F_i	Y_i	True limit	R.f	p.R.f
20 _ 39	6	29.5	19.5 _ 39.5	0.12	12
40 _ 59	14	49.5	39.5 _59.5	0.28	28
60 _ 79	18	69.5	59.5 _ 79.5	0.36	36
80 _ 99	10	89.5	79.5 _ 99.5	0.20	20
100 _ 119	2	109.5	99.5 _ 119.5	0.04	4
	$\sum f_i = 50$				

Example (3) .

The following table presented weight (kg) frequency distribution of 65 students : debit , make less than and more than cumulative frequency table and from both conclude following :

(A) . how many student their weight less than 70 kg .

(B) . percentage of students their weight less than 70 kg .

(C) . how many students their weight not less than 60 kg .

(D) How many students their weight not less than 60 kg but it is less than 80 kg .

Solution : we find less and more than cumulative distribution table :

classes	F_i	Less than	More than
50 – 54	8	Less than 50 = 0	More than 50 = 65
55 – 59	10	Less than 55 = 8	More than = 57
60 – 64	16	Less than 60 = 18	More than = 47
65 – 69	14	Less than 65 = 34	More than = 31
70 – 74	10	Less than 70 = 48	More than = 17
75 – 79	5	Less than 75 = 58	More than = 7
80 – 84	2	Less than 80 = 63	More than = 2
		Less than 85 = 65	More than 85 = 0
Total	$\sum f_i = 65$		

(A). From the less than frequency table :

number of student their weight less than 70 kg = 48

(B). The percentage of students their weight less than 70 kg = $\frac{48}{65}*100 = 73.8$

(C). From the more than frequency table :

number of student their weight not less than 60 kg = 47

(D). Number of students their weight not less than 60 kg but it is less than 80 kg :

= 47 – 2 = 45

Graphical Presentation : التمثيل البياني

(A). Histogram المدرج التكراري

This is a graphic consist of more than vertical rectangular , where the vertical axis represents frequencies while the horizontal axis represented the classes lengths .

to draw a histogram must be follow these steps :

1. draw the vertical axis and horizontal axis .

2. graduation the horizontal axis to equal partitions represents the true class limits , and leave small interspace between the zero point and true limit of first class and divided the vertical axis to equal partitions looks so the largest frequency .

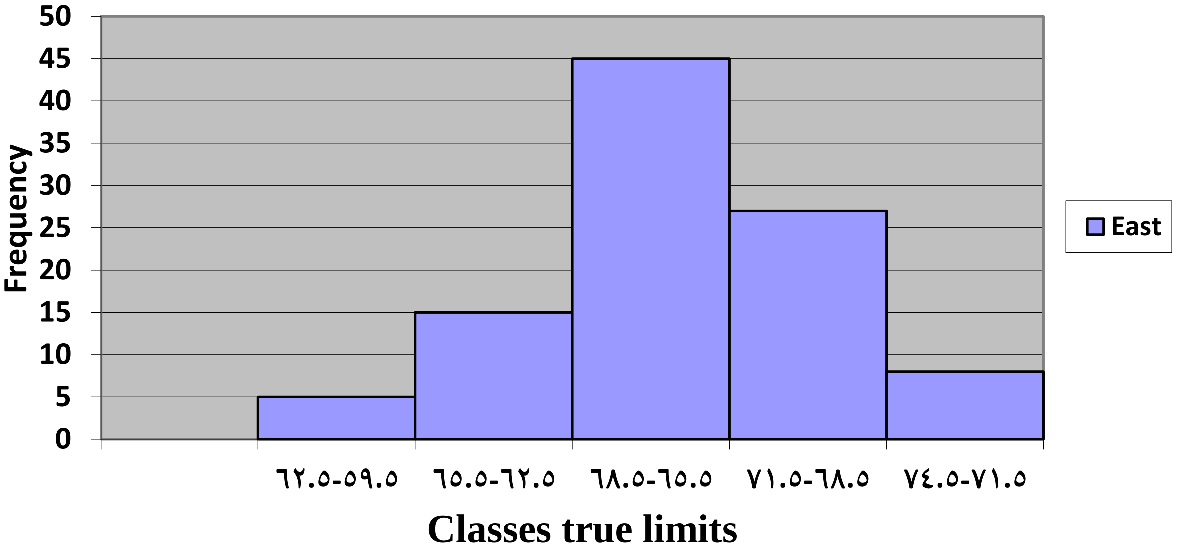
3. over each class, draw vertical rectangular it's base represents the length of that class, and it's rise represents the frequency of that class .

Example : From the following frequency table draw a histogram to explain it's data .

Classes	Number of students	True limits
60-62	5	59.5-62.5
63-65	15	62.5-65.5
66-68	45	65.5-68.5
69-71	27	68.5-71.5
72-74	8	71.5-74.5
total	100	

Solution :

(1). We must estimate the true class limits and write it on the horizontal axis which represented the classes length , then determined a range of observation number which represented the frequency and write it on the vertical axis .



(b) . Frequency Polygon المضلع التكراري

It is a surf rectum lines connected between points , each one located over a class mark at rise represent frequency of that class .

To draw the frequency polygon we followed these steps :

1. draw the vertical axis and horizontal axis .

2. graduation the horizontal axis to equal partitions represents all classes marks , and leave about small interspace between the zero point and the first class mark and divided the vertical axis to equal partitions looks so the largest frequency .

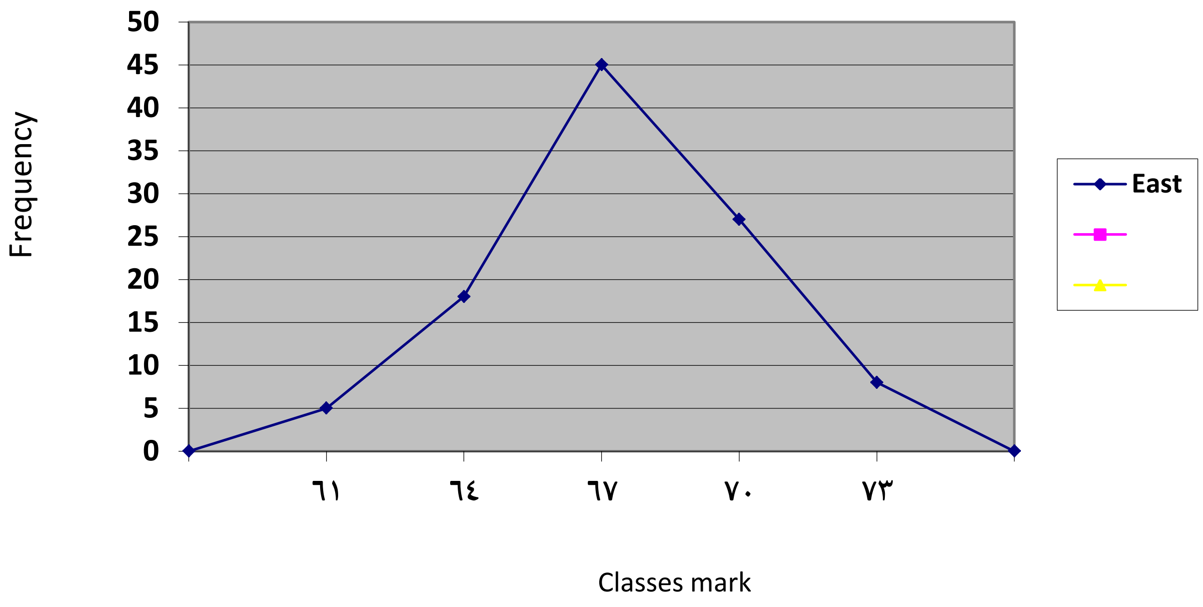
3. put a point in front of each class mark it's rise equivalent frequency of that class .

4. connect between these points by rectum lines .

Example : From the following frequency table draw a frequency polygon to explain it's data

Classes	Frequency (f_i)	Class mark (y_i)
60 – 62	5	61
63 – 65	15	64
66 – 68	45	67
69 – 71	27	70
72 – 74	8	73
Total	100	

Solution : We must estimate the classes mark and write it on the horizontal axis , then determined a range of observation number which represented the frequency and write it on the vertical axis .



Graphical Presentation Of Cumulative Distribution : **التمثيل البياني لجدول التوزيع التكراري التجمعي**

(A). Less than Frequency Polygon : المضلع التكراري التجميعي التصاعدي

To draw less than frequency polygon , we must follow these steps :

1. draw the vertical axis and horizontal axis .

2. graduation the horizontal axis to equal partitions represents all classes limits , and leave about small interspace between the zero point and the first class limit and divided the vertical axis to equal partitions looks so the largest cumulative frequency .

3. put a point in front of each class limit it's rise equivalent less than cumulative frequency of that class .

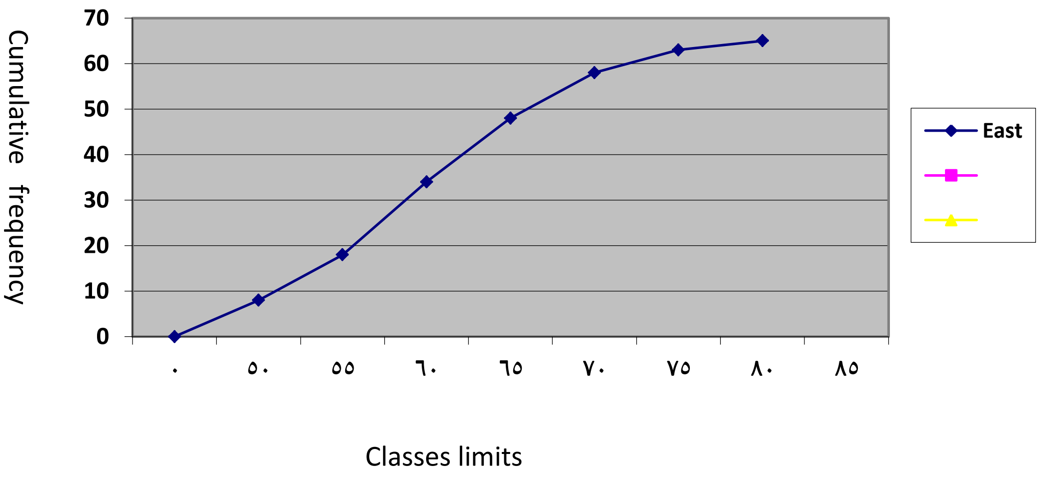
4. connect between these points by rectum lines .

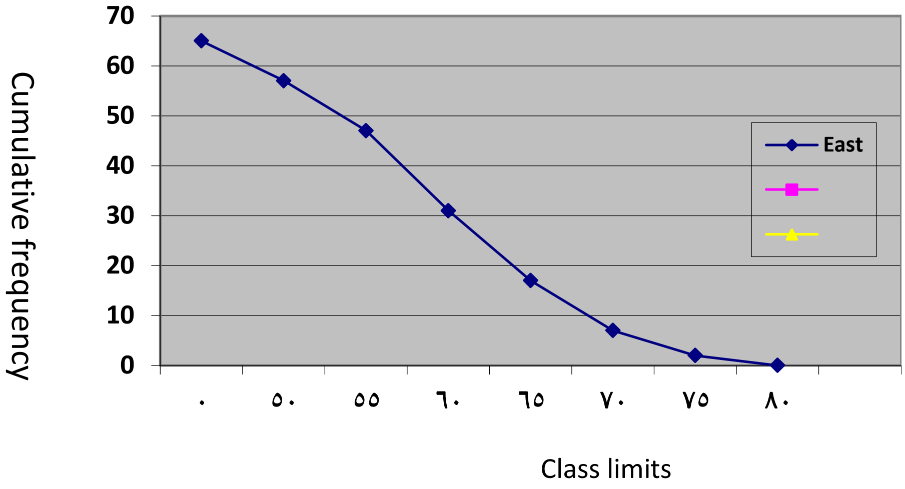
(B). More than Frequency Polygon :   المضلع التكراري التجميعي التنازلي

To draw more than frequency polygon we follow the same above steps .

Example : From this table draw the less and the more than frequency polygon :

classes	F_i	Less than	More than
50 – 54	8	Less than 50 = 0	More than 50 = 65
55 – 59	10	Less than 55 = 8	More than = 57
60 – 64	16	Less than 60 = 18	More than = 47
65 – 69	14	Less than 65 = 34	More than = 31
70 – 74	10	Less than 70 = 48	More than = 17
75 – 79	5	Less than 75 = 58	More than = 7
80 – 84	2	Less than 80 = 63	More than = 2
		Less than 85 = 65	More than 85 = 0
Total	$\sum f_i = 65$		




70
60
50
40
30
20
10
0
Cumulative frequency
East
٠
٥٠
٥٥
٦٠
٦٥
٧٠
٧٥
٨٠
Class limits

The fifth lecture Simple Regression & Correlation الأنحدار والأرتباط البسيط

If we have two variables x_i and y_i , and for example the variable (X_i) represent egg production and variable (y_i) represent egg weight , so there are two relationships between them :

(A). There is equation we can put it to prediction of which about (y) as dependent variable through (x) as independent variable and this called (Regression الأنحدار) .

(B). There is a relationship between the two variables (x) and (y) called (Correlation الأرتباط) to measured the the correlation between two independent variables .

Simple Linear Regression : الأنحدار الخطي البسيط

It is the relationship between two variables , one of them independent variable it's code (y) and the other dependent variable it's code (x) and defined as { the average change in variable (y) which accompanied the change by one unit in variable (x) }.

Regression coefficient which it's code (b) was measured through this equation :

$$b = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum x_i^{\,2} - \frac{(\sum x_i)^2}{n}}$$

But , the linear regression equation is : $\bar{y} = a + bx$ and , $a = \bar{y} - b\bar{x}$

Example : The variable (x) represents the body live weight of five sheeps at the beginning of feeding trail = 32 , 31 , 45 , 38 , 36 and the variable (y) represents their body live weight at the end of feeding trail = 39 , 42 , 52 , 45 , 43 . calculate the linear regression equation ?

Solution :

X_i	Y_i	X_i^2	Y_i^2	$X_i y_i$
32	39	1024	1521	1248
31	42	961	1764	1302
45	52	2025	2704	2340
38	45	1444	2025	1710
36	43	1296	1849	1548
$\sum x_i = 182$	$\sum y_i = 221$	$\sum x_i^{\,2} = 6750$	$\sum y_i^{\,2} = 9863$	$\sum x_i y_i = 8148$

$$b = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum x_i^{2} - \frac{(\sum x_i)^2}{n}} = \frac{8148 - \frac{(182)(221)}{5}}{6750 - \frac{(182)^2}{5}} = \frac{8148 - 8044.4}{6750 - 6624.8} = \frac{103.6}{125.2} = 0.83$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{221}{5} = 44.2 \qquad \bar{x} = \frac{\sum x_i}{n} = \frac{182}{5} = 36.4$$

a = ȳ - bx̄

= 44.2 – 0.83 (36.4) = 14 so that → $\bar{y}_x = a + b\bar{x}$

So , $\boxed{\bar{y}_x = 14 + 0.83x}$ is the linear regression equation .

If you are asked to find the expected value of (y) if you know that (x) value was (40) :

$\bar{y}_x = 14 + 0.83\ (40) = 47.2$

Simple Correlation الأرتباط البسيط

It is the relationship between two independent variables . Therefore , the simple correlation coefficient define as { it is a measure of the degree of correlation or bonding between two independent variables } .

The symbol of correlation coefficient is (r) and it's value ranged between (-1) and (+1) :

As : $-1 \leq r \leq 1$.

If the correlation coefficient value was positive , that mean that the increase in one of the variable accompanied by increase in the other variable . But , when the correlation coefficient value was negative , that mean that the increase in one of the variable accompanied by decrease in the other variable .

Correlation coefficient calculated by this equation :

$$r = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sqrt{(\sum x_i^{2} - \frac{(\sum x_i)^2}{n})(\sum y_i^{2} - \frac{(\sum y_i)^2}{n})}}$$

Example : Calculate the correlation coefficient (r) between length and width plant’s leaves :

Width leaves X_i	Length leaves Y_i	x_iy_i	X_i^2	Y_i^2
16	18	288	256	324
15	15	225	225	225
19	22	418	361	484
17	20	340	289	400
13	15	195	169	225
$\sum x_i = 80$	$\sum y_i = 90$	$\sum x_iy_i = 1466$	$\sum x_i^2 = 1300$	$\sum y_i^2 = 1658$

Solution : $r = \dfrac{\sum x_i y_i - \dfrac{(\sum x_i)(\sum y_i)}{n}}{\sqrt{(\sum x_i^2 - \dfrac{(\sum x_i)^2}{n})(\sum y_i^2 - \dfrac{(\sum y_i)^2}{n})}}$

$$= \frac{1466 - \dfrac{(80)(90)}{5}}{\sqrt{(1300 - \dfrac{(80)^2}{5})(1658 - \dfrac{(90)^2}{5})}} \quad = \quad \frac{1466 - 1440}{\sqrt{(1300 - 1280)(1658 - 1620)}}$$

$$= \frac{26}{\sqrt{(20)(38)}} = \frac{26}{\sqrt{760}} = \frac{26}{27.6} = 0.9$$

First Lecture

1. **Statistic : علم الإحصاء**

   A science which involves collecting data or observations by scientific method and tabulating, summarizing , analyzing , and presenting these data to gain a results and then to take decisions which are right and good

2. **Variable : المتغير**

   It is any phenomenon or a trait which shows differentiations between their observations , and has a symbol Y or X or Z .

3. **Qualitative variables : المتغيرات الوصفية**

   Traits cannot be measured by numbers directly, like eyes color or hair , social state .

4. **Quantitative variables : المتغيرات الكمية**

   Traits can be measured by numbers directly, like weight , length , milk and wool yield .

5. **Continuous variables : متغيرات مستمرة**

   The observations of these traits were take a value within a range , like the length of the students are ranged from 150 – 185 cm. ($150 \leq Y \leq 185$) .

6. **Discrete variables : متغيرات متقطعة**

   The observations of these traits were take non Continuous value (integer value) , like the number of hens in flock , the number of books in the library .

7. **Population : المجتمع**

   All possible values of variable .

8. **Sample : العينة**

   It is part of a population .

Statistical notations الرموز الإحصائية

As we said later , any variable has a symbol Y or X or Z , and any value has a symbol y_i or x_i or z_i , so if we have ages for 5 students as : 20, 18, 24, 22, 16 years we write them as :

Y_i = 20, 18, 24, 22, 16 .   so that :

$y_1= 20$ is the first value or observation .

$y_2= 18$ is the second value or observation . and ect…….

$Y_n= 16$ is the last value or observation .

And the total values of the variable are symbolic as ($\sum_{i=1}^{n} y_i$) which name (sigma) or (summation of ….)

Therefore , the symbol ($\sum_{i=1}^{n} y_i$) read as : summation of y value from y_1 to y_n ,

as : $\sum y_i = y_1 + y_2 + \ldots\ldots.. + y_n$

There are also a partial summation as : $\sum_{i=3}^{5} y_i$ which mean :

$\sum y_i = y_3 + y_4 + y_5$

Summation of all observations square , symbolic as : $\sum y_i^2$ which mean :

$\sum y_i^2 = y_1^2 + y_2^2 + \ldots\ldots. + y_n^2$

But square of observations summation , symbolic as : $(\sum y_i)^2$ which mean :

$(\sum y_i)^2 = (y_1 + y_2 + \ldots\ldots. + y_n)^2$

The symbol $\sum x_i y_i = x_1y_1 + x_2y_2 + \ldots\ldots\ldots + x_ny_n$

while the symbol $(\sum x_i)(\sum y_i) = (x_1+ x_2+\ldots..+x_n)(y_1+y_2+\ldots\ldots.+y_n)$

There are sum rules that are useful in collection process as following:

(1) .

If we have (C) as a fixed number , so :

$\sum c = nc$.

The proof : $\sum c = c_1 + c_2 + \ldots.. + c_n = nc$

(2) .

If we have (C) as a fixed number , so :

$\sum cy_i = c \sum y_i$

The proof : $\sum cy_i = cy_1 + cy_2 + \ldots\ldots + cy_n$

$= c (y_1 + y_2 + \ldots\ldots + y_n)$

$= c\sum y_i$

(3).

If we have two variables and we wanted to collected their values , as : $\sum(X_i + Y_i) = \sum X_i + \sum Y_i$

The proof :

$\sum(X_i + Y_i) = (x_1 + y_1) + (x_2 + y_2) + \ldots\ldots.. + (x_n + y_n)$

$= (x_1 + x_2 + \ldots\ldots +x_n) + (y_1 + y_2 + \ldots\ldots+ y_n)$

$= \sum x_i + \sum y_i$

Therefore , we must know there are deferent result between some statistical notations , for example:

$$\sum\frac{x_i}{y_i} = \frac{x_1}{y_1} + \frac{x_2}{y_2} + ... + \frac{x_n}{y_n}$$

While : $\frac{\sum x_i}{\sum y_i} = \frac{x_1 + x_2 + ... + x_n}{y_1 + y_2 + ... + y_n}$

As well as , $\sum(x_i - 3) = \sum x_i - n(3)$ is differ than $\sum x_i - 3$

<u>Ex</u>. If you know the value of both variables X and Y are :

X_i = 2, 6, 3, 1

Y_i = 3, 9, 6, 2 find value of these limits : $\sum(x_i - 3)$, $\sum x_i - 3$, $\sum\frac{x_i}{y_i}$, $\frac{\sum x_i}{\sum y_i}$,

$\sum(y_i + x_i)^2$, $\sum y_i^{\,2} - \frac{(\sum y_i)^2}{n}$ and $\sum(y_i - x_i)^2$.

(1). $\sum(x_i - 3) = \sum x_i - 3(n)$

$= (x_1+ x_2+x_3 + x_4) - (4*3)$

$= (2+6+3+1) – (12)$

$=12 – 12 = 0$

(2). $\sum x_i - 3 = (x_1 + x_2 + x_3 +x_4) – 3$

$= (2 + 6 + 3 + 1) – 3 = 12-3 = 9$

(3). $\sum\frac{x_i}{y_i} = \frac{x_1}{y_1} + \frac{x_2}{y_2} + \frac{x_3}{y_3} + \frac{x_4}{y_4}$

$= \frac{2}{3} + \frac{6}{9} + \frac{3}{6} + \frac{1}{2} = \frac{12+12+9+9}{18} = \frac{42}{18} = 2.33$

(4). $\frac{\sum x_i}{\sum y_i} = \frac{x_1 + x_2 + x_3 + x_4}{y_1 + y_2 + y_3 + y_4} = \frac{2+6+3+1}{3+9+6+2} = \frac{12}{20} = 0.\ 6$

(5). $\sum(y_i + x_i)^2 = \sum(y_1 + x_1)^2 + (y_2 + x_2)^2 + (y_3 + x_3)^2 + (y_4 + x_4)^2$

$$= (3+2)^2 + (9+6)^2 + (6+3)^2 + (2+1)^2$$

= 25 + 225 + 81 + 9 = 340

(6). $\sum (y_i - x_i)^2 = \sum (y_i{}^2 - 2x_i y_i + x_i{}^2)$

$$= \sum y_i{}^2 - 2\sum x_i y_i + \sum x_i{}^2$$

$$= (3^2 + 9^2 + 6^2 + 2^2) - 2(2*3 + 6*9 + 3*6 + 1*2) + (2^2 + 6^2 + 3^2 + 1^2)$$

= (9+81+36+4) – 2(6+54+18+2) + (4+36+9+1)

= (130) - 2(80) + (50) = 130-160+ 50 = 20

(7). $\sum y_i{}^2 - \frac{(\sum y_i)^2}{n} = \left(y_1{}^2 + y_2{}^2 + y_3{}^2 + y_4{}^2\right) - \frac{(y_1 + y_2 + y_3 + y_4)^2}{4}$

$$= (3^2 + 9^2 + 6^2 + 2^2) \ - \frac{(3+9+6+2)^2}{4}$$

= (9+81+36+4) $\frac{(20)^2}{4}$ = 130 – 100 = 30

The forth lecture

Measures Of Dispersion Or Variation مقاييس التشتت أو الاختلاف

It means the spacing or convergence between the observational values of a variable. And dispersion measures are measures of how the observations are dispersed from their mean . Whenever the dispersion is a large it indicates the heterogeneity between observations and whenever, if the dispersion was small so that indicates a few differences between observations values .

There are several measures of dispersion, the most important of them :

(First) Absolute dispersion measures : **مقاييس التشتت المطلق**

It units has the same units of the original values, and it is the most important measures :

1. The range. المدى

2 .The mean deviation. الأنحراف المتوسط

3. The variance and the standard deviation. التباين والأنحراف القياسي

(Second)The relative dispersion measures: **مقاييس التشتت النسبي**

which have no units of measurement, the most important is coefficient of variation (C.V) .

Absolute dispersion measures : مقاييس التشتت المطلق

(1). The Range (R) : المدى

It is the difference between upper value and lower value in that group .

Example : Find the range of these following groups :

Y_i = 12 , 6 , 7 , 3 , 15 , 10 , 18 and 5 .

X_i = 9 , 3 , 8 , 8 , 9 , 8 , 9 and 18 .

Solution : R of $y_i = y_{max} - y_{min} = 18 - 3 = 15$

R of $x_i = x_{max} - x_{min} = 18 - 3 = 15$

The range of both groups is similar but , actually we notice that the difference in group (y_i) is more than it in group (x_i) . Therefore , the range sometimes is mistaken because it dependent on both values side only .

(2). The Mean Deviation (M.D) : الأنحراف المتوسط

(A). Ungrouped data : بيانات غير مبوبة

If we have (n) of observations y_1 , y_2, ………y_n , so their mean deviation is the absolute mean deviation (without signal) from their arithmetic mean .

M.D = $\frac{\sum|y_i-\bar{y}|}{n}$

Example : Find the mean deviation of these values : $Y_i = 9 , 8 , 6 , 5 , 7$.

Solution :

y_i	$y_i - \bar{y}$	$\lvert y_i-\bar{y}\rvert$
9	2	2
8	1	1
6	- 1	1
5	- 2	2
7	0	0
$\bar{y}=\frac{\sum y_i}{n}=\frac{35}{5}=7$	0	6

So,M.D.= $\frac{\sum|y_i-\bar{y}|}{n}=\frac{6}{5}=1.2$

(B). Grouped data : بيانات مبوبة

If y_1 , y_2 , ………y_k represent classes marks in frequency table with their frequency f_1 , f_2 , ………f_k respectively , فأن الأنحراف المتوسط هو

M.D. = $\frac{\sum f_i|y_i-\bar{y}|}{\sum f_i}$

Example : Find the mean deviation for this frequency table :

Classes	f_i	y_i	f_iy_i	$\lvert y_i-\bar{y}\rvert$	$f_i\lvert y_i-\bar{y}\rvert$
60 – 62	5	61	305	6.45	32.25
63 – 65	18	64	1152	3.45	62.10
66 – 68	42	67	2814	0.45	18.90
69 – 71	27	70	1890	2.55	68.85
72 – 74	8	73	584	5.55	44.40
	100		6745		226.50

Solution :

$\bar{y}=\frac{\sum f_iy_i}{\sum f_i}=\frac{6745}{100}=67.45$

M.D. = $\frac{\sum f_i|y_i-\bar{y}|}{\sum f_i}=\frac{226.50}{100}=2.265$

(3). Variance and Standard Deviation التباين والأنحراف القياسي

(A). Ungrouped data : If we have (n) of observations y_1 , y_2 , y………y_n so , the variance (S^2) is :

$$s^2 = \frac{\sum (y_i - \bar{y})^2}{n-1} = \frac{\sum y_i^{\,2} - \frac{\left(\sum y_i\right)^2}{n}}{n-1}$$

Notice : sum of square (ss) = $\sum(y_i - \bar{y})^2$ so, $S^2 = \frac{ss}{n-1}$

Standard deviation (S) is the square root for the variance of that sample :

$$s = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{\sum y_i^{\,2} - \frac{\left(\sum y_i\right)^2}{n}}{n-1}}$$

Example : From this data , calculate the standard deviation ?

y_i = 9 , 8 , 6 , 5, 7 .

Solution : 1. lengthy method : الطريقة المطوّلة للحل

y_i	$y_i - \bar{y}$	$(y_i - y)^2$
9	9-7 = 2	4
8	8-7 =1	1
6	6-7 = -1	1
5	5-7 = -2	4
7	7-7 = 0	0
$\bar{y} = \frac{\sum y_i}{n} = \frac{35}{5} = 7$		10

$$s = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{10}{4}} = \sqrt{2.5} = 1.58$$ الأنحراف القياسي

1. Abbreviated method : الطريقة المختصرة

y_i	y_i^2
9	81
8	64
6	36
5	25
7	49
$\sum y_i = 35$	$\sum y_i^2 = 255$

$$ss = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 255 - \frac{(35)^2}{5} = 10$$

$$S = \sqrt{\frac{10}{4}} = \sqrt{2.5} = 1.58$$ but the variance is :

$$S^2 = \frac{10}{4} = 2.5$$

(B). Grouped data : If y_1 , y_2 , ...y_k represent classes marks in frequency table with their frequency f_1 , f_2 , ...f_k respectively , so their standard deviation is :

$$s = \sqrt{\frac{\sum f_i (y_i - \bar{y})^2}{\sum f_i - 1}} = \sqrt{\frac{\sum f_i y_i^2 - \frac{(\sum f_i y_i)^2}{\sum f_i}}{\sum f_i - 1}}$$

example : Calculate the standard deviation and the variance of this frequency table :

(1). lengthy method : الطريقة المطولة

Classes	f_i	y_i	f_iy_i	y_i-ȳ	$(y_i$-ȳ$)^2$	$f_i(y_i$-ȳ$)^2$
60 – 62	5	61	305	- 6.45	41.6	208.01
63 – 65	18	64	1152	- 3.45	11.9	214.24
66 – 68	42	67	2814	- 0.45	0.2	8.505
69 – 71	27	70	1890	2.55	6.5	175.5
72 – 74	8	73	584	5.55	30.8	246.4
	100		6745			852.75

$$s = \sqrt{\frac{\sum f_i (y_i - \bar{y})^2}{\sum f_i - 1}} = \sqrt{\frac{852.75}{99}} = 2.9$$ $s^2 = (2.9)^2 = 8.6$

(2). Abbreviated method : الطريقة المختصرة

Classes	f_i	y_i	f_iy_i	y_i^2	$f_iy_i^2$
60 – 62	5	61	305	3721	18605
63 – 65	18	64	1152	4096	73728
66 – 68	42	67	2814	4489	188538
69 – 71	27	70	1890	4900	132300
72 – 74	8	73	584	5329	42632
	100		6745		455803

$$s = \sqrt{\frac{\sum f_i y_i^2 - \frac{\left(\sum f_i y_i\right)^2}{\sum f_i}}{\sum f_i - 1}} = \sqrt{\frac{455803 - \frac{(6745)^2}{100}}{99}} = 2.9$$

$s^2 = (2.9)^2 = 8.6$

Relative dispersion measures : **مقاييس التشتت النسبي**

Relative dispersion measure are important when comparing the dispersion of two groups that differ in units of measurement for their values, because measures of relative dispersion are free of units of measure . The most important measures of relative dispersion are :

Coefficient of variation : معامل الاختلاف

If (S) and (y) is the standard deviation and the arithmetic mean for any data respectively , so their coefficient of variation (C.V.) is : C.V. = $\frac{s}{\bar{y}}$*100

Example : If the finishing results of statistics and chemistry lessens for the first class was :

	statistic	Chemistry
Arithmetic mean	78	73
Standard deviation	8	7.6

Where is the dispersion more in both lessens ?

Solution : to know that we find the coefficient of variation :

C.V. of statistic = $\frac{s}{\bar{y}}$*100 = $\frac{8}{78}$*100 = 10.25 %

C.V. of chemistry = $\frac{s}{\bar{y}}$*100 = $\frac{7.6}{73}$*100 = 10.41 % that mean the dispersion in chemistry was the more .

The seventh lecture **Test Of Hypothesis** **اختبار الفرضيات**

Is the most important subjects to take decisions with regard to any phenomenon in a population . Where are taken a sample from the population and we used all the information to get a decision to accept or reject the hypothesis statistical .

Hypothesis Statistical الفرضية الإحصائية

It is a clam or permit may be correct or wrong about a parameter or more in many or one population . So , the hypothesis will be accepted when the data of the sample were matching to the hypothesis , and will be rejected when the data of the sample were not matching to the hypothesis, therefore the researcher always trying to formulated the hypothesis in hopes rejected for example if the researcher wanted to make a compared between an importer chicken strain and a local chicken strain , so he put a hypothesis which said : there is no significant differ between both strains .

The hypothesis which formulated by the researcher with hoping to reject it is called (Null Hypothesis) which it's symbol (H_0) .

If we reject the null hypothesis , we must accept another hypothesis which called (Alternative hypothesis) it's symbol (H_i) , so it is the hypothesis which the researcher accepted it when he reject the null hypothesis .

The method which we follow to take the decision may be fill us in two error :

(1). Type 1 Error :

If the researcher reject the null hypothesis when it is right .

(2). Type π Error :

If the researcher accept the null hypothesis when it is wrong .

Steps to test hypotheses خطوات اختبار الفرضيات

1. Determine type of population distribution : if the variable is random and follow a normal distribution or another distribution .
2. Formulation the null and alternative hypotheses :

   If we tested that mean population (μ) equal to a certain value (μ_0) against the alternative hypothesis which said that $\mu \neq \mu_0$, so the null hypothesis (H_0) and the alternative hypothesis (H_i) would be as following :

   $H_0 : \mu = \mu_0$

   $H_i : \mu \neq \mu_0$

   For example , when we tested the degree of filling cans paste tomato by Karbala factory which it weight = 250 g , so we put the two hypotheses as following :

   $H_0 : \mu = 250$

   $H_i : \mu \neq 250$ that mean the average packing may be more or less than 250 g .
3. Selection of level of significance :

   It is the probability which we reject the null hypothesis when it is right , or it is the probability to fall in to the type I error .

   The researcher determine the degree of probability , in the agriculture sciences we always choice $\alpha = 0.05$ or 0.01 probably . Usually , we code to ($\alpha = 0.05$) by one star (*) and to

($\alpha = 0.01$) by two stars (**) . Therefore , if we used significant level (0.05) , that mean if we repeated the traits 100 times , the probability to fall in the error is five times , as well as (0.01) .

4. Determined the rejection and acceptance rejoins :

   Rejection Region :

   If the value of the statistical test fall on it caused to refuse the null hypothesis (H_0). We determined this region after we choice the significant level (α) and dedepending on the formulation of alternative hypothesis (H_i) .

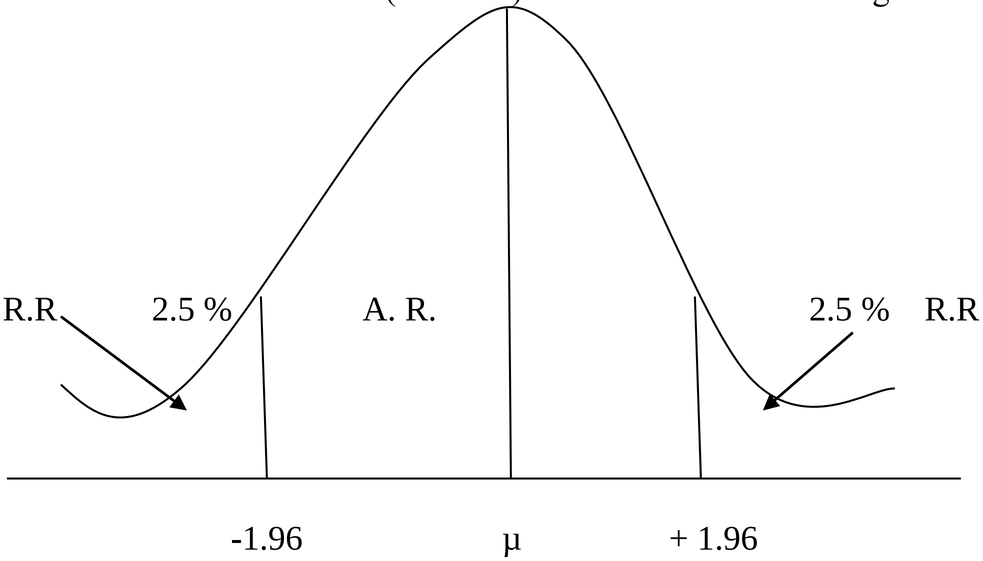
   Acceptance Region :

   If the value of the statistical test fall on it caused to accept the null hypothesis (H_0).

   There are three cases to determined the rejection and the acceptance regions as following :

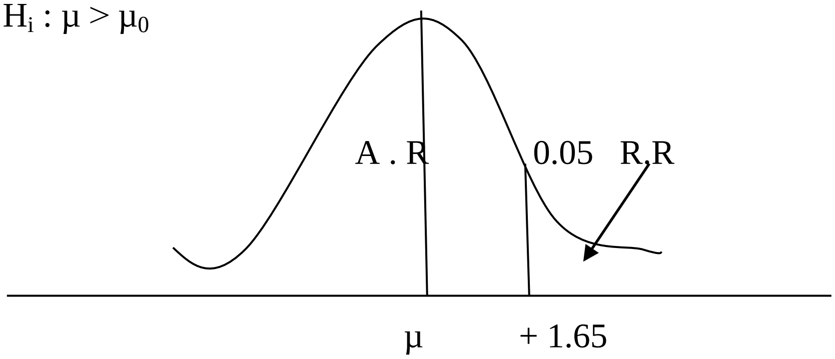
   A. If the alternative hypothesis (H_i) provided that the mean of sample don't equal the mean of population :

   $H_i : \mu \neq \mu_0$ that mean the value of mean sample either more or less than mean of Population . so, the reject region in this case will be on the both sides of mean population , and the significant level (α) will be (2.5%) on both sides of the mean when the researcher select ($\alpha = 0.05$) as shown in the following drawing :



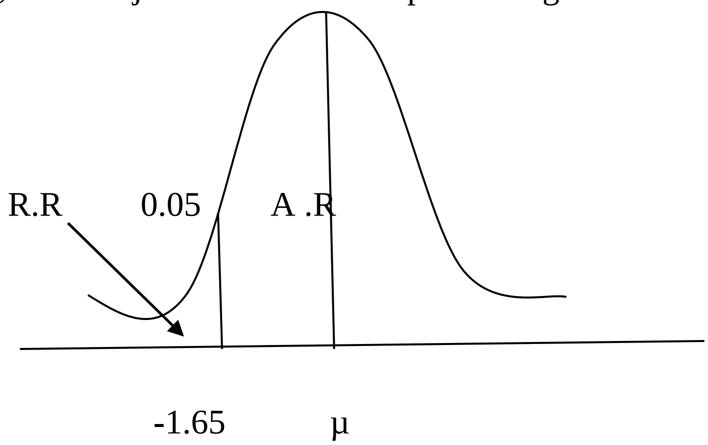
   B. If the alternative hypothesis (H_i) provided that the mean of sample more than a certain value (μ_0) , so the rejection and the acceptance regions will be as following :

$H_i : \mu > \mu_0$



C. If the alternative hypothesis (H_i) provided that the mean of sample less than a certain value (μ_0) , so the rejection and the acceptance regions will be as following :

$H_i : \mu < \mu_0$



5. Choice test – statistic :

This depended on the type of relationship between the theory values of population and their calculated values from the sample . The value of test-statistic which symbol is (Z) was determined through this equation :

$$Z = \frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

As the : $\bar{Y}$ = sample mean
μ = population mean
∂ = standerd error and n = sample number .

5. The Decision :
If the calculated value of (Z) was fall in rejection region , so we refused the null hypothesis (H_0) and accepted alternative hypothesis (H_i) , that mean the differences was significant between the theory values of population and the calculated values from the sample . But if the calculated value of (Z) was fall in acceptance region , so we accepted the null hypothesis (H_0) and refused alternative hypothesis (H_i) , that mean the differences was not significant between the theory values of population and the calculated values from the sample and it is not real but resulting from the chance .

Table show the limits of reject region when the test $H_0 : \mu = \mu_0$ against three cases of the alternative hypothesis (H_i) :

Test Cases	Refuse H_0 if $\alpha = 0.05$	Refuse H_0 if $\alpha = 0.01$
H_0: $\mu = \mu_0$ H_i $\mu \neq \mu_0$	$Z \geq 1.96$ and $Z \leq -1.96$	$Z \geq 2.58$ and $Z \leq -2.58$
$H_i : \mu > \mu_0$	$Z \geq 1.65$	$Z \geq 2.33$
$H_i : \mu < \mu_0$	$Z \leq -1.65$	$Z \leq -2.33$

First : Testes concerning one mean :

In this test we compared between the sample mean and the population mean to see if the sample belong to this population or not . so , if the result was opposite , that mean the calculated sample mean do not difference significantly from the population mean .

Example 1.

A company produced animal protein , claimed that the protein percentage in it's yield is at least (45 %) with standard deviation (S.D = 8 %) . To test this claim , we take a sample of (49) bag which their average protein-percentage was find (42.5 %) . therefore , is the company was honest in it claim at ($P \leq 0.01$) ?

Solution :

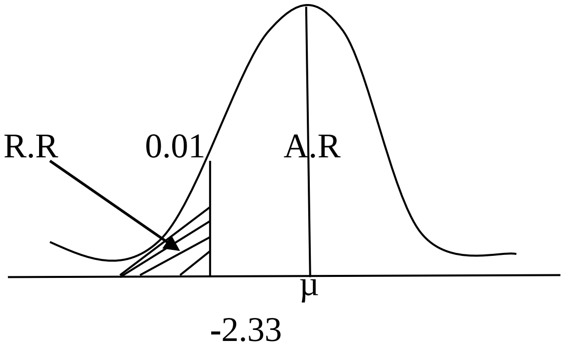
First we must formulate the null and alternative hypotheses as following :

$H_0 : \mu \geq 45$ %

$H_i : \mu < 45$ %

From the question , ($\alpha = 0.01$)

Second , we must determine the rejection and acceptance region as show in following drawing :



Third , find the calculate test-statistic value through this equation :

$$Z = \frac{\bar{Y}-\mu}{\frac{\partial}{\sqrt{n}}} = \frac{42.5-45}{\frac{8}{\sqrt{49}}} = -2.19$$

Forth , the decision : AS the calculated Z (- 2.19) is laying at the acceptance region , therefore we accept H_0 and refuse the H_i and that mean the company claim is true .

Example 2.

A certain company sales tomato juice , claimed that the percentage of vitamin C in each box is equal to (23)mg/100g with standard deviation of (2) mg/100g . If the average ratio of vitamin C in a sample of (64) pack was (20)mg/100g , is the company was true on her claim at probability 0.05 ?

Solution :

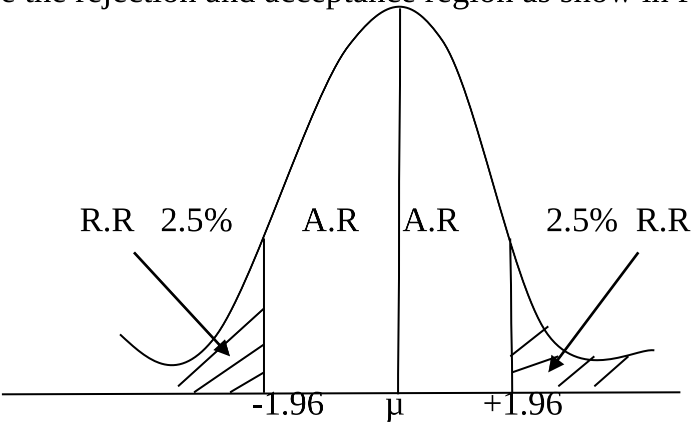
First we must formulate the null and alternative hypotheses as following :

$H_0 : \mu = 23$

$H_i : \mu \neq 23$

From the question , ($\alpha = 0.05$)

Second , we must determine the rejection and acceptance region as show in following drawing :



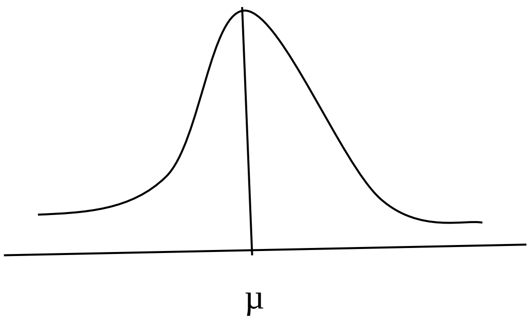
Third , find the calculate test-statistic value through this equation :

$$Z = \frac{\bar{Y} - \mu}{\frac{\partial}{\sqrt{n}}} = \frac{20-23}{\frac{2}{\sqrt{64}}} = -12$$

Forth , the decision : AS the calculated Z (- 12) is laying at the rejection region , therefore we refused H_0 and accepted the H_i and that mean the company claim is not true .

<u>The sixth lecture</u> **Normal Distribution** **التوزيع الطبيعي**

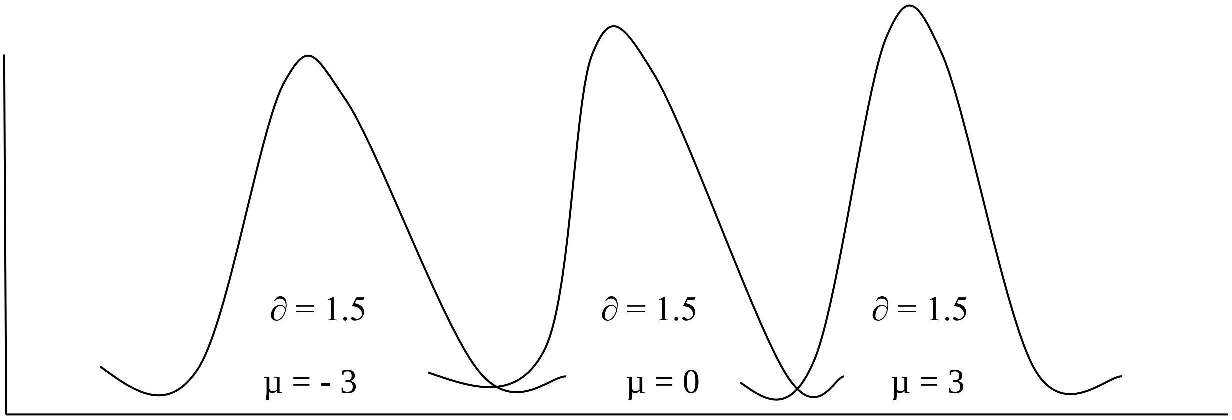
A lot of variables are distributed as normal distribution , including the biological , physiological , social and other important traits . The values of the variable which is distributed naturally take the form of the inverted bell or inverted cup . for this, sometimes we called it (Gaussian Curve) which derived it's equation , as it is evident in the following draw:



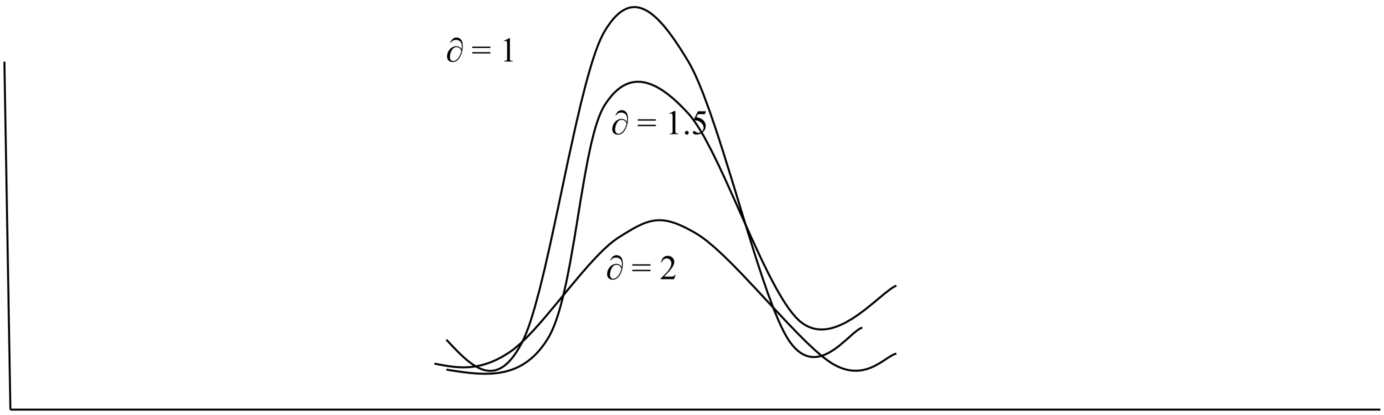
The normal distribution function depend on two things are :

The arithmetic mean (μ) and the variance (∂^2) which determine site and the form of the normal curve .

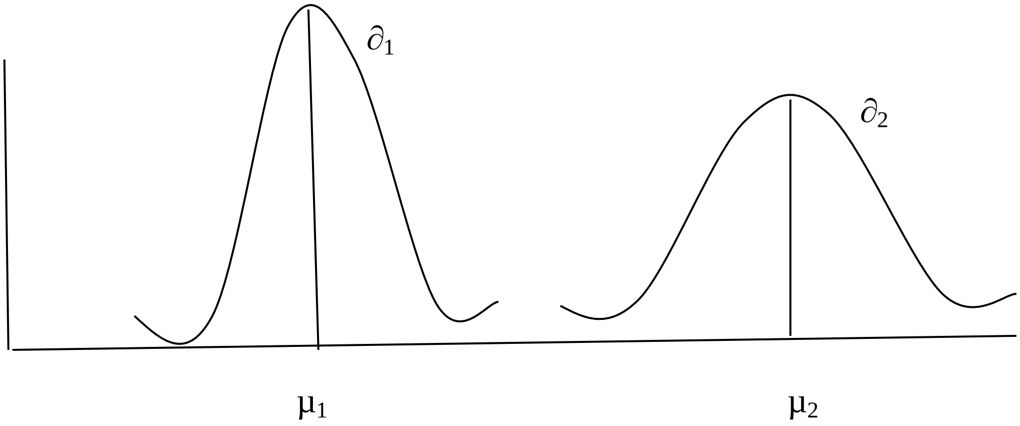
The following figure shows three normal distributions which have same standard deviation (∂) but their arithmetic means is different .



While the following figure shows three normal distributions which have same arithmetic means but their standard deviations is different .



While the following figure shows two normal distributions which have two different arithmetic mean and standard deviations .



Than previously can be summarized **<u>properties of normal distribution</u>** :

(1). The shape in the form of alarm or an inverted bell .

(2). The values are concentrated around the mean which divides the shap in to two equal parts .

(3). The shape has a peak point which decreased gradually when we move to both sides .

(4). The shape has a mean = 0 and a variance = 1.

(5). The total area under the normal curve = 1 .

To calculate the area under the normal curve which represent the degree of probability through change the random variables (y) which distributed naturally to standard normal distribution (Z) by this method

$$Z = \frac{y-\mu}{\sigma}$$

For example , if the (y) was between two limits (y_1) and (y_2) so, (Z) for it was between Z_1 and Z_2 where the :

$$z_1 = \frac{y_1-\mu}{\sigma}$$

$$Z_{2=}\frac{y_2-\mu}{\partial}$$

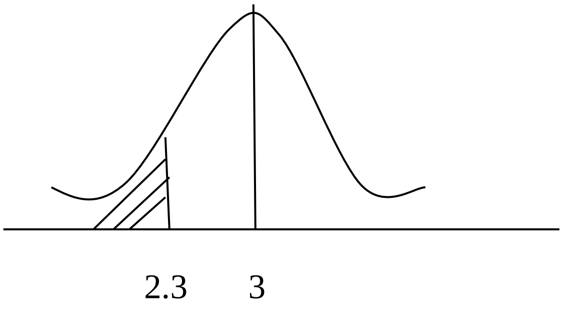
Therefore , the area that located between the two limits y_1 and y_2 equal to the area which located between Z_1 and Z_2 , in other words :

$P(y_1 < y < y_2)$ = $P(z_1 < Z < z_2)$

Note that the statistics books have a table give the areas under standard normal distribution which presents P (Z< z) values of Z which confined between (-3.4) and (+ 3.4) .

<u>Example 1. :</u> A certain type of car's battery , their average consumption is (3) years with a standard deviation (0.5) year , if the duration of consumption follow a normal distribution , what is the probability that a particular battery will consume less than (2.3) years ?

Solution :



$Z = \frac{y-\mu}{\sigma} = \frac{2.3-3}{0.5} = -1.4$

$\therefore P(y < 2.3) = P(z < -1.4)$

From the normal curve table:

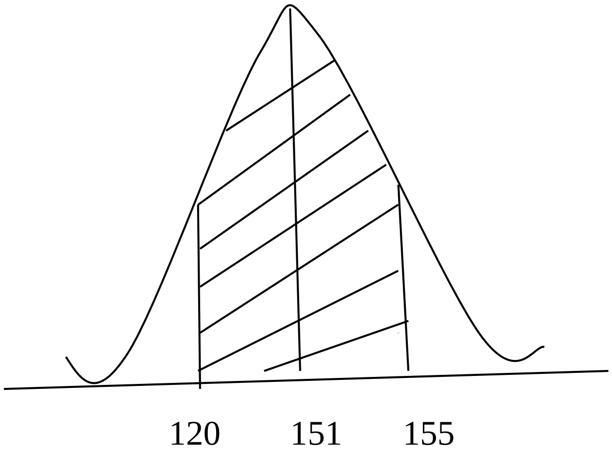
The area which meet (Z) value (-1.4) is = 0.0808 .

Therefore , the probability of this battery equal 8% .

<u>Example 2.</u> If the length's average of (500) students in a secondary school was (151 cm) with standard (15 cm) . Suppose that the length trait distributed naturally distribution , Find the number of students whose :

1. Their length between 120 – 155 cm .
2. Their length more than 181 cm .
3. Their length less than 128 cm .

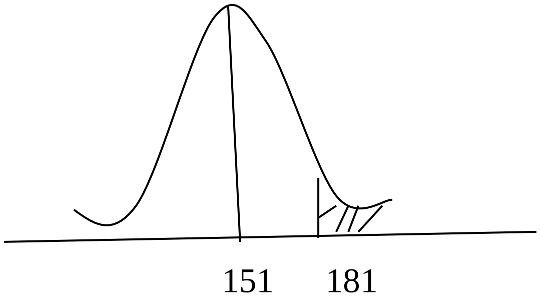
Solution:



1. $Z_1 = \frac{y_1-\mu}{\sigma} = \frac{120-151}{15} = -2.07$

   $Z_2 = \frac{y_2-\mu}{\sigma} = \frac{155-151}{15} = 0.27$

   $P(120 < y < 155) = P(-2.07 < z < 0.27)$
   $= P(z < 0.27) - p(z < -2.07)$ from the normal curve table :
   = 0.6064 – 0.0192 = 0.5872
   This means that about 60 % of pupils their length between 120 – 155 .
   Therefor , 0.60 * 500 = 300 pupils .

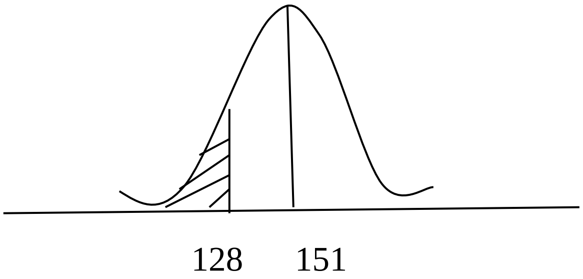


2. $z = \frac{y-\mu}{\sigma} = \frac{181-151}{15} = 2$

   P $(y > 181) = p(z > 2)$
   P$(y > 181)$ = 1 – p $(z < 2)$
   = 1 – 0.9772 = 0.228
   This means that about 2 % of pupils their length are more than (181 cm) .

Therefore , 0.02 * 500 = 10 pupils .

3. $z = \frac{y-\mu}{\sigma} = \frac{128-151}{15} = -1.53$



$p(y < 128) = p(y < -1.53)$

From the normal curve table , the area which meet (z) value (- 1.53) is = 0.0630

Therefore , the ratio of pupils their length are less than (128 cm) = 6.3 %

0.063 * 500 = 32 pupils .

Lecture third

Measures Of Central Tendency

Those measures are looking for estimate a value around which most of these data are centered, and this average value is number that expresses or represents all of these data.

The most important measures of central tendency are: -

(1): The arithmetic mean. الوسط الحسابي

(2): The median . الوسيط

(3): The mode . المنوال

(1): <u>The Arithmetic Mean</u> : It is the value resulting from dividing the sum of the observations values on its numbers and its symbol (ȳ). It's calculation methods :

(A). From ungrouped data : if we have (n) of observations y_1 , y_2 ………y_n so , the arithmetic mean for it is : ȳ = $\frac{\sum y_i}{n}$

Example (1). The following data represents amount of rain was falling yearly (ml) on the Mosul city during the five-year period ago, 520, 350, 450 380 400 .find the average amount of rainfall .

$$\bar{y} = \frac{\sum y_i}{n} = \frac{520+350+450+380+400}{5} = \frac{2100}{5} = 420 \text{ ml}$$

(B). From grouped data : if we have y_1 , y_2 …….. y_n represent class marks in frequency table with their frequencies f_1 , f_2 …….f_n respectively , so the arithmetic mean is :

$$\bar{y} = \frac{\sum f_i y_i}{\sum f_i}$$

Example (2). From the following frequency table of plantes length , calculate the arithmetic mean :

classes	F_i
31-40	1
41-50	2
51-60	5
61-70	15
71-80	25
81-90	20
91-100	12
total	$\sum f_i = 80$

Solution : follow these steps :

(1) estimate classes marks .

class mark of 1st class = lower limit + upper limit / 2 $= \frac{31+40}{2} = 35.5$ وهكذا بقية الفئات

(2)multiplied each class mark by it's frequency 35.5 * 1= 35.5, 45.5 * 2 = 91 وهكذا

(3)divided summation of (each class mark * it's frequency) by total frequency.

So, ȳ = $\frac{\sum f_i y_i}{\sum f_i} = \frac{6130}{80} = 76.62$ cm

classes	F_i	Y_i	F_i*y_i
31-40	1	35.5	35.5
41-50	2	45.5	91
51-60	5	55.5	277.5
61-70	15	65.5	982.5
71-80	25	75.5	1887.5
81-90	20	85.5	1710
91-100	12	95.5	1146
Total	$\sum f_i = 80$		$\sum f_i y_i = 6130$

Properties of arithmetic mean : **خواص الوسط الحسابي**

(A). The sum of deviations of values from their arithmetic mean = zero .

$\sum (y_i - \overline{y}) = 0$ for (ungrouped data)

$\sum f_i (y_i - \overline{y}) = 0$ for (grouped data)

The proof :

$\sum (y_i - \overline{y}) = \sum y_i - \sum \overline{y}$ for un grouped data

$= \sum y_i - n\overline{y} = \sum y_i - \sum y_i = 0$

$\sum f_i (y_i - \overline{y}) = \sum f_i y_i - \overline{y} \sum f_i$ for grouped data

$$= \sum f_i y_i - (\frac{\sum f_i y_i}{\sum f_i}) \sum f_i$$

$$= \sum f_i y_i - \sum f_i y_i = 0$$

Look for follow table to explanation that :

If we have these values : 8 , 3 , 5 , 12 , 10 .

Y_i	$(y_i - \bar{y})$
8	$8 - 7.6 = 0.4$
3	$3 - 7.6 = - 4.6$
5	$5 - 7.6 = - 2.6$
12	$12 - 7.6 = 4.4$
10	$10 - 7.6 = 2.4$
$\sum y_i = 38$, $\bar{y} = \frac{\sum y_i}{n}$ $\bar{y} = 7.6$	$\sum (y_i - y) = 0$

(B). The sum of the squares of the deviations from arithmetic mean is at least possible, that mean less than the sum of the squares of the deviates from any value except the value of arithmetic mean . so ,

$\sum (y_i - \overline{y})^2 =$ at least . since , we explained that through the following example :

Example : $y_i = 9 , 8 , 6 , 5 , 7$

$$\therefore \bar{y} = \frac{\sum y_i}{n} = \frac{9+8+6+5+7}{5} = 7$$

$$\sum (y_i - \overline{y})^2 = (9-7)^2 + (8-7)^2 + (6-7)^2 + (5-7)^2 + (7-7)^2 = 10$$

If we minus from the previous values any number (it symbol as A) ,(except arithmetic mean) as A=10 so, the sum of squares of the deviates will be :

$$\sum (y_i - A)^2 = \sum (y_i - 10)^2$$

$$= (9-10)^2 + (8-10)^2 + (6-10)^2 + (5-10)^2 + (7-10)^2 = 55$$

And it is normal that 55 is more than 10 .

(C). when we plus fixed number (k) to any value from observation sough that :

The arithmetic of a new values = the arithmetic of the original values + the fixed number (k) .

$X_i = y_i + k$

$\bar{X} = \bar{y}_i + k$

Example :If we have $y_i = 8 , 2 , 3 , 12 , 10$. so, their arithmetic mean $\bar{y} = \frac{\sum y_i}{n} = \frac{35}{5} = 7$

Since , if we add a fixed number (k=3) to each previous value :

So , the new values will be : $x_i = 11 , 5 , 6 , 15 , 13$. and their arithmetic mean is

$X = \frac{\sum x_i}{n} = \frac{50}{5} = 10$ which is in the fact x = ȳ + 3 = 7 + 3 = 10

(D). If we multiplied each value of the observation by fixed number (k) :

The arithmetic mean of new values = the arithmetic mean of original values * k

$Z_i = k\, y_i$ and $\bar{Z} = k\, \bar{y}$

Example : $y_i = 8 , 3 , 2 , 12 , 10$. and $\bar{y} = 7$ if : $z_i = 5y_i$ find value of $\bar{z}$.

Solution : $z_i = 40 , 15 , 10 , 60 , 50$.

$\bar{z} = \frac{\sum z_i}{n} = \frac{175}{5} = 35$ and this value is equal = (5)(ȳ) so , z̄ = (5)(7) = 35

We can generalization the two pervious properties by this following law :

If we have : $x_i = a + by_i$

So that : $\bar{x} = a + b\bar{y}$

(E). The arithmetic mean of sum two variables value = sum of two arithmetic variables .

$\bar{z} = \bar{x} + \bar{y}$ Example :

X_i	Y_i	$Z_i = x_i + y_i$
2	5	7
4	10	14
4	8	12
8	7	15
7	10	17
$\bar{x} = \frac{\sum x_i}{n} = \frac{25}{5} = 5$	$\bar{y} = \frac{\sum y_i}{n} = \frac{40}{5} = 8$	$\bar{z} = 13$

(F). If each value from the observations (y_i) has special weight fitting with her importance (w_i) , so the arithmetic mean (metrical) for these values is : $\bar{y} = \frac{\sum w_i y_i}{\sum w_i}$

Example :

The following values represent the examinations results of one student in statistics although we know each exam has weight or importance or certain percentage :

exam	Result (y_i)	Weight (w_i)	W_iy_i
1st	70	10 %	700
2nd	60	30 %	1800
3rd	75	10 %	750
4th	55	50 %	2750
		$\sum w_i = 100$	$\sum w_iy_i = 6000$

The arithmetic mean is ȳ = $\frac{\sum w_i y_i}{\sum w_i} = \frac{6000}{100} = 60$

(2). <u>The Median (M̄e)</u> : the value for which 50% of the observations , when arranged in order of magnitude , lie on each side .

(A). Ungrouped data :

If we have (n) of values or observations y_1, y_2 , ……..y_n and we arranged it cumulatively or descending :

(1). If the number of (n) is singular , so the median value is the value which their arrange is $\frac{n+1}{2}$ as M̄e = $y_{(n+1)}$ / 2

(2). If the number of (n) is even , so the median is the mean of the two values which their arrange is n/2 , n/2+1 as M̄e = $\frac{y_{n/2} + y_{(n/2)+1}}{2}$

<u>Example 1.</u> find the median of following values : 84 , 87 , 76 , 82 and 80 .

<u>Solution</u> : first , we must arranged the values cumulatively : 76 , 80 ,82 , 84 and 87

As , a number of (n) is singular (n=5) so , M̄e value is their arrange is :

$\frac{n+1}{2} = \frac{5+1}{2} = 3$ أي القيمة الثالثة حسب الترتيب التصاعدي اعلاه so , the M̄e = y_3 = 82

<u>Example 2.</u> find the median of following values : 5 , 4 , 8 , 7 , 3 , 12 , 9 and 2 .

<u>Solution :</u> we must arranged the values cumulatively : 2 , 3 , 4 , 5 , 7 , 8 , 9 and 12

As a number of (n) is even (n=8) so , the M̄e value is the mean of n/2 , n/2+1

$\frac{n}{2} = \frac{8}{2} = 4 (y_4)$ and $\frac{n}{2} + 1 = \frac{8}{2} + 1 = 5 (y_5)$ So , M̄e = $\frac{y_4 + y_5}{2} = \frac{5+7}{2} = 6$

(B). Grouped data : if y_1 , y_2 , ……..y_n represent classes marks in frequency table with frequency f_1 , f_2, ………f_k respectively , so the median value (be helped by less than cumulative frequency table) is :

$$\bar{\text{M}}\text{e} = L_1 + \left[\frac{(\sum f_i / 2) - F_i}{f_i} \right] w$$

Whereas : L_1 = the lower true limit of median class الحد الأدنى لفئة الوسيط

$\sum f_i$ = total frequency مجموع التكرارات

F_i = cumulative frequency at median class beginning التكرار المتجمع عند بداية فئة الوسيط

f_i = median class frequency = cumulative frequency at the ending of median class - cumulative frequency at the beginning of median class التكرار المتجمع عند نهاية فئة الوسيط – التكرار المتجمع عند بداية فئة الوسيط .

w = median class length طول الفئة

to find median value we make these steps : لأيجاد قيمة الوسيط نقوم بالخطوات التالية

1. make less than cumulative frequency table . عمل جدول توزيع تكراري تجمعي تصاعدي
2. find median arrange ($\sum f_i/2$) . أيجاد ترتيب الوسيط
3. determine class median by find two recessive values in the less than cumulative frequency which arrange median located between them . نحدد فئة الوسيط وهي الفئة التي تقع قيمة الوسيط بين حديها
4. apply the above law . نطبق القانون أعلاه

Example . From this frequency table find the median :

classes	f_i	Less than
		F_i
60 – 62	5	0
63 – 65	18	5
66 – 68	42	23
69 – 71	27	65
72 – 74	8	92
74	$\sum$fi = 100	100

Median arrange $= \frac{\sum f_i}{2} = \frac{100}{2} = 50$

L_1 = 65.5 the lower true limit of class median

F_i = 23 the cumulative frequency at the beginning of median class

f_i = 65 – 23 = 42 median class frequency

W = 68.5 – 65.5 = 3 median class length

So , M̄e $= L_1 + \left[\frac{(\sum f_i / 2) - F_i}{f_i} \right] w$

$= 65.5 + \left[\frac{50 - 23}{42} \right] (3) = 67.43$

We can also find the median value by using graphical presentation for both less and more than curves by rebate column (line) from their cross point to the horizontal axis to make cut on point which represent median value .

(C). <u>The Mode (M̄o):</u>

(A). Ungrouped data : If we have (n) of observations y_1 , y_2 , ...y_n so the mode is the observation which has more frequency . From that my be there one mode or two mode or more than two mode or may have no mode .

Example : Find the mode for these data : a.3 , 5 , 2 , 6 , 5 , 9 , 5 , 2 , 8 and 6 .

M̄o = 5

b. 51.6 , 48.7 , 50.3 , 49.5 and 48.9 .

M̄o = there is no mode .

(B). Grouped data : If y_1 , y_2 , ...y_n represent classes marks in frequency table with their frequency f_1 , f_2 , ...f_n respectively , so the mode is :

M̄o = $L_1 + \left(\frac{d_1}{d_1 + d_2} \right) w$

As : L_1 = The upper true limit of mode class . الحد الأعلى الحقيقي لفئة المنوال

d_1 = the difference between mode class and previous class . الفرق بين تكرار فئة المنوال والفئة السابقة لها

d_2 = the difference between mode class and follow class . الفرق بين تكرار فئة المنوال والفئة اللاحقة لها

w = class length طول الفئة

Example : Find the mode of this frequency table :

classes	f_i
60 – 62	5
63 – 65	18
66 – 68	42
69 – 71	27
72 – 74	8
	∑fi =100

Solution : the mode class is (66 – 68) because it has the largest frequency (42) . so,

$L_1 = 65.5$

$d_1 = 42 - 18 = 24$

$d_2 = 42 - 27 = 15$

$w = 3$

$\bar{M}o = 65.5 + \left(\frac{24}{24+15}\right)(3) = 67.35$