MEASURES OF CENTRAL TENDENCY

In order to describe a data set, we frequently seek out a representative or typical numerical value. It's common to refer to this value as the average. Since such typical values tend to lie centrally within a set of data arranged according to magnitude, averages are also known as measures of central tendency.

THE ARITHMETIC MEAN

One of the most widely used measures of central tendency is the arithmetic mean or briefly the mean. For a set of data, the mean can be defined as the sum of the data entries divided by the number of entries.

The mean for ungrouped data: For a set of n items $x_1, x_2, x_3, \ldots., x_n$, the mean $\bar{X}$ (read x bar) is calculated by the following formula:

$$\bar{X} = \frac{\sum x_i}{n}$$

Example

The weights (in kilograms) for a sample boxes are listed below.

15 17 22 16 20 25

What is the mean weigh of the boxes?

To find the mean weigh, divide the sum of the weighs by the number of weighs in the sample.

$$\bar{X} = \frac{15 + 17 + 22 + 16 + 20 + 25}{6} = 19.16$$

So, the mean weigh of the boxes is about 19.16 kg

The mean for grouped data (The weighted mean): A weighted mean is the mean of a data set whose entries have varying weights.

The weighted mean for discrete variable: If $x_1, x_2, x_3, \ldots., x_k$, are data points and $n_1, n_2, x_3, \ldots., n_k$, represent their respective frequencies, then the weighted mean is given by

$$\bar{X} = \frac{\sum(n_i * x_i)}{\sum n_i}$$

Example

In order to study the housing occupancy rate, we have the following data

Number of residents x_i	2	3	4	5	6	7
Number of houses n_i	6	6	8	7	3	3

Find the mean of the data

$$\bar{X} = \frac{2*6+3*6+4*8+5*7+6*3+7*3}{6+6+8+7+3+3} = \frac{136}{33} = 4.12 \approx 4$$

The value of the mean is **abstract**. The number of residents per house is about 4.12, number which does not correspond to a concrete fact.

The weighted mean for continuous variable: If a continuous variable data are presented in a frequency distribution, then the mean of the frequency distribution for a sample is approximated by

$$\bar{X} = \frac{\sum(n_i * c_i)}{\sum n_i}$$

where c_i is the midpoint of a class.

Example

Monthly salaries sample (in thousands of dinars) for administration employees.

Class Range	[20 30[	[30 40[	[40 50[	[50 60[	[50 60[	[60 70[

Frequency	7	11	6	9	5	4

Find the mean of the frequency distribution.

To find the mean of the frequency distribution, the computational procedure is as follows.

- Find the sum of the frequencies
- Find the midpoint of each class (c_i).
- Find the sum of the products of the midpoints and the frequencies ($n_i\, c_i$).
- Calculate the weighted mean of the frequency distribution.

Class Range	Frequency	Midpoints	$n_i\, c_i$
[20 30[	7	25	175
[30 40[	11	35	385
[40 50[	6	45	270
[50 60[	9	55	495
[60 70[	5	65	325
[70 80[	4	75	300
SUM	**42**	/	**1950**

$$\bar{X} = \frac{1950}{42} = 46.428$$

So, the mean salary is about 46.428(10^3) DA

Properties of the arithmetic mean

1- The arithmetic mean is fully representative since it considers all items observed.

2- The arithmetic mean is affected by outliers.

3- The algebraic sum of the deviations of a set of numbers x_i from their arithmetic mean is zero.

$\sum(x_i - \bar{x}) = 0$

4- The sum of the squares of the deviations of a set of numbers x_i from their arithmetic mean is a minimum.

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

5- If k_1 numbers have mean m_1, k_2 numbers have mean m_2 , . . . , k_j numbers have mean m_j, then the mean of all the numbers is the weighted arithmetic mean of all the means given by

$$\bar{X} = \frac{k_1 m_1 + k_2 m_2 + \cdots + k_j m_j}{k_1 + k_2 + \cdots + k_j}$$

GEOMETRIC MEAN

The geometric mean is a measure of the central tendency of a positive random variable and is used in a very broad range of natural and social science disciplines. For instance, the geometric mean is used in economics and finance to compute growth rate average and cumulative compounding rates.

The geometric mean for ungrouped data: The geometric mean G of a set of N positive numbers x_1, x_2 , x_3 , . . . , x_n is the N^th root of the product of the numbers:

$$G = \sqrt[n]{x_1 x_2 x_3 \ldots . x_n}$$

We can also compute The geometric mean by logarithms as follow

$$Ln(G) = \frac{1}{n}\sum Lnx_i \Rightarrow G = e^{Ln(G)} = e^{\left(\frac{1}{n}\sum Lnx_i\right)}$$

The above formula shows that the logarithm of the geometric mean of a set of data is the weighted arithmetic mean of the logarithms of the data set entries.

Example: find the geometric mean of the numbers 3, 8 and 9

$G = \sqrt[n]{(3)(8)(9)} = 6$

$Ln(G) = \frac{1}{3}\big(Ln(3) + Ln(8) + Ln(9)\big)$

$= 1.79175947 \Rightarrow G = e^{1.79175947}$

$\Rightarrow G = 6$

The geometric mean for grouped data: For grouped data If x₁, x₂, x₃, …., xₖ, are data points (or c₁, c₂, c₃, …., cₖ, are midpoints) and n₁, n₂, x₃, …., nₖ, represent their respective frequencies, then the weighted geometric mean is given by the following formula

$$G = \sqrt[N]{(x_1)^{n_1}(x_2)^{n_2}(x_3)^{n_3}\ldots.(x_k)^{n_k}}$$

Where $N = \sum_{i=1}^{k} n_i$

And also in logarithms form

$$Ln(G) = \frac{1}{N}\sum n_i\, Lnx_i$$

Example (a): find the geometric mean for the data in the table below

Number of residents x_i	Number of houses n_i	Ln x_i	n_i Ln x_i
2	1	0,69314718	0,69314718
3	3	1,09861229	3,29583687
4	2	1,38629436	2,77258872
5	1	1,60943791	1,60943791
∑	7	/	**8,37101068**

1st method

$G = \sqrt[7]{(2)^1(3)^3(4)^2(5)^1}$

$= \sqrt[7]{4320}$

$= 3.306$

2nd method

$Ln(G) = \frac{1}{7}\,(8.37101068) = 1{,}19585867$

$G = e^{1.19585867} = 3.306$

Example (b): find the geometric mean for the following frequency distribution

Class Range	**Frequency**	**Midpoints (c_i)**	**Ln c_i**	**n_i Ln c_i**
[20 30[	7	25	3,218875825	22,53213077
[30 40[	11	35	3,555348061	39,10882868
[40 50[	6	45	3,80666249	22,83997494
[50 60[	9	55	4,007333185	36,06599867
[60 70[	5	65	4,17438727	20,87193635
[70 80[	4	75	4,317488114	17,26995245
SUM	**42**	/		**158,688822**

$Ln(G) = \frac{1}{42}\,(158{,}688822)$

$= 3{,}77830528$

$G = e^{3{,}77830528}$

$= 43{,}7418488$

The geometric mean is used in particular to calculate the growth rate average of some phenomena, such as economic growth rates, population growth, etc.

THE HARMONIC MEAN

The harmonic mean is used in cases where the variable being studied is a ratio of two variables, as in calculating a speed average (kilometers divided by the number of hours) or that of an average density (number of inhabitants divided by the area in m^2), etc.

The harmonic mean for ungrouped data: The harmonic mean H of a set of non-zero numbers x_1, x_2, x_3, . . ., x_n is the reciprocal arithmetic mean of the reciprocals of these numbers:

$$H = \frac{n}{\sum \frac{1}{x_i}} \qquad H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \cdots + \frac{1}{x_n}}$$

Example (a): find the harmonic mean of the numbers 4,5, 8 and 10

$$H = \frac{4}{\frac{1}{4} + \frac{1}{5} + \frac{1}{8} + \frac{1}{10}}$$

$$= \frac{4}{0.25 + 0.2 + 0.125 + 0.1}$$

$$= 5.926$$

The harmonic mean for grouped data: If x_1, x_2, x_3, …., x_k, are non-zero data points (or c_1, c_2, c_3, …, c_k, are midpoints) and n_1, n_2, x_3, ..., n_k, represent their respective frequencies, then the weighted harmonic mean is given by the following

$$H = \frac{\sum n_i}{\sum \frac{n_i}{x_i}}$$

Example (b): find the harmonic mean for the following frequency distribution

Class Range	Frequency	Midpoints (c_i)	n_i /c_i
[20 30[	7	25	0,28
[30 40[	11	35	0,314
[40 50[	6	45	0,133
[50 60[	9	55	0,164
[60 70[	5	65	0,077
[70 80[	4	75	0,053
SUM	**42**	/	1,022

$$H = \frac{42}{1.022} = 41.115$$

Example (c):

A car drives for 200 kilometers at 80 km/h, then for 150 kilometers at 100 km/h.
What is the average speed during its trip?

$$H = \frac{200 + 150}{\frac{200}{80} + \frac{150}{100}} = \frac{350}{4} = 87.5$$

So, the average speed during the trip is 87.5 km/h.

THE QUADRATIC MEAN

The quadratic mean of a set of numbers is the square root of the arithmetic mean of the squares of the numbers.

The quadratic mean is frequently used in physical applications and also intervenes in the definition of other statistical indicators, such as standard deviation.

THE QUADRATIC MEAN FOR UNGROUPED DATA

in the case of a set of (n) numbers $x_1, x_2, ..., x_n$, the quadratic mean, is defined by

$$Q = \sqrt{\frac{\sum x_i^2}{n}}$$

Example (c): find the quadratic mean of the set 3, 4, 8, 10

$$Q = \sqrt{\frac{3^2 + 4^2 + 8^2 + 10^2}{4}} = 6.874$$

THE QUADRATIC MEAN FOR GROUPED DATA

For grouped data If $x_1, x_2, x_3, …., x_k$, are data points (or $c_1, c_2, c_3, …., c_k$, are midpoints) and $n_1, n_2, x_3, …., n_k$, represent their respective frequencies, then the weighted quadratic mean is given by

$$Q = \sqrt{\frac{\sum n_i x_i^2}{\sum n_i}}$$

Example: find the quadratic mean for the following frequency distribution

Class Range	Frequency	Midpoints (c_i)	c_i^2	$n_i\,c_i^2$
[20 30[	7	25	625	4375
[30 40[	11	35	1225	13475
[40 50[	6	45	2025	12150
[50 60[	9	55	3025	27225
[60 70[	5	65	4225	21125
[70 80[	4	75	5625	22500
SUM	**42**	/	/	100850

$$Q = \sqrt{\frac{100850}{42}} = 49$$

Note: The previously calculated averages satisfy the following relationship

$$H < G < \bar{X} < Q$$

THE MEDIAN

The median of a set of data is a value that divides the data set into two equal halves. This will therefore be the value of the variable such that 50% of the population is above and 50% is below.

THE MEDIAN OF UNGROUPED DATA

To find the median of a data set, the computational procedure is as follows:
arrange the data in order of magnitude in increasing or decreasing order.
If the number of observations is odd, the median is the middle value of the ordered list. $Me = X_{\frac{n+1}{2}}$

If the number of observations is even, the median is the arithmetic mean of the two values closest to the middle of the ordered list.

$$Me = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}$$

Example: find the median for the following set of numbers
(i) 3, 5, 6, 14, 18, 10, 7
(ii) 12, 22, 18, 15, 19, 10, 24, 20

solution

(i) Re-arranging the numbers in ascending order, we have

X_1	X_2	X_3	X_4	X_5	X_6	X_7
3	**5**	**6**	**7**	**10**	**14**	**18**

Because there are seven entries (an odd number), the median is the middle data entry.

$$Me = X_{\frac{7+1}{2}} = X_4 = 7$$

(ii) Re-arranging the numbers in ascending order, we have

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
10	**12**	**15**	**18**	**19**	**20**	**22**	**24**

Since the data set has an even number of entries (8), the median is the mean of the two middle data entries.

$$Me = \frac{X_{\frac{8}{2}} + X_{\frac{8}{2}+1}}{2} = \frac{X_4 + X_5}{2} = \frac{18 + 19}{2} = 18.5$$

THE MEDIAN OF GROUPED DATA

The median for discrete variable: in the case of a discrete variable, the median can be obtained using the cumulative frequencies or the cumulative relative frequencies; the median is then the value of the variable with which a cumulative frequency of N/2 is associated or a cumulative relative frequency of 50% of the data set.

If the median does not fall on an exact value of the variable, for convention, we retain the value of the immediately superior variable.

Example: the following table represent the housing occupancy rate in a city

Number of residents x_i	2	3	4	5	6	7
Number of houses n_i	6	6	8	7	3	3

Find the median of the data

Solution

To find the median, first we find the cumulative frequency (or cumulative relative frequency) and then determine the rank of the median (N/2) (or 0.5).

$$\frac{N}{2} = \frac{33}{2} = 16.5$$

The rank of the median N/2 = 16.5

The median

$M_e = 4$

x_i	frequencies n_i	Relative frequencies (f_i)	Cumulative frequencies n_i	Relative frequencies (f_i)
2	6	0,182	6	0,182
3	6	0,182	12	0,364
4	8	0,242	20	0,606
5	7	0,212	27	0,818
6	3	0,091	30	0,909
7	3	0,091	33	1
	33	1		

The rank of the median = 0.5

Since the rank of the median lies between the respective cumulative values 12 and 20 corresponding to the variable values 3 and 4, respectively, the value of the median is the largest value, i.e. Me = 4

The median for continuous variable: for continuous variable the median can be obtained either by interpolation or graphically.

a) The interpolation approach: To find the median, the computational procedure is as follows.
- find the cumulative frequency (or cumulative relative frequency)
- determine the rank of the median (N/2) (or 0.5).
- determine the median class, which is the class containing the median
- calculate the median by the formula

$$Me = L_l + \frac{\frac{N}{2} - (\sum n_i)_{l-1}}{n_e} * k$$

L_l: lower class boundary of the median class
N: total frequencies
$(\sum n_i)_{l-1}$: sum of frequencies of all classes lower than the median class
n_e: frequency of the median class
k: length of the median class interval

Example: find the median for the following frequency distribution of workers according to their age.

x_i	[20 25[	[25 30[	[30 35[	[35 40[	[40 45[	[45 50[	[50 55[	[55 60[
n_i	4	10	24	34	14	8	4	2

Solution

1- First we find the cumulative frequency (or cumulative relative frequency)

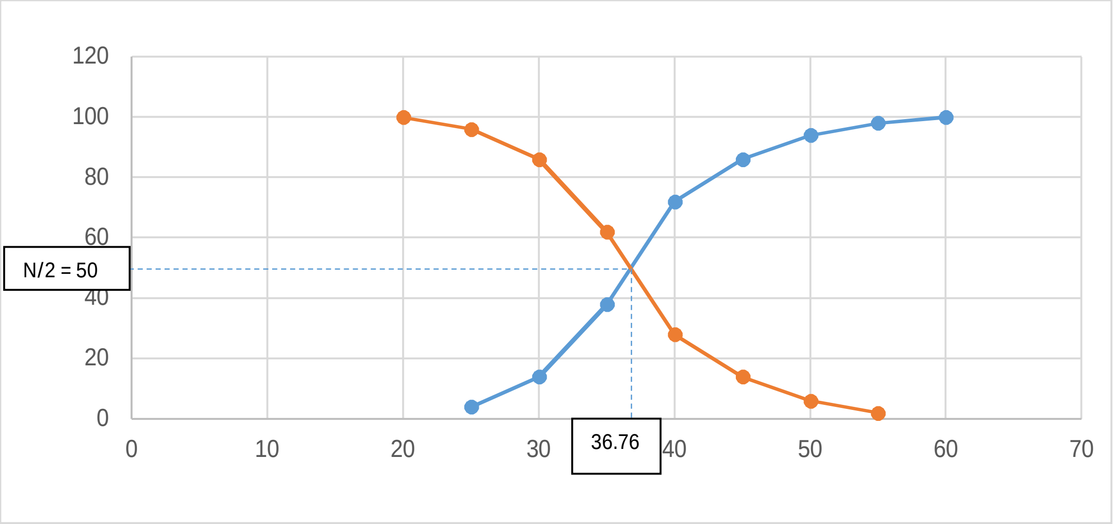
xi	n_i	n_i ↗	n_i ↘	f_i	f_i ↗	f_i ↘
[20 25[	4	4	100	0,04	0,04	1
[25 30[	10	14	96	0,1	0,14	0,96
[30 35[	24	38	86	0,24	0,38	0,86
[35 40[	34	72	62	0,34	0,72	0,62
[40 45[	14	86	28	0,14	0,86	0,28
[45 50[	8	94	14	0,08	0,94	0,14
[50 55[	4	98	6	0,04	0,98	0,06
[55 60[	2	100	2	0,02	1	0,02
	100			1		

50

2- We determine the rank of the median N/2 = 100/2 = 50 (50% of the data lie between 38 and 72)
3- The median class is [35 40[
4 - We calculate the median as follow

$$Me = 35 + \frac{\frac{100}{2} - 38}{34} * 5 = 36.76$$

b) The graphical approach: The median of a grouped data can be obtained using the cumulative frequency curve (ogive) and finding from it the value ‘x’ at the 50% point (abscissa), which corresponds to the intersection of the less then and the more then ogives.



QUANTILES

All quantities that are defined as partitioning a set of data arranged in order of magnitude into a number of equal portions are called quantiles. Examples include the quartiles, deciles and the percentiles.

QUARTILES: the quartiles are the values that divides a set of data into four equal parts. These values, denoted by Q_1, Q_2, and Q_3, are called the first, second, and third quartiles.
The first quartile Q_1 is the value below which we find 25% of the data.
The second quartile Q_2 is the median.
The third quartile Q_3 is the value below which we find 75% of the data.
The calculation of the quartiles for both ungrouped and grouped data is similar to parallel calculations of the median for ungrouped and grouped data using appropriately modified versions.

QUARTILES OF UNGROUPED DATA

The calculation of the quartiles for ungrouped data is similar to parallel calculations of the median for ungrouped data using appropriately modified versions.
- arrange the data in order of magnitude in increasing order.
- If N is odd

$$Q_i = X_{\frac{i(n+1)}{4}}$$

If N is even

$$Q_i = X_{\frac{i\,n}{4}}$$

QUARTILES OF GROUPED DATA

The calculation of the quartiles for grouped data is similar to parallel calculations of the median for grouped data using appropriately modified versions.
quartiles for discrete variable: The Q_i quartile is the value of the variable with which a cumulative frequency of (i*N/4) is associated.

quartiles for continuous variable: the quartiles can be obtained by the formula

$$Q_i = L_l + \frac{\frac{i\,N}{4} - (\sum n_i)_{l-1}}{n_{Q_i}} * k$$

DECILES: the deciles are the values that spilt a set of data into ten equal parts. These values, denoted by D_1, D_2, … D_9, are called the first decile, second decile, up to ninth decile.

DECILES OF UNGROUPED DATA

Deciles are determined in the same manner as quartiles
- arrange the data in order of magnitude in increasing order.
- If N is odd

$$D_i = X_{\frac{i(n+1)}{10}}$$

If N is even

$$D_i = X_{\frac{i\,n}{10}}$$

DECILES OF GROUPED DATA

The calculation of the deciles for grouped data is similar to parallel calculations of the quartiles for grouped data using appropriately modified versions.

The deciles for discrete variable: The D_i decile is the value of the variable with which a cumulative frequency of (i*N/10) is associated.

The quartiles for continuous variable: deciles can be obtained by the formula

$$D_i = L_l + \frac{\frac{i\,N}{10} - (\sum n_i)_{l-1}}{n_{D_i}} * k$$

PERCENTILES: percentiles are the values that spilt a set of data into one hundred equal parts. These values, denoted by P_1, P_2, … P_{99}, are called the first percentile, second percentile, up to ninety ninth percentile.

PERCENTILES OF UNGROUPED DATA

percentiles are determined in the same manner as quartiles and deciles

- arrange the data in order of magnitude in increasing order.
- If N is odd

$$P_i = X_{\frac{i(n+1)}{100}}$$

If N is even

$$P_i = X_{\frac{i\,n}{100}}$$

PERCENTILES OF GROUPED DATA

The calculation of the percentiles for grouped data is similar to parallel calculations of the quartiles for grouped data using appropriately modified versions.

percentiles for discrete variable: The P_i percentile is the value of the variable with which a cumulative frequency of (i*N/100) is associated.

percentiles for continuous variable: percentiles can be obtained by the formula

$$P_i = L_l + \frac{\frac{i\,N}{100} - (\sum n_i)_{l-1}}{n_{P_i}} * k$$

The graphical approach: quantiles of a grouped data can be obtained using the cumulative frequency or the cumulative relative (or percent) frequency curve (ogive) and finding from it the value 'x' at the appropriate % point (abscissa).

Example: the following table gives the distribution of workers according to their age.

x_i	[20 25[	[25 30[	[30 35[	[35 40[	[40 45[	[45 50[	[50 55[	[55 60[
n_i	4	10	24	34	14	8	4	2

For the above frequency distribution, find the following

The first and third quartile

The ninth decile

The ninety fifth percentile

Solution

1- First we find the cumulative frequency (or cumulative relative frequency)

xi	n_i	$n_i \nearrow$	f_i	$f_i \nearrow$
[20 25[	4	4	0,04	0,04
[25 30[	10	14	0,1	0,14
[30 35[	24	38	0,24	0,38
[35 40[	34	72	0,34	0,72
[40 45[	14	86	0,14	0,86
[45 50[	8	94	0,08	0,94
[50 55[	4	98	0,04	0,98
[55 60[	2	100	0,02	1
	100		1	

For quartiles

2- We determine the rank of the first quartile N/4 = 100/4 = 25 (25% of the data lie between 14 and 38)

3- The first quartile class is [30 35[

4 - We calculate the first quartile as follow

$$Q_1 = 30 + \frac{\frac{100}{4} - 14}{24} * 5 = 32.29$$

By the same way we find the third quartile

$$Q_3 = 40 + \frac{\frac{3(100)}{4} - 72}{14} * 5 = 41.07$$

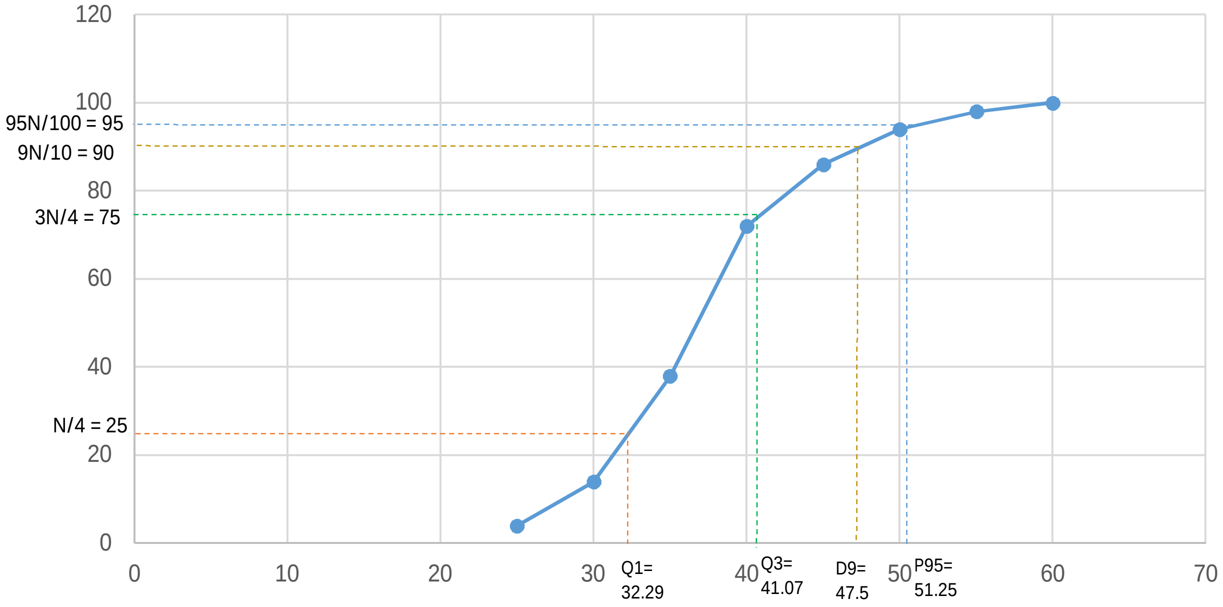
For the ninth decile

$$D_9 = 45 + \frac{\frac{9(100)}{10} - 86}{8} * 5 = 47.5$$

For the ninety fifth percentile

$$P_{95} = 50 + \frac{\frac{95(100)}{100} - 94}{4} * 5 = 51.25$$

Graphically



THE MODE

The most frequent value of a distribution is called the mode.

A distribution having one mode, two modes, or more than two modes are called respectively Unimodal, bimodal or multi – modal distribution. the mode sometimes does not exist If no value is repeated, or if all classes have the same frequency.

THE MODE OF UNGROUPED DATA

The mode of a set of data is that value which occurs with the greatest frequency

Example: find the mode for the following set of numbers

(i) 3, 5, 6, 5, 18, 5, 7
(ii) 1, 4, 2, 4, 6, 6, 8, 4, 6, 3
(iii) 12, 22, 18, 15, 19, 10, 24, 20

solution

(i) The mode is 5, with frequency 3
(ii) The set is bimodal, it has two modes 4 and 6
(iii) There is no mode.

THE MODE OF GROUPED DATA

The mode of a grouped distribution is the value at the point around which the items tend to be most heavily concentrated.

The mode for discrete variable: The mode is the value of the variable most frequently observed

Example: find the mode of the following distribution

Number of residents x_i	2	3	4	5	6	7
Number of houses n_i	6	6	8	7	3	3

Solution

The mode is 4, with frequency 8

The mode for continuous variable: for continuous variable the mode can be obtained either by interpolation or graphically.

- The mode by formula is as follow

$$M_o = L_l + \frac{\Delta_1}{\Delta_1 + \Delta_2} * k$$

L_l: lower class boundary of the modal class
Δ_1: the difference between the frequency of the modal class and the frequency of the class immediately before the modal class.
Δ_2: the difference between the frequency of the modal class and the frequency of the class immediately after the modal class,
k: length of the median class interval

If the class widths are different, in order to calculate the mode, it is necessary to find the corrected frequencies as mentioned above.

- Graphical method: The mode for continuous variable data can be obtained using the histogram.

Example (a): Class widths are equal

find the mode of the following distribution using the formula and the graphical method.

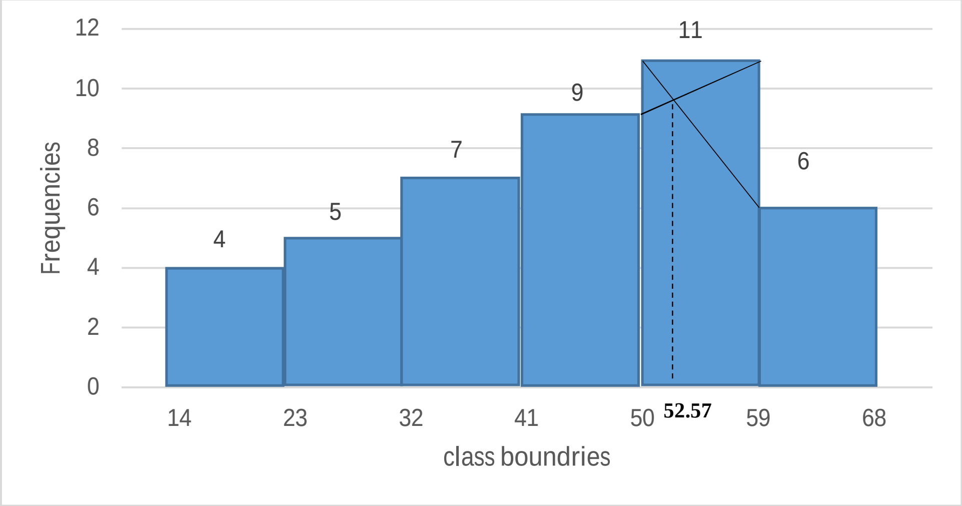
Class Range	[14 23[	[23 32[	[32 41[	[41 50[	[50 59[	[59 68[
Frequency	4	5	7	9	11	6

Solution

$$M_o = L_l + \frac{\Delta_1}{\Delta_1 + \Delta_2} * k$$

$$= 50 + \frac{11-9}{(11-9)+(11-6)} * 9 = 52.57$$

graphical method



Example (b): Class widths are different

find the mode of the following distribution using the formula and the graphical method.

x_i	[20 25[	[25 35[	[35 45[	[45 60[	[60 80[	[80 85[
n_i	10	30	40	45	50	10

In order to find the mode, we have to calculate the corrected frequencies (n_i*).

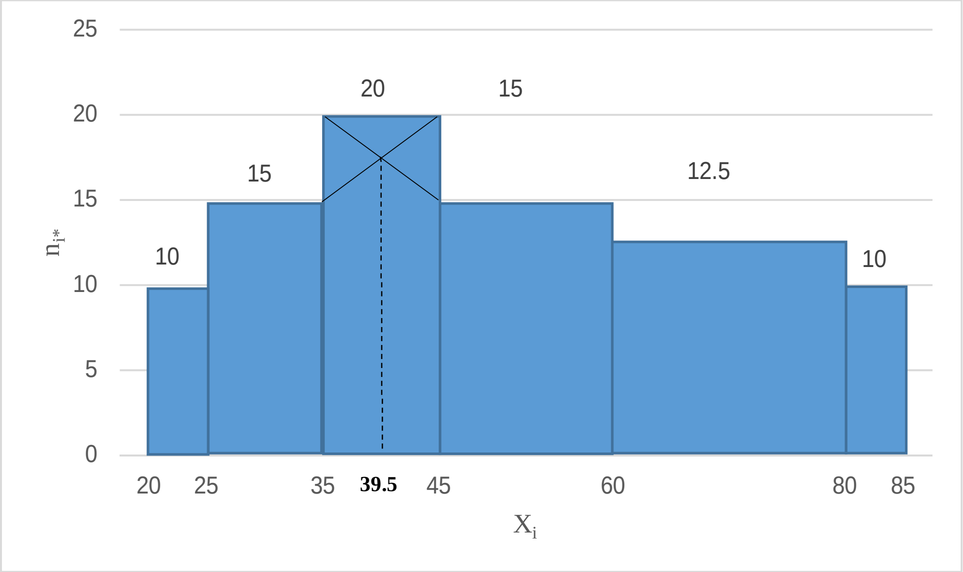
x_i	n_i	L_i	d_i	n_i*
[20 25[	10	5	2	2 * 5 = 10
[25 35[	30	10	3	3 * 5 = 15
[35 45[	40	10	4	4 * 5 = 20
[45 60[	45	15	3	3 * 5 = 15
[60 80[	50	20	2.5	2.5 * 5 = 12.5
[80 85[	10	5	2	2 * 5 = 10
Sum	210	/	/	

Solution

$$M_o = L_l + \frac{\Delta_1}{\Delta_1 + \Delta_2} * k$$

$$= 35 + \frac{20 - 15}{(20 - 15\,) + (20 - 15\,)} * 9 = 39.5$$

graphical method


25
20
15
10
5
0
n_{i^*}
10
15
20
15
12.5
10
20
25
35
39.5
45
60
80
85
X_i