Larbi Ben M'hidi-Oum El Bouaghi University

Faculty of Exact Sciences and Natural and Life Sciences

Departement of Mathematics and Computer Science

First year Licence **Introduction to probability and descriptive statistics**

Answers of series N° 2 : Graphs and measures of position and variability

Exercise 02 (quantitative discrete data) : The frequency table :

Values x_i	1	2	3	4	5	$\sum$
Frequency n_i	84	29	3	3	1	$n = 120$
ICF $N_{x=x_i} \uparrow$	84	113	116	119	120	////
$n_i\ x_i$	84	58	9	12	5	168
$n_i\ x_i^2$	84	116	27	48	25	300

1. The sample of intrest is the subset of vehicles,

   The sample size : $n = 120 = \sum\ n_i$

   The variable X of interest is the number of passengers in each vehicle,

   The type of X : quantitative discrete.

2. Draw the frequency diagram (bar chart) such as the x-axis for the values x_i (line 1) and the y-axis for the n_i (line 2).

3. Plot the increasing cumulative frecuency curve (or the frequency curve) such as the x-axis for the values xi (line 1) and the y-axis for the $N_x \uparrow$ (line 3).

$$N_x \uparrow = \sum_{i\,:\,x_i \leq x} n_i\ , \qquad x \in \mathbb{R}$$

4. **Measures of position (or central tendency)**

   The mean :

$$\overline{x} = \frac{\sum_i\ n_i \times x_i}{n} = \frac{168}{120} = 1.4$$

   The median : notice that $n = 120$ an even number, so

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

   from the line $N_{x=x_i} \uparrow$, we obtain : $Me = \dfrac{1+1}{2} = 1.$

   The first quartile q_1 :

$$q_1 = x_{\frac{n}{4}} = 1$$

   The third quartile q_3 :

$$q_3 = x_{\frac{3n}{4}} = 2$$

   The mode : From the line of n_i, we notice that the the most frequent is equal to $n_1 = 84$, then $Mo = 1$.

5. **Measures of despersion (or variability or spread)**

   The rang : $R = max - min = 5 - 1 = 4$.

   The variance :

$$Var(X) = \frac{\sum_i \ n_i \times x_i^2}{n} - \overline{x}^2 = \frac{300}{120} - (1.4)^2 = 0.54$$

   The standard deviation : $\sigma_X = \sqrt{Var(X)} = 0.73$.

   The coefficient of variation : $CV = \frac{\sigma_X}{\overline{x}} = 0.52$.

Answer 03 :

1. We have the range $R = max - min = \alpha - 800 = 3200$, so $\alpha = 4000$.

2. We have

$$\overline{x} = 2012 = \frac{\sum_i n_i \ c_i}{n} = \frac{48400 + 48000 + \frac{100+\beta}{2} \ 52 + \frac{\beta+2400}{2} \ 18 + 172800}{200}$$

$$\frac{332400 + 35 \ \beta}{200} = 2012 \quad \Rightarrow \quad \beta = 2000.$$

3. Complete the table.

Classes $[e_{i-1}, e_i[$	$[800, 1400[$	$[1400, 1600[$	$[1600, 2000[$	$[2000, 2400[$	$[2400, 4000[$	$\sum$
Centre of classes c_i	1100	1500	1800	2200	3200	////
Frequency n_i	44	32	52	18	54	n=200
ICF $N_{x=e_i} \uparrow$	44	76	128	146	200	////
RF f_i	0.22	0.16	0.26	0.09	0.27	1
ICRF $F_{x=e_i} \uparrow$	0.22	0.38	0.64	0.73	1	////
$a_i = e_i - e_{i-1}$	600	200	400	400	1600	////
u_i	3	1	2	2	8	////
$d_i = \frac{n_i}{u_i}$	14.67	32	26	9	6.75	////

   Line 2 : $c_i = \frac{e_{i-1} + e_i}{2}$.

   Line 5 : we have $f_1 = F_{x=e_1=1400} \uparrow$ and $f_i = F_{e_i} \uparrow - F_{e_{i-1}} \uparrow, \quad i = 2, ..., 5$.

   Line 3 : $n_i = f_i \times n$.

   Line 4 : $N_{x=e_i} \uparrow = \sum_{e<e_i} \ n_i$ such as $e \in [800, 4000[$. Or $N_{x=e_i} \uparrow = F_{x=e_i} \uparrow \times n$.

   $\sum_i \ n_i \times \ c_i^2 = 93380 \ \times 10^4$ (we need this sum to calculate the variance).

   Or $\sum_i \ f_i \times \ c_i^2 = 4669000$.

4. **- The frequency (or relative frequency) curve :**

   Step 01 : Calculate $N_x \uparrow$ (or $F_x \uparrow$).

   Step 02 : Determine the points $(e_i, \ N_{x=e_i} \uparrow)$ (or $(e_i, \ F_{x=e_i} \uparrow)$), such as the x-axis for the classes and the y-axis for the $N_x \uparrow$ (or $F_x \uparrow$).

   Step 03 : Draw the curve.

We can deduce the median and the quartiles graphically.

- The frequency (or relative frequency) histogram :

Step 01 : We add a new line for calculating the amplitude (width) of classes a_i (line 6). According to this line, note that the width a_i **are not equal**, so

Step 02 : We add two new lines, the first one to determine the unit u_i (line 7) such as $u_2 = 1$ because the width $a_2 = 200$ is the minimum of the widths a_i (see the table), and the second one for calculating the density $d_i = \dfrac{n_i}{u_i}$ (or $d_i = \dfrac{f_i}{u_i}$) (line 8).

Step 03 : Draw the histogram such as the x-axis for the classes and the y-axis for the densities d_i.

5. **Measures of position**

- **The mode :** from the line 9, note that the most density is $d_2 = 32$, so

the mode class : $[e_1, e_2[= [1400, 1600[$

- **The median** is the solution to the equation :

$$N_{x=Me} \uparrow = \frac{n}{2}$$

so we have

$$\begin{array}{rcccl} 76 & \leq & \dfrac{n}{2} = 100 & < & 128 \quad (from\ the\ line\ 4) \\ 1600 & \leq & Me & < & 2000 \quad (from\ the\ line\ 1) \end{array}$$

so the median class is : $[1600, 2000[\Rightarrow Me \in [1600, 2000[$. Then

$$Me = 1600 + (2000 - 1600)\ \frac{\frac{n}{2} - 76}{128 - 76} = 1784.615$$

The second method : the median is the solution to the equation

$$F_{Me} \uparrow = 0.5$$

so we have

$$\begin{array}{rcccl} 0.38 & \leq & 0.5 & < & 0.64 \quad (from\ the\ line\ 6) \\ 1600 & \leq & Me & < & 2000 \quad (from\ the\ line\ 1) \end{array}$$

Then, we obtain :

$$Me = 1600 + (2000 - 1600)\ \frac{0.5 - 0.38}{0.64 - 0.38} = 1784.615$$

- **The first quartile** q_1 is the solution to the equation

$$N_{q_1} \uparrow = \frac{n}{4}$$

so we have

$$\begin{aligned} 44 \;&\leq\; \frac{n}{4} = 50 \;<\; 76 \quad (from\ the\ line\ 4) \\ 1400 \;&\leq\; q_1 \;<\; 1600 \quad (from\ the\ line\ 1) \end{aligned}$$

so

$$q_1 = 1400 + (1600 - 1400)\ \frac{\frac{n}{4} - 44}{76 - 44} = ...$$

The second method : the first quartile q_1 is the solution to the equation

$$F_{q_1} \uparrow = 0.25$$

so we have

$$\begin{aligned} 0.22 \;&\leq\; 0.25 \;<\; 0.38 \quad (from\ the\ line\ 6) \\ 1400 \;&\leq\; q_1 \;<\; 1600 \quad (from\ the\ line\ 1) \end{aligned}$$

then

$$q_1 = 1400 + (1600 - 1400)\ \frac{\frac{1}{4} - 0.22}{0.38 - 0.22} = ...$$

- **The third quartile** q_3 is the solution to the equation

$$N_{q_3} \uparrow = \frac{3\,n}{4}$$

so we have

$$\begin{aligned} 146 \;&\leq\; \frac{3\ n}{4} = 150 \;<\; 200 \quad (from\ the\ line\ 4) \\ 2400 \;&\leq\; q_3 \;<\; 4000 \quad (from\ the\ line\ 1) \end{aligned}$$

so

$$q_3 = 2400 + (4000 - 2400)\ \frac{\frac{3n}{4} - 146}{200 - 146} = ...$$

The second method : the third quartile q_3 is the solution to the equation

$$F_{q_3} \uparrow = 0.75$$

so we have

$$\begin{aligned} 0.73 \;&\leq\; 0.75 \;<\; 1 \quad (from\ the\ line\ 6) \\ 2400 \;&\leq\; q_1 \;<\; 4000 \quad (from\ the\ line\ 1) \end{aligned}$$

so

$$q_3 = 2400 + (4000 - 2400)\ \frac{\frac{3}{4} - 0.73}{1 - 0.73} = ...$$

Measures of variability (or dispersion, or spread)

- The variance :

$$Var(X) = \left[\frac{1}{n}\sum_i n_i \times c_i^2\right] - \overline{x}^2 = 620856$$

or

$$Var(X) = \left[\sum_i f_i \times c_i^2\right] - \overline{x}^2 = 620856$$

- The standard deviation : $\sigma_X = \sqrt{Var(X)} = 787.94.$

- The coefficient of variation : $CV = \dfrac{\sigma_X}{\overline{x}} = 0.39.$

Answer 04 : Construct the box plot for :

A. The first set of data :

$$32, 32, 45, 55.5, 56, 56, 59, 68, 70, 72, 77, 78, 79, 80, 81, 84, 84.5, 90, 90, 99$$

We have

-The smallest value : $min = 32$

-The first quartile : $q_1 = x_{\frac{n}{4}} = 56$

-The median : $Me = \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} = \dfrac{72+77}{2} = 74.5$

-The third quartile : $q_3 = x_{\frac{3\,n}{4}} = 81$

-The largest value : $max = 99$

B. The second set of data :

$$25.5, 45, 65, 68, 76, 78, 78, 79, 79, 80, 81, 81, 83, 84.5, 85, 88, 89, 90, 90, 98, 98, 98$$

We have

-The smallest value : $min = 25.5$

-The first quartile : $q_1 = x_{\frac{n}{4}} \backsimeq x_6 = 78$

-The median : $Me = \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} = \dfrac{81+81}{2} = 81$

-The third quartile : $q_3 = x_{\frac{3\,n}{4}} \backsimeq x_{16} = 88$

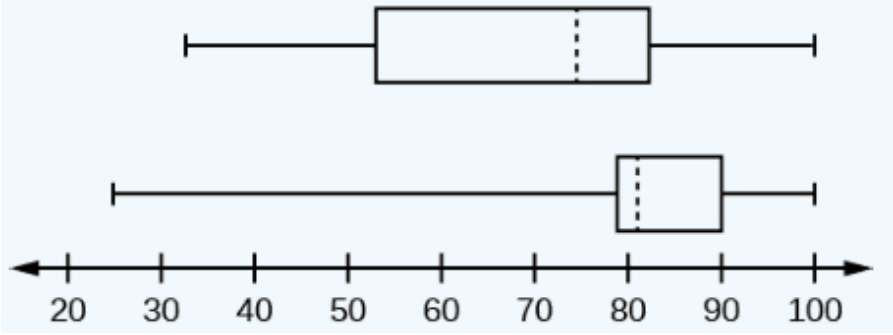
-The largest value : $max = 98$

We have :

The interquartile range for the first data is : $IQR_A = q_3 - q_1 = 82.5 - 56 = 26.5.$

The interquartile range for the second data is : $IQR_B = q_3 - q_1 = 89 - 78 = 11.$

So, the first data set has the wider spread for the middle 50% of the data, because the IQR_1 is greater than the IQR_2. This means that there is more variability in the middle 50% of the first data set.



Answer 06 :

The age distribution of a group of persons is given in the following table :

Classes $[e_{i-1}, e_i[$	< 9	$[9, 11[$	$[[11, 13[$	$[13, 15[$	$[15, 17[$	$[[17, 21[$	$\sum$
Age	< 9	< 11	< 13	< 15	< 17	< 21	////
Number $N_{x=e_i} \uparrow$	0	12	25	33	37	n= 40	////
Frequencies n_i	0	12	13	8	4	3	n=40
$c_i = \dfrac{e_i + e_{i-1}}{2}$							////
$n_i \times c_i$							
$n_i \times c_i^2$							
$a_i = e_i - e_{i-1}$							////
Units u_i							////
Densities d_i							////

1. Draw the diagram for this data. (i.e. draw the frequency curve $N_x \uparrow$)
2. Calculate graphically the median and the quartiles. Draw the box plots.
3. We have $N_{x=12} \uparrow$ the number of persons with age < 12 so :

$$\begin{array}{rcl} 11 & \leq & x = 12 \quad < \quad 13 \quad (from\ the\ line\ 1) \\ 12 & \leq & N_{x=12} \uparrow \ < \quad 25 \quad (from\ the\ line\ 3) \end{array}$$

   Then, we obtain :

$$N_{x=12} \uparrow = 12 + (25 - 12)\ \frac{12 - 11}{13 - 11} = 18.5 \backsimeq 19\ persons$$

   Then, the proportion of persons with age between 12 and 15 years :

$$P_{[12,15]} = \frac{N_{[12,15]}}{40} \times 100 = \frac{N_{15} \uparrow\ - N_{12} \uparrow}{40} \times 100 = \frac{33\ -\ 19}{40} \times 100 = 35\%$$

4. $\star$ For the mean, we add in the table the lines n_i, c_i and $n_i \times c_i$.

   For calculate the variance, we add in the table the line $n_i \times c_i^2$.

   For find the mode class, we add in the table the lines a_i, u_i and the line of densities $d_i = \dfrac{n_i}{u_i}$.

5. ⋆ On note

$$Y = \frac{X - 14}{2} = \frac{X}{2} - \frac{14}{2} = 0.5X - 7 = aX + b$$

where $a = 0.5$, $b = -7$ and X is the variable studied. So :

The mean of a new data Y is given by :

$$\overline{y} = a\overline{X} + b$$

and the variance of a new data Y is given by :

$$Var(Y) = a^2\ Var(X).$$