Лекція 13. Перевірка правильності непараметричних гіпотез про довільний розподіл генеральної сукупності

– критерій Колмогорова;

– критерій Мізеса $\left(\omega^2\right)$;

13.2. Критерій згоди А. Н. Колмогорова

Нехай $F(x)$ - теоретична функція розподілу імовірностей неперервної випадкової величини X, що є кількісною ознакою об'єктів генеральної сукупності. При цьому згідно із гіпотезою H_0 вид цієї функції відомий. Нехай $F^*(x)$ — емпірична функція розподілу, отримана внаслідок опрацювання вибірки обсягом n. Порівняємо емпіричну функцію розподілу $F^*(x)$ із гіпотетичною $F(x)$ і якщо міра розходженості між ними мала, то вважати правильною гіпотезу H_0. Найбільш природною і простою із таких мір є рівномірна віддаль

$$D_n = \max_{-\infty<x<+\infty}\left|F^*(x)-F(x)\right|. \tag{13.1}$$

Оскільки вибірка утворюється випадковим чином, то D_n є випадковою, при цьому величина $D_n\sqrt{n}$ має граничний при $n\to\infty$ розподіл, обчислений в припущенні, що гіпотеза H_0 є правильною. При умові неперервності $F^*(x)$ цей розподіл має такий вид (таблиця значень функцій подана у конспекті цієї лекції):

$$P\left(D_n<\lambda\right)=P\left(D\sqrt{n}<\lambda\right)\underset{n\to\infty}{\to}K(\lambda)\begin{cases}0, & при \quad \lambda\le 0;\\ \sum\limits_{k=-\infty}^{\infty}(-1)^k\cdot e^{-2k^2\lambda^2}, & при \quad \lambda>0.\end{cases} \tag{13.2}$$

За умови, що теоретична функція розподілу $F(x)$ неперервна.

Функція розподілу імовірностей $K(\lambda)$ називається *функцією Колмогорова.*

Задаючи рівень значущості α, з співвідношення

$$P(\lambda > \lambda_\alpha) = P\left(D\sqrt{n} \ge \lambda_\alpha\right) = 1 - \sum_{k=-\infty}^{\infty} (-1)^k \cdot e^{-2k^2\lambda_\alpha^2} = \alpha,$$

можна знайти критичні значення розподілу Колмогорова за дод. 9.

Критерій Колмогорова приписує прийняти гіпотезу H_0, якщо $D_n < K_{n,\alpha}$ і відхиляти у випадку виконання нерівності $D_n(x) \ge K_{n,\alpha}$, де α — рівень значущості, $K_{n,\alpha}$ — критичне значення критерію, яке знаходиться за дод. 11 (із файлу додатків). Відмітимо, що в деяких посібниках і довідниках наводяться таблиці критичних значень для статистики $D_n\sqrt{n}$ **(дод. 9)**.

Для обсягів вибірки $n > 100$ з урахуванням співвідношення (13.2) випливає асимптотичне співвідношення

$$K_{n,\alpha} \approx \frac{k_{1-\alpha}}{\sqrt{n}}, \tag{13.3}$$

де $K(k_{1-\alpha}) \approx 1 - \alpha$. Значення $k_{1-\alpha}$ знаходяться за дод. 9 (із файлу додатків).

Нехай статистичний розподіл вибірки задається таблицею:

x_i	x_1	x_2	$\cdots$	x_{k-1}	x_k
n_i	n_1	n_2	$\cdots$	n_{k-1}	n_k

де $\sum_{i=1}^{k} n_i = n$, $x_1 < x_2 < \ldots < x_k$. Тоді емпірична функція розподілу $F^*(x)$ має наступний вигляд:

$$F^*(x) = \begin{cases} 0, & \textit{якщо} \quad x \le x_1; \\ \dfrac{n_1}{n}, & \textit{якщо} \quad x_1 < x \le x_2; \\ \dfrac{n_1 + n_2}{n}, & \textit{якщо} \quad x_2 < x \le x_3; \\ \quad\cdots, & \textit{якщо} \quad\quad \cdots \quad; \\ \dfrac{n_1 + n_2 + \ldots + n_{k-1}}{n}, & \textit{якщо} \quad x_{k-1} < x \le x_k; \\ 1, & \textit{якщо} \quad x > x_k. \end{cases} \tag{13.4}$$

При знаходженні значення статистики D_n, що визначається за формулою (13.1), потрібно враховувати сходинковий характер графіка функції $F^*(x)$. Найбільша за абсолютною величиною різниця між $F^*(x)$ і $F(x)$ буде досягатися в одній із точок розподілу (рис. 13.1).

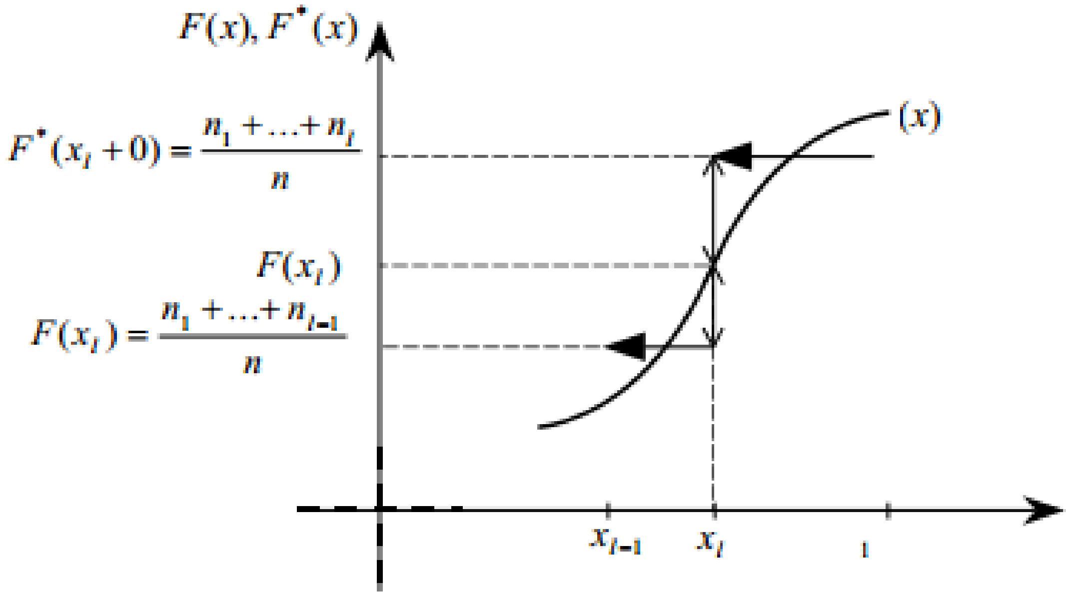

Рис. 13.1.

Тобто

$$D_n = \max\left\{D_n^{(1)}; D_n^{(2)}\right\}, \tag{13.5}$$

де

$$\begin{aligned} D_n^{(1)} &= \max_{i=\overline{1,k}}\left|\frac{n_1+n_2+\ldots+n_i}{n} - F(x_i)\right| = \max_{i=\overline{1,k}}\left|F^*(x_i+0) - F(x_i)\right|, \\ D_n^{(2)} &= \max_{i=\overline{1,k}}\left|F(x_i) - \frac{n_1+n_2+\ldots+n_{i-1}}{n}\right| = \max_{i=\overline{1,k}}\left|F(x_i) - F^*(x_i)\right|. \end{aligned} \tag{13.6}$$

Якщо у статистичному розподілі вибірки ***всі частоти дорівнюють одиниці***, тоді D_n знаходиться з такого співвідношення

$$D_n = \max_{1\le i\le n}\left|F(x_i) - \frac{2\cdot i-1}{2\cdot n}\right| + \frac{1}{2\cdot n}. \tag{13.7}$$

Формули (13.5), (13.6) або (13.7) дають практичну реалізацію двостороннього критерія Колмогорова для випадку ***дискретного розподілу частот або відносних частот вибірки***.

Розглянемо ***інтервальний статистичний розподіл вибірки***:

$[x_i; x_{i+1})$	$[x_1; x_2)$	$(x_2; x_3)$	$\dots$	$(x_{k-1}; x_k]$
n_i	n_1	n_2	$\dots$	n_k

Ескізи графіків емпіричної та теоретичної функцій розподілу в цьому випадку (рис. 13.2) вказують на суттєву відмінність при знаходженні відхилень $\left|F^*(x) - F(x)\right|$ в порівнянні із дискретним розподілом (рис. 13.1).

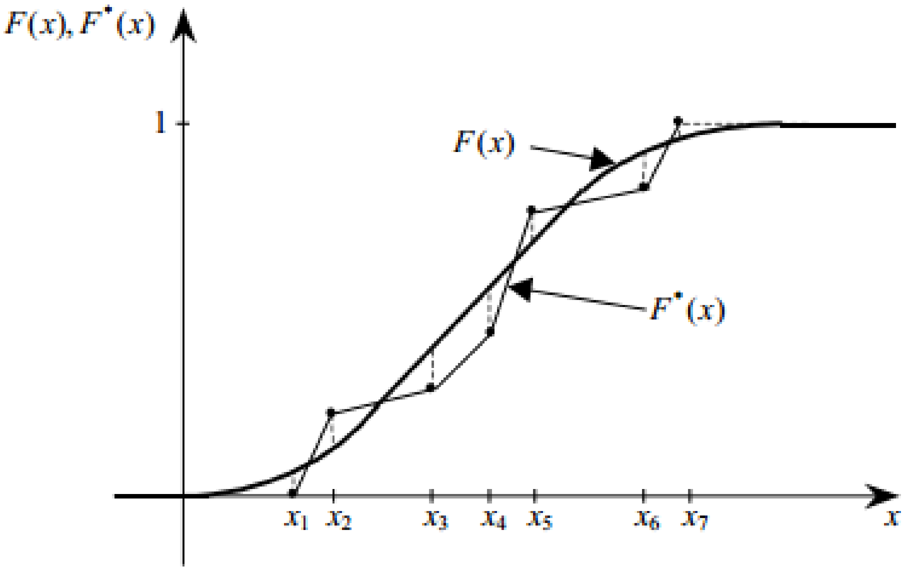

Рис. 13.2.

В цьому випадку слід вибирати односторонню критичну область. Тоді статистика критерію Колмогорова задається формулою

$$D_n^+ = \max_{1 \le i \le k} \left|F^*(x) - F(x)\right|. \tag{13.8}$$

У випадку використання одностороннього критерія на рівні значущості гіпотеза H_0, відкидається, якщо $D_n^+ \ge K_{n;\,\alpha}^+$, де $K_{n;\,\alpha}^+$ — критичне значення. При $n \to \infty$ має місце асимптотичне співвідношення

$$K_{n;\,\alpha}^+ \approx \sqrt{\frac{-\ln\alpha}{2n}} \tag{13.9}$$

Додаток (лекція)

Таблиця значень функції

$$P(\lambda) = 1 - K(\lambda) = P(D \ge \lambda) = 1 - \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k\lambda}$$

λ	P(λ)	λ	P(λ)	λ	P(λ)	λ	P(λ)	λ	P(λ)
<0,29	1,00000	0,76	0,6104	1,23	0,0970	1,70	0,0062	2,17	0,0002
0,30	0,99999	0,77	0,5936	1,24	0,0924	1,71	0,0058	2,18	0,0001
0,31	0,99998	0,78	0,5770	1,25	0,0879	1,72	0,0054	2,19	0,0001
0,32	0,99995	0,79	0,5606	1,26	0,0836	1,73	0,0050	2,20	0,0001
0,33	0,99991	0,80	0,5441	1,27	0,0794	1,74	0,0047	2,21	0,0001
0,34	0,99983	0,81	0,5280	1,28	0,0755	1,75	0,0044	2,22	0,0001
0,35	0,9997	0,82	0,5120	1,29	0,0717	1,76	0,0041	2,23	0,0001
0,36	0,9995	0,83	0,4962	1,30	0,0681	1,77	0,0038	2,24	0,0001
0,37	0,9992	0,84	0,4806	1,31	0,0646	1,78	0,0035	2,25	0,0001
0,38	0,9987	0,85	0,4653	1,32	0,0613	1,79	0,0033	2,26	0,0001
0,39	0,9981	0,86	0,4503	1,33	0,0582	1,80	0,0031	2,27	0,0001
0,40	0,9972	0,87	0,4355	1,34	0,0551	1,81	0,0029	2,28	0,0001
0,41	0,9960	0,88	0,4209	1,35	0,0522	1,82	0,0027	2,29	0,0001
0,42	0,9945	0,89	0,4067	1,36	0,0495	1,83	0,0025	2,30	0,0001
0,43	0,9926	0,90	0,3927	1,37	0,0469	1,84	0,0023	2,31	0,000046
0,44	0,9903	0,91	0,3791	1,38	0,0444	1,85	0,0021	2,32	0,000042
0,45	0,9874	0,92	0,3657	1,39	0,0420	1,86	0,0020	2,33	0,000038
0,46	0,9840	0,93	0,3527	1,40	0,0397	1,87	0,0019	2,34	0,000035
0,47	0,9800	0,94	0,3399	1,41	0,0375	1,88	0,0017	2,35	0,000032
0,48	0,9753	0,95	0,3275	1,42	0,0354	1,89	0,0016	2,36	0,000030
0,49	0,9700	0,96	0,3154	1,43	0,0335	1,90	0,0015	2,37	0,000027
0,50	0,9639	0,97	0,3036	1,44	0,0316	1,91	0,0014	2,38	0,000024
0,51	0,9572	0,98	0,2921	1,45	0,0298	1,92	0,0013	2,39	0,000022
0,52	0,9497	0,99	0,2809	1,46	0,0282	1,93	0,0012	2,40	0,000020
0,53	0,9415	1,00	0,2700	1,47	0,0266	1,94	0,0011	2,41	0,000018
0,54	0,9325	1,01	0,2594	1,48	0,0250	1,95	0,0010	2,42	0,000016
0,55	0,9228	1,02	0,2492	1,49	0,0236	1,96	0,0009	2,43	0,000014
0,56	0,9124	1,03	0,2392	1,50	0,0222	1,97	0,0009	2,44	0,000013
0,57	0,9019	1,04	0,2296	1,51	0,0209	1,98	0,0008	2,45	0,000012
0,58	0,8896	1,05	0,2202	1,52	0,0197	1,99	0,0007	2,46	0,000011
0,59	0,8772	1,06	0,2111	1,53	0,0185	2,00	0,0007	2,47	0,000010
0,60	0,8643	1,07	0,2024	1,54	0,0174	2,01	0,0006	2,48	0,000009
0,61	0,8508	1,08	0,1939	1,55	0,0164	2,02	0,0006	2,49	0,000008
0,62	0,8368	1,09	0,1857	1,56	0,0154	2,03	0,0005	2,50	0,0000075
0,63	0,8222	1,10	0,1777	1,57	0,0145	2,04	0,0005	2,55	0,0000044
0,64	0,8073	1,11	0,1700	1,58	0,0136	2,05	0,0004	2,60	0,0000026
0,65	0,7920	1,12	0,1626	1,59	0,0127	2,06	0,0004	2,65	0,0000016
0,66	0,7764	1,13	0,1555	1,60	0,0120	2,07	0,0004	2,70	0,0000010
0,67	0,7604	1,14	0,1486	1,61	0,0112	2,08	0,0004	2,75	0,0000006
0,68	0,7442	1,15	0,1420	1,62	0,0105	2,09	0,0003	2,80	0,0000003
0,69	0,7278	1,16	0,1356	1,63	0,0098	2,10	0,0003	2,85	0,00000018
0,70	0,7112	1,17	0,1294	1,64	0,0092	2,11	0,0003	2,90	0,00000010
0,71	0,6945	1,18	0,1235	1,65	0,0086	2,12	0,0002	2,95	0,00000006
0,72	0,6777	1,19	0,1177	1,66	0,0081	2,13	0,0002	3,00	0,00000003
0,73	0,6609	1,20	0,1122	1,67	0,0076	2,14	0,0002		
0,74	0,6440	1,21	0,1070	1,68	0,0071	2,15	0,0002		
0,75	0,6272	1,22	0,1019	1,69	0,0066	2,16	0,0002		

Приклад 13.1. На основі вибірки: 0,6917; 0,1794; 0,7410; 0,3094; 0,1174; 0,5424; 0,0834; 0,6288; 0,9401; 0,6606 для рівня значущості $\alpha = 0,05$ за допомогою критерія Колмогорова перевірити гіпотезу H_0 про те, що дана генеральна сукупність на проміжку [0; 1] має рівномірний розподіл, тобто $H_0 : F(x) = x$, $(0 \le x \le 1)$.

Розв'язання:

Таблиця 13.1

l	x_l	$F(x_l)$	$\frac{2l-1}{2n}$	$F(x_l) - \frac{2l-1}{2n}$	$\left\lvert F(x_l) - \frac{2l-1}{2n} \right\rvert$
1	0,0834	0,0834	0,05	0,0334	0,0334
2	0,1174	0,1174	0,15	-0,0326	0,0326
3	0,1794	0,1794	0,25	-0,0706	0,0706
4	0,3094	0,3094	0,35	-0,0406	0,0406
5	0,5424	0,5424	0,45	0,0924	0,0924
6	0,6288	0,6288	0,55	0,0788	0,0788
7	0,6606	0,6606	0,65	0,0106	0,0106
8	0,6917	0,6917	0,75	-0,0583	0,0583
9	0,7410	0,7410	0,85	-0,1090	**0,1090**
10	0,9401	0,9401	0,95	-0,0099	0,0099

$D_{10} = 0,1090 + 0,05 = 0,1590$. За дод. 11 при рівні значущості $\alpha = 0,05$ і $n = 10$ знаходимо критичне $K_{10,0;\,0,05} = 0,409$. Оскільки $D_{10} = 0,1590 < 0,409 = K_{10,0;\,0,05}$, то робимо висновок, що результати спостережень не суперечать гіпотезі H_0.

Приклад 13.2. За останні 10 місяців деякою фірмою були укладені угоди на суми l_i, які відбулися n_i раз:

l_i	40,26	40,28	40,30	40,32	40,34	40,36	40,38	40,40	40,42	40,44
n_i	1	4	6	11	15	16	12	7	5	3

Перевірити за допомогою критерію Колмагорова гіпотезу про те, що дана вибірка добута з генеральної сукупності, що рівномірно розподілена в інтервалі $\left[40,24-40,44\right]$.

Розв'язання: Гіпотеза H_0 - статистичний розподіл є рівномірним.

Знайдемо емпіричну функцію розподілу згідно формули:

$$F^*(x)=\frac{n_x}{n},$$

де n_x - накопичені частоти варіант; n - обсяг вибірки і дорівнює 80.

$$F^*(x_1)=\frac{0}{80}=0;$$

$$F^*(x_2)=\frac{0+1}{80}=0,0125;$$

$$F^*(x_3)=\frac{0+1+4}{80}=0,0625;$$

$$F^*(x_4)=\frac{0+1+4+6}{80}=0,1375;$$

$$F^*(x_5)=\frac{0+1+4+6+11}{80}=0,275;$$

$$F^*(x_6)=\frac{0+1+4+6+11+15}{80}=0,4625;$$

$$F^*(x_7)=\frac{0+1+4+6+11+15+16}{80}=0,6625;$$

$$F^*(x_8)=\frac{0+1+4+6+11+15+16+12}{80}=0,8125;$$

$$F^*(x_9)=\frac{0+1+4+6+11+15+16+12+7}{80}=0,9;$$

$$F^*(x_{10})=\frac{0+1+4+6+11+15+16+12+7+5}{80}=0,9625;$$

$$F^*(x_{11})=\frac{0+1+4+6+11+15+16+12+7+5+3}{80}=1.$$

Теоретичну функцію шукаємо за інтегральною функцією рівномірного розподілу:

$$F(x)=\frac{x-a}{b-a},\text{ якщо } x\in[a;b],$$

де $a=40,24$ та $b=40,44$.

Отже $F(x)=\frac{x-40,24}{40,44-40,24}$; $F(x)=\frac{x-40,24}{0,2}$ - це буде пряма.

При $x=40,24$, $F(x_1)=\frac{40,24-40,24}{0,2}=0$;

при $x=40,26$, $F(x_2)=\frac{40,26-40,24}{0,2}=0,1$;

при $x=40,28$, $F(x_3)=\frac{40,28-40,24}{0,2}=0,2$;

при $x=40,30$, $F(x_4)=\frac{40,30-40,24}{0,2}=0,3$;

при $x=40,32$, $F(x_5)=\frac{40,32-40,24}{0,2}=0,4$;

при $x=40,34$, $F(x_6)=\frac{40,34-40,24}{0,2}=0,5$;

при $x=40,36$, $F(x_7)=\frac{40,36-40,24}{0,2}=0,6$;

при $x=40,38$, $F(x_8)=\frac{40,38-40,24}{0,2}=0,7$;

при $x=40,40$, $F(x_9)=\frac{40,40-40,24}{0,2}=0,8$;

при $x=40,42$, $F(x_{10})=\frac{40,42-40,24}{0,2}=0,9$;

при $x=40,44$, $F(x_{11})=\frac{40,44-40,24}{0,2}=1.$

Занесемо наші обчислення до таблиці:

i	x_i	n_i	n_x	$F^*(x_i)$	$F(x_i)$	$\left\|F^*(x_i)-F(x_i)\right\|$
1	40,24	0	0	0	0	0
2	40,26	1	1	0,0125	0,1	0,0875
3	40,28	4	5	0,0625	0,2	0,1375
4	40,30	6	11	0,1375	0,3	**0,1625**
5	40,32	11	22	0,276	0,4	0,125
6	40,34	15	37	0,4625	0,5	0,0375
7	40,36	16	53	0,6625	0,6	0,0625
8	40,38	12	65	0,8125	0,7	0,1125
9	40,40	7	72	0,9	0,8	0,1
10	40,42	5	77	0,9625	0,9	0,0625
11	40,44	3	80	1	1,0	0

Побудуємо графіки $F^*(x)$ та $F(x)$.

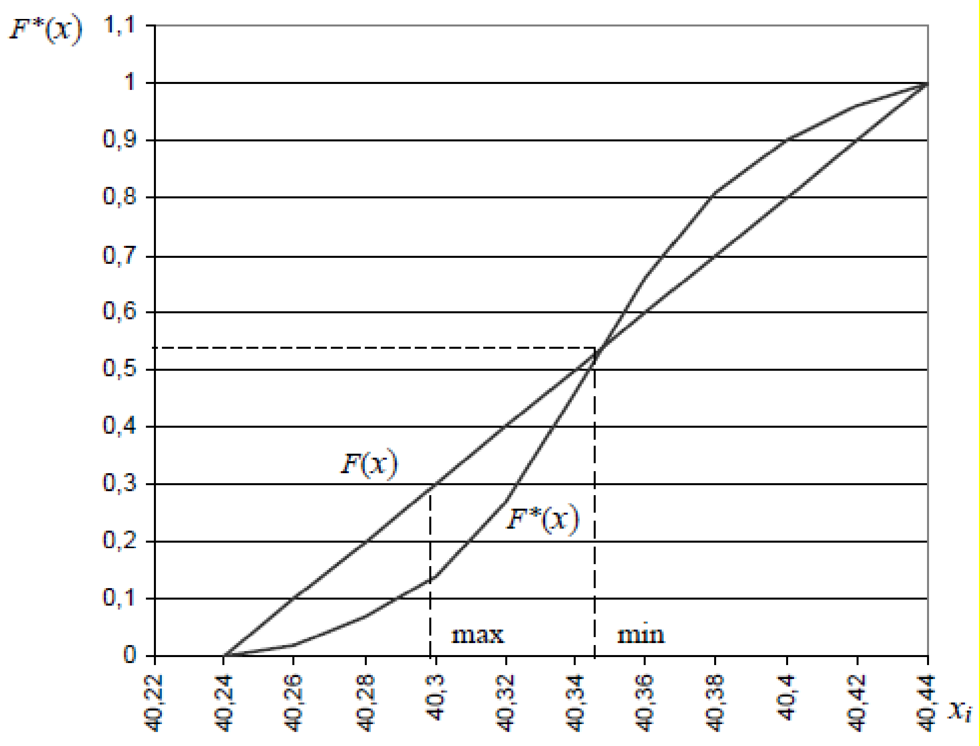

Рис. 13.3.

$D^* \max\left|F^*(x) - F(x)\right| = 0{,}1625$.

$\lambda_{спост} = D^*\sqrt{n} = 0{,}1625 \cdot \sqrt{80} = 1{,}453$.

а) згідно дод. 9 критичних значень розподілу Колмогорова для заданого рівня значущості (ймовірності) $\alpha = 0{,}05$ знаходимо $\lambda_{0,05} = 1{,}358$. Оскільки $\lambda_{спост} = 1{,}453 > \lambda_{крит} = 1{,}358$, то нульова гіпотеза про рівномірний закон розподілу генеральної сукупності відкидається.

б) За табл. лекції знаходимо, що цьому значенню відповідає ймовірність β що дорівнює $P(\lambda \geq 1{,}45) \approx 0{,}0298$.

Дана ймовірність β є мала і оскільки $0{,}0298 < \alpha = 0{,}05$ то розбіжність між емпіричним і теоретичним розподілами не є випадковою. Тому генеральна сукупність не розподілена за рівномірним законом.

в) $D^* = 0{,}1625$, за дод. 11 або табл. 20 $K_{n;\,\alpha} = K_{80;\,0,05} = 0{,}15$. Оскільки $D^* = 0{,}1625 > K_{80;\,0,05} = 0{,}15$, то нульова гіпотеза відкидається.

13.2. Критерій згоди омега квадрат $\left(\omega^2\right)$ Смірнова–Крамера-фон Мізеса

Як критерій перевірки нульової гіпотези про вигляд розподілу використовується випадкова величина

$$\omega^2 = \int_{-\infty}^{\infty} \left(F^*(x) - F(x)\right)^2 f(x)dx,$$

де $F^*(x)$ - емпірична функція розподілу, $F(x)$ - теоретична функцію розподілу, $f(x)$ - теоретична функція щільності ймовірності. Враховуючи, що $F^*(x_i) = \dfrac{n_x}{n}$ та інтегруючи, можна отримати граничне співвідношення наступного виду

$$n\omega^2 = F_{спост} = \frac{1}{12n} + \sum_{i=1}^{n}\left(F(x_i) - \frac{2i-1}{2n}\right)^2,$$

для $F\left(x_i\right)$ - теоретична функція статистичного розподілу; x_i - випадкова величина з вибірки; n - обсяг вибірки.

Більш точний критерій обчислюється згідно формули:

$$\left(n\omega^2\right)' = F_{спост} = \left[n\omega^2 - \frac{0,4}{n} + \frac{0,6}{n^2}\right] \cdot \left(1 + \frac{1}{n}\right),$$

Критичне значення можна отримати скориставшись таблицею:

α	0,5	0,4	0,3	0,2	0,1	0,05	0,01	0,001
$F_{кр}$	0,1184	0,1467	0,1843	0,2414	0,3473	0,4614	0,7435	1,168

де α - рівень значущості. Якщо $F_{спост} < F_{кр}$, то нульова гіпотеза про прийнятий закон розподілу приймається, інакше відхиляється.

Наприклад 13.3. Візьмемо приклад 13.1 та за допомогою критерію згоди Мізеса визначимо чи вірна гіпотеза про рівномірний закон розподілу генеральної сукупності на основі нашої вибірки.

i	$\left\vert F\left(x_i\right) - \frac{2i-1}{2n}\right\vert$	$\left(F\left(x_i\right) - \frac{2i-1}{2n}\right)^2$
1	0,0334	0,00111556
2	0,0326	0,00106276
3	0,0706	0,00498436
4	0,0406	0,00164836
5	0,0924	0,00853776
6	0,0788	0,00620944
7	0,0106	0,00011236
8	0,0583	0,00339889
9	0,1090	0,011881
10	0,0099	0,00009801
	Сума:	0,04005974

$$n\omega^2 = F_{спост} = \frac{1}{12n} + \sum_{i=1}^{n}\left(F\left(x_i\right) - \frac{2i-1}{2n}\right)^2 = \frac{1}{12\cdot 10} + 0,04005974 = 0,04839;$$

$$\left(n\omega^2\right)' = F_{спост} = \left[0,04839 - \frac{0,4}{10} + \frac{0,6}{10^2}\right]\cdot\left(1 + \frac{1}{10}\right) = 0,0543.$$

Згідно таблиці знайдемо критичне значення: $F_{кр} = 0,4614$.

Оскільки, $F_{спост} < F_{кр}$, то гіпотеза про рівномірний закон розподілу генеральної сукупності на основі нашої вибірки приймається.

Наприклад 13.4. Візьмемо приклад 13.2 та за допомогою критерію згоди Мізеса визначимо чи вірна гіпотеза про рівномірний закон розподілу генеральної сукупності на основі нашої вибірки.

i	x_i	$F^*\left(x_i\right)$	$F\left(x_i\right)$	$\frac{2i-1}{2n}$	$F\left(x_i\right) - \frac{2i-1}{2n}$	$\left(F\left(x_i\right) - \frac{2i-1}{2n}\right)^2$
1	40,24	0	0	0,0063	-0,0063	$0,3969\cdot 10^{-4}$
2	40,26	0,0125	0,1	0,0187	0,0813	0,0066
3	40,28	0,0625	0,2	0,0312	0,1688	0,0285
4	40,30	0,1375	0,3	0,0437	0,2562	0,0657
5	40,32	0,276	0,4	0,0563	0,3448	0,1182
6	40,34	0,4625	0,5	0,0688	0,4313	0,1860
7	40,36	0,6625	0,6	0,0813	0,5187	0,2691
8	40,38	0,8125	0,7	0,0938	0,6062	0,3675
9	40,40	0,9	0,8	0,1062	0,6938	0,4813
10	40,42	0,9625	0,9	0,1187	0,7812	0,6104
11	40,44	1	1,0	0,1313	0,8688	0,7547
Сума:						2,8879

$$n\omega^2 = F_{спост} = \frac{1}{12n} + \sum_{i=1}^{n}\left(F\left(x_i\right) - \frac{2i-1}{2n}\right)^2 = \frac{1}{12\cdot 80} + 2,8879 = 2,907;$$

$$\left(n\omega^2\right)' = F_{спост} = \left[2,8879 - \frac{0,4}{80} + \frac{0,6}{80^2}\right] \cdot \left(1 + \frac{1}{80}\right) = 2,9201.$$

Згідно таблиці знайдемо критичне значення: $F_{кр} = 0,4614$.

Оскільки, $F_{спост} > F_{кр}$, то гіпотеза про рівномірний закон розподілу генеральної сукупності на основі нашої вибірки відхиляється.