Chapter 4

Descriptive Statistics

4.1 Statistical Vocabulary

1. **Population:** is the set of individuals or objects of the same nature on which the study relates.
   The components of the population are called individuals, elements, or statistical units.

2. **Sample:** A subset of the population selected for study

3. **Statistical Variable:**The character X is the subject under statistical study.

4. **Statistical Modality:**The category or the different possible situations (levels) of a statistical variable.

Example 4.1.1 *Scores of six students in a statistics test can be* $4, 6, 8, 3, 2$ *and* 9 *marks.*

- *The variable is:* ***marks****.*
- *The modalities (values) are:* $4, 6, 8, 3, 2, 9$*.*

4.1.1 Types of Statistical Variables

There are two types of statistical variables

4.1.1.1 Quantitative Variables

Quantitative Variables are the variables that can be measured or give us numbers representing counts, or variables whose modalities are numbers.
Moreover the quantitative variables are divided into two main types distrete and continous.

1. **Discrete :** A variable is discrete if it takes isolated values (usually integers).

   Example 4.1.2
   - *The number of children in a family* $(1, 2, 3)$.
   - *The number of students in a classroom.*
   - *The number of accidents in a city.*

2. **Continuous:** A variable is continuous when it can take numbers from an interval of real numbers.

Example 4.1.3
- ***Temperature:*** *For example, the temperature in Mila city last summer was between* 15 *and* 56, *i.e.,* $[15, 56]$.

4.1.1.2 Qualitative Variables

Qualitative Variables are variables that are not measurable numerically.
The qualitative variables can be

1. **Nominal:** Variables whose modalities cannot be ordered according to their meaning

   Example 4.1.4
   - *Gender: male, female.*
   - *Eye color: Black, Brown.*
   - *Religion: Muslim, Christian.*

2. **Ordinal:** Variables whose modalities are ordered according to their meaning (hierarchy).

   Example 4.1.5
   - ***Grade:*** $\{A, B, C, D, E\}$.
   - ***Rating scale:*** *(bad, good, excellent).*
   - ***Ranking of football players:*** *(first grade, second grade, third grade).*

4.2 Organizing data

- **Statistical Series:**A statistical series is the simplest form of presenting statistical data.

- **Absolute Frequency n_i:** The absolute frequency n_i is the number of statistical elements corresponding to a given modality.
- **Cumulative Absolute Frequency $n_i^{c\uparrow}$:** The cumulative absolute frequency $n_i^{c\uparrow}$ is the number of individuals corresponding to the same modality and all previous modalities.
- **Relative Frequency f_i:** The relative frequency is defined by

$$f_i = \frac{n_i}{N},$$

  where N is the total number of observations.
- **Cumulative Relative Frequency $f_i^{c\uparrow}$:** The cumulative relative frequency is defined by

$$f_i^{c\uparrow} = \frac{n_i^{c\uparrow}}{N}.$$

Example 4.2.1 *The table below shows the number of children per family, with absolute frequency (n_i), cumulative absolute frequency (N_i^-), relative frequency (f_i), and cumulative relative frequency (F_i^-):*

Number of Children x_i	n_i	$n_i^{c\uparrow}$	f_i	$f_i^{c\uparrow}$
0	16	16	0.250	0.250
1	18	34	0.281	0.531
2	14	48	0.218	0.749
3	11	59	0.172	0.921
4	3	62	0.047	0.968
5	2	6	0.031	1.000
Total	64	–	1.000	–

Continuous Statistical Variables

- **Class Limits:** For each class, we define:
  - a lower limit,
  - an upper limit.
- **Class Amplitude:** The amplitude of a class is

$$a = \text{upper limit} - \text{lower limit}.$$

- **Class Center** The center of a class is

$$c_i = \frac{\text{lower limit} + \text{upper limit}}{2}.$$

Example 4.2.2 *Suppose we record the daily study hours of 30 students. We group the*

data into classes and calculate class limits, class amplitude, and class centers.

Class	*Lower Limit*	*Upper Limit*	*Amplitude a*	*Class Center* c_i
[0, 2[	0	2	2	1
[2, 4[	2	4	2	3
[4, 6[	4	6	2	5
[6, 8[	6	8	2	7
[8, 10[	8	10	2	9

4.3 Position Parameters

4.3.1 Mean

- Cas of discrete statistical variable:

  Let X be a discrete statistical variable and $x_1, x_2, \ldots, x_k$ its valeurs with frequencies $n_1, n_2, \ldots, n_k$, and $N = \sum_{i=1}^{k} n_i$ The mean is

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

Example 4.3.1

x_i	n_i
0	1
1	4
2	5
3	2

Calculation of the mean:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{k} x_i n_i = \frac{0\cdot 1 + 1\cdot 4 + 2\cdot 5 + 3\cdot 2}{12} = \frac{0+4+10+6}{12} = \frac{20}{12} \approx 1.67$$

- Cas of Continuous statistical variable:

  For a continuous variable grouped into classes, the mean is

  $$\bar{x} = \frac{1}{N}\sum_{i=1}^{k} n_i c_i,$$

  where c_i is the class center.

Example 4.3.2

Class	n_i	c_i
[0, 1[	0	0.5
[1, 3[	1	2
[3, 5[	3	4
[5, 7[	5	6
[7, 9[	7	8

Calculation of the mean:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{5} n_i c_i = \frac{2\cdot 0.5 + 5\cdot 2 + 6\cdot 4 + 4\cdot 6 + 3\cdot 8}{20} = \frac{1+10+24+24+24}{20} = \frac{83}{20} = 4.15$$

4.3.2 Mode

- Cas of discrete statistical variable:

  The mode M_0 is the value that occurs most frequently.

Example 4.3.3

x_i	n_i
0	2
1	3
2	4
3	1

The highest frequency is $n_3 = 4$ *corresponding to* $x_3 = 2$*, so the mode is:*

$$M_0 = 2$$

- Cas of continuous variable:

  In this case the mode is calculated by the formula:

$$Mo = L_i + \left(\frac{d_1}{d_1 + d_2}\right) a$$

  - L_i: the lower limit of the **modal class** (the class that has the highest frequency)
  - d_1: the absolute frequency of the modal class - the absolute frequency of the previous class $(n_i - n_{i-1})$.
  - d_2: the absolute frequency of the modal class - the absolute frequency of the next class $(n_i - n_{i+1})$.
  - a: the amplitude of the modal class.

Example 4.3.4

class	n_i
$[1,60 - 1,65[$	*3*
$[1,65 - 1,70[$	*8*
$[1,70 - 1,75[$	*2*

- *The modal class is:* $[1,65-1,70[$.
- $L_i = 1,65$.
- $d_1 = 8-3 = 5$.
- $d_2 = 8-2 = 6$.
- $a = 1,70-1,65 = 0,05$

then $Mo = 1,65 + \left(\frac{5}{5+6}\right) 0,05 = 1,67$

4.3.3 Median

- Cas of discrete variable:

  The median Me is the value at the center of a series of numbers arranged in ascending order.

  – If n is even, then
  $$Me = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$$

  – If n is odd, then
  $$Me = x_{\frac{n+1}{2}}$$

Example 4.3.5 *The number of children of 6 families is as follows*

$$7, 3, 1, 1, 5, 2$$

We first order the values:

$$1, 1, 2, 3, 5, 7$$

We have $n = 6$ *is even so* $Me = \frac{x_3+x_4}{2} = \frac{2+3}{2} = 2.5$.

Example 4.3.6 *The number of children of 7 families is as follows*

$$3, 2, 1, 0, 0, 1, 2$$

We first order the values:

$$\underbrace{0,0,1}_{3},\ \underbrace{\mathbf{1}}_{Me=x_4=1},\ \underbrace{2,2,3}_{3}$$

We have $n = 7$ is odd so $Me = x_4 = 1$.

- Cas of continuous variable:

  In this case the median is given by

$$Me = L_i + \left(\frac{\frac{N}{2} - \sum_{i=1}^{<Me} n_i}{n_{Me}} \right) a$$

  - L_i: the lower limit of the median class
  - $\sum_{i=1}^{<Me} n_i$ = the sum of the absolute frequencies corresponding to all classes below the median class.
  - n_{Me} = the absolute frequency of the median class.
  - a: the amplitude of the median class.

Example 4.3.7

Class	n_i	c_i
[0,1[	2	0.5
[1,3[	5	2
[3,5[	6	4
[5,7[	4	6
[7,9[	3	8

The median class

$$\frac{N}{2} = \frac{20}{2} = 10$$

the median class is [3,5[

$$L_i = 3, \quad a = 2, \quad n_{Me} = 6, \quad \sum_{i=1}^{<Me} n_i = 7$$
$$Me = 3 + \frac{10-7}{6} \cdot 2 = 3 + \frac{3}{6} \cdot 2 = 3 + 1 = 4$$

4.4 Quartiles

Quartiles are the three values that divide the distribution into four equal parts.

- The first quartile Q_1 represents 25% of the sample i.e. Q_1 is the value x_i whose position is the smallest integer following $\frac{n}{4}$.
- The second quartile Q_2 represents 50% of the sample.
- The third quartile Q_3 represents 75% of the sample i.e. Q_3 is the value x_i whose position is the smallest integer following $\frac{3n}{4}$.

4.4.1 Interquartile Range

The interquartile range is

$$I_Q = Q_3 - Q_1.$$

Example 4.4.1 *In the example of the following observations:*

x_i	*1*	*3*	*5*	*7*	*9*
n_i	*1*	*2*	*1*	*2*	*2*
n_i^c	*1*	*3*	*4*	*6*	*8*

- *We have* $n = 8$ *and* $\frac{n}{4} = 2$ *so* $Q_1 = x_2 = 3$.
- *We have* $n = 8$ *and* $\frac{3n}{4} = 6$ *so* $Q_3 = x_6 = 7$.

4.5 Dispersion parameters

Dispersion parameters are the parameters that summarize the dispersion of values around the central value.

4.5.1 Range

The difference between the largest value and the smallest value observed is called the range e.

$$e = x_{max} - x_{min}$$

Example 4.5.1 *The marks of 10 students are as follows:*

$$2, 3, 10, 10, 11, 12, 15, 18, 19, 20$$

Then $e = x_{max} - x_{min} = 20 - 2 = 18$.

4.5.2 Variance

A variance is the arithmetic mean of the squares of the differences between the values of a variable and the arithmetic mean.

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

4.5.3 Standard deviation

We call standard deviation denoted σ_X the square root of the variance.

$$\sigma_X = \sqrt{V(X)}$$

4.5.4 Coefficient of variation

The coefficient of variation, CV, is defined by:

$$CV = \frac{\sigma_X}{\bar{x}}$$

Example 4.5.2 *Given the data:*

x_i	*0*	*1*	*2*	*3*	*4*
n_i	*2*	*3*	*1*	*1*	*1*

$\bar{x} = 1.5$

$$V(X) = \frac{1}{8}\sum_{i=1}^{5} n_i x_i^2 - (1.5)^2 = \frac{1}{8}(2(0)^2 + 3(1)^2 + 1(2)^2 + 1(3)^2 + 1(4)^2) - 2.25$$

$$V(X) = \frac{32}{8} - 2.25 = 1.75$$

The standard deviation: $\sigma_X = \sqrt{1.75} = 1.3$.

The coefficient of variation: $CV = \frac{1.3}{1.5} = 0.87$.