19.10.2024

Seminar W4 – 832

Exercise 1. User ratings on the website **imdb.com** range from 1 to 10. The following is an (imaginary) dataset of the first 100 ratings for the Netflix series **Squid Game (2021)**:

4	4	9	4	3	8	9	10	3	4
9	8	6	10	10	9	9	8	5	6
1	5	6	10	8	4	10	9	9	9
8	9	10	8	7	7	7	7	2	10
8	10	6	2	7	9	8	6	5	8
6	1	8	5	3	8	4	10	9	8
3	3	10	10	10	10	10	9	7	8
7	7	7	8	7	7	9	8	10	9
7	4	4	10	7	5	2	7	8	8
10	4	6	10	8	8	9	8	6	9

(a) Build the (ungrouped) frequency distribution table for this data (both absolute and relative frequencies);

(b) Compute the arithmetic, geometric and harmonic means of the data;

(c) Find the median, the mode and the range of the data;

(d) Find the quartiles, the interquartile range and the outliers of the data;

(e) Find the moment of order 2, the variance and the coefficient of variation of the data.

Sol.:

x_i	f_i	r_i	F_i
1	2	0.02	2
2	3	0.03	5
3	5	0.05	10
4	9	0.09	19
5	5	0.05	24
6	8	0.08	32
7	14	0.14	46
8	20	0.20	66
9	16	0.16	82
10	18	0.18	100

(b) $x_1, \dots, x_n$

$$AM = \frac{x_1 + x_2 + \dots + x_n}{n} \qquad GM = \sqrt[n]{x_1 x_2 \cdots x_n}$$

$$HM = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}} \qquad (AM \geq GM \geq HM)$$

$$AM = \frac{2\cdot 1 + 3\cdot 2 + 5\cdot 3 + 9\cdot 4 + 5\cdot 5 + 8\cdot 6 + 14\cdot 7 + 20\cdot 8 + 16\cdot 9 + 18\cdot 10}{100}$$

$$= \frac{724}{100} = 7.24$$

$$GM = \sqrt[100]{1^2 \cdot 2^3 \cdot 3^5 \cdot 4^9 \cdot 5^5 \cdot 6^8 \cdot 7^{14} \cdot 8^{20} \cdot 9^{16} \cdot 10^{18}}$$

$$HM = \frac{100}{\frac{2}{1} + \frac{3}{2} + \frac{5}{3} + \frac{9}{4} + \frac{5}{5} + \frac{8}{6} + \frac{14}{7} + \frac{20}{8} + \frac{16}{9} + \frac{18}{10}}$$

X = "selection from the dataset"

x_{me} is the value for which

$$P(X < x_{me}) \leq \frac{1}{2}, \quad P(X \geq x_{me}) \geq \frac{1}{2}$$

$$\hookrightarrow \frac{\#\text{ of dataset entries smaller than } x_{me}}{\#\text{ of dataset entries}}$$

$$F_k = \sum_{i=1}^{k} f_i$$

$F_7 = 46 \qquad F_8 = 66 \qquad F_7 < 50 < F_8$

$$\Rightarrow \quad x_{me} = 8$$

$$x_{mo} = 8$$

$$\text{range} = x_{max} - x_{min} = 10 - 1 = 9$$

Percentile: Q_i, $\quad P(X < Q_i) \leq \frac{i}{100}$

$$P(X \geq Q_i) \geq \frac{100 - i}{100}$$

Quartiles: $\quad Q_1 : \quad P(X < Q_1) \leq \frac{1}{4}$

$$P(X \geq Q_1) \geq \frac{3}{4}$$

$Q_2 = x_{me}$ $\qquad P(X < Q_2) \leq \frac{1}{2} \quad \left(P(X \geq Q_2) \geq \frac{1}{2}\right)$

Q_3 $\qquad P(X < Q_3) \leq \frac{3}{4} \quad \left(P(X \geq Q_3) \geq \frac{3}{4}\right)$

$Q_1 = 6 \qquad Q_2 = x_{me} = 8 \qquad Q_3 = 9$

$IQR = Q_3 - Q_1 = 9 - 6 = 3$

Finding the outliers:

α is an outlier $\iff \alpha \notin \left[Q_1 - \frac{3}{2} \cdot IQR,\ Q_3 + \frac{3}{2} IQR\right]$

$\iff \alpha \notin \left[6 - \frac{3}{2} \cdot 3,\ 9 + \frac{3}{2} \cdot 3\right]$

$\iff \alpha \notin [1.5,\ 13.5]$

outliers : 1

x_i	n_i	r_i	F_i
1	2	0.02	2
2	3	0.03	5
3	5	0.05	10
4	9	0.09	19
5	5	0.05	24
6	8	0.08	32
7	14	0.14	46
8	20	0.20	66
9	16	0.16	82
10	18	0.18	100

e) $\overline{y_k} = \frac{1}{N} \sum_{i=1}^{N} x_i^k$

$\overline{y_2} =$

$= \frac{1}{100} \cdot \left(2 \cdot 1^2 + 3 \cdot 2^2 + 5 \cdot 3^2 + 9 \cdot 4^2 + 5 \cdot 5^2 + 8 \cdot 6^2 + 14 \cdot 7^2 + 20 \cdot 8^2 + 16 \cdot 9^2 + 18 \cdot 10^2\right)$

$= \frac{5678}{100} = 56.78$

$\overline{\sigma^2} = \overline{y_2} - \overline{x}^2$ $\qquad \left(V(X) = E(X^2) - E(X)^2\right)$

$= 56.78 - (7.24)^2 = 4.3624$

$\overline{CV} = \frac{\overline{\sigma}}{\overline{x}} \approx 0.28729 = 28\%$

Exercise 3. The following dataset represents the average net monthly income (in thousands of RON) of 30 Romanian families.

4.5	8.9	4.4	6.5	3.5	7.8	9.4	10.2	23	9.8
7.6	4.9	7.3	19	10	9.4	3.9	9.8	4.5	3.6
12	5.3	16.4	10.5	6.8	4.2	10	9.7	4.9	9.8

(a) Group the data into classes and construct the grouped frequency distribution table using Sturges' rule;

(b) Group the data into classes and construct the grouped frequency distribution table using the formula for the class width;

(c) Find the range, the mean and the median of the data;

(d) Find the quartiles, the interquartile range and the outliers of the data.

Sol. : (a)

Sturges' rule.

n = number of classes =

$= 1 + \frac{10}{3} \log_{10}(N)$

$\simeq 1 + \lceil \log_2(N) \rceil$

$n = 1 + \frac{10}{3} \log_{10}(30) \simeq 5.9237 \qquad n = 6$

$x_{max} - x_{min} = 23 - 3.5 = 19.5$

$C = \frac{x_{max} - x_{min}}{n} = \frac{19.5}{6} = 3.25$

No	Class	Mark	Frequency	Cum. Frequency	Rel. Frequency	Cum. Rel. Frequency
1	$[3.5, 6.75)$	5.125	11	11	$\frac{11}{30} = 36.66\%$	36.66 %
2	$[6.75, 10)$	8.375	11	22	$\frac{11}{30} = 36.66\%$	73.33 %
3	$[10, 13.25)$	11.625	5	27	16.66 %	90 %
4	$[13.25, 16.5)$	14.875	1	28	3.33 %	93.33 %
5	$[16.5, 19.75)$	18.125	1	29	3.33 %	96.66 %
6	$[19.75, 23]$	21.375	1	30	3.33 %	100 %

range : $23 - 3.5 = 19.5$

mean :

median : 8.375

$Q_1 = 5.125$ $\qquad$ $IQR = 11.625 - 5.125 = 6.5$

$Q_3 = 11.625$ $\qquad$ α outlier $\Leftrightarrow$ $\alpha \notin [5.125 - \frac{3}{2} \cdot 6.5, 11.625 + \frac{3}{2} \cdot 6.5]$

x outlier $\iff$ $x \notin [-4.6250, 21.375]$

outlier : 23