19.10.2021

Seminar VII – 831

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	rf_i	F_i	rF_i
1	2	0.02	2	0.02
2	3	0.03	5	0.05
3	5	0.05	10	0.1
4	9	0.09	19	0.19
5	5	0.05	24	0.24
6	8	0.08	32	0.32
7	14	0.14	46	0.46
8	20	0.2	66	0.66
9	16	0.16	82	0.82
10	18	0.18	100	1

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

(b) $x_1, \dots, x_n$

$$\overline{x_a} = \frac{x_1 + \dots + x_n}{n} \qquad \overline{x_g} = \sqrt[n]{x_1 x_2 \cdots x_n} \qquad x_h = \frac{n}{\frac{1}{x_1} + \dots + \frac{1}{x_n}}$$

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

$$= 7.24$$

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

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

$X =$ "choosing a number from the table"

c) x_{me} is the number so that

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

$x_{me} = 8$

$x_{mo} = 8$

$x_{max} - x_{min} = 10 - 1 = 9$

$$\frac{\#\text{ of elements} < x_{me}}{\text{total number of elements}}$$

d) quartiles: Q_1, Q_2, Q_3

Q_1 the number s.t. $P(X < Q_1) \leq \frac{1}{4}$

$Q_2 = x_{me}$

Q_3 the number s.t. $P(X < Q_3) \leq \frac{3}{4}$

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

$IQR = Q_3 - Q_1 = 3$

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

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

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

$\Rightarrow$ outliers: 1

(e) $\overline{\nu_k} = \frac{1}{N}\sum_{i=1}^{N} x_i^k$ $\qquad \overline{\nu_1} = \overline{x}$

$\overline{\sigma}^2 = \overline{\nu_2} - \overline{x}^2$ $\qquad (V(X) = E(X^2) - E(X)^2)$

$\overline{\nu_2} = \frac{2\cdot 1^2 + 3\cdot 2^2 + 5\cdot 3^2 + 9\cdot 4^2 + 5\cdot 5^2 + 8\cdot 6^2 + 14\cdot 7^2 + 20\cdot 8^2 + 16\cdot 9^2 + 8\cdot 10^2}{100} =$

$= 56.68$

$\Rightarrow \overline{\sigma}^2 = 56.68 - (7.24)^2 = 4.3624 \Rightarrow \overline{\sigma} = 2.0886$

$CV = \frac{\overline{\sigma}}{\overline{x}} = \frac{2.0886}{7.24} = 0.28729$

Exercise 2. The following data represents the number of days of sick leave taken by each of the 50 employees of a given company over the last 6 weeks:

2, 2, 0, 0, 5, 8, 3, 4, 1, 0, 0, 7, 1, 7, 1, 5, 4, 0, 4.0, 1, 8, 9, 7, 0

1, 7, 2, 5, 5, 4, 3, 3, 0, 0, 2, 5, 1, 3, 0, 1, 0, 2, 4, 5, 0, 5, 7, 5, 1.

(a) Construct the grouped frequency distribution table for this dataset;

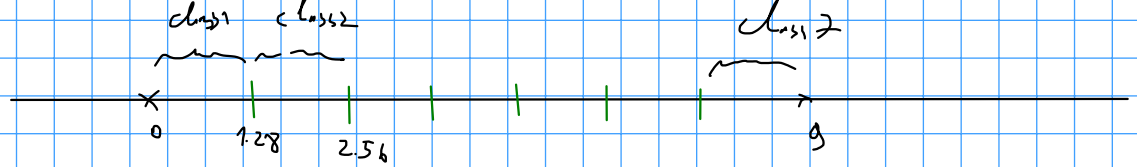
(b) Plot the data using a line graph;

(c) Plot the data as a histogram and relative frequency polygon.

(a) Sturges' rule: $n = 1 + \frac{10}{3}\log_{10}(N) \approx 6.66 \approx 7$

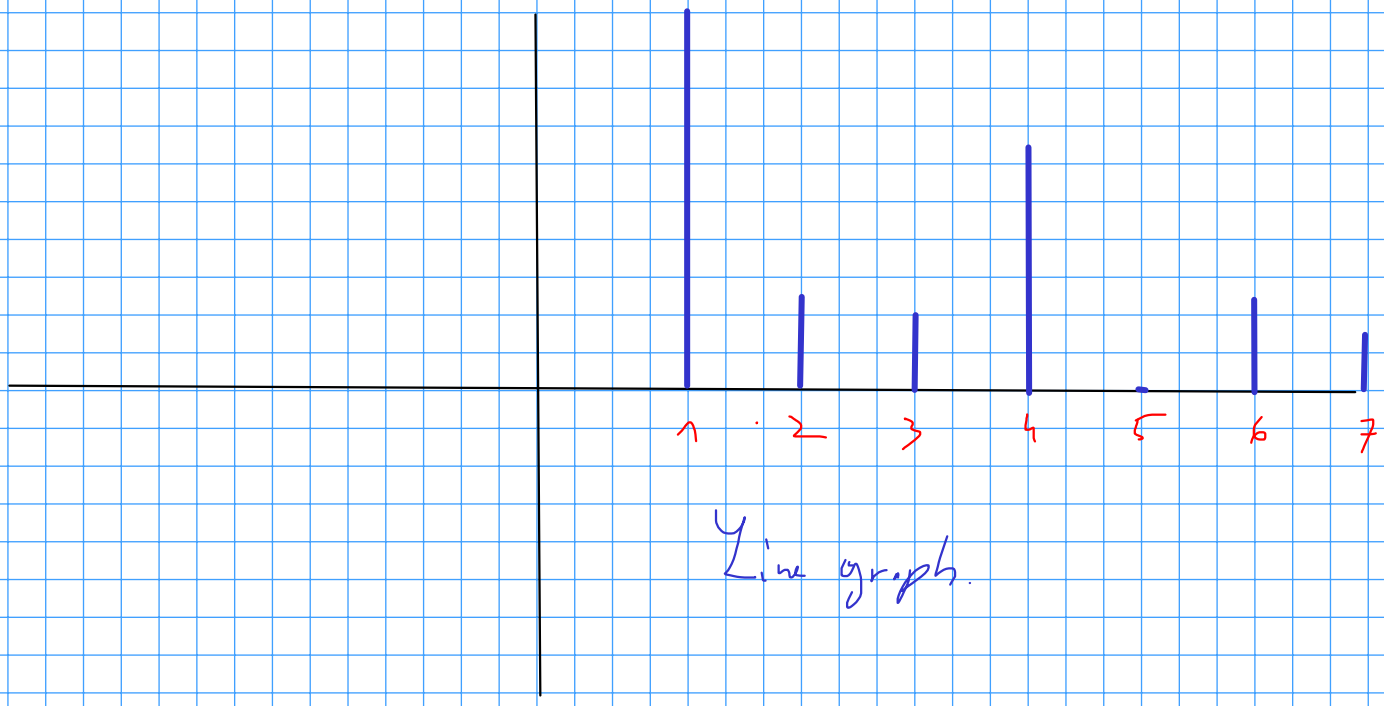
Range: $x_{max} - x_{min} = 9 - 0 = 9$

Length: $\frac{x_{max} - x_{min}}{7} = \frac{9}{7} \simeq 1.2857 \simeq 1.28$



No	Class	Mark	f_i	F_i	rf_i	rF_i
1	$[0, 1.28)$	0.64	20	20	0.4	0.4
2	$[1.28, 2.56)$	1.92	5	25	0.1	0.5
3	$[2.56, 3.84)$	3.2	4	29	0.08	0.58
4	$[3.84, 5.12)$	4.48	13	42	0.26	0.84
5	$[5.12, 6.4)$	5.76	0	42	0	0.84
6	$[6.4, 7.68)$	7.04	5	47	0.1	0.94
7	$[7.68, 9]$	8.34	3	50	0.06	1

$l = 1.28$



Line graph.

No	Class	Mark	f_i	F_i	rf_i	rF_i
1	[0, 1.28)	0.64	20	20	0.4	0.4
2	[1.28, 2.56)	1.92	5	25	0.1	0.5
3	[2.56, 3.84)	3.2	4	29	0.08	0.58
4	[3.84, 5.12)	4.48	13	42	0.26	0.84
5	[5.12, 6.4)	5.76	0	42	0	0.84
6	[6.4, 7.68)	7.04	5	47	0.1	0.94
7	[7.68, 9]	8.34	3	50	0.06	1

Histogram

The areas of the rectangles are proportional to the f_i

$\Rightarrow$ height of the bin i =

$$= \frac{f_i}{\text{length of the class}} \cdot \alpha$$

$\alpha \in \mathbb{R}$