Mohamed Boudiaf University - M'sila
Faculty of Technology
Department of Civil Engineering-Department of Electrical Engineering
Module: Probability-Statistics

Chapter 1: Basic definitions-One variable statistical series

Merini Abdelaziz*

October 20, 2024

Contents

1 Descriptive statistic **2**
- 1.1 Introduction 2
- 1.2 Statistical Vocabulary 2
  - 1.2.1 Population-Sample-Variables-Measurement 2
  - 1.2.2 Different types of statistical variables 3

2 Statistical Series with One Variable **4**
- 2.1 Raw Data 4
- 2.2 Organize and graph Qualitative data 4
  - 2.2.1 Organize Qualitative data 5
  - 2.2.2 Graphical Presentation of Qualitative Data 5
- 2.3 Organizing and Graphical Presentation Quantitative Data 7
  - 2.3.1 Organize Quantitative data 7
  - 2.3.2 Graphical Presentation of Quantitative Data 9

3 Measures of Central Tendency **10**
- 3.1 Arithmetic mean 11
  - 3.1.1 Properties of the mean 11
- 3.2 The Mode 12

*

3.2.1 The mode for a continuous quantitative variable 12
3.3 The median 13
3.3.1 The Coefficient of Variation 14

4 Measures of Variation 14
4.1 Range 14
4.2 The Variance and Standard Deviation 15
4.3 Coefficient of Variation 15

1 Descriptive statistic

1.1 Introduction

Descriptive statistics are procedures used to summarize, organize, and make sense of a set of scores or observations.

1.2 Statistical Vocabulary

We will begin by defining the terms used in statistics to designate numerical observations.

1.2.1 Population-Sample-Variables-Measurement

Definition 1. *A **A population** is any specific collection of objects of interest. The components of the population are called **individuals** or **statistical unity**.*

Remark 1. *Note that a population can be a collection of any things, like Ipad set, Books, animals or inanimate, therefore it does not necessary deal with people.*

Definition 2. *A **Sample** is the subset of the population. When the sample consists of the entire population,it is called a **census**.*

Example 1. *Population: Students at the University of M'sila.*
Statistical units: Students.
Sample: Students in the second year of technology.

Definition 3. *A variable is a characteristic under study that takes different values for different elements.*

Example 2. *Civil Status. Place of residence Age. Socio-professional categories. Eye colour. Number of languages spoken. Height. Temperature. Nationality...*

Definition 4. *The value of a variable for an element is called an observation or **measurement**.*

Example 3. *1- The variable is "marital status".*
The measurement are "single, married, divorced, Widowed".

2-The variable is "socio-professional categories".
The measurement are "Employees, workers, retirees,...".

1.2.2 Different types of statistical variables

In statistics, we have two types of variables according to their elements; first type is called **quantitative variable** and the second one is called **qualitative variable.**

Definition 5. *Qualitative variable (or categorical data) gives us names or labels that are not numbers representing the observations. It can be of one of the following types:*

a) Nominal: the measurement are not ordered, for example: eye colour.

b) Ordinal: the measurements are ordered, for example: bBAC mention (passable, fairly good, good).

Quantitative variable:

Definition 6. *Quantitative variable gives us numbers representing counts or measurements. It can be of one of the following types:*

a)Discrete variables assume values that can be counted.

, for example: number of children per family.

b) Continuous variables assume all values between any two specific values, i.e. they take all values in an interval.

For example: height, weight.

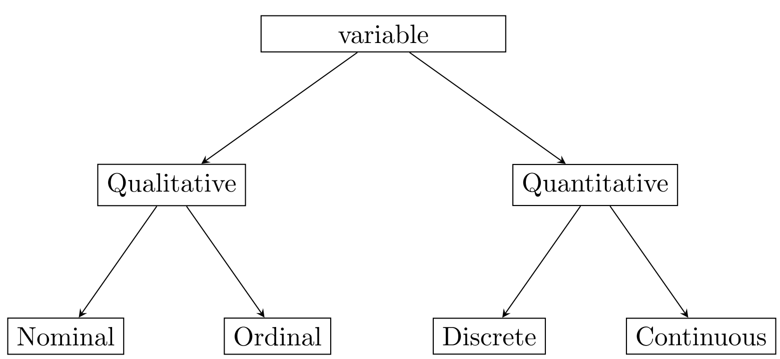


Figure 1: Types of variables

2 Statistical Series with One Variable

In this sectionn we will learn how to organize and display the Qualitative and Quantitative data.
A statistical series is the sequence of observations of one (or more) variable(s), taken from the individuals in a population.
The values of the variable X are denoted $x_1, ..., x_i..., x_N$.

2.1 Raw Data

Definition 7. *Data recorded in the sequence in which they are collected and before they are processed or ranked are called **raw data**. Consider the following three examples to discuss the concept of raw data.*

Example 4. *Suppose that a sample of 50 second year technology students at the University of M'sila were selected and these students were asked about their degree of satisfaction with of the exam results. The answers of these students are recorded below where*
(v) means very highly satisfied,
(s) means somewhat satisfied, and
(n) means not satisfied.

n	*n*	*n*	*v*	*s*	*n*	*n*	*n*	*v*	*v*	*n*	*s*	*v*	*v*	*v*	*n*	*n*	*s*	*n*	*s*	*v*	*n*	*s*	*v*	*v*
v	*s*	*v*	*n*	*n*	*v*	*s*	*n*	*v*	*v*	*v*	*v*	*s*	*s*	*v*	*v*	*n*	*s*	*s*	*v*	*v*	*v*	*n*	*s*	*n*

Example 5. *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.*

Example 6. *In a city, 45 families were surveyed for the number of cell phones they used. Prepare a discrete frequency array based on their replies as recorded below. 1; 3; 2 ;2; 2; 2; 1; 2; 1; 2; 2 ;3; 3; 3; 3; 3; 3; 2 ;3 ;2 ;2 ;6; 1; 6; 2; 1; ;5 1; 5; 3; 2; 4; 2; 7; 4; 2; 4; 3; 4; 2; 0; 3; 1; 4; 3.*

2.2 Organize and graph Qualitative data

In this section we will learn how to organize and display the Qualitative and Quantitative data.

Definition 8. ***The total frequency*** *noted N, is the number of individuals that make up the population.* $card(\Omega) = N$.

Definition 9. *A **frequency** of a modalitiy is the number of times a data value occurs. noted* n_i

Definition 10. ***A relative frequency*** *of a modalitiy noted* $f_i = \dfrac{n_i}{N}$

Definition 11. *The percentage of a modalitiy is* $p_i = fi \times 100$

Properties 1) $\Sigma n_i = N$
2) $\Sigma f_i = 1$
3) $0 f_i \leqslant 1$

2.2.1 Organize Qualitative data

In this section we will study some methods that used to organize qualitative data set.

a) Frequency table

Definition 12. *A frequency table for qualitative data lists all categories , names or labels and the number of elements that belong to each of the categories, names or labels.*

Example 7. *Refer to example 4 we can view the frequency table as follow:*

Variable	*Frequency* n_i
v	20
s	12
n	18
$Sum = \Sigma$	50

b)Relative Frequency and Percentage Distributions

Example 8. *Let's take the previous example. Applying the definition of relative frequency and the percentage of each category we get the following table:*

Variable	*Frequency* n_i	*Relative Frequency* f_i	*Percentage* $p_i\%$
v	20	*0.40*	*40*
s	12	*0.24*	*24*
n	18	*0.36*	*36*
$Sum = \Sigma$	50	*1*	*100*

2.2.2 Graphical Presentation of Qualitative Data

There are many types of graphs that are used to display qualitative data; in this part we will study and graph two of such graphs which they are commonly used to display the qualitative data, these graphs are the **Bar chart** and the **Pie chart**.

1)Bar Graph (Chart)

Definition 13. *A graph made of bars whose heights represent of respective categories is called a* ***bar graph****.*

Construct a Bar Graph (Chart)

1. Represent the categories on the horizontal axis (All categories are represented by intervals of the same width).
2. Mark the frequencies on the vertical axis.
3. Draw one bar for each category .

Example 9. *Refer to the example 7, we construct bar graph*

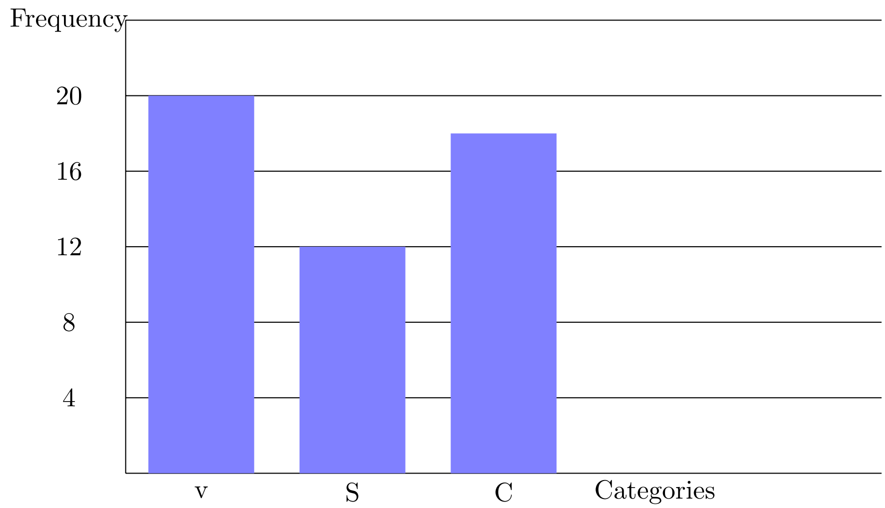


Figure 2: Bar graph of satisfaction with exam results

2)Pie Chart

Definition 14. *A circle divided into portions that represents the relative frequencies or percentages of a population or a sample of different categories is called a pie chart.*

Construct a Pie Graph (Chart)

1. Draw a circle. 2. Find the central angle for each category by the following equation: $\alpha_i = f_i 360^0$.
3. Draw sectors corresponding to the angles that obtained in step 2.

Example 10.

Let's take example 7,

Variable	Frequency n_i	Relative Frequency f_i	angle α_i
v	20	0.40	144
s	12	0.24	86.4
n	18	0.36	129.6
$Sum = \Sigma$	50	1	360

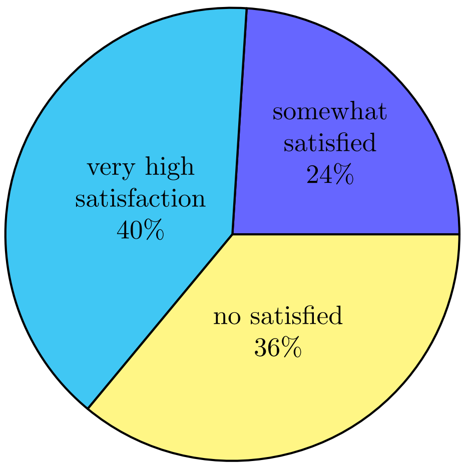


Figure 3: Pie chart of satisfaction with exam results

2.3 Organizing and Graphical Presentation Quantitative Data

In this section we will study some methods that used to organize quantitative data set.

Definition 15. *The ascending cumulative frequency (* $N_{i\nearrow}$ *) of a value* x_i *can be found by adding all the frequencies less than* x_i*.*

Definition 16. *The descending cumulative frequency (* $N_{i\searrow}$ *) of a value* x_i *indicates the frequency of values that are greater than or equal to* x_i*.*

2.3.1 Organize Quantitative data

a) Frequency Table:

Definition 17. *A frequency table for quantitative data is an effective way to summarize or organize a dataset. It's usually composed of two columns:*
The values or class intervals.
Their frequencies

For a discrete quantitative variable, we take the following example

Example 11. *Refer to example 6, we can view the frequency table as follow::*

modalities x_i	*Frequency* n_i
0	1
1	7
2	15
3	12
4	5
5	2
6	2
7	1
$Sum = \Sigma$	45

Example 12. *Refer to the previous example,*

modalities x_i	*Frequency* n_i	$N_i \nearrow$	$N_i \searrow$
0	1	*1*	*45*
1	7	*8*	*44*
2	15	*23*	*37*
3	12	*35*	*22*
4	5	*40*	*10*
5	2	*42*	*5*
6	2	*44*	*3*
7	1	*45*	*1*
$Sum = \Sigma$	45	/	/

For a continuous quantitative variable ; we need to apply the following steps.

Construct a Frequency Distribution Table:

If a number of classes k are not given, we follow the steps below:
1-Calculate the range:$w = x_{max} - x_{min}$
2-The number of classes, k, is calculated using one of two formulas :
Sturge's rule $k = 1+3.322LogN \in N$, Yule's rule $k = 2.5(N)^{\frac{1}{4}}$, with $5 \leq k \leq 15$.
3-Determine the Width of classse $L = \dfrac{w}{N_c}$.
4-Find the class limits: $[x_{min}, x_{min} + L[$,.......,$[x_{max} - L, x_{max}]$
-Calculate the Midpoints of each classse,$m_i = \dfrac{Lowerlimit + Upperlimit}{2}$
6-Count the number of data entries for each class, and record the number in the row of the table for that class.

Class	Midpoint	Frequency n_i
$[x_{min}, x_{min} + L[$	m_1	n_1
$\vdots$	$\vdots$	$\vdots$
$[x_{max} - L, x_{max}]$	m_k	n_k
Sum	/	N

Example 13. *Refer to example 5, we can view the frequency table as follow::*
Step1 *Calculate the range:* $w = x_{max} - x_{min} = 58 - 18 = 40$
Step2 *The number of classes* $k = 1 + 3.322Log50 = 6.6 \simeq 7$

Step3 *the Width of classse* $L = \frac{w}{N_c} = \frac{40}{7} \simeq 5.7 \simeq 6.$

Class	*Frequency* n_i	*Midpoint*
[18; 24[	3	21
[24; 30[	*7*	*27*
[30; 36[	12	33
[36; 42[	*13*	39
[42; 48[	7	45
[48; 54[	4	51
[54; 60]	4	57
Sum	*50*	/

2.3.2 Graphical Presentation of Quantitative Data

1)Case of a Discrete quantitative data :

1ine gragh This graph has two axes, a horizontal axis representing the values of the variable, and a vertical axis representing the frequencies or relative frequencies. Each value is associated with a segment whose height is proportional to the frequencies or the frequency of this modality.

Example 14.

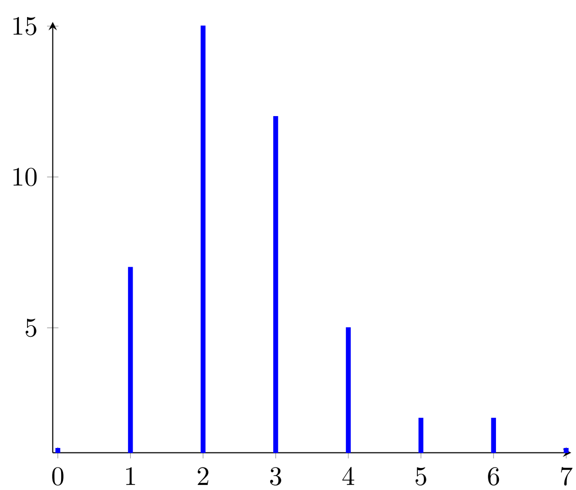


Figure 4: 1ine gragh of the number of cell phones

2)Case of a Continuous quantitative data :

Histogram A histogram of grouped data in a frequency distribution table with equal class widths is a graph in which class boundaries are marked on the horizontal axis and the frequencies, relative frequencies, or percentages are marked on a vertical axis.

Remark 2.

Histograms in case of unequal classes widths: we calculate the adjusted frequency h_i using the formula,

$$h_i = \frac{a}{a_i} \times n_i$$

were, $a = PGCD(a_1, ...a_k)$

Example 15. *Refer to example 5,*

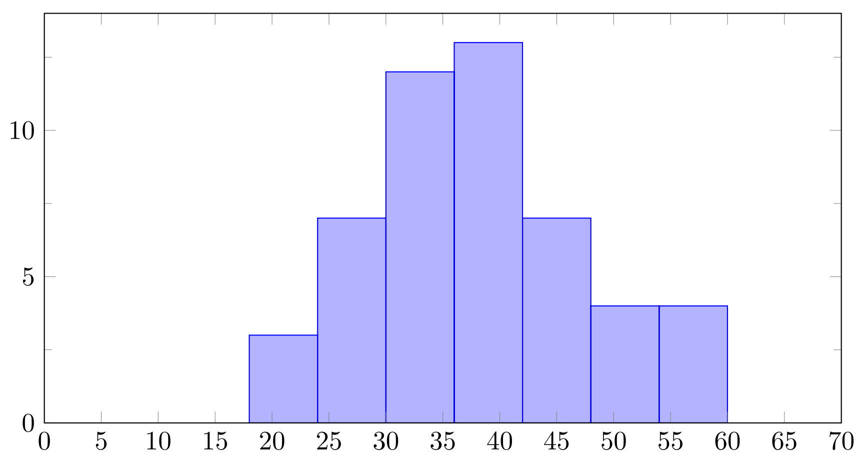


Figure 5: Histogram of the age of the employees

3 Measures of Central Tendency

A measure of central tendency is very important tool that refer to the centre of a histogram or a frequency distribution curve. In This section we will discuss three measures of central tendency and learn how to calculate it. Such measures are the mean, the median, and the mode for the two cases (continuos and discete quantitative variables).

3.1 Arithmetic mean

Let the following statistical series $x_1, ..., x_i..., x_N$.

Definition 18. *The arithmetic mean (or the average) is the sum of all the statistical data divided by the number of data items :*

$$\bar{x} = \frac{x_1 + x_2 + ... + x_N}{N} \tag{1}$$

Example 16. *The numbers of children in 8 families are 0, 0, 1, 1, 1, 2, 3, 4. The mean is*
$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{0+0+1+1+1+2+3+4}{8} = \frac{12}{8} = 1.5$$

Definition 19.

(Weighted arithmetic mean) If a value x_i is observed n_i times, formula (1) becomes:
$$\bar{x} = \Sigma_{i=1}^{k} \frac{n_i x_i}{N} = \Sigma_{i=1}^{k} f_i x_i \tag{2}$$

Example 17. *Consider the table:*

x_i	$y\ n_i$
0	2
1	3
2	1
3	1
4	1
$Sum = \Sigma$	8

The mean is
$$\bar{x} = \Sigma_{i=1}^{k} \frac{n_i x_i}{N} = \frac{3 \times 0 + 2 \times 1 + 1 \times 2 + 1 \times 3 + 1 \times 4}{8} = \frac{12}{8} = 1.5$$

Remark 3. *. In the continuous case, we choose the value xi equal to the centre of the of the corresponding class ci , i.e.* $\bar{c} = \Sigma_{i=1}^{k} \frac{n_i c_i}{N} = \Sigma_{i=1}^{k} f_i c_i$

3.1.1 Properties of the mean

i) The sum of the deviations from the mean is 0 :
$\Sigma_{i=1}^{k} n_i(x_i - \bar{x}) = 0$

ii) Linearity property: If
$\forall i \in \{1, 2, ..., k\}, a, b \in \mathbb{R}$ we have: $y_i = ax_i + b$, then :$\bar{y} = a\bar{x} + b$.

iii) Property of partial means: Consider two statistical series with respective means $\bar{x_1}$ and $\bar{x_2}$ and respective numbers N_1 and N_2.
The mean of the two series is
$\bar{x} = \frac{N_1\bar{x_1} + N_2\bar{x_2}}{N_1 + N_2}$

Proof. See WS. □

3.2 The Mode

Definition 20. *The mode (Mo) is the value that occurs most often in a data set.*

Remark 4. *Refer to example 5,*

Example 18.

Variable	*Frequency* n_i
v	20
s	12
n	18
$Sum = \Sigma$	50

The mode $Mo = v$.

Example 19. *Refer to example 9,*

modalities x_i	*Frequency* n_i
13	2
14	4
15	5
16	8
17	6
$Sum = \Sigma$	25

The mode $Mo = 16$.

1) The mode can be calculated for all types of quantitative and qualitative variable.
2) The mode is not necessarily unique.
3) For a series divided into classes, we speak of a modal class.

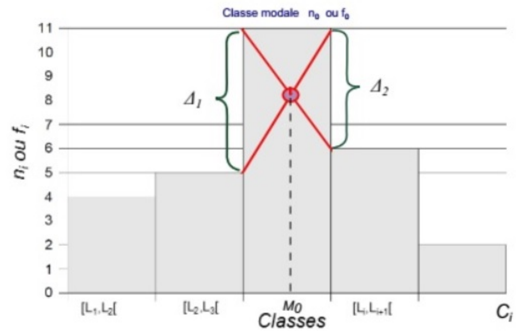
3.2.1 The mode for a continuous quantitative variable

can be calculated by the following formula :

$$Mo = L_{i-1} + a_i \frac{\Delta_1}{\Delta_1 + \Delta_2} \quad (3)$$

with
L_{i-1} : The lower bound of the modal class.
a_i :is the class width of the mode class.
Δ_1 is the deference between the frequency of the mode class and the frequency of the previous class.
Δ_2 is the deference between the frequency of the mode class and the frequency of the next class.

Example 20. *In example 13 :*



Class	*Frequency n_i*	*Midpoint*
[18; 24[	3	21
[24; 30[	7	27
[30; 36[	12	33
[36; 42[	13	39
[42; 48[	7	45
[48; 54[	4	51
[54; 60]	4	57
Sum	50	/

The modal class $= [36; 42[$. *Therefore,* $Mo = 36 + 6\dfrac{13-12}{(13-12)+(13-7)} \simeq 36.8$

3.3 The median

Definition 21. *The median (Me) is the value of the middle term in a data set that has been ranked in increasing or decreasing order.*

The Median for Ungrouped Data Note that, to find the median of a given data we need the following three steps
1. Rank the given data sets (in increasing or decreasing order)
2. Find the middle term for the ranked data set that obtained in step 1.

3. The value of this term represents the median.
The median of the ranked data $x_1, ..., x_i..., x_N$. is given by

$$Mo = \begin{cases} x_{\frac{n+1}{2}}, & \text{if } n \text{ is odd} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, & \text{if } n \text{ is even} \end{cases}$$

Example 21. *Find the median for the data set:*
312, 257, 421, 289, 526, 374, 497.
Solution: First, the data set after we have ranked in increasing order is:
257 ,289 ,312, 374, 421, 497, 526.
$N = 7$ *is odd, then* $Mo = x_4 = 374$

Example 22. *Find the median for the data set:*
8, 12, 7, 17, 14, 45, 10, 13, 17, 13, 9, 11
Solution: First, we rank the data in increasing order:
7, 8 ,9 ,10, 11, 12, 13, 13, 14, 17, 17, 45.
$N = 12$ *is even, then* $Mo = \frac{x_6 + x_7}{2} = \frac{12 + 13}{2} = 12.5$

The Median for grouped Data Suppose that we have a frequency distribution table with k classes, then one calculate the median of this grouped data by the following relation:

$$Me = L_i + \frac{\left(\frac{N}{2} - N_i\right)}{N_{i+1} - N_i} \left(L_{i+1} - L_i\right)$$

where, l_i is the lower bound of the median class.
l_{i+1} is the upper bound of the median class.
N_{i+1} is the ascending cumulative frequency of the median class.
N_i is the ascending cumulative frequency of the previous class.

3.3.1 The Coefficient of Variation

4 Measures of Variation

The usual dispersion characteristics are range, variance and standard deviation.

4.1 Range

Definition 22. *The range for ungrouped data is defined by: Range = Largest value – Smallest value*

Example 23. *Find the range for the data set: 40, 10, 20, 30, 35, 40, 50, 60.*
Solution:
The largest value is 60, and the smallest value is 10. Therefore
Range = Largest value – Smallest value = 60 – 10 = 50

4.2 The Variance and Standard Deviation

Definition 23. *The variance is the quantity:*

$$V(x) = \Sigma_{i=1}^{k} \frac{n_i(x_i - \bar{x})^2}{N} = \Sigma_{i=1}^{k} f_i(x_i - \bar{x})^2 \quad (4)$$

Definition 24. *Standard deviation The standard deviation is the quantity*

$$\sigma_x = \sqrt{V(x)} \quad (5)$$

Properties

1) $V(x) = \Sigma_{i=1}^{k} \frac{n_i x_i^2}{N} - \bar{x}^2 = \Sigma_{i=1}^{k} f_i x_i^2 - \bar{x}^2$

2) If $\forall i \in \{1, 2, ..., k\}, a, b \in \mathbb{R}$ we have: $y_i = ax_i + b$, $a, b \in R$, then :
$V(y) = a^2 V x$,
$\sigma_y = |a|\sigma_x$.

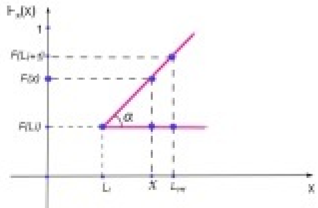
3)

4.3 Coefficient of Variation

Definition 25. *The coefficient of variation (C_v) is a relative measure of the dispersion of the data around the mean. The coefficient of variation is given by*

$$C_v = \frac{\sigma}{\bar{x}} \quad (6)$$

Properties 1) The greater the coefficient of variation, the greater the dispersion.
2) Generally expressed as a percentage. Without unit.
3) Used to compare 2 statistical series.


1
F(x)
α
X