PhD Misiura Ie. (доцент Місюра Є.Ю.)

Theme 7. Limit theorems of probability theory. Preprocessing of statistical data

Граничні теореми теорії ймовірностей. Первинне опрацювання статистичних даних

DISCRETE STATISTIC SERIES OF A SAMPLE

Values x_i	x_1	x_2	…	x_k	**SUM**
Frequencies m_i	m_1	m_2	…	m_k	$\sum_{i=1}^{k} m_i = n$
Relative frequencies $w_i = \frac{m_i}{n}$	w_1	w_2	…	w_k	$\sum_{i=1}^{k} w_i = 1$

CONTINUOUS STATISTIC SERIES OF A SAMPLE

Intervals (classes) $[x_i; x_{i+1})$	$[x_1; x_2)$	$[x_2; x_3)$	...	$[x_i; x_{i+1})$	...	$[x_{k-1}; x_k)$	**SUM**
Frequencies m_i	m_1	m_2	...	m_i	...	m_k	$\sum_{i=1}^{k} m_i = n$
Class midpoint x_i'	x_1'	x_2'	...	x_i'	...	x_k'	
Relative frequencies $w_i = \frac{m_i}{n}$	w_1	w_2	…	w_i	…	w_k	$\sum_{i=1}^{k} w_i = 1$

A midpoint of the corresponding interval is: $x_i' = \frac{x_i + x_{i+1}}{2}$, the length of each interval is: $h = x_{i+1} - x_i$.

Empirical distribution function: $F^{cum}(X < x_i) = \frac{m_i^{cum}}{n} = w_i^{cum}$, where m_i^{cum} is a number of values less than x_i, n is a sample size. A graph of this function is called a cumulative function.

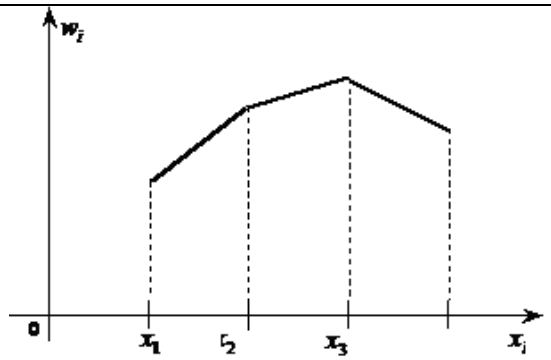

Fig. 1. Polygon of relative frequencies (for a discrete series)

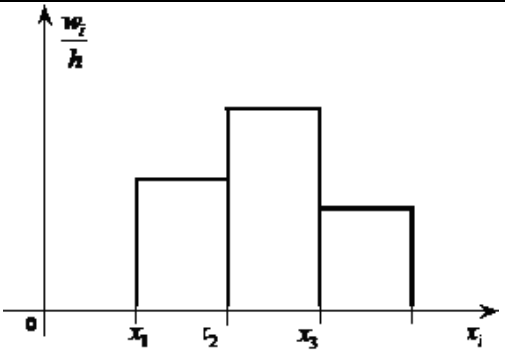

Fig. 2. Histogram of relative frequencies (for a continuous series)

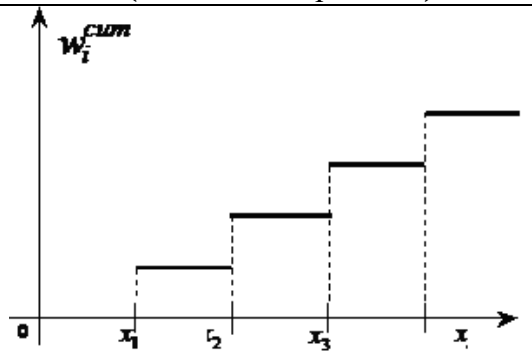

Fig. 3. An empirical function

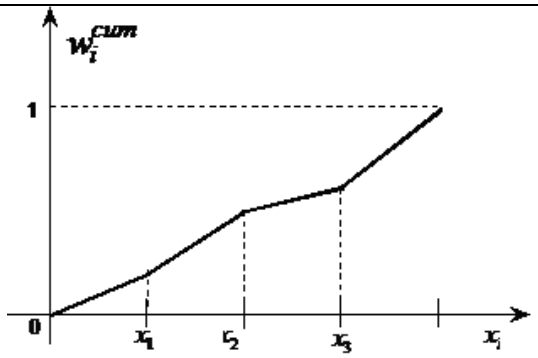

Fig. 4. A cumulative function

NUMERICAL CHARACTERISTICS

For a discrete series	For a continuous series
1. Sample mean: $\overline{x_s} = \sum_{i=1}^{k} x_i w_i$	**1. Sample mean:** $\overline{x_s} = \sum_{i=1}^{k} x_i' \cdot w_i$
2. Sample variance: $S_x^2 = \sum_{i=1}^{k} x_i^2 w_i - \left(\overline{x_s}\right)^2 = \overline{x_s^2} - \left(\overline{x_s}\right)^2$	**2. Sample variance:** $S_x^2 = \sum_{i=1}^{k} \left(x_i'\right)^2 w_i - \left(\overline{x_s}\right)^2 = \overline{x_s^2} - \left(\overline{x_s}\right)^2$
3. A root-mean-square deviation of a sample: $S_x = \sqrt{S_x^2}$	
3*. A corrected root-mean-square deviation: $S_{cor} = \sqrt{\frac{n}{n-1} S_x^2}$	
4. A coefficient of variation: $\nu = \frac{S_x}{\overline{x_s}} \cdot 100\ \%$	
5. Initial moments of the l-th order: $\nu_l = \overline{x_s^l} = \sum_{i=1}^{k} x_i^l w_i$. $\nu_1 = \overline{x_s}$, $\nu_2 = \overline{x_s^2} = \sum_{i=1}^{k} x_i^2 w_i$, $\nu_3 = \overline{x_s^3} = \sum_{i=1}^{k} x_i^3 w_i$.	
6. Central moments of the l-th order: $\mu_l = \overline{\left(x - \overline{x_s}\right)^l} = \sum_{i=1}^{k} \left(x_i - \overline{x_s}\right)^l w_i$ $\mu_1 = 0$, $\mu_2 = \sum_{i=1}^{k} \left(x_i - \overline{x_s}\right)^2 w_i = S_x^2 = \nu_2 - \nu_1^2$, $\mu_3 = \sum_{i=1}^{k} \left(x_i - \overline{x_s}\right)^3 w_i = \nu_3 - 3\nu_2\nu_1 + 2\nu_1^3$, $\mu_4 = \sum_{i=1}^{k} \left(x_i - \overline{x_s}\right)^4 w_i = \nu_4 - 4\nu_3\nu_1 + 6\nu_2\nu_1^2 - 3\nu_1^4$	

7. Mode: $M_o = x_i \left(\underset{1 \le i \le k}{m_i = \max} \right)$ **(DSS) and** $M_o = x_i + \frac{m_{M_0} - m_{M_0 - 1}}{2m_{M_0} - m_{M_0 - 1} - m_{M_0 + 1}} \cdot h$ **(CSS)**

where x_i is the lower limit of the modal class (interval); m_{M_0} is the frequency of the modal class; $m_{M_0 - 1}$ is the frequency of the class preceding the modal interval; $m_{M_0 + 1}$ is the frequency of the class following the modal interval; h is the length of the modal interval

8. Median: $M_e = \begin{cases} x_{m+1}, & if \quad k = 2m+1 \\ \frac{x_{m+1} + x_m}{2}, & if \quad k = 2m \end{cases}$ **(DSS)**

and $M_e = x_i + \frac{0.5 - F^{cum}(x_{i-1})}{F^{cum}(x_i) - F^{cum}(x_{i-1})} \cdot h$ **(CSS),**

where $F^{cum}(x_i) > 0.5, F^{cum}(x_{i-1}) < 0.5,\ h = x_{i+1} - x_i$

x_i is the lower limit of the median class (interval)

9. Range: $R = x_{\max} - x_{\min}$

10. Asymmetry: $A_S = \frac{\mu_3}{S_x^3}$ (shows a deviation of a value from its central position on the left ($A_S > 0$) or on the right ($A_S < 0$))

11. Excess: $E_S = \frac{\mu_4}{S_x^4} - 3$ (characterizes a deviation of a value from its central position up ($E_S > 0$) or down ($E_S < 0$)).

CONSTRUCTION OF CONTINUOUS STATISTIC SERIES OF A SAMPLE

1. Sturges' formula: $k = 1 + 3.322 \cdot \lg n$ (a number of intervals)

2. A length of each interval: $h = \frac{R}{k}$

3. Find: $x_{beggining} = x_{\min} - \frac{h}{2}$ and $x_{end} = x_{\max} + \frac{h}{2},\ x_{i+1} = x_i + h.$

Task 1. A random variable *X* has the sample of 20 elements:

a) 12, 14, 19, 15, 14, 18, 13, 16, 17, 12
18, 17, 15, 13, 17, 14, 14, 13, 14, 16

b)

2	4	4	7	6	5	2	2	3	4
4	3	6	5	4	7	6	6	5	3
2	4	2	3	5	7	4	3	3	2
4	5	6	6	10	4	3	3	2	3

PhD Misiura Ie. (доцент Місюра Є.Ю.)

1) Construct a discrete statistic series for this sample, plot a polygon of relative frequencies and $F(x)$;

2) calculate numerical characteristics, find M_o and M_e.

Task 2. A random variable *X* has the sample of 40 elements:

10, 13, 10, 9, 9, 12, 12, 6, 7, 9, 8, 9, 11, 9, 14, 13, 9, 8, 8, 7, 10,

10, 11, 11, 11, 12, 8, 7, 9, 10, 14, 13, 8, 8, 9, 10, 11, 11, 12, 12.

1) Construct a discrete statistic series for this sample, plot a polygon of relative frequencies and $F(x)$;

2) Calculate $\overline{x_s}, S_x, R, v, A_s, E_s$; 3) find M_o and M_e.

Task 3. A random variable *X* has the sample of 40 elements:

24	41	39	38	28	33	17	40
20	38	20	11	43	24	38	23
22	29	49	12	36	23	35	40
20	29	38	23	40	49	47	34
48	40	35	31	30	47	25	20

1) Construct a continuous statistical series.

2) Plot a histogram.

3) Calculate numerical characteristics, find M_o and M_e.

Task 4. A random variable *X* has the sample:

a)

18,8	16,0	12,6	20,0	30,0	16,4	14,6	18,4	11,6	17,4
10,4	26,4	16,2	15,0	23,6	29,2	17,0	15,6	21,0	12,0
10,2	13,6	16,6	15,4	15,8	18,0	20,2	16,0	24,0	28,0
16,4	19,6	27,0	24,8	11,0	15,8	18,4	21,6	24,2	24,8
25,8	25,2	13,4	19,4	16,6	21,6	30,0	14,0	26,0	19,0

b)

14,7	19,0	24,5	20,8	12,3	24,6	17,0	14,2	19,7	18,8
18,1	20,5	21,0	20,7	20,4	14,7	25,1	22,7	19,0	19,6
19,0	18,9	17,4	20,0	13,8	25,6	13,0	19,0	18,7	21,1
13,3	20,7	15,2	19,9	21,9	16,0	16,9	15,3	21,4	20,4
12,8	20,8	14,3	18,0	15,1	23,8	18,5	14,4	14,4	21,0

1) Construct a continuous statistical series.

2) Plot a histogram.

3) Calculate numerical characteristics, find M_o and M_e.

PhD Misiura Ie. (доцент Місюра Є.Ю.)

Task 5. For fifty students the number of points (discipline is computer science) is given:

12,	14,	19,	15,	14,	18,	13,	16,	17,	12,
20,	17,	15,	13,	17,	16,	20,	14,	14,	13,
17,	16,	15,	19,	16,	15,	18,	17,	15,	14,
16,	15,	15,	18,	15,	15,	19,	14,	16,	18,
18,	15,	15,	17,	15,	16,	16,	14,	14,	17.

1) Construct a discrete statistic series for this sample, plot a polygon of relative frequencies and $F(x)$; 2) Calculate $\overline{x_s}, S_x, R, v, A_s, E_s$; 3) find M_o and M_e.

Task6. A voltage (volts) in an electricity network for each hour is given with the help of series: 222, 219, 224, 220, 218, 217, 221, 220, 215, 218, 223, 225, 220, 226, 221, 216, 211, 219, 220, 221, 222, 218, 221, 219. 1) Construct a discrete statistic series for this sample, plot a polygon of relative frequencies and $F(x)$; 2) Calculate $\overline{x_s}, S_x, R, v, A_s, E_s$; 3) Find M_o and M_e.

Task 7. The productivity x_i the area of hectares is given:

x_i, centner/hectare	*25*	30	35	40	45
m_i	2	3	8	4	3

1) Plot a polygon of relative frequencies and $F(x)$;

2) Calculate $\overline{x_s}, S_x, R, v, A_s, E_s$;

3) Find M_o and M_e.

Task 8. Plot a polygon, a cumulative function, a histogram of relative frequencies for this series.

$x_i - x_{i+1}$	10–15	15–20	20–25	25–30	30–35
m_i	5	15	20	35	25

Find an empirical distribution function and plot its graph. Find $\overline{x_s}, S_x, R, v, A_s, E_s$, M_o and M_e.

Task 9. Plot a polygon, a cumulative function, a histogram of relative frequencies for this series.

$x_i - x_{i+1}$	0-100	100–200	200–300	300-400	400-500	500-600
m_i	10	15	20	25	18	12

Find an empirical distribution function and plot its graph. Find $\overline{x_s}, S_x, R, v, A_s, E_s$, M_o and M_e.

Task 10. Plot a polygon, a cumulative function, a histogram of relative frequencies for this series.

$x_i - x_{i+1}$	2-7	7-12	12-17	17-22	22-27
m_i	5	10	25	6	4

Find an empirical distribution function and plot its graph. Find $\overline{x_s}, S_x, R, v, A_s, E_s, M_o$ and M_e.

Task 11. Plot a polygon, a cumulative function, a histogram of relative frequencies for this series.

$x_i - x_{i+1}$	3-5	5-7	7-9	9-11	11-13
m_i	5	20	35	30	10

Find an empirical distribution function and plot its graph. Find $\overline{x_s}, S_x, R, v, A_s, E_s, M_o$ and M_e.

Task 12. Plot a polygon, a cumulative function, a histogram of relative frequencies for this series.

$x_i - x_{i+1}$	10-20	20-30	30-40	40-50	50-60
m_i	5	25	30	25	15

Find an empirical distribution function and plot its graph. Find $\overline{x_s}, S_x, R, v, A_s, E_s, M_o$ and M_e.

Task 13. Plot a polygon, a cumulative function, a histogram of relative frequencies for this series.

$x_i - x_{i+1}$	190-200	200–210	210–220	220-230	230-240	240-250
m_i	10	26	56	64	30	14

Find an empirical distribution function and plot its graph. Find $\overline{x_s}, S_x, R, v, A_s, E_s, M_o$ and M_e.

Task 14. A continuous series of X is given. 1) Find an empirical distribution function and plot its graph. 2) Find $\overline{x_s}, S_x, R, v, A_s, E_s, M_o$ and M_e. 3) Calculate the initial v_k and central μ_k moments *k*-th order (*k*=1,2,3,4).

a) X is the number of agreements, $n = 400$ (investors)

$x_i - x_{i+1}$	0	1	2	3	4	5	6	7	8	9	10
m_i	146	97	73	34	23	10	6	3	4	2	2

b) X is the profit (units of money), the number of companies is $n = 1000$

$x_i - x_{i+1}$	0-500	500-1000	1000-1500	1500-2000	2000-2500	2500-3000
m_i	58	96	239	328	147	132

c) X is the productivity increment (in %), the number of companies is $n = 100$

$x_i - x_{i+1}$	4-6	6-8	8-10	10-12	12-14	14-16	16-18	18-20	20-22	22-24	24-26
m_i	1	3	6	11	15	20	14	12	10	6	2

d) Loading of a telephone line between two points from 9 till 16 hours (one day) is characterized with the help of table:

Hours	9–10	10–11	11–12	12–13	13–14	14–15	15–16
Number of calls	16	21	18	10	12	15	13

e) Results of measuring of a height of 100 students are given:

Height	158–162	162–166	166–170	170–174	174–178	178–182	182–186
Number of students	4	12	18	24	19	13	10

Task 15. Number of details, produced by 100 workers, is given:

338	336	312	322	381	302	296	360	342	334	348	304	323
310	368	314	298	312	322	350	304	302	336	334	304	292
324	331	324	334	314	338	324	292	298	342	338	331	325
324	326	314	312	362	368	321	352	304	302	332	314	304
312	381	290	322	326	316	328	340	324	320	364	304	340
330	314	342	322	362	298	316	298	332	342	316	326	308
321	302	304	322	296	322	338	324	323	298	312	322	350
290	318	332	354	324	332	304	282	304	321	356	366	328

1) Construct a continuous statistical series.

2) Construct a discrete statistical series.

3) Plot a polygon, a histogram and and a cumulative function.

4) Calculate $\overline{x_s}, S_x, R, v$ and A_s, E_s, M_o, M_e.