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)**

   <u>The mean :</u>

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

   <u>The median :</u> notice that $n = 120$ an even number, so

$$Me = \frac{\left(\frac{n}{2}\right)^{th}\ value\ +\ \left(\frac{n}{2}+1\right)^{th}\ value}{2}$$

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

   <u>The first quartile q_1 :</u>

$$q_1 = (\frac{n}{4})^{th}\ value = 1$$

   <u>The third quartile q_3 :</u>

$$q_3 = (\frac{3n}{4})^{th}\ value = 2$$

   <u>The mode :</u> 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

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

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

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
FC $N_{x=e_i} \uparrow$	44	76	128	146	200	////
RF f_i	0.22	0.16	0.26	0.09	0.27	1
RFC $F_{x=e_i} \uparrow$	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 :** draw the curve such as the x-axis for the classes and the y-axis for the $N_x \uparrow$ (or $F_x \uparrow$). 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 $d_i = \frac{n_i}{u_i}$ (or $d_i = \frac{f_i}{u_i}$) (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 amplitude of the mode class : $a_2 = e_2 - e_1 = 200$

$m_1 = d_2 - d_1 = 32 - 14.67$

$m_2 = d_2 - d_3 = 32 - 26$

so, the mode is given by :

$$Mo = e_1 + a_2 \frac{m_1}{m_1 + m_2} = 1548.56$$

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

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

so we have

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

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

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

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

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

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

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

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

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

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

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

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 : $q_1 = \left(\frac{n}{4}\right)^{th}\ value = 56$

-The median : $Me = \dfrac{\left(\frac{n}{2}\right)^{th}\ value\ +\ \left(\frac{n}{2}+1\right)^{th}\ value}{2} = \dfrac{72+77}{2} = 74.5$

-The third quartile : $q_3 = \left(\frac{3\,n}{4}\right)^{th}\ value = 81$

-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 : $q_1 = \left(\frac{n}{4}\right)^{th}\ value \backsimeq 6^{th}\ value = 78$

-The median : $Me = \dfrac{\left(\frac{n}{2}\right)^{th}\ value\ +\ \left(\frac{n}{2}+1\right)^{th}\ value}{2} = \dfrac{81+81}{2} = 81$

-The third quartile : $q_3 = \left(\frac{3\,n}{4}\right)^{th}\ value \backsimeq 16^{th}\ value = 88$

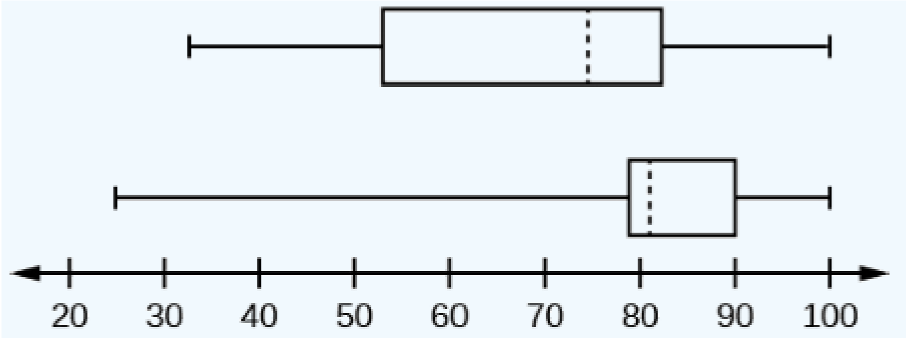
-The largest value : $max = 98$

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 05 : Consider a following data set $\{X_1, ..., X_n\}$ of a quantitative variable X. Let

$$\overline{X} = \frac{\sum_i X_i}{n} \quad and \quad Var(X) = \frac{\sum_i X_i^2}{n} - \overline{X}^2$$

the mean and the variance of X respectively. We define a new data set $\{Y_1, ..., Y_n\}$ such as

$$Y_i = \alpha X_i + \beta \quad i = 1, ..., n$$

a) We have

$$\begin{aligned}
\overline{Y} = \frac{\sum_i Y_i}{n} &= \frac{\sum_i \left(\alpha X_i + \beta\right)}{n} \\
&= \frac{\alpha \sum_i X_i + \sum_i \beta}{n} \\
&= \alpha \frac{\sum_i X_i}{n} + \frac{n\,\beta}{n} \\
&= \alpha\,\overline{X} + \beta.
\end{aligned}$$

b)

$$\begin{aligned}
Var(Y) = \frac{\sum_i Y_i^2}{n} - \overline{Y}^2 &= \frac{\sum_i \left(\alpha X_i + \beta\right)^2}{n} - \left(\alpha\,\overline{X} + \beta\right)^2 \\
&= \frac{\sum_i \left(\alpha^2 X_i^2 + \beta^2 + 2\,\alpha\beta\, X_i\right)}{n} - \left(\alpha^2\overline{X}^2 + \beta^2 + 2\alpha\,\beta\,\overline{X}\right) \\
&= \alpha^2 \frac{\sum_i X_i^2}{n} + \frac{\sum_i \beta^2}{n} + \frac{\sum_i 2\,\alpha\beta\, X_i}{n} - \alpha^2\overline{X}^2 - \beta^2 - 2\alpha\,\beta\,\overline{X} \\
&= \alpha^2 \left(\frac{\sum_i X_i^2}{n} - \overline{X}^2\right) + \frac{n\beta^2}{n} + 2\,\alpha\beta\,\frac{\sum_i X_i}{n} - \beta^2 - 2\alpha\,\beta\,\overline{X} \\
&= \alpha^2\, Var(X).
\end{aligned}$$

The second method

$$\begin{aligned}
Var(Y) = \frac{1}{n}\sum_i \left(Y_i - \overline{Y}\right)^2 &= \frac{1}{n}\sum_i \left(\alpha X_i + \beta - \alpha\overline{X} - \beta\right)^2 \\
&= \frac{1}{n}\sum_i \alpha^2 \left(X_i - \overline{X}\right)^2 \\
&= \alpha^2\, Var(X).
\end{aligned}$$