Contents

Chapter 1

Univariate Statistical Series

0.1 Basic Concepts and Statistical Vocabulary

0.1.1 Introduction

The purpose of descriptive statistics is to understand collective phenomena. It aims to gather and process information, which is generally very numerous. The description is done through the presentation of data (as synthetically as possible), their graphical representation, and the calculation of numerical summaries. Then, it analyzes and interprets the data to draw conclusions. Mathematically, a statistic is a function X from a set denoted by Ω to another set C,

$$X : \Omega \to C$$

$$\omega \mapsto X(\omega)$$

Example 1. *We study the family situation of teachers in the computer science department. Therefore,*

1. Ω *is the set of computer science department teachers,*
2. $C = \{$*single, married, divorced, widowed*$\}$*,*
3. $X(\omega) =$ *the family situation of teacher* ω*.*

Example 2. *We are interested in the age of each of the 50 employees in a company. We have the following raw data:*

$$36, 30, 30, 56, 58, 47, 30, 45, 47, 18, 47, 33, 26, 51, 41, 33, 45,$$
$$39, 36, 41, 51, 21, 33, 30, 18, 56, 24, 26, 41, 26, 37, 26, 33, 39,$$
$$51, 56, 33, 24, 51, 37, 24, 37, 41, 41, 45, 33, 45, 33, 30, 37.$$

Therefore: Ω is the set of the 50 employees of the company, $C = \mathbb{R}$, and $X(\omega)$ = the age of employee ω.

0.1.2 Statistical Vocabulary

Definition 1. ***Population****: A set of numerous elements that share a common property to be studied.*

Example 3.
1. *Consider the set of all students in section A. We are interested in the number of brothers and sisters of each student. In this case, Ω set of students.*
2. *If we now turn our attention to automobile traffic in a city, the population is then made up of all vehicles likely to circulate in this city on a given date. In this case, Ω set of vehicles.*

Definition 2. ***Individual (or statistical unity)****: The element of the population on which the study is conducted. An individual can be a person, an animal, or an object.*

Sample*: A representative part of the population. It is generally impossible to gather information relative to all individuals of the population.*

Example 4.
- *In the example mentioned above, an individual is any student from the section.*
- *If we study the annual production of a factory producing metal beverage cans. The population is the set of all cans produced during the year, and a can constitutes an individual.*

Definition 3. ***Characteristic****: It is the particular and common aspect that one proposes to study in individuals.*

Example 5. *Height, temperature, nationality, eye color, socio-professional category...*

Definition 4. ***Modalities***

The modalities of a statistical variable are the different values that it can take.

Example 6.

- *The variable is "marital status".*
- *The modalities are "single, married, divorced".*

The variable is "switch status".

- *The modalities are "0 and 1".*

The variable is "socio-professional categories".

- *The modalities are "Employees, workers, retirees,...".*

0.2 Types of Characters

Definition 5. ***Qualitative Characteristic****: It is non-measurable and describes a state. Generally, it answers the question: How...*

Example 7. ***The state of a house:*** *the following modalities can be considered*

- *Old.*
- *Dilapidated.*
- *New.*
- *Renovated.*

A qualitative characteristic is classified as *ordinal* when its modalities can be arranged in a specific order. Conversely, it is termed *nominal* when such ordering is not applicable.

Definition 6. *1. **Nominal Characteristic:** This category includes information groups without an inherent order. For instance, the names of different types of fruit (e.g., apples, bananas, oranges) do not require a sequential arrangement. Another example is the colors of a rainbow, where each color (e.g., red, green, blue) is distinct without an implied hierarchy.*

*2. **Ordinal Characteristic:** This type involves groups that necessitate a particular sequence, which can be either ascending or descending. For example, the stages of educational degrees (e.g., bachelor's, master's, doctorate) suggest a progression in level of education. Another example is the rating of hotels by stars (e.g., one-star, two-star, five-star), where the number of stars indicates the quality or range of services provided.*

Definition 7. ***Quantitative Characteristic**: Measurable, when the data are numerical. Generally, it answers the question: How much... A quantitative characteristic is classified as*

- A **discrete statistical variable** is a variable that takes only isolated values. Examples include the number of children, the number of languages spoken, height, weight, salary, etc.

– **-A continuous statistical**, the data are grouped into classes that are intervals disjoint from one another, and whose union encompasses all observations. Each class is considered to be a single modality.Examples of continuous statistical variables include measurements of length (height, width, length, thickness, diameter, etc.), time, weight (or mass), and measurements derived from these (area, volume, speed, density, etc.), salary is generally considered as a continuous statistical variable.

Frequency Let the statistic $X : \Omega \to C$, $\omega_i \mapsto X(\omega_i) = x_i$. The partial frequency (or partial count) of the modality $x_i \in X(\Omega) \subseteq C$, is the cardinality of the set $X^{-1}(\{x_i\})$, denoted n_i. It is the number of individuals who have the same modality x_i.

Example 8. *Consider the example of the family situation of teachers in the computer science department.* $x_i = X(\omega_i)$ = *the family situation of the teacher* ω_i. *If there are 100 teachers in the computer science department, we obtain a statistical series of 100 values: Married, married, single, married, married, married, married, single, single, single, single, married, married, widowed, married, married, single, ... Simply listing the 100 values does not make the information practical.*

A convenient way to represent the results is to create a statistical distribution of frequencies. We consider the set of the 4 observed modalities (the 4 family situations) and for each, we provide the number n_i of individuals (teachers) who have that situation.

x_i (Modalities)	n_i
Married	n_1
Single	n_2
Divorced	n_3
Widowed	n_4
Total	100

We have: $\Sigma n_i = N$ = total frequency.

We can establish the distribution of l relative frequencies $f_i = \frac{n_i}{N}$ in which each relative frequency is expressed as a proportion (between 0 and 1) or as a percentage (between 0 and 100) of the total frequency. $\Sigma_{i=1}^{k} f_i = 1$

If the characteristic is quantitative, we define the cumulative frequency CIn_i of the modality x_i by $CIn_i = \Sigma_{j=1}^{i} n_j$.

Similarly, we define the cumulative relative frequency F_i of the modality x_i by $F_i = \Sigma_{j=1}^{i} f_j = \frac{CIn_i}{N}$.

0.3 Representation of a Statistical Series

0.3.1 Representation in a Table:

1. The title is labeled as follows: distribution of such individuals according to a certain characteristic. At the bottom of the table, the source of the information is indicated, and the date and place can be added.

2. The body of the table:

   (a) For a qualitative statistical series, it includes 3 columns: the modalities x_i in the first, the frequencies n_i in the second, and the relative frequencies in percentages $(100 \times f_i)$ in the third.

   (b) For a discrete quantitative statistical series $(x_i, n_i)_{i=1,\dots,k}$, the body of the table is similar to that of a qualitative series, with a fourth column added for cumulative frequencies and a fifth for cumulative relative frequencies.

   (c) For a continuous quantitative statistical series, the number of classes and their positioning must be defined beforehand. Some useful rules are:

      i. The classes $([a_1, a_2[, [a_2, a_3[, \dots, [a_k, a_{k+1}[)$ are mutually disjoint sets, and their union encompasses the entire dataset.

      ii. The number of classes k should be neither too small nor too large and should depend on the number of data N: $5 \leq k \leq 15$.

      iii. The average number of data per class $= N/k \geq 5$.

      iv. If possible, for practical reasons, classes of the same amplitude (length) e are chosen.

      v. In this case $e = \frac{w+1}{k} = \frac{(x_{\max} - x_{\min})+1}{k}$ and thus $k = \frac{w+1}{e}$.

      vi. In the first column, we list the classes; the other columns are the same as for a discrete series. We can add a column for the class centers.

Example 9. ***Age Distribution of Employees***

For the statistical series on the age of employees from Example 2,

36, 30, 30, 56, 58, 47, 30, 45, 45, 39, 45, 33, 30, 37,
47, 18, 47, 33, 26, 51, 41, 33, 41, 41, 45, 33, 24, 37,
39, 36, 41, 51, 21, 33, 30, 18, 51, 56, 33, 24, 51, 37,
56, 24, 26, 41, 26, 37, 26, 33.

This can be studied as a discrete statistical series (as there are only 15 modalities of the age characteristic. Age (x_i) and frequency (n_i):

x_i	18	21	24	26	30	33	36	37	39	41	45	47	51	56	58	Total
n_i	2	1	3	4	5	7	2	4	2	5	4	3	4	3	1	50

Grouping values into classes for a continuous statistical series:

X	n_i
$[18, 25[$	6
$[25, 32[$	9
$[32, 39[$	13
$[39, 46[$	11
$[46, 53[$	7
$[53, 60[$	4
Total	50

0.3.2 Graphical Representation

I-Representation of a Qualitative Series

1. **The pie chart representation**: each modality is represented by a sector (a portion) of the disk. The area (and thus the central angle α_i) of the sector is proportional to the frequency of the modality. $\alpha_i = 360^\circ \times f_i$

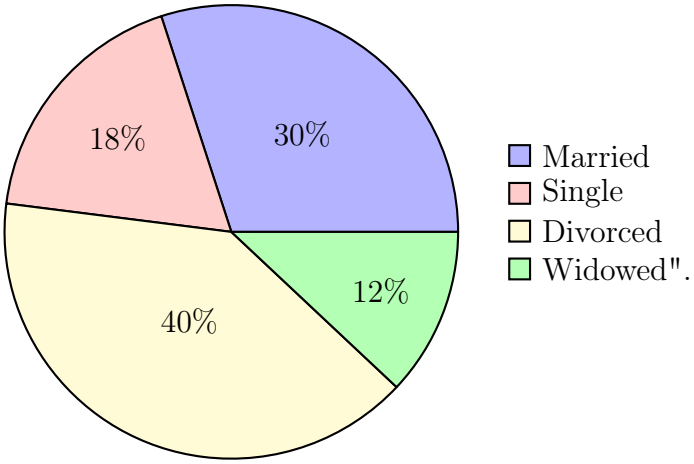


Figure 1: Distribution of Marital Statuses

2. **The representation by the bar graph**: the modalities are represented on a cartesian coordinate system by rectangles with a constant base and heights proportional to the frequencies.

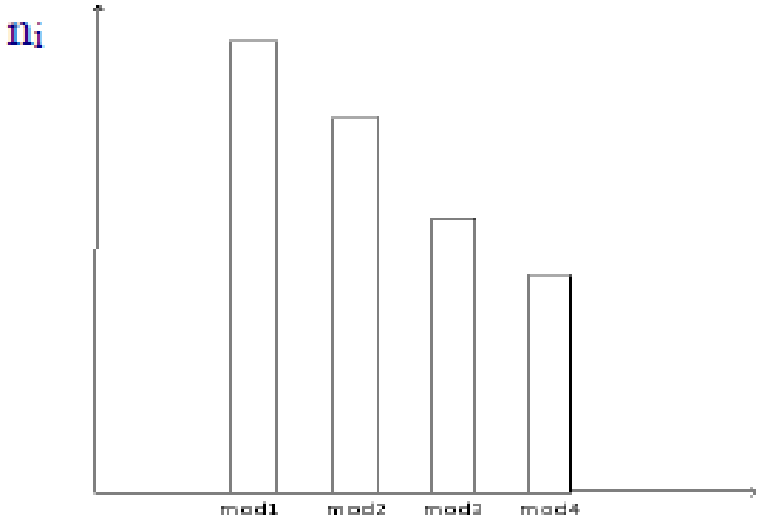


Figure 2: representation by bar chart

II-Representation of a Quantative Series

: There are two types of representations:

1. **differential diagram**: It corresponds to the representation according to frequencies (or relative frequencies).

   (a) The differential diagram of a discrete series is a bar diagram. On a cartesian coordinate system, from each point with coordinates $(x_i, 0)$, a bar is drawn with length proportional to n_i or f_i.

   (b) The differential diagram of a continuous series is called a histogram: it is the figure obtained by drawing from each base $[a_i, a_{i+1}[$ a rectangle with area (and not height) proportional to n_i or f_i.

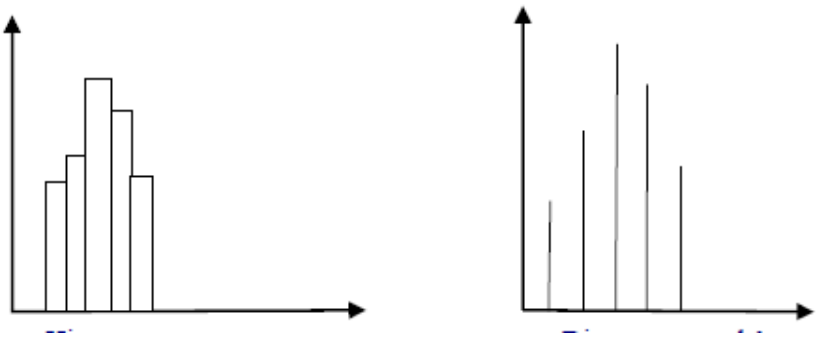


Figure 3: histogram and bar graph

2. **integral diagram (or the cumulative curve)**

   (a) For a discrete series, the graphical representation (stepwise) of the distribution function is defined by:

$$F(x) = \sum_j f_j = f_1 + f_2 + \cdots + f_i \quad \text{if } x_i \leq x < x_{i+1}$$

   or

$$CIn_i = \sum_j n_j = n_1 + n_2 + \cdots + n_i \quad \text{if } x_i \leq x < x_{i+1}$$

   (b) The cumulative frequency diagram for a continuous series: on a Cartesian coordinate system, represent each class $[a_i, a_{i+1}[$ by a point with coordinates

(a_{i+1}, CIn_i). Connect the successive points with straight line segments to obtain the cumulative frequency polygon. This polygon is then smoothed to obtain the cumulative curve (the integral diagram).

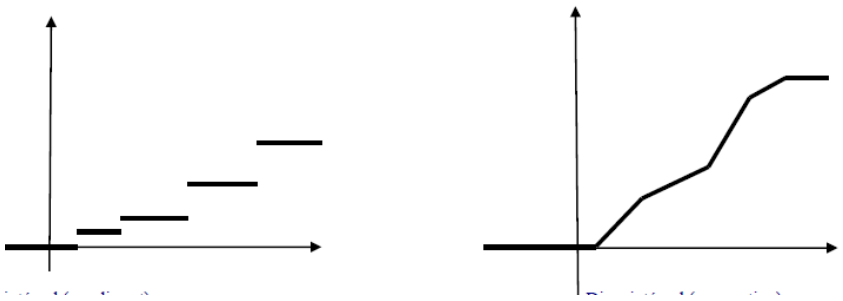


Figure 4: cumulative frequency curve

Example 10. *1. Practical Steps Involved in Drawing Histogram in Case of Equal Class Intervals*

Step 1: *Take a variable on the X-axis and frequency on the Y-axis.*

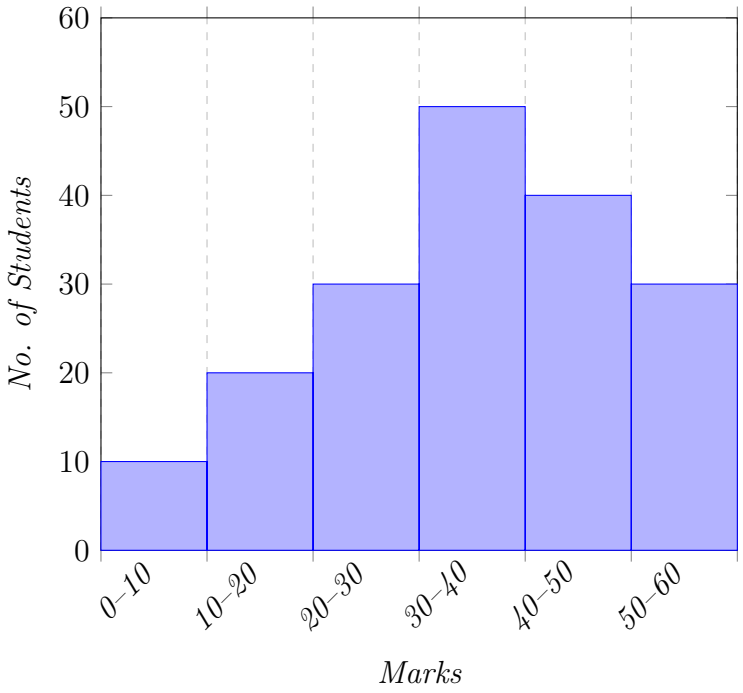
Step 2: *Draw adjacent vertical rectangles for each class interval. The series of such adjacent vertical rectangles so formed gives the histogram of frequency distribution, and its area represents the total frequency of distribution as spread throughout various classes.*

From the following data, draw Histogram.

Solution

Marks	No. of Students
0–10	10
10–20	20
20–30	30
30–40	50
40–50	40
50–60	30

Table 1: Distribution of Student Marks



2. *Practical Steps Involved in Drawing Histogram in Case of Unequal Class Intervals*

   Step 1: *Take a variable on the X-axis and frequency on the Y-axis.*

   Step 2: *Draw adjusted frequencies for unequal class intervals on the assumption that all frequencies are equally distributed throughout the class as follows:*

   Adjusted Frequency of any class =

$$\left(\frac{\textit{Width of the Lowest Class Interval}}{\textit{Width of the given Unequal Class Interval}}\right) \times \textit{Frequency of the Unequal Class Interval}$$

Step 3: *Draw adjacent vertical rectangles for each class interval. The series of such adjacent vertical rectangles so formed gives the histogram of frequency distribution, and its area represents the total frequency of distribution as spread throughout various classes.*

Reason for Adjustment of Frequencies: If histogram is constructed before adjusting frequencies of unequal class intervals, the area of rectangles would not be proportionate to the frequency and it will give undue weightage to those having larger class intervals.

For example Draw a histogram for the following distribution relating to the marks secured by the students of a class in Accountancy:

Marks	No. of Students
0–5	5
5–10	15
10–15	25
15–20	50
20–25	40
25–30	30
30–40	20
40–50	16

Table 2: Distribution of Student Marks

Solution:

Step 1: *Calculation of adjusted frequencies of unequal class intervals.*

Take 5 (the smallest width) as the normal width, the frequencies of the classes, i.e., 30–40 and 40–50, will be adjusted as follows:

Adjusted Frequency of any Class =

$$\left(\frac{\textit{Width of the Lowest Class Interval}}{\textit{Width of the given Unequal Class Interval}}\right) \times \textit{Frequency of the Unequal Class Interval}$$

For the Class 30–40, the adjusted frequency or frequency density will be:

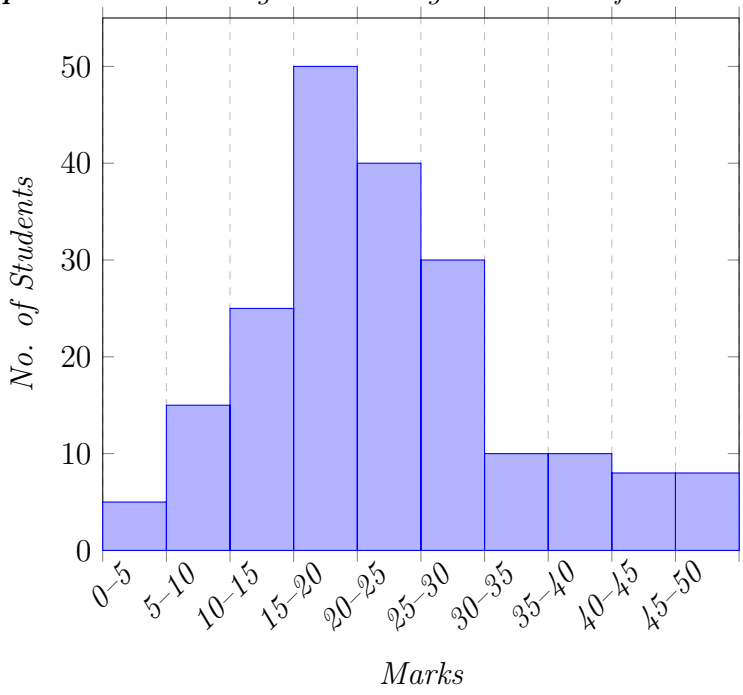
$\frac{5}{10} \times 20 = 10$

For the Class 40–50, the adjusted frequency or frequency density will be:

$\frac{5}{10} \times 16 = 8.$

Step 2: *Take marks on the X-axis and frequency on the Y-axis.*

Step 3: *Draw a Histogram showing the Marks of Students in Accountancy.*



0.4 Statistical Indicators

0.4.1 Measures of central tendency and position

1. **The Mode (Mo)**

   This is the value of the variable that has the highest frequency. If the variable is continuous, the modal class is first defined. This is the class that has the highest frequency per unit interval.

$$Mo = a_i + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) \cdot e_i$$

where:

- a_i: lower boundary of the modal class
- e_i: width of the modal class
- Δ_1: mean frequency of the modal class - mean frequency of the previous class
- Δ_2: mean frequency of the modal class - mean frequency of the following class

2. **The Median (Me)** This is the value of the variable that divides the observations of the series, which have been previously arranged in ascending or descending order, into two equal parts.

   For a discrete statistical series $x_1, x_2, \ldots, x_N$ where N is the total size:

   - If N is odd: $Me = x_{\frac{N+1}{2}}$
   - If N is even: $Me = \frac{1}{2}\left(x_{\frac{N}{2}} + x_{\frac{N}{2}+1}\right)$

   For a continuous statistical series, the median class is determined. The ith class $[a_i, a_{i+1}[$ is the median class if $F_{i-1} \leq \frac{1}{2} < F_i$ (or $n(i-1)c \leq \frac{N}{2} < n_i c$)

$$Me = a_i + \left(\frac{\frac{N}{2} - CIn(i-1)}{n_i}\right) e_i \quad \text{or} \quad Me = a_i + \left(\frac{\frac{1}{2} - F_{i-1}}{f_i}\right) e_i$$

Example 11. *A cereal manufacturer conducts a survey to verify if the cereal boxes indeed contain 500 grams as indicated on the packaging. A sample of 1000 boxes produced in one day is checked. The following data is obtained:*

Weight Range (g)	(n_i)	(CIn_i)
$[494, 496[$	*33*	*33*
$[496, 498[$	*168*	*201*
$[498, 500[$	*415*	*616*
$[500, 502[$	*293*	*909*
$[502, 504[$	*75*	*984*
$[504, 508[$	*16*	*1000*

Calculate the mode and the median for the cereal boxes.

(a) ***Mode*** *The hiest frequency is 415, so we have: Modal class* $[498, 500[$ *(class with the highest frequency,* h_i*)*
$Mode(Mo) = 498 + \left(\frac{415 - 168}{(415 - 168) + (415 - 293)}\right) \times 2 = 499.34g$

(b) ***Median*** *The third class* $[498, 500[$ *(i=3) is the median class (it is the first class with a cumulative frequency greater than* $N/2$*)*
$Median(Me) = 498 + \left(\frac{415/2 - 201}{293}\right) \times 2 = 499.44g$

3. **The Arithmetic Mean ($\bar{m}$)** The arithmetic mean is the value that all the data would take for their total sum to remain unchanged.

For a discrete variable:

$$\bar{m} = \frac{1}{N} \sum_{i=1}^{k} n_i x_i$$

For a continuous variable, the x_i are the midpoints of the classes c_i.

Variable Type	Mean Calculation Formula
Discrete Variable	$\bar{m} = \frac{1}{n} \sum_{i=1}^{n} x_i$
Continuous Variable	$\bar{m} = \frac{1}{n} \sum_{i=1}^{n} c_i$

Table 3: Calculation of Mean for Discrete and Continuous Variables

Example 12. *Calculate the mean $\bar{m}$ weight of a box of cereal:*

Interval	n_i	c_i	$n_i \times c_i$
$[490, 496[$	33	493	33×493
$[496, 498[$	168	497	168×497
$[498, 500[$	415	499	415×499
$[500, 502[$	293	501	293×501
$[502, 504[$	75	503	75×503
$[504, 510[$	16	507	16×507

The total n_i is 1000 and the total $n_i \times c_i$ is 499480.

The arithmetic mean weight $\bar{m}$ is given by:

$$\bar{m} = \frac{1}{N} \sum n_i \cdot c_i = \frac{499480}{1000} = 499.48\, g$$

4. **Quartils** A quartile is one of the three values that divide the sorted data into four equal parts, such that each part represents one-fourth (1/4) of the population sample.

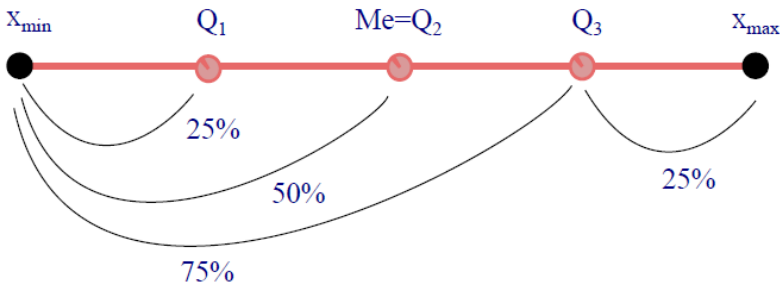


Figure 5: The Quartiles

For the discrete data $Q_1 =$ size of $\left(\frac{n}{4}\right)^{\text{th}}$ term

$Q_3 =$ size of $\left(\frac{3n}{4}\right)^{\text{th}}$ term

Q_2 is the median For the continuous data

$$Q_1 = a_i + \left(\frac{\frac{N}{4} - CIn(i-1)}{n_i} \right) e_i \quad \text{or} \quad Me = a_i + \left(\frac{\frac{1}{2} - F_{i-1}}{f_i} \right) e_i$$

Example 13. *(a)* ***discrete variable*** *calculate the quartils of this data*

x_i	18	21	24	26	30	33	36	37	39	41	45	47	51	56	58	Total
n_i	2	1	3	4	5	7	2	4	2	5	4	3	4	3	1	50
CIn_i	2	3	6	10	15	22	24	28	30	35	39	42	46	49	50	

Table 4: Data distribution with cumulative frequencies

Given a discrete data set with cumulative frequencies, we calculate the first Q_1 and third Q_3 quartiles as follows:

i. *For the first quartile (Q1), locate the value at the position corresponding to $\frac{50}{4} = 12.5$. Q_1 corresponds to the value just above the 12.5th position in the cumulative frequency, which is 30.*

ii. *For the third quartile (Q3), locate the value at the position corresponding to $\frac{3\times 50}{4} = 37.5$. Q_3 corresponds to the value just above the 37.5th position in the cumulative frequency, which is 45.*

Thus, for our dataset, the calculated quartiles are:

- $Q1 = 30$
- $Q3 = 45$

(b) ***continuous variable*** *we take the example of a box of cereal*

Weight Range (g)	Number of Boxes (n_i)	Cumulative Frequency (CIn_i)
$[494, 496[$	33	33
$[496, 498[$	168	201
$[498, 500[$	415	616
$[500, 502[$	293	909
$[502, 504[$	75	984
$[504, 508[$	16	1000

Table 5: Distribution of Box Weights with Cumulative Frequencies

$$Q_1 = a_i + \left(\frac{\frac{N}{4} - CIn_{i-1}}{n_i} \right) e_i \quad and \quad Q_3 = a_i + \left(\frac{\frac{3N}{4} - CIn_{i-1}}{n_i} \right) e_i$$

So

$$Q_1 = 498 + \left(\frac{250 - 201}{415} \right) 2$$

$$Q_1 = 498.24$$

and

$$Q_3 = 500 + \left(\frac{750 - 616}{75} \right) 2$$

$$Q_3 = 503.57$$

0.4.2 Measures of Dispersion

They measure the fluctuations (variation) of values around the central value and assess the spread of the series.

1. **the interquartile range**;The difference between the third quartile and the first quartile it is a measure of dispersion of the series

2. **Range (W)**: $W = x_{\max} - x_{\min}$

3. Variance V: It is the arithmetic mean of the squares of the deviations from the mean.

Variable	General Variance Form	Simplified Form (Koenig's theorem)
Discrete Variable	$V = \frac{1}{N}\sum_{i=1}^{n} n_i(x_i - \bar{m})^2$	$V = \frac{1}{N}\sum_{i=1}^{n} n_i x_i^2 - \bar{m}^2$
Continuous Variable	$V = \frac{1}{N}\sum_{i=1}^{n} n_i(c_i - \bar{m})^2$	$V = \frac{1}{n}\sum_{i=1}^{n} n_i c_i^2 - \bar{m}^2$

Table 6: Calculation of Variance for Discrete and Continuous Variables

4. **Standard Deviation**:

$$\sigma_X = \sqrt{V}$$

5. **Absolute Deviation**:

$$AD(X) = \frac{1}{N}\sum_{i=1}^{n} |x_i - \bar{m}|$$

6. **Coefficient of Variation**: The coefficient of variation is a relative measure of the dispersion of data around the mean. It is used to compare the degree of variation from one sample to another, even if the means are different.

$$CV = \frac{\sigma_X}{\bar{m}} \times 100\%$$

. If $CV > 0.15$ (or 15%), then the series is considered dispersed.