Mohamed Khider Biskra University

Faculty of Exact Sciences and SNV

Biology department

Module: Math and Stat

Level: Licence 1

Prof: Amel Chine

University year:2023/2024

Chapter 1

Descriptive Statistics

1.1 Basic notions

Population: It is a set of individuals or objects such as people, animals, plants, things.

Sample: it is a subset of the population.

Statistical units or objects: are the elements of the population or sample.

Character or variable: a character is the variable that characterizes the members of the population. we distinguish two types of variables:

- Qualitative variables: they have no numeric value. They described by a word or phrase for example: blood group (A+,O+, ect), color of eyes (brown, blue,...), gender (male, female)

- Quantitative variables: are variables that are counted or measured on a numerical scale. They described by a number for example: exam scor (11,15,18,....), number of children in a family (0,1,2,..), Height of students (160,175,170,....)

avearge temperature(18,20,25,...). Quantitative data can be:

- Discrete: are variables that are counted. For example: number of children in a family (0,1,2,..). number of people in a city,.....
- Continuous: are numeric variables that are measured on a continuous scale (interval). for example: height of persons ([160-165[, [165,170[, [170,175])

Modalities or values data : are the values taken by the character or the variable, as example: number of children per family (0,1,2,4,3.) so the modalities are {0,1,3,4}.

Size of population or sample size : it is the number of statistical units in the population, denoted by N or in the sample and denoted by n.

Statistical series: We call a statistical series the sequence of valuestaken by a variable X on the staistical units.The values of the variable are noted

$$X_1, X_2, ..., X_n$$

Example 1 *In a study of* 10 *families we calculate the number of children per family. The values of variable are:*

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

so: the variable X is the number of children per family and the modalities are:{0,1,2,3}. Then

$$X_1 = 0, X_2 = 1, X_3 = 2, X_4 = 3$$

and the sample size is $n = 10$.

1.2 Qualitative variable

A qualitative variable has distinct valueswhich cannot be ordered. Note i the number of distinct values or modalities. We call absolute frequency of a modality the number of times that this modality is repeated. Note n_i the absolute frequency of modality x_i. The relative Frequency noted f_i is the absolute frequency divided by the number of statistical units. It is defined by:

$$f_i = \frac{n_i}{n}$$

Example 2 *Let a statistical series about civil status of 30 persons*

M	M	D	U	M	U	W	D	W	D
D	M	M	M	M	U	U	U	M	M
D	D	U	W	U	M	M	M	D	M

where M: married,U: unmarried, D:divorced and W:widower

So the qualitative variable X is: civil status and its modalities are {M,D,U,W} where $k = 4$ *(number of modalities). The sample* $n = 30$.

With this statistical series we orgnize the frequency table as follows:

x_i	n_i	f_i	*percent*=f_i.100
M	13	$\frac{13}{30} = 0.433$	43.3
D	7	$\frac{7}{30} = 0.233$	23.3
U	7	$\frac{7}{30} = 0.233$	23.3
W	3	$\frac{3}{30} = 0.1$	10

We conclude that the highest percentage is that of married people with 43.33%.

1.3 Quantitative variable

As we saw in the previous section, a quantitative variable is a variable that can be measured and counted and can take two types: discrete: the variable in this case is countable and takes isolated values, or continuous: its values are represented by intervals called classes.

1.3.1 Discrete quantitative variable

Frequency, relative frequency and cumulative frequency

1- **Frequency:** the frequency noted n_i is the number of times that the modalities are repeated.

2- **Relative frequency**: The relative Frequency noted f_i is the frequency divided by the sample size or the number of statistical units. It is defined by:

$$f_i = \frac{n_i}{n}$$

Remark 3 *1.*

$$\sum_{i=1}^{k} n_i = n, \quad \sum_{i=1}^{k} f_i = 1.$$

2.

$$0 < f_i < 1$$

3- **Cumulative frequency:** The cumulative frequency is the frequency of the first modality is added to the frequency of the second modality, and this sum is added to the third modality and so on then, frequencies that are obtained this way are known as cumulative frequency. Cumulative frequency is used to know the number of observations that lie above (or below) a particular frequency in a given data statistics. There are two types of cumulative frequency less than type and more than type, noted by $N_i \nearrow$ and $N_i \searrow$ respectively. The cumulative frequencies are defined by:

$$N_i \nearrow = \sum_{i=1}^{j} n_i, \ j = 1, 2, ...k$$

and

$$N_i \searrow = n - \sum_{j=1}^{i} n_j.$$

4- **Cumulative relative frequency:** The cumulative relative frequency less than, noted $F_i \nearrow$ is the proportion between the The cumulative frequency $N_i \nearrow$ and n.The cumulative relative frequency more than, noted $F_i \searrow$ is the proportion between the The cumulative frequency $N_i \searrow$ and n.

$$F_i \nearrow = \frac{N_i \nearrow}{n},\ F_i \searrow = \frac{N_i \searrow}{n}$$

5- **Frequency table:** Is given by:

X_i	n_i	f_i	$N_i \nearrow$	$N_i \searrow$	$F_i \nearrow$	$F_i \searrow$
X_1	n_1	$f_1 = \frac{n_1}{n}$	n_1	$n_1 + n_2 + ... + n_k$	$\frac{N_1 \nearrow}{n}$	$\frac{N_1 \searrow}{n}$
X_2	n_2	$f_2 = \frac{n_2}{n}$	$n_1 + n_2$	$n_2 + ... + n_k$	$\frac{N_2 \nearrow}{n}$	$\frac{N_2 \searrow}{n}$
$\vdots$						
X_{k-1}				$n_k + n_{k-1}$	$\frac{N_{k-1} \nearrow}{n}$	$\frac{N_{k-1} \searrow}{n}$
X_k	n_k	$f_k = \frac{n_k}{n}$	$n_1 + n_2 + ... + n_k$	n_k	$\frac{N_k \nearrow}{n}$	$\frac{N_k \searrow}{n}$
Total	n	1				

Example 4 *Given the following series of data on notes for 20 students*

11	*12*	*15*	*14*	*14*	*10*	*10*	*12*	*10*	*10*
15	*11*	*11*	*10*	*14*	*11*	*12*	*12*	*15*	*12*

Question: *Fill the freauency table*

Answer: *The variable is the notes of students and the sample size* $n = 20$. *The modalities are* $\{10, 11, 12, 14, 15\}$ *so* $k = 5$. *First we sort the modalities from the samallest value to the largets and we present the frequency table*

X_i	n_i	f_i	$N_i \nearrow$	$N_i \searrow$	$F_i \nearrow$	$F_i \searrow$
$X_1 = 10$	5	$\frac{5}{20} = 0.25$	5	20	$\frac{5}{20} = 0.25$	$\frac{20}{20} = 1$
$X_2 = 11$	4	$\frac{4}{20} = 0.2$	9	15	$\frac{9}{20} = 0.45$	$\frac{15}{20} = \frac{3}{4}$
$X_3 = 12$	5	$\frac{5}{20} = 0.25$	14	11	$\frac{14}{20} = 0.7$	$\frac{11}{20} = 0.55$
$X_4 = 14$	3	$\frac{3}{20} = 0.15$	17	6	$\frac{17}{20} = 0.85$	$\frac{6}{20} = \frac{3}{10}$
$X_5 = 15$	3	$\frac{3}{20} = 0.15$	20	3	$\frac{20}{20} = 1$	$\frac{3}{20} = 0.15$
$Total$	20	1				

1.3.2 Descriptive statstics measures

Central tendency: A measure of central tendancy is a value that represents typical, or central , entry of a data set. The measures of central tendancy are: mean, mode, mediane

and quartiles.

1. **Mean:** The mean of a data set is the sum of the product between modality and its frequency divided by sample size n,or it is the sum of product between the modality and its relative frquency.Tthe mean is noted by $\bar{X}$ and defined by

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{k} n_i X_i$$

or

$$\bar{X} = \sum_{i=1}^{k} f_i X_i$$

Example 5 *For the follwing data on ages of 10 persons:*

$$35, 32, 30, 30, 30, 30, 22, 22, 25, 25$$

we calculate the mean. First we sort the data and we determine the values of the variable. X =age, $n = 10$, modalities are:$\{22, 25, 30, 32, 35\}$ and $k = 5$. Second step we fill the frequency table

X_i	n_i	f_i	$N_i \nearrow$	$F_i \nearrow$	$n_i X_i$	$f_i X_i$
22	2	0.2	2	$\frac{5}{20} = 0.25$	44	4.4
25	2	0.2	4	$\frac{9}{20} = 0.45$	50	5.0
30	4	0.4	8	$\frac{14}{20} = 0.7$	120	12
32	1	0.1	9	$\frac{17}{20} = 0.85$	32	3.2
35	1	0.1	10	$\frac{20}{20} = 1$	35	3.5
Total	10	1			$\sum_{i=1}^{5} n_i X_i = 281$	$\sum_{i=1}^{5} f_i X_i = 28.1$

so $\bar{X} = 28.1$ or $\bar{X} = \frac{1}{10}(281) = 28.1$.

2. **Mode:** The mode is the modality that has the greatest frequency, it is noted by Mo. For the following data set

$$1, 1, 2, 3, 3, 3, 4, 4, 5$$

we have 5 modalities $\{1, 2, 3, 4, 5\}$. The mode is $Mo = 3 = X_3$ because $n_3 = 3$ is the greatest frequency and $n_1 = 2$, $n_2 = 1$, $n_4 = 2$ and $n_5 = 1$.

Remark 6 *we can find more than mode in a data set. For this data set*

$$1, 1, 1, 2, 3, 3, 3, 4, 4, 5$$

we find two modes $X_1 = 1$ *and* $X_3 = 3$ *because they has the same greatest frequency* $n_1 = n_3 = 3$. *in that case the dataset called bimodal.*

3. **Median:** The median of a data set is the modality that lies in the middle of the data when the data set is ordred, or it is the modality which divides the statistical series into two equal parts (50%).It is noted by Me. We have two cases to detrmine the mediane:

   - If the sample size n is an odd number, the meadian is the modality in position $\frac{n+1}{2}$

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

     For example. The following data set of 9 numbers

     $$1, 1, 2, 3, 3, 3, 4, 4, 5$$

     we observe that sample size $n = 9$, so The mediane value is $Me = X_{\frac{9+1}{2}} = X_5 = 3$.

   - If the sample size n is an even number, the meadian is the mean of the two middle values in positions $\frac{n}{2}$and$\frac{n}{2} + 1$

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

We take the following example. This data set of 10 numbers

$$1, 1, 2, 3, 3, 4, 4, 5, 5, 6$$

The mediane value is:

$$\begin{aligned} Me &= \frac{X_{\frac{10}{2}} + X_{\frac{10}{2}+1}}{2} \\ &= \frac{X_5 + X_6}{2} \\ &= \frac{3+4}{2} \\ &= 3.5 \end{aligned}$$

4. **Quartiles**: The quartile is a measure which divides the data set or the statistical series into four parts, or quarters, of more-or-less equal size and the data must be ordered from smallest to largest to compute quartiles. There are three quartiles as follows:

   - **The first quartile Q_1:** It is the value which cuts off the first quarter of the sample. It is also known as the lower quartile, as 25% of the data is below this point. It can be calculated as
   $$Q_1 = X_{\frac{n}{4}}.$$
   - **The second quartile (Q_2):** Is the median of a data set.
   - **The third quartile** (Q_3): It is the value which cuts off the third quarter of the sample. It is known as the upper quartile, as 75% of the data lies below this point. It can be calculated as
   $$Q_3 = X_{\frac{3n}{4}}.$$

we take the follwing example. The following data set of 9 numbers

$$1, 1, 2, 3, 3, 3, 4, 4, 5$$

then

$$\begin{aligned} Q_1 &= X_{\frac{10}{4}} = X_{2.25} \backsimeq X_3 = 2 \\ Q_3 &= X_{\frac{30}{4}} = X_{7.5} \backsimeq X_8 = 4 \end{aligned}$$

We write the median and the quartiles in function of cumulative frequency as follows:

$$\begin{aligned} Me &= N^{-1} \nearrow \left(\frac{n}{2}\right) \\ Q_1 &= N^{-1} \nearrow \left(\frac{n}{4}\right) \\ Q_3 &= N^{-1} \nearrow \left(\frac{3n}{4}\right) \end{aligned}$$

1. **Example 7** *We take the example (5), and according the frequency table and the column of* $N_i \nearrow$ *we determine the median,* Q_1 *and* Q_3

   the sample size $n = 10$ *then*

$$\begin{aligned} Me &= \frac{N^{-1} \nearrow \left(\frac{10}{2}\right)}{2} \\ &= N^{-1} \nearrow (5) \\ &= 30 \end{aligned}$$

   and

$$\begin{aligned} Q_1 &= N^{-1} \nearrow \left(\frac{10}{4}\right) = N_{2.5}^{-1} \backsimeq N_3^{-1} = 25 \\ Q_3 &= N^{-1} \nearrow \left(\frac{30}{4}\right) = N_{7.5}^{-1} \backsimeq N_8^{-1} = 30 \end{aligned}$$

We can calculate the meadian and the quartiles by the cumulative relative frequency and we read the values in frequency table in the column :$F_i \nearrow$

$$\begin{aligned} Me &= F^{-1} \nearrow (0.5) \\ Q_1 &= F^{-1} \nearrow (0.25) \\ Q_3 &= F^{-1} \nearrow (0.75) \end{aligned}$$

In this example, we obtain:

$$\begin{aligned} Me &= F^{-1} \nearrow (0.5) = 30 \\ Q_1 &= F^{-1} \nearrow (0.25) = 25 \\ Q_3 &= F^{-1} \nearrow (0.75) = 30 \end{aligned}$$

To determine the mode we use th column n_i, the mode in our example is $Mo = 30$ because the highest frequency is 4.

Measures of variation:

1. **Variance:** The variance is the mean square deviation between each value in the data set and the center of the distribution represented by the mean. Its defined by

$$var(X) = \frac{1}{n} \sum_{i=1}^{k} n_i (X_i - \bar{X})^2$$

or

$$var(X) = \frac{1}{n} \sum_{i=1}^{k} n_i X_i^2 - \bar{X}^2$$

and because $\frac{n_i}{n} = f_i$, the variance can be calculated by

$$var(X) = \sum_{i=1}^{k} f_i X_i^2 - \bar{X}^2$$

2. **Standard deviation:** The standard deviation is the square root of the variance. Its denoted by σ :

$$SD = \sqrt{var(X)}$$

Example 8 *We take the same example (5), in the frquency table we can add a new column to calculate* $f_i X_i^2$ *and then we calculate the variance and standard deviation*

X_i	n_i	f_i	$N_i \nearrow$	$N_i \searrow$	$n_i X_i$	$f_i X_i$	$f_i X_i^2 = (f_i X_i) X_i$
22	2	0.2	*2*	*10*	44	4.4	$22 * 4.4 = 96.8$
25	2	0.2	*4*	*8*	50	5.0	$25 * 5 = 125.0$
30	4	0.4	*8*	*6*	120	12	$30 * 12 = 360.0$
32	1	0.1	*9*	*2*	32	3.2	$32 * 3.2 = 102.4$
35	1	0.1	*10*	*1*	35	3.5	$35 * 3.5 = 122.5$
Total	10	1		*0*	$\sum_{i=1}^{5} n_i X_i = 281$	$\sum_{i=1}^{5} f_i X_i = 28.1$	$\sum_{i=1}^{5} f_i X_i^2 = 806.7$

and $\bar{X}^2 = 789.61$, *so*

$$\begin{aligned} var(X) &= 806.7 - 789.61 \\ &= 17.09 \end{aligned}$$

and the standard deviation

$$\begin{aligned} D &= \sqrt{17.09} \\ &= 4.134 \end{aligned}$$

3. **Range:** The range of a data set is the diffrence between maximum value and minimum value in the set. It s defined by

$$Range = X_{\max} - X_{\min}$$

Example 9 *According the example (5), the range is*

$$\begin{aligned} Range &= 35 - 22 \\ &= 13 \end{aligned}$$

4. **Interquartile range:** The interquartile range of a data set is the diffrence between the first quartile Q_1 and the third quartile Q_3. It s defined by

$$IRange = Q_3 - Q_1$$

Remark 10 *Unlike range and interquartile range, variance is a measure that accounts for the dispersion of all values in a data set.*