

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ДЕРЖАВНИЙ ВИЩИЙ НАВЧАЛЬНИЙ ЗАКЛАД
«ДОНЕЦЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ»
КАФЕДРА ПРИКЛАДНОЇ МАТЕМАТИКИ І ІНФОРМАТИКИ



МЕТОДИЧНІ ВКАЗІВКИ І ЗАВДАННЯ
до виконання лабораторних робіт
з дисципліни «ЕМПІРИЧНІ МЕТОДИ ПРОГРАМНОЇ ІНЖЕНЕРІЇ»
для студентів спеціальності
121 «Інженерія програмного забезпечення» всіх форм навчання

Луцьк -2022

УДК 681.3.06(071)

Методичні вказівки і завдання до виконання лабораторних робіт з дисципліни “Емпіричні методи програмної інженерії” для студентів спеціальності 121 «Інженерія програмного забезпечення» всіх форм навчання / укладач: І.А. Назарова. – Покровськ: ДВНЗ «ДонНТУ», 2022. – 47с.

Методичні вказівки містять вимоги і рекомендації до виконання лабораторних робіт, зокрема, методи обробки одновимірної вибірки, застосування критеріїв узгодження та методів вирівнювання статистичних рядів розподілу.

Укладач: Назарова І.А., к.т.н., доцент кафедри прикладної математики та інформатики

Рецензент: Ковальов С.О., к.т.н., доцент кафедри комп’ютерної інженерії

Відповідальний за випуск: Маслова Н.О., доцент, к.т.н., завідувач кафедри "Прикладна математика і інформатика"

Розглянуто на засіданні кафедри прикладної математики і інформатики, протокол №8 від 1.09.2022р.

Затверджено навчально-методичним відділом ДВНЗ «ДонНТУ», протокол № від . .2022р.

© І.А.Назарова, 2022

ЗМІСТ

Вступ	4
Лабораторна робота №1	5
Дослідження способів обробки одновимірної вибірки	
Лабораторна робота №2	14
Визначення точкових статистичних оцінок параметрів розподілу за одновимірною вибіркою	
Лабораторна робота №3	25
Визначення інтервальних статистичних оцінок параметрів розподілу за одновимірною вибіркою	
Лабораторна робота №4	33
Згладжування (вирівнювання) статистичних рядів	
Лабораторна робота №5	42
Побудова теоретичного закону розподілу за експериментальними даними	
Рекомендована література	47

ВСТУП

Методичні вказівки і завдання до лабораторних робіт за курсом “Емпіричні методи програмної інженерії” призначаються для бакалаврів, що навчаються за спеціальністю 121 Інженерія програмного забезпечення денної та заочної форм навчання.

Методичні вказівки містять завдання на лабораторні роботи, індивідуальні варіанти та рекомендації до виконання лабораторних робіт за базовими темами курсу, що направлені на обробку та аналіз одновимірного статистичного матеріалу. Окремі розділи присвячено визначенню точкових та інтервальних оцінок параметрів розподілу, теоретичного закону розподілу, що описує експериментальні дані на основі критеріїв узгодження Пірсона та Колмогорова.

Методичні вказівки та завдання містять у повному обсязі всю необхідну інформацію для ефективної організації роботи студентів в процесі виконання лабораторних робіт за дисципліною «Емпіричні методи програмної інженерії», але не дублюють відповідний лекційний курс. Вони будуть корисними при виконанні та оформленні результатів лабораторного практикуму, тому, що написані українською мовою із використанням загальноприйнятої математичної, а також статистичної термінології.

Лабораторна робота №1

Дослідження способів обробки одновимірної вибірки

Мета роботи: формування уявлень та практичних навичок у первинній статистичній обробці одновимірних вибірок.

Теоретична довідка

Генеральна сукупність (ГС) – сукупність всіх мислимих і реально існуючих спостережень, вироблених за цього реальному комплексі умов. ГС – всі об'єкти, що підлягають дослідженню.

Вибіркова сукупність (ВС) – підмножина об'єктів генеральної сукупності, що безпосередньо беруть участь у дослідженні.

Обсяг генеральної сукупності – кількість об'єктів ГС.

Обсяг вибіркової сукупності – кількість об'єктів ВС.

Способи подання вибірки:

- 1) простий статистичний ряд;
- 2) варіаційний ряд;
- 3) частотно-варіаційний ряд;
- 4) інтервальний чи статистичний ряд, згрупований ряд

1. Простий статистичний ряд – це первинна форма запису статистичного матеріалу, оформлена у вигляді таблиці, де n – обсяг вибірки, i – номер досліду, $i = \overline{1, n}$ та $x_i, i = \overline{1, n}$ – спостережене вибіркове значення (табл. 1.1).

Таблиця 1.1 – Простий статистичний ряд

i	1	2	...	n
x_i	x_1	x_2	...	x_n

Варіанти – елементи вибірки, що різняться числовим значенням.

2. Варіаційний ряд – вибірка, впорядкована у порядку неспадання або зростання:

$$\min_{i=1}^n x_i = x_1^* \leq x_2^* \leq \dots \leq x_n^* = \max_{i=1}^n x_i.$$

3. **Частото - варіаційний ряд** – таблиця (табл. 1.2), в якій задані різні варіанти (x_i) та частота їх з'явлення у виборці (n_i) або відповідна статистична ймовірність (p_i^*).

Таблиця 1.2 – Частотно-варіаційний ряд

x_i	Різні варіанти
n_i	Частота варіанту
p_i^*	Статистична ймовірність

Частота – величина, що показує скільки разів зустрічається варіант з певним значенням у варіаційному ряді.

Частість (відносна частота, статистична ймовірність) – відношення частоти до загальної кількості спостережених значень (обсягу вибірки):

$$p_i^* = n_i / n.$$

4. Угрупування та інтервальний статистичний ряд

Алгоритм побудови:

1. Побудувати варіаційний ряд, знайти мінімальне та максимальне значення в ряді, порахувати розмах (амплітуду) вибірки:

$$R = X_{\max} - X_{\min}.$$

2. Визначити необхідну кількість інтервалів k за формулою Стерджеса (найближче ціле зверху):

$$k = [1 + 3.32 \cdot \log_{10} n].$$

Формула дає осереднене значення кількості інтервалів.

3. Визначити ширину інтервалу $\Delta x = h$ за формулою:

$$h = R/k.$$

Інтервальний ряд (угрупування) представляється у вигляді таблиці 1.3.

Таблиця 1.3 – Інтервальний ряд (угрупкування)

1	2	...	k
$x_{\min}, x_{\min} + h$	$x_{\min} + h, x_{\min} + 2h$	...	$x_{\min} + (k - 1)h, x_{\max}$
m_1	m_2	...	m_k
$p_1^* = \frac{m_1}{n}$	$p_2^* = \frac{m_2}{n}$	...	$p_k^* = \frac{m_k}{n}$

$\sum_{i=1}^k m_i = n$, де m_i – частота попадання до i -го інтервалу.

$\sum_{i=1}^k p_i^* = 1$, де p_i^* – відносна частота попадання до i -го інтервалу.

Якщо деякий інтервал – незначущий, тобто в нього не попав жоден елемент з вибірки, то його об'єднують із сусіднім інтервалом, інтервали стають не однакової ширини.

Графічні способи зображення вибірки

1. **Полігон** (частот чи частостей) може бути побудований за варіаційним, частотно - варіаційним або за інтервальним рядом (рис. 1.1). За **ОХ** відкладаються або вибіркові значення, або варіанти, або інтервали. За **ОУ** – частота або статистична ймовірність.

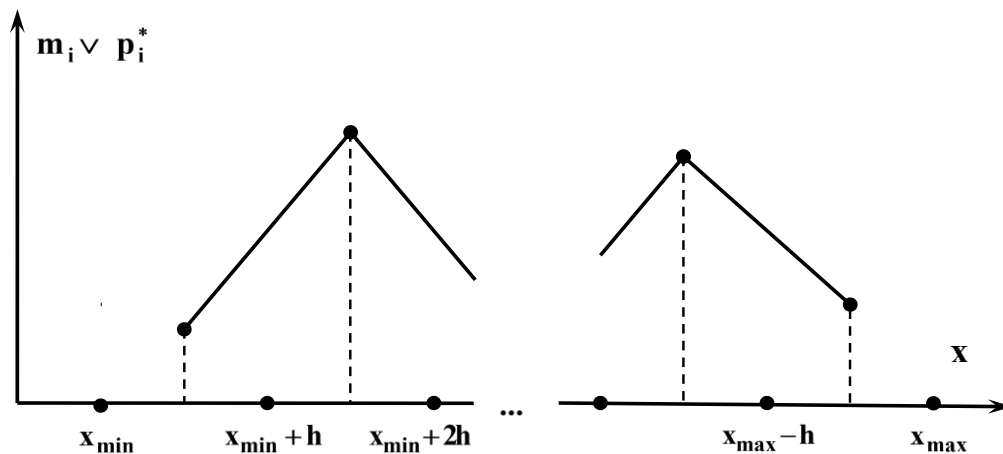


Рисунок 1.1 – Полігон частот/частостей

2. **Гістограма** – це стовпчикова діаграма, емпіричний аналог функції щільності розподілу (рис. 1.2). По осі **OX** – інтервальний ряд, по осі **OY** – величина H_i .

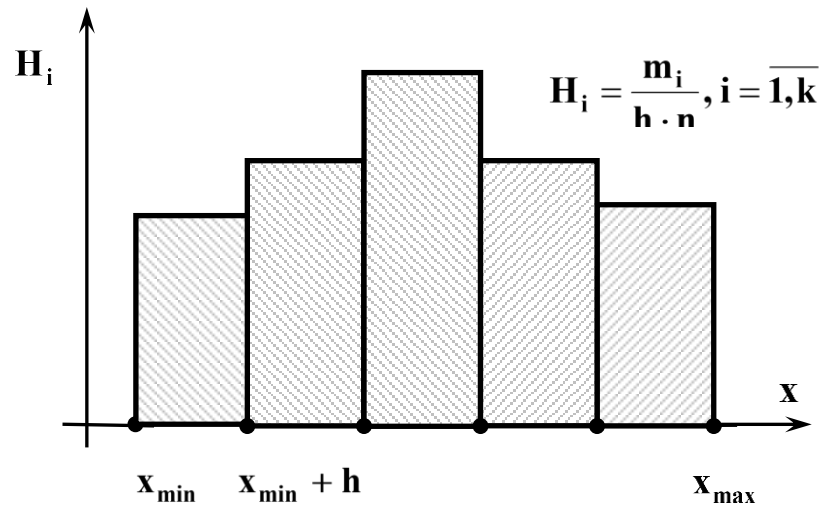


Рисунок 1.2 – Гістограма за інтервальним рядом

3. Статистична (емпірична) функція розподілу $F^*(x)$

Статистичною функцією розподілу називається функція:

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

де n_x – кількість елементів вибірки, менших за x ,

n – обсяг вибірки.

Статистична функція розподілу у даній точці x дорівнює накопленій відносній частоті або статистичній ймовірності того, що ВВ приймає значення менше за x :

$$F^*(x) = \begin{cases} 0, & x \leq x_1 \\ p_1^*, & x_1 < x \leq x_2 \\ p_1^* + p_2^*, & x_2 < x \leq x_3 \\ \dots & \\ p_1^* + p_2^* + \dots + p_n^* = 1, & x > x_n \end{cases}.$$

Графічне зображення емпіричної функції представлено рисунку 1.3. За **ОХ** значення елементів вибіркової сукупності (частотно-варіаційний або інтервальний ряд). За **ОУ** – відносні частоти накопиченим порядком або накопичена статистична ймовірність.

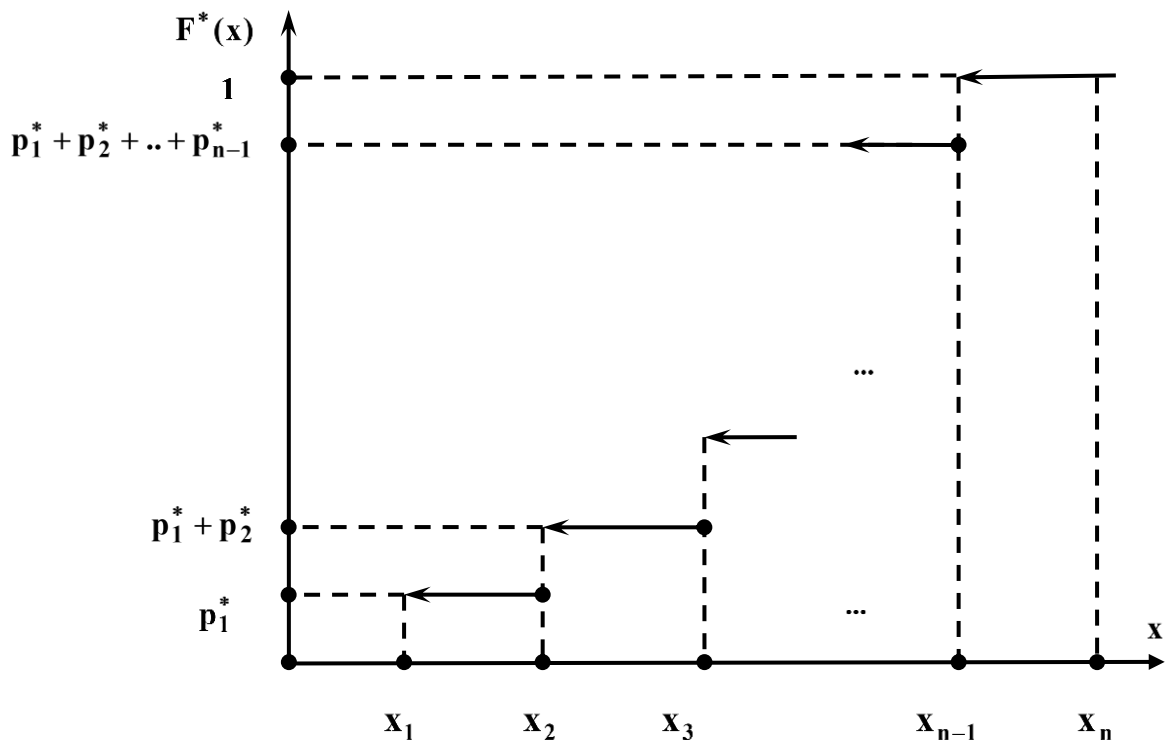


Рисунок 1.3 – Емпірична функція розподілу

Завдання до лабораторної роботи

За заданим варіантом завдання вибрати статистичні дані для первинної обробки.

Провести статистичний аналіз:

- 1) побудувати варіаційний та частотно-варіаційний ряд;
- 2) знайти розмах варіювання;
- 3) визначити кількість груп за формулою Стерджеса;
- 4) побудувати угруповання та інтервальний ряд;
- 5) навести графічне зображення полігону частот, гістограми, емпіричної функції розподілу.

Варіанти завдань

№ 1. Час розв'язання контрольної завдання учнями 4-го класу (хв):

38 60 41 51 33 42 45 21 53 60 68 52 47 46 49
49 14 57 54 59 11 47 28 48 58 32 42 58 61 90

№ 2. Тривалість роботи електронних ламп одного типу в годинах:

13.4 17.7 15.2 15.1 13.0 21.9 14.0 17.9 15.1 16.5 16.6
14.2 16.3 14.6 11.7 16.4 15.1 17.6 14.1 18.8 11.6 13.9
18.0 12.4 17.2 14.5 16.3 13.7 15.5 16.2 8.4 14.7 13.4

№ 3. Вимірювання ємності у 80 польових транзисторів дали такі результати:

1.9 3.1 1.3 0.7 3.2 1.1 2.9 2.7 2.7 4.0
1.7 3.2 0.9 0.8 3.1 1.2 2.6 1.9 2.3 3.2
4.1 1.3 2.4 4.5 2.3 0.9 1.4 1.6 2.2 3.1
1.5 1.1 2.3 4.3 2.1 0.7 1.2 1.5 1.8 2.9

№ 4. Час відновлення діодів з однієї партії (в наносекундах):

69 73 70 68 61 73 70 72 67 70 66 70 71
76 68 71 71 68 70 64 65 72 70 70 69 67
66 70 77 69 71 74 72 72 72 68 70 67 72

№ 5. Час реакції (у секундах):

8.5 7.1 6.7 6.2 2.9 4.4 6.0 5.8 5.4 8.2 6.9 6.5
3.8 6.0 6.0 5.6 5.3 7.7 6.8 6.5 6.1 4.2 4.7 5.6
5.3 7.4 6.7 6.4 6.1 4.5 6.0 5.8 5.8 5.6 6.1 5.4

№ 6. Витрати на харчування 66 сімей (США), кожна з яких складається з 4 осіб (у доларах):

48 44 40 51 44 18 46 57 57 34 38 47 48 52
39 41 39 38 43 29 45 54 38 28 48 28 47 52

№ 7. Показники часу для 24 бігунів на одній дистанції (в секундах):

33,5; 20,1; 23,6; 26,3; 19,9; 16,7; 23,2; 31,4; 28,2; 35,3; 29,3; 30,5; 25;
24,2; 19; 20; 28; 17,9; 24; 25; 20,2; 19,5; 20,5; 23.

№ 8. Дані, що зібрані по сім'ям, дають тижневе споживання хліба на сім'ю з 4 чоловік (в кг):

6, 5, 7, 6.5, 4, 8, 4.5, 6.5, 5.5, 4.8, 9, 8.5, 7.2, 7, 6.5, 5, 5.5, 4, 5, 6.25, 5.1, 4.4,
4.3, 6, 5, 8.5, 4, 9.8.

№9. Результату біохімічного аналізу крові 30 осіб – кількість еритроцитів (10^{12} /л):

3,4; 3,6; 4; 3,5; 3,7; 4,1; 3,4; 4,2; 3,7; 3,4; 4,2; 2,8; 4; 3,5; 3,6; 4; 3,7; 4,3;
3,6; 3,9; 3,6; 3,5; 3,1; 3; 3,6; 3,7; 3,2; 3,6; 3,9; 3,3.

№ 10 Число пасажирів деякого авіарейсу за 30 польотів склало:

128; 121; 134; 118; 123; 109; 120; 116; 125; 128; 121; 129; 130; 131; 127;
119; 114; 124; 110; 126; 134; 125; 128; 123; 128; 133; 132; 136; 134; 129.

№ 11 Дані являють собою результати вимірювання питомого опору мікросхем після легування полікремнію:

2,2; 33; 76; 32,2; 49,5; 32,5; 191,5; 112,5; 32,9; 114,8; 33,7; 69,1; 112,5; 48,5;
16,5; 50; 51; 39; 66; 40; 41.5, 20.4, 67.3, 80.1, 70.

№ 12 Відхилення від номінального розміру, отримані при вимірюванні деталей, мають таке значення (в мм):

0,02; 0,07; 0,13; 0,05; 0,11; 0,17; 0,17; 0,05; 0,03; 0,11; 0,04; 0,14; 0,10; 0,11;
0,13; 0,14; 0,16; 0,07; 0,13; 0,04; 0,03; 0,13; 0,11; 0,06; 0,10; 0,13; 0,16; 0,06;
0,14; 0,06; 0,04; 0,08; 0,14; 0,08; 0,08; 0,14; 0,13; 0,07; 0,16; 0,08; 0,02; 0,10.

№ 13 Дані являють собою врожайність зернових культур (ц / га):

28,0; 21,0; 27,6; 16,2; 29,7; 26,8; 30,3; 15,7; 25,5; 15,8; 40,6; 27,2; 16,1; 19,8;
25,3; 16,9; 31,4; 26,5; 20,4; 15,6; 20,3; 37,4; 23,2; 38,3; 28,0; 29,0; 37,7; 34,7;
30,0; 20,6.

№ 14 Якість ґрунту на 30 різних полях (в балах):

79; 70; 80; 71; 77; 77; 84; 66; 74; 67; 89; 69; 70; 75; 80; 63; 80; 73; 75; 79;
69; 90; 75; 86; 85; 77; 92; 87; 75; 69.

№ 15 Споживання вовни в 25 регіонах (у кг на душу населення):

1,86; 1,99; 4,18; 3,14; 2,98; 1,88; 1,57; 1,6; 1,64; 3,1; 2,36; 2,5; 2,62; 1,54;
2,15; 1,7; 1,9; 2,46; 2,27; 2,24; 2,3; 2,17; 1,8; 1,9; 3,1. 0.916 0.917 0.918
0.919 0.919 0.916 0.917 0.923 0.920 0.916 0.917 0.922 0.915 0.917
0.916 0.912.

Контрольні питання

1. Дати визначення генеральної та вибіркової сукупностей.
2. Дати визначення обсягу ГС та ВС.
3. Способи подання одновимірної вибірки: простий статистичний ряд, варіаційний ряд, частотно - варіаційний ряд, інтервальний або статистичний ряд, згрупований ряд.
4. Графічні способи опису одновимірної вибірки: полігон, гістограма, емпірична функція розподілу.
5. Варіаційний та частотно-варіаційний ряди.
6. Варіант. Частота, частість. Угрупування.
7. Кількість інтервалів. Формула Стерджеса.
8. Інтервальний або статистичний ряд.
9. Ймовірність попадання в інтервал.
10. Гістограма.

11. Полігон частот або частостей.
 12. Емпірична функція розподілу.
- Література [1-3,6-7]

Лабораторна робота № 2

Визначення точкових статистичних оцінок параметрів розподілу за одновимірною вибіркою

Мета роботи: формування уявлень та практичних навичок у визначенні точкових статистичних оцінок параметрів розподілу генеральних сукупностей за статистичними даними для одновимірних вибірок.

Теоретична довідка

Оцінка параметра генеральної сукупності, що отримана за вибірковою сукупністю й така, що виражається одним числом називається **точковою оцінкою**. Розглянемо деяку ВВ X , що описує досліджувану ГС, закон розподілу якої має невідомий параметр a . Необхідно знайти підходящу оцінку параметра a за результатами n незалежних випробувань, у кожному з яких ВВ X прийняла певне значення: $x_1, x_2, \dots, x_n$.

Властивості оцінок параметрів ГС

Довільна оцінка параметра a , що обчислена за вибірковою сукупністю, позначається – a^* . Щоб підкреслити, що ця оцінка параметра a є функцією вибірових значень, введемо в позначення індекс n :

$$a^* = a_n^* = a^*(x_1, x_2, \dots, x_n).$$

a^* – функція значень $x_1, x_2, \dots, x_n$,, тому вона є ВВ (рис. 2.1).

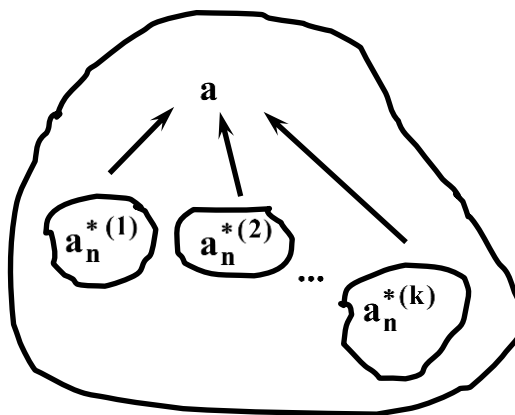


Рисунок 2.1 – Точкові оцінки параметрів ГС

Закон розподілу оцінки параметра може залежати:

- від закону розподілу самої ВВ X ;
- від якісного значення параметра a ;
- від числа вимірів ВВ X (n – обсягу вибірки).

Оцінка параметра ГС є доброякісною, якщо вона задовольняє трьом вимогам:

- обґрунтованість (слухність, змістовність);
- незміщеність;
- ефективність.

Оцінка a_n^* параметра a називається **обґрунтованою**, якщо вона при збільшенні числа дослідів сходиться за ймовірністю до оцінюваного параметру:

$$a_n^* \xrightarrow{n \rightarrow \infty} a,$$
$$\lim_{n \rightarrow \infty} P(|a_n^* - a| \leq \varepsilon) = 1.$$

У випадку обґрунтованості помилок має значення збільшення обсягу вибірок, тобто при цьому стають малоймовірними значні помилки при оцінюванні.

При досить великих n практично достовірно:

$$a_n^* = a^* \approx a.$$

Оцінка a_n^* параметра a називається **незміщеною**, якщо її математичне очікування дорівнює оцінюваному параметру:

$$M[a_n^*] = a,$$

інакше оцінка називається **зміщеною** відносно параметра ГС.

Якщо ця рівність не виконується, то оцінка параметра, отримана за різними вибірками, буде в середньому або завищувати $M[a_n^*] > a$, або занижувати $M[a_n^*] < a$ значення параметра a . Вимога, щоб оцінка параметра

розподілу була незміщеною, гарантує відсутність систематичних помилок при оцінюванні.

Незміщена оцінка a_n^* параметра a називається **ефективною**, якщо вона має мінімальну дисперсію серед всіх можливих незміщених оцінок параметра a , обчислених за вибірками того ж самого обсягу n .

У якості статистичної оцінки параметра ГС бажано мати оцінку, що одночасно задовольняє трьом вимогам:

- незміщеність,
- обґрунтованість,
- ефективність.

Така оцінка позначається, як $\tilde{a}$.

Досягти цього вдається не завжди, тому для спрощення розрахунків використовують або незначно зміщені оцінки, або оцінки з більшою, ніж мінімальна, дисперсією. Щоразу необхідно критичне вивчення характеру поводження оцінок.

Оцінка математичного очікування

Постановка задачі: маємо ВВ X з деяким законом розподілу та параметрами – математичне очікування й дисперсія. Обидва параметри невідомі.

Над випадковою величиною X зроблено n незалежних випробувань. Знайти доброякісну оцінку математичного очікування цієї ВВ.

У якості оцінки математичного очікування приймаємо середнє арифметичне спостережених значень:

$$m^* = \frac{\sum_{i=1}^n x_i}{n}.$$

Потрібно показати, що ця оцінка є: обґрунтованою, незміщеною, ефективною.

1. Обґрунтованість.

За законом великих чисел (теорема Чебишева) при досить великій кількості незалежних випробувань середнє арифметичне спостережених значень ВВ X сходиться за ймовірністю до її математичного очікування.

$$\frac{\sum_{i=1}^n x_i}{n} = m^* \xrightarrow{n \rightarrow \infty} m.$$

2. Незміщеність: $M[m^*] = m$.

$$M[m^*] = M\left[\frac{\sum_{i=1}^n x_i}{n}\right] = \frac{M\left[\sum_{i=1}^n x_i\right]}{n} = \frac{1}{n} \cdot \sum_{i=1}^n M[x_i] = \frac{n \cdot m}{n} = m.$$

3. Ефективність.

$$D[m^*] = \min,$$

$$D[m^*] = D\left[\frac{\sum_{i=1}^n x_i}{n}\right] = \frac{D\left[\sum_{i=1}^n x_i\right]}{n^2} = \frac{1}{n^2} \cdot \sum_{i=1}^n D[x_i] = \frac{n \cdot D}{n^2} = \frac{D}{n}.$$

Ефективність характеристики середнього арифметичного залежить від закону розподілу ВВ X . Дисперсія розглянутої оцінки є міра її точності.

$$D[m^*] = \frac{D}{n}.$$

У практичній статистиці істинне значення дисперсії заміняють на її вибіркове значення для максимального наявного обсягу вибірки.

Для нормального закону розподілу оцінка m^* є ефективною.

Таким чином, оцінка математичного очікування для нормального закону є доброякісною:

$$m^* = \frac{\sum_{i=1}^n x_i}{n} = \tilde{m}.$$

Для інших законів: $m^* \approx \tilde{m}$.

Як оцінку m за частотно-варіаційним рядом приймають середньозважену оцінку:

$$m^* = \sum_{i=1}^k x_i \cdot p_i^* = \sum_{i=1}^k x_i \cdot \frac{n_i}{n},$$

де k – кількість варіантів у ряді.

Для інтервального ряду:

$$m^* = \sum_{i=1}^k x'_i \cdot p_i^* = \sum_{i=1}^k x'_i \cdot \frac{m_i}{n}$$

x'_i – середина i -го інтервалу, k – кількість інтервалів.

Оцінка дисперсії

Постановка задачі: маємо випадкову величину X , що описує генеральну сукупність.

За результатами n незалежних випробувань отримана вибірка значень $x_1, x_2, \dots, x_n$, x_i – екземпляр ВВ із характеристиками m, D . Знайти доброякісну оцінку дисперсії для цієї випадкової величини.

У якості оцінки дисперсії приймається статистичне середнє центрованої випадкової величини:

$$D^* = \frac{1}{n} \cdot \sum_{i=1}^n (x_i - m^*)^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - (m^*)^2.$$

Нагадаємо, що дисперсія це:

$$D[X] = M\left[\left(\overset{\cdot}{X}\right)^2\right] = M[(X - m_X)^2], D[X] = M[X^2] - m_X^2.$$

Визначимо, чи є оцінка D^* – доброякісною.

1. Незміщеність:

$$M[D^*] = D.$$

Використаємо штучний прийом:

$$D^* = \frac{1}{n} \cdot \sum_{i=1}^n (x_i - m^*)^2 = \frac{1}{n} \cdot \sum_{i=1}^n (x_i - m + m - m^*)^2 =$$

$$= \frac{1}{n} \cdot \sum_{i=1}^n ((x_i - m) - (m^* - m))^2.$$

Розкриємо квадрат двочлена й суму:

$$D^* = \frac{1}{n} \cdot \sum_{i=1}^n ((x_i - m)^2 - 2 \cdot (x_i - m) \cdot (m^* - m) + (m^* - m)^2) =$$

$$= \frac{1}{n} \cdot \sum_{i=1}^n (x_i - m)^2 - \frac{2}{n} \cdot (m^* - m) \sum_{i=1}^n (x_i - m) + (m^* - m)^2 =$$

$$= \frac{1}{n} \cdot \sum_{i=1}^n (x_i - m)^2 - 2 \cdot (m^* - m)^2 + (m^* - m)^2 = \frac{1}{n} \cdot \sum_{i=1}^n (x_i - m)^2 - (m^* - m)^2.$$

За визначенням незміщеності:

$$M[D_X^*] = M\left[\frac{1}{n} \cdot \sum_{i=1}^n (x_i - m_X)^2 - (m_X^* - m_X)^2\right] =$$

$$= \frac{1}{n} \cdot \sum_{i=1}^n M[(x_i - m_X)^2] - M[(m_X^* - m_X)^2] =$$

$$= \frac{1}{n} \cdot \sum_{i=1}^n M[(x_i - m_X)^2] - D[m_X^*] =$$

$$= \frac{1}{n} \cdot \sum_{i=1}^n M[(x_i - m_X)^2] - \frac{D_X}{n} =$$

$$= \frac{1}{n} \cdot n \cdot D_X - \frac{D_X}{n} = D_X \cdot \frac{(n-1)}{n}.$$

Отримана оцінка не є незміщеною:

$$M[D_X^*] = D_X \cdot \frac{(n-1)}{n}.$$

Поправлена або поліпшена оцінка дисперсії (незміщена оцінка):

$$\tilde{D} = \frac{D^* \cdot n}{n-1} = \frac{\sum_{i=1}^n (x_i - m^*)^2}{n-1}.$$

2. Обґрунтованість:

$$D^* \xrightarrow{n \rightarrow \infty} D.$$

Дисперсії та її оцінки за вибіркою:

$$D_X = D[X] = M[X^2] - m_X^2,$$

$$D^* = \frac{1}{n} \sum_{i=1}^n x_i^2 - (m^*)^2.$$

Обидва члени правої частини другої рівності – обґрунтовані оцінки математичного очікування випадкової величини та її квадрата.

$$\begin{aligned} \tilde{D} &= \frac{\sum_{i=1}^n x_i^2}{n} - (m^*)^2, \quad \frac{\sum_{i=1}^n x_i^2}{n} \xrightarrow[n \rightarrow \infty]{P} M[X^2], \\ (m^*)^2 &\xrightarrow[n \rightarrow \infty]{P} m^2, \\ M[X^2] - m_X^2 &= D_X. \end{aligned}$$

Отже, оцінка дисперсії обґрунтована.

3. Ефективність.

Доказ ефективності оцінки дисперсії в загальному випадку складний та залежить від виду закону розподілу ВВ X .

Як міра точності обчислення дисперсії для довільного закону приймається наступна величина:

$$D[\tilde{D}] = \frac{\mu_4}{n} - \frac{(n-3)}{n \cdot (n-1)} \cdot D^2.$$

Окремі випадки формули $D[\tilde{D}]$ для деяких відомих законів розподілу.

Нормальний закон: центральні моменти нормального розподілу пов'язані співвідношенням:

$$\begin{aligned} \mu_s &= (s-1)\sigma^2\mu_{s-2}, \\ \mu_4 &= 3D^2. \end{aligned}$$

Тоді,

$$\begin{aligned} D[\tilde{D}] &= \frac{\mu_4}{n} - \frac{(n-3)}{n \cdot (n-1)} \cdot D^2 = \frac{3}{n} D^2 - \frac{n-3}{n(n-1)} D^2 = \frac{2n}{n(n-1)} D^2, \\ D[\tilde{D}] &= \frac{2}{n-1} \cdot D^2. \end{aligned}$$

Рівномірний закон розподілу:

$$\mu_4 = 1.8D^2,$$

$$D = \mu_2 = \frac{(b-a)^2}{12},$$

$$\mu_4 = \frac{(b-a)^4}{80},$$

$$D[\tilde{D}] = \frac{0.8n + 1.2}{n(n-1)} D^2.$$

Оцінки генеральної дисперсії за частотно-варіаційним рядом:

$$D^* = \sum_{i=1}^k x_i^2 \cdot p_i^* - \tilde{m}^2,$$

$$D^* = \sum_{i=1}^k x_i^2 \cdot \frac{n_i}{n} - \tilde{m}^2,$$

або

$$D^* = \frac{\sum_{i=1}^k (x_i - \tilde{m})^2 \cdot n_i}{n} = \sum_{i=1}^k (x_i - \tilde{m})^2 \cdot p_i^*,$$

де k – кількість варіантів у ряді.

Для інтервального ряду:

$$D^* = \sum_{i=1}^k (x'_i)^2 \cdot p_i^* - \tilde{m}^2 = \sum_{i=1}^k x_i'^2 \cdot \frac{m_i}{n} - \tilde{m}^2,$$

x'_i – середина i -го інтервалу,

k – кількість інтервалів.

Особливості оцінок моментів більше високих порядків

Принципових труднощів у визначенні оцінок моментів високих порядків немає. Це пов'язано з тим, що вони виражаються як математичне очікування ВВ, що представляють собою ступінь ВВ або їхніх центрованих значень.

$$\alpha_s^*[X] = \frac{\sum_{i=1}^n x_i^s}{n},$$

$$\mu_s^*[X] = \frac{\sum_{i=1}^n (x_i - m^*)^s}{n}.$$

Відносна помилка для математичного очікування при $n = 50$:

$$\frac{\sigma_{\tilde{m}}}{\sigma_X^*} = \frac{1}{\sqrt{n}}, \quad \frac{\sigma_{\tilde{m}}}{\sigma_X^*} = 14\%.$$

При тому ж числі вимірів у нормальному законі розподілу ВВ відносна помилка дисперсії:

$$\frac{\sigma_{\tilde{D}}}{\tilde{D}} \cdot 100\% = \sqrt{\frac{2}{n-1}} \cdot 100\% = 20\%.$$

Для моментів більш високих порядків досягнення тієї ж точності тобчислень забезпечується при істотному зростанні числа вимірів ВВ, тим більше, чим вище порядок моментів.

Задача 1.

Є дані щодо кількості розрядів коду, які були використані при симетричному шифруванні для обміну повідомленнями. Потрібно порахувати оцінки математичного очікування та дисперсії.

Вхідні дані зведено в таблицю 2.1.

Таблиця 2.1

i	Кількість розрядів	Частота (кількість повідомлень)	Частість	Накопичена частість
1	94-100	3	0.03	0.03
2	100-106	7	0.07	0.1
3	106-112	11	0.11	0.21
4	112-118	20	0.2	0.41
5	118-124	28	0.28	0.69
6	124-130	19	0.19	0.88
7	130-136	10	0.1	0.98
8	136-142	2	0.02	1

Розв'язання.

Оцінка математичного очікування за заданим інтервальним рядом:

$$\tilde{m} = m^* = \sum_{i=1}^8 x'_i \cdot p_i^* = \frac{97 \cdot 3 + \dots + 139 \cdot 2}{100} = 119.2 .$$

Оцінка дисперсії за заданим інтервальним рядом:

$$D^* = \sum_{i=1}^8 (x'_i - m^*)^2 \cdot p_i^* =$$
$$\frac{(97 - 119.2)^2 \cdot 3 + \dots + (139 - 119.2)^2 \cdot 2}{100} = 87.48 ,$$
$$\tilde{D} = \frac{87.48 \cdot 100}{99} = 88.36$$

Завдання до лабораторної роботи

1. Для дослідження точкових характеристик параметрів розподілу скористаємося вибіркою з лабораторної роботи №1.
2. Обчислити та проаналізувати точкові оцінки математичного очікування (m) та дисперсії (D , еквівалентне позначення – σ^2) для простого та інтервального рядів:
 - обчислити та проаналізувати точкову оцінку математичного очікування за варіаційним та інтервальним рядом;
 - обчислити та проаналізувати точкову оцінку дисперсії за варіаційним та інтервальним рядом;
 - одержані точкові оцінки проілюструвати графічно (наприклад, на гістограмі).
3. Обчислити та проаналізувати точкові оцінки коефіцієнтів асиметрії та ексцесу за варіаційним рядом.

Контрольні питання

1. Точкова оцінка параметра розподілу.
2. Вимоги до точкових оцінок розподілу: спроможність, незміщеність, ефективність.

3. Точкові оцінки математичного очікування за варіаційним, частотно-варіаційним та інтервальним рядами.

4. Точкові оцінки дисперсії за варіаційним, частотно-варіаційним та інтервальним рядами. Поліпшена чи виправлена дисперсія.

5. Оцінка стандартного чи середньоквадратичного відхилення.

6. Вибіркові коефіцієнти асиметрії та ексцесу.

Література [1-3, 5-7]

Лабораторна робота № 3

Визначення інтервальних статистичних оцінок параметрів розподілу за одновимірною вибіркою

Мета роботи: формування уявлень та практичних навичок у визначенні інтервальних статистичних оцінок параметрів розподілу генеральних сукупностей за статистичними даними для одновимірних вибірок.

Теоретична довідка

Точкова оцінка навіть, якщо вона незміщена, обґрунтована, ефективна дає наближене значення параметра генеральної сукупності θ , особливо для вибірок малого обсягу, відрізняється від істинного значення параметра, тобто від θ .

Потрібно знати, до яких помилок може привести заміна параметра θ на його точкову оцінку $\hat{\theta}$ та з яким ступенем упевненості можна чекати, що ці помилки не вийдуть за відомі межі.

Уявлення про точність і надійність оцінки параметра θ дає так звана **міра ймовірності**, а саме **інтервальна оцінка**.

Приклад інтервальної оцінки. Середній зріст 175-180см (не 178 см).

Як міру ймовірності приймають:

1) **довірчу ймовірність β** , з якої істинне значення параметра θ буде перебувати в заданому відносно статистичної оцінки інтервалі;

2) **інтервал довіри** відносно статистичної оцінки, в який з заданою ймовірністю β потрапить істинне значення параметра θ .

Нехай для параметра θ генеральної сукупності отримана доброякісна оцінка $\hat{\theta}$. Потрібно оцінити отриману при цьому помилку.

Призначимо деяку достатньо велику ймовірність ($\beta = 0,9; 0,95; 0,997$) таку, що подію з цією ймовірністю β можна вважати практично достовірною.

Потрібно знайти **інтервал довіри**:

$$P(a_n^{*(1)} < a < a_n^{*(2)}) = \beta.$$

Межі інтервалу – **надійні межі**: $a_n^{*(1)}$; $a_n^{*(2)}$.

Інтервальна оцінка параметра а (інтервал довіри)

$I_\beta[a] = (a_n^{*(1)}, a_n^{*(2)})$ – числовий інтервал відносно статистичної оцінки параметра, який з заданою імовірністю β покриває реальне значення параметра a (рис. 3.1).

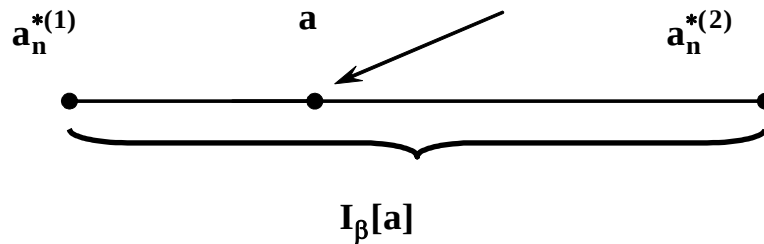


Рисунок 3.1 – Довірчий інтервал для параметру ГС a

Найчастіше інтервал довіри вибирають **симетричним** відносно статистичного параметру (рис. 3.2).

$$I_\beta[a] = (a_n^{*(1)}, a_n^{*(2)}),$$

$$a_n^{*(1)} = a_n^* - \varepsilon; a_n^{*(2)} = a_n^* + \varepsilon.$$

$$P(|a_n^* - a| < \varepsilon) = \beta;$$

$$P(a_n^* - \varepsilon < a < a_n^* + \varepsilon) = \beta;$$

$$I_\beta[a] = (a_n^* - \varepsilon; a_n^* + \varepsilon).$$

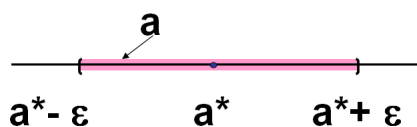


Рисунок 3.2 – Симетричний довірчий інтервал

$\alpha = 1 - \beta$ – рівень значущості, мала ймовірність того, що розбіжності між параметром і його оцінкою більше або дорівнює абсолютній величині інтервалу довіри:

$$P(|a - a_n^*| \geq \varepsilon) = \alpha.$$

Найчастіше $\alpha = 0.05; 0.1$.

Інтервал довіри – числовий інтервал значень параметра **a** генеральних сукупності, які не суперечать дослідним даним або сумісні з дослідними даними.

Межі інтервалу і його величина отримані за вибірковими даними і тому випадкові на відміну від самого параметра **a**.

Величина інтервалу довіри істотно залежить:

- 1) від обсягу вибірки: зі зростанням **n** величина інтервалу зменшується;
- 2) від величини довірчої ймовірності: чим більше довірна ймовірність **β**, тим більше ε .

Оцінка інтервалу довіри для математичного очікування

- 1) “Трубий” метод;
- 2) “Точний” метод.

Визначення I_β можливо, якщо відомий закон розподілу статистичної оцінки, який залежить від закону розподілу самої випадкової величини і від конкретного значення параметра генеральної сукупності.

“Трубий” метод оцінки довірчого інтервалу для математичного очікування

“Трубий” метод використовується при наступних допущеннях:

- допущення нормальності закону розподілу випадкової величини;
- заміна параметрів цього закону їхніми статистичними оцінками.

Постанова задачі: випадкова величина **X** – описує генеральну сукупність з невідомими параметрами **m, D**.

Знайти інтервал довіри $I_\beta[m]$ для m , якщо задано довірчу ймовірність і отримані результати експерименту.

Дано: $x_1, x_2, \dots, x_n; \beta$.

Знайти: $I_\beta[m]: a = m_X; a^* = m^*$.

Відомо, що статистична оцінка математичного очікування дорівнює:

$$m^* = \tilde{m} = \frac{\sum_{i=1}^n x_i}{n}.$$

В якості оцінки реального математичного очікування за вибіркою приймається середнє арифметичне n незалежних спостережуваних значень.

x_i – деякий екземпляр випадкової величини X з параметрами m_X, σ_X .

Оцінка m_X – це сума n незалежних однаково розподілених випадкових величин, тоді, за центральною граничною теоремою при досить великому n закон розподілу цієї суми наближається до нормального.

У практичній статистиці навіть при відносно невеликому числі випробувань від 10 до 20 вважається, що закон розподілу прямує до нормального. Тоді, ймовірність попадання в інтервал для нормального закону дорівнює:

$$P(a < X < b) = \Phi^*\left(\frac{b - m_X}{\sigma_X}\right) - \Phi^*\left(\frac{a - m_X}{\sigma_X}\right).$$

Аналогічно, в симетричний інтервал $\pm \varepsilon$ відносно m_X :

$$\begin{aligned} P(m_X - \varepsilon < X < m_X + \varepsilon) &= \Phi^*\left(\frac{\varepsilon}{\sigma_X}\right) - 1 + \Phi^*\left(\frac{\varepsilon}{\sigma_X}\right) = \\ &= 2\Phi^*\left(\frac{\varepsilon}{\sigma_X}\right) - 1 = 2\Phi_0\left(\frac{\varepsilon}{\sigma_X}\right). \end{aligned}$$

Розглянута випадкова величина X – це оцінка математичного очікування:

$$\tilde{m} = m^* \approx N\left(m, \frac{D}{n}\right).$$

$$P(|\tilde{m} - m| < \varepsilon) = \beta = 2\Phi^*\left(\frac{\varepsilon}{\sigma_{m^*}}\right) - 1 \Rightarrow$$

$$\begin{aligned} \Rightarrow \Phi^* \left(\frac{\varepsilon}{\sigma_{m^*}} \right) &= \frac{1+\beta}{2} \Rightarrow \\ \varepsilon &= \sigma_{m^*} \arg \Phi^* \left(\frac{1+\beta}{2} \right) = \sigma_{m^*} \arg \Phi_0 \left(\frac{\beta}{2} \right) \\ \sigma_{m^*} &= \frac{\sigma_X}{\sqrt{n}} \Rightarrow \\ \Rightarrow \varepsilon &= \frac{\sigma_X}{\sqrt{n}} \arg \Phi^* \left(\frac{1+\beta}{2} \right) = \frac{\sigma_X}{\sqrt{n}} \arg \Phi_0 \left(\frac{\beta}{2} \right) \end{aligned}$$

Величина інтервалу довіри для математичного очікування дорівнює (“грубий метод“):

$$\begin{aligned} u_\beta &= \arg \Phi^* \left(\frac{1+\beta}{2} \right), \\ I_\beta[m] &= \left(m^* - u_\beta \sqrt{\frac{D^*}{n}}; m^* + u_\beta \sqrt{\frac{D^*}{n}} \right). \end{aligned}$$

Величина інтервалу довіри для математичного очікування дорівнює (“грубий метод“):

$$\begin{aligned} \varepsilon &= \sigma_{m^*} \underbrace{\arg \Phi^* \left(\frac{1+\beta}{2} \right)}_{u_{1-\frac{\alpha}{2}}}, \\ \varepsilon &= \frac{\sigma_X}{\sqrt{n}} \cdot u_{1-\frac{\alpha}{2}}. \end{aligned}$$

“Точний” метод оцінки довірчого інтервалу для математичного очікування

Якщо σ_X невідомо, то використовують σ_X^* і замість нормального розподілу – **t**-розподіл (Стюдента):

$$\varepsilon = \frac{\sigma^*}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1),$$

$t_{1-\frac{\alpha}{2}}(n-1)$ – табличне значення, квантиль **t**-розподілу.

Задача 2.

З великої партії деталей взяли **25** штук. Знайшли середній діаметр і дисперсію діаметра: $m^* = 1.9$ см та $D^* = 0.001$ см².

Знайти інтервал довіри для середнього діаметру (з ймовірністю 0.9), тобто інтервал, у якому з 0,9 перебуває середнє всієї партії деталей.

$$\sigma[m^*] = \sqrt{\frac{D^*}{25}} = \sqrt{\frac{0.001}{25}} = 0.002,$$

$$\beta = 0.9, \alpha = 1 - \beta = 0.1, 1 - \frac{\alpha}{2} = 0.95.$$

Квантиль: $t_{1-\frac{\alpha}{2}}(n-1) = t_{0.95}(24) = 1.711$ (за таблицями квантилей t -розподілу). Відповідь: $\varepsilon = 0.002 * 1.711 = 0.003$.

Довірчий інтервал для математичного очікування:

$$I_{0.9}[m] = (1.897; 1.903).$$

Зауваження. Чим більше довірча ймовірність β , тим ширше інтервал довіри I_β і навпаки (рис. 3.3).

β	I_β	
0,9	1,897	1,903
0,99	1,895	1,905
0,999	1,893	1,907

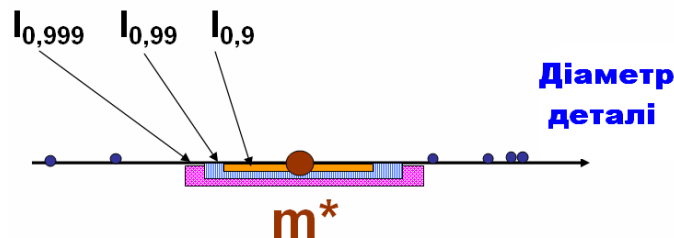


Рисунок 3.3 – Ілюстрація залежності величини інтервалу довіри від довірчої ймовірності

“Грубий” метод оцінки довірчого інтервалу для дисперсії D_x

Постанова задачі: дана випадкова величина X з нормальним законом розподілу та невідомими параметрами m і D . Зроблено n незалежних випробувань.

Потрібно із заданою довірчою ймовірністю знайти інтервал довіри для **D**.

Дано: $x_1, x_2, \dots, x_n; \beta$.

Знайти: $I_\beta[D]: a = D_X; a^* = D^*$ або $\tilde{a} = \tilde{D}$.

В якості оцінки **D** приймаємо:

$$\tilde{D} = \frac{\sum_{i=1}^n (x_i - \tilde{m})^2}{n - 1}.$$

За аналогією з математичним очікуванням грубий метод:

$$I_\beta[D] = (\tilde{D} - \varepsilon; \tilde{D} + \varepsilon),$$

$$\varepsilon = u_\beta \cdot \sigma[\tilde{D}].$$

Щоб скористатися цими формулами замість реальних **D** і користуються їхніми оцінками:

$$D[\tilde{D}] = \frac{\mu_4}{n} - \frac{n-3}{n(n-1)} D_X^2,$$

$$\mu_4^*[X] = \frac{\sum_{i=1}^n (x_i - m^*)^4}{n}.$$

Нормальний закон:

$$D[\tilde{D}] = \frac{2}{n-1} \cdot D_X^2.$$

Рівномірний:

$$D[\tilde{D}] = \frac{0,8n+1,2}{n(n-1)} D_X^2.$$

“Точний” метод оцінки довірчого інтервалу для дисперсії

Якщо **m** відомо, довірчий інтервал за “точним” методом дорівнює:

$$I_\beta[D] = \left(\frac{nD^*}{\chi_{1-\frac{\alpha}{2}}^2(n)}; \frac{nD^*}{\chi_{\frac{\alpha}{2}}^2(n)} \right).$$

Якщо **m** невідомо (беруть **m***), тоді:

$$I_\beta[D] = \left(\frac{nD^*}{\chi_{1-\frac{\alpha}{2}}^2(n-1)}; \frac{nD^*}{\chi_{\frac{\alpha}{2}}^2(n-1)} \right).$$

Завдання до лабораторної роботи

1. Для досліджень інтервальних оцінок параметрів розподілу скористаємося вибіркою із лабораторної роботи №1.

2. Обчислити та проаналізувати інтервальні оцінки математичного очікування (**m**) та дисперсії (**D**) для простого ряду:

– побудувати довірчі інтервали для **m** та **D** за допомогою “грубого” та “точного” методів;

– використовуючи різні значення довірчої ймовірності (наприклад, **0,8; 0,9; 0,95; 0,975; 0,995; 0,999**) проаналізувати залежність довжини довірчого інтервалу від величини довірчої ймовірності;

– побудувати графіки залежності величини довірчого інтервалу від вірогідності для кожного параметра.

Контрольні питання

1. Визначення інтервальної оцінки параметра розподілу чи довірчого інтервалу.

2. Довірча ймовірність, довірчі межі.

3. “Грубий” та “точний” прийоми визначення інтервальних оцінок.

4. Інтервальні оцінки для математичного очікування.

5. Інтервальні оцінки дисперсії.

6. Характер залежності величини довірчого інтервалу від обсягу вибірки та довірчої ймовірності.

Література [1-2, 6-7]

Лабораторна робота № 4

Згладжування (вирівнювання) статистичних рядів

Мета роботи: встановлення відповідності між статистичними і теоретичним розподілами шляхом побудови кривої розподілу, що виражає істотні риси емпіричного матеріалу.

Теоретична довідка

У всякому статистичному розподілі неминуче присутні елементи випадковості, пов'язані з обмеженим числом спостережень, попаданням в вибірку сукупність певних значень і т.п. Тільки при дуже великому числі спостережень ці елементи випадковості згладжуються, і випадкове явище виявляє в повній мірі властиву йому закономірність. На практиці майже ніколи не мають справи з таким великим числом спостережень і змушені зважати на те, що будь-якому статистичному розподілу властиві в більшій чи меншій мірі риси випадковості. Тому при обробці статистичного матеріалу часто доводиться вирішувати питання про те, як підібрати для даного статистичного ряду теоретичну криву розподілу, яка має лише істотні риси статистичного матеріалу, а не випадковості, пов'язані з недостатнім обсягом експериментальних даних. Таке завдання називається завданням вирівнювання (згладжування) статистичних рядів. Завдання вирівнювання полягає в тому, щоб підібрати теоретичну плавну криву розподілу, яка з тієї чи іншої точки зору найкращим чином описує даний статистичний розподіл.

Припущення про вид закону розподілу може бути зроблено виходячи з:

- теоретичних передумов, пов'язаних з фізичною суттю завдання;
- з досвіду попередніх аналогічних досліджень;
- на підставі графічного зображення емпіричного розподілу, найчастіше гістограми.

Як правило, принциповий вид теоретичної кривої вибирається заздалегідь з міркувань, пов'язаних з сутністю завдання, а в деяких випадках

просто із зовнішнім виглядом статистичного розподілу. Аналітичний вираз обраної кривої розподілу залежить від деяких параметрів, а завдання вирівнювання статистичного ряду самостійно стає у задачу раціонального вибору тих значень параметрів, при яких відповідність між статистичними і теоретичним розподілами виявляється найкращою.

Завдання знаходження оцінок параметрів розподілу формалізується наступним чином. Задана вибірка обсягу n : $x_1, x_2, \dots, x_n$ з генеральної сукупності, що описується ВВ X . Передбачається деякий теоретичний закон розподілу ВВ X з функцією щільності розподілу виду: $f(x_1, x_2, \dots, x_n; a_1, a_2, \dots, a_k)$, де $a_1, a_2, \dots, a_k$ – невідомі параметри розподілу, які підлягають визначенню.

Метод моментів

Ідея методу моментів: для отримання невідомих оцінок параметрів $a_1, a_2, \dots, a_k$ розподілу генеральної сукупності X певна кількість вибірових початкових і/або центральних моментів (α_s^* і/або μ_s^*) прирівнюються до відповідних теоретичних аналогів (α_s і/або μ_s), отриманих для передбачуваного теоретичного закону розподілу. З отриманої системи рівнянь знаходяться оцінки параметрів $a_1, a_2, \dots, a_k$.

Алгоритм методу моментів

1. Визначаємо k теоретичних моментів для передбачуваного теоретичного закону розподілу. Припустимо, що $f(x_1, x_2, \dots, x_n; a_1, a_2, \dots, a_k)$ – щільність розподілу випадкової величини X . Визначимо за допомогою цієї щільності k будь-яких моментів випадкової величини, наприклад, перші k початкових/центральных моментів за формулами:

$$\begin{cases} \alpha_s = M[x^s] = \int_{-\infty}^{+\infty} x^s \cdot f(x) dx \\ \mu_s = M[(x - m_x)^s] = \int_{-\infty}^{+\infty} (x - m_x)^s \cdot f(x) dx \end{cases}.$$

Кількість теоретичних моментів дорівнює кількості невизначених параметрів розподілу.

2. За вибіркою спостережень ВВ X визначаємо значення k вибіркових моментів та прирівнюємо до відповідних теоретичних аналогів.

$$\begin{cases} \alpha_s^* = \sum_{i=1}^n \frac{x_i^s}{n} \\ \mu_s^* = \sum_{i=1}^n \frac{(x_i - m_x^*)^s}{n} \end{cases}.$$

3. Отримаємо систему із k рівнянь із k невідомими:

$$\begin{cases} \alpha_s^* \approx \alpha_s \\ \mu_s^* \approx \mu_s \end{cases} \Rightarrow a_1, a_2, \dots, a_k - ?,$$

розв'язуючи яку отримуємо шукані значення $a_1, a_2, \dots, a_k$.

Примітка: якщо є можливість використовувати доброякісні оцінки параметрів розподілу, тобто $\tilde{\alpha}_s$ і $\tilde{\mu}_s$.

Згідно з методом моментів, параметри $a_1, a_2, \dots, a_k$ вибираються з таким розрахунком, щоб кілька найважливіших числових характеристик (моментів) теоретичного розподілу дорівнювали відповідним статистичним характеристикам. Наприклад, якщо теоретична крива $f(x_1, x_2, \dots, x_n; a_1, a_2)$, залежить тільки від двох параметрів a_1 і a_2 , ці параметри вибираються так, щоб математичне очікування $M[X]$ і дисперсія $D[X]$ теоретичного розподілу збігалися з відповідними статистичними характеристиками m_x^* і D_x^* .

Якщо крива $f(x_1, x_2, \dots, x_n; a_1, a_2, a_3)$ залежить від трьох параметрів, можна підібрати їх так, щоб збіглися перші три моменти і т.д.

Слід зауважити, що при вирівнюванні статистичних рядів нераціонально користуватися моментами з порядком вище четвертого, так як точність обчислення моментів різко падає зі збільшенням їх порядку.

Приклад: знайти методом моментів оцінки параметрів **a** і **b** рівномірного закону розподілу.

Теоретичні значення першого початкового та другого центрального моментів рівномірного закону розподілу виражаються формулами:

$$\alpha_1 = m_x = \frac{a + b}{2},$$
$$\mu_2 = D_x = \frac{(b - a)^2}{12}.$$

Вибіркові моменти:

$$\tilde{m}_x = \tilde{\alpha}_1 = \frac{\sum_{i=1}^n x_i}{n}; \tilde{D}_x = \tilde{\mu}_2 = \frac{\sum_{i=1}^n (x_i - \tilde{m}_x)^2}{n}.$$

Рівняння методу моментів:

$$\begin{cases} \alpha_1 = \tilde{\alpha}_1 \\ \mu_2 = \tilde{\mu}_2 \end{cases} \Rightarrow \Rightarrow \begin{cases} a = ? \\ b = ? \end{cases}.$$

Метод найменших квадратів

Ідея методу найменших квадратів (МНК): при згладжуванні емпіричних залежностей дуже часто виходять з так званого принципу або методу найменших квадратів, вважаючи, що найкращим наближенням до емпіричної залежності в даному класі функцій є таке, при якому сума квадратів відхилень звертається в мінімум.

При цьому питання про те, в якому саме класі функцій слід шукати найкраще наближення, вирішується вже не з математичних міркувань, а з міркувань, пов'язаних з фізикою розв'язуваної задачі, з урахуванням характеру отриманої емпіричної кривої і ступеня точності проведених спостережень. Часто принциповий характер функції, що виражає досліджувану залежність, відомий заздалегідь з теоретичних міркувань, з досвіду ж потрібно отримати лише деякі чисельні параметри, що входять у вираз функції; саме ці параметри підбираються за допомогою методу найменших квадратів.

МНК застосовується при згладжуванні довільних експериментальних залежностей і є загально-математичним методом.

Алгоритм методу найменших квадратів (МНК)

Задача відновлення аналітичного виразу деякої функції $y = f(x)$ методом найменших квадратів вимагає, щоб міра відхилення експериментальних значень від обраної функції була мінімальною в заданих n точках $(x_i, y_i), i = \overline{1, n}$.

Дано: результати спостережень у n точках: $(x_i, y_i), i = \overline{1, n}$.

Знайти аналітичну залежність $f(x)$ максимально наближену до заданих точок $(x_i, y_i), i = \overline{1, n}$. За МНК функція $f(x)$ може не проходити через задані точки, потрібно лише мінімізувати квадратичну помилку відхилення експериментальних значень від аналітичних (рис. 4.1).

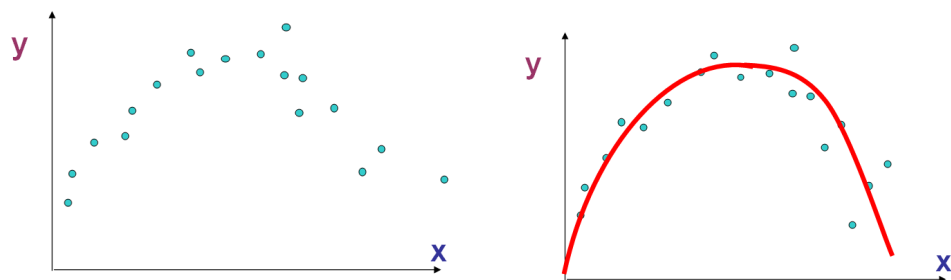


Рисунок 4.1 – Ілюстрація методу найменших квадратів

В загальному випадку функція $f(x)$ буде залежати від вектору невідомих параметрів, тобто $\bar{a} = (a_1, a_2, \dots, a_k): f(x; \bar{a})$. Наприклад, якщо вибір аналітичної залежності буде у класі лінійних функцій, $y = ax + b$, то маємо два невідомих параметри: $a_1 = a, a_2 = b$. Якщо вибір аналітичної залежності буде у класі квадратичних функцій, $y = ax^2 + bx + c$, то маємо 3 невідомих параметри: $a_1 = a, a_2 = b, a_3 = c$.

Будується функція Q – міра відхилення експериментальних значень від обраної функції $f(x; a_1, a_2, \dots, a_k)$:

$$Q(a_1, a_2, \dots, a_k) = \sum_{i=1}^n [y_i - f(x_i; a_1, a_2, \dots, a_k)]^2$$

Вибираються такі значення вектору параметрів $\bar{a} = (a_1, a_2, \dots, a_k)$, щоб функція Q досягала мінімуму.

Продиференціюємо функцію Q по кожному з параметрів і прирівняємо часткові похідні до нуля – необхідні умови існування екстремуму для функції багатьох змінних:

$$\begin{cases} \frac{\partial Q(x_i; a_1, a_2, \dots, a_k)}{\partial a_1} = 0, \\ \frac{\partial Q(x_i; a_1, a_2, \dots, a_k)}{\partial a_2} = 0, \\ \dots \\ \frac{\partial Q(x_i; a_1, a_2, \dots, a_k)}{\partial a_k} = 0, \end{cases}$$

$$\begin{cases} \sum_{i=1}^n [y_i - f(x_i; a_1, a_2, \dots, a_k)] \left(\frac{\partial f}{\partial a_1} \right)_i = 0, \\ \sum_{i=1}^n [y_i - f(x_i; a_1, a_2, \dots, a_k)] \left(\frac{\partial f}{\partial a_2} \right)_i = 0, \\ \dots \\ \sum_{i=1}^n [y_i - f(x_i; a_1, a_2, \dots, a_k)] \left(\frac{\partial f}{\partial a_k} \right)_i = 0. \end{cases}$$

Отримана система називається системою нормальних рівнянь, складається з k рівнянь та має k невідомих. Розв'язанням системи є вектор оцінок параметрів розподілу.

Таким чином, найкраще узгодження теоретичної кривої та експериментальних даних за МНК: сума квадратів відхилень експериментальних точок від обираємої кривої, досягає мінімуму.

Переваги МНК:

- порівняно простий математичний метод визначення параметрів;
- існує теоретичне обґрунтування з імовірнісної точки зору.

Метод максимальної правдоподібності

Цей метод є основним для отримання оцінок параметрів ГС за статистичними даними.

Ідея методу максимальної правдоподібності: будується функція правдоподібності, що виражає щільність ймовірності спільної появи результатів вибірки: $x_1, x_2, \dots, x_n$ та знаходиться таке значення вектору параметрів $\bar{a} = (a_1, a_2, \dots, a_k)$, яке доставляє максимум функції правдоподібності.

Алгоритм методу максимальної правдоподібності

Передбачається, що щільність розподілу випадкової величини X залежить від деякої множини невідомих параметрів $a_1, a_2, \dots, a_k$, і має вигляд $f(x_1, x_2, \dots, x_n; a_1, a_2, \dots, a_k)$.

1. Будується функція правдоподібності, що виражає щільність ймовірності спільної появи результатів вибірки, і має вигляд:

– для безперервних ВВ:

$$L(x_1, x_2, \dots, x_n; a_1, a_2, \dots, a_k) = \prod_{i=1}^n f(x_i; a_1, a_2, \dots, a_k);$$

– для дискретних ВВ:

$$L(x_1, x_2, \dots, x_n; a_1, a_2, \dots, a_k) = \prod_{i=1}^n p(x_i; a_1, a_2, \dots, a_k).$$

2. Згідно з методом максимальної правдоподібності в якості оцінки невідомого вектору $\bar{a} = (a_1, a_2, \dots, a_k)$ приймається значення, яке максимізує функцію правдоподібності, тобто забезпечує:

$$L(x_1, x_2, \dots, x_n; a_1, a_2, \dots, a_k) \rightarrow \max$$

або (операція логарифмування не змінює точку максимуму)

$$\ln L(x_1, x_2, \dots, x_n; a_1, a_2, \dots, a_k) \rightarrow \max.$$

3. Необхідні умови екстремуму функції багатьох змінних:

$$\begin{cases} \frac{\partial L(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n; \mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k)}{\partial \mathbf{a}_1} = 0, \\ \frac{\partial L(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n; \mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k)}{\partial \mathbf{a}_2} = 0, \\ \dots \\ \frac{\partial L(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n; \mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k)}{\partial \mathbf{a}_k} = 0 \end{cases}$$

або

$$\begin{cases} \frac{\partial \ln L(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n; \mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k)}{\partial \mathbf{a}_1} = 0, \\ \frac{\partial \ln L(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n; \mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k)}{\partial \mathbf{a}_2} = 0, \\ \dots \\ \frac{\partial \ln L(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n; \mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k)}{\partial \mathbf{a}_k} = 0 \end{cases}$$

дають систему рівнянь, з якої знаходиться невідомий вектор оцінок параметрів розподілу $\bar{\mathbf{a}} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k)$.

Оцінки параметрів $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k$ отримані за цим методом, називають – МП-оцінками або оцінками максимальної правдоподібності.

Приклад: визначити методом максимальної правдоподібності оцінку параметра експоненціального закону розподілу.

Дано: $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$; $\lambda - ?$.

Функція щільності розподілу експоненти дорівнює:

$$f(\mathbf{x}) = \begin{cases} \lambda \cdot e^{-\lambda \mathbf{x}}, & \mathbf{x} \geq 0 \\ 0, & \mathbf{x} < 0 \end{cases}.$$

Запишемо функцію правдоподібності:

$$L = \prod_{i=1}^n f(\mathbf{x}_i, \lambda) = \lambda^n \cdot e^{-\lambda \sum_{i=1}^n \mathbf{x}_i}.$$

Прологарифмуємо її:

$$\ln L = \ln \left(\lambda^n \cdot e^{-\lambda \sum_{i=1}^n \mathbf{x}_i} \right) = n \ln \lambda - \lambda \cdot \sum_{i=1}^n \mathbf{x}_i.$$

Необхідна умова екстремуму:

$$\frac{dL}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^n x_i = 0 \Rightarrow \lambda = \frac{1}{\frac{\sum_{i=1}^n x_i}{n}} = \frac{1}{\tilde{m}_x}.$$

Завдання до лабораторної роботи

1. За видом побудованої в п.1 лабораторної роботи № 1 гістограми зробити припущення про закон розподілу.
2. Знайти параметри передбачуваного розподілу методом моментів і максимальної правдоподібності.
3. Уточнити параметри розподілу методом найменших квадратів..
4. Виконати порівняльний аналіз отриманих результатів.
5. Всі знайдені аналітичні залежності зобразити графічно

Контрольні питання

1. Статистичні оцінки параметрів розподілу.
2. Найбільш вживані розподіли ВВ та їх властивості, Нормальний (закон Гауса), рівномірний (закон постійної щільності) та експоненціальний закони.
3. Задача згладжування (вирівнювання) статистичних рядів.
4. Методи оцінок параметрів розподілу.
5. Метод моментів.
6. Метод максимальної правдоподібності. Функція максимальної правдоподібності.
7. Метод найменших квадратів. Система нормальних рівнянь.

Література [4-7]

Лабораторна робота № 5
Побудова теоретичного закону розподілу
за експериментальними даними

Мета роботи: формування уявлень та практичних навичок у застосуванні критеріїв узгодження Колмогорова та Пірсона, що перевіряють, чи відповідає гіпотетичний теоретичний закон розподілу статистичним даним.

Теоретична довідка

Загальна характеристика критеріїв узгодження

Критерієм узгодження називають критерій перевірки гіпотези про передбачуваний закон розподілу випадкової величини. Висування і перевірка гіпотези про вид закону розподілу генеральної сукупності зазвичай є завершальним етапом обробки одновимірної вибірки. На практиці для висунення такої гіпотези попередньо виконується побудова гістограми, робиться припущення про закон розподілу і оцінюються параметри розподілу.

Для перевірки гіпотези про закон розподілу генеральної сукупності застосовуються критерії Колмогорова, Пірсона, Смирнова і т.п. Всі ці критерії ґрунтуються на порівнянні емпіричних (отриманих дослідним шляхом) і теоретичних (обчислених в припущенні теоретичного закону розподілу генеральної сукупності) частот або частостей.

Критерії узгодження дозволяють визначити характер розбіжностей між передбачуваним теоретичним і статистичним розподілом: чи пов'язані розбіжності з випадковими факторами (обмежене число спостережень, **n**) або вони істотні і пов'язані з тим, що підібрана крива погано вирівнює статистичний розподіл.

Нульова гіпотеза **H₀** полягає в тому, що випадкова величина **X** підпорядкована деякому певному закону розподілу. Для перевірки цієї гіпотези вибирають деяку випадкову величину, що характеризує ступінь розбіжності теоретичного та емпіричного розподілу.

Випадкова величина **U** вибирається таким чином, щоб при досить великих обсягах вибірки **n** був відомий закон її розподілу і він практично не залежав від закону розподілу випадкової величини **X**. Знаючи закон розподілу випадкової величини **U**, можна порахувати ймовірність того, що вона не менше деякого фактичного значення, що спостерігається: $P(U \geq u) = \alpha$.

Якщо ця ймовірність мала, то гіпотезу **H₀** слід відкинути, як мало правдоподібну.

Якщо ця ймовірність – велика, то кажуть що експериментальні дані не суперечать основній гіпотезі **H₀**.

Мала ймовірність означає, що за рахунок чисто випадкових розбіжностей відмінності між теоретичним і емпіричним розподілом не можуть бути пояснені, тобто основна гіпотеза відкидається, як така, що суперечить дослідним даним.

Завдання встановлення теоретичного закону розподілу **ВВ** за емпіричним розподілом, описаним одним з відомих способів, вимагає вирішення 2 підзадач:

- 1) визначити вид передбачуваного теоретичного закону розподілу;
- 2) визначити оцінки параметрів передбачуваного теоретичного закону розподілу.

Припущення про вид закону розподілу може бути зроблено виходячи з:

- теоретичних передумов, пов'язаних із фізичною сутністю завдання;
- з досвіду попередніх аналогічних досліджень;
- на підставі графічного зображення емпіричного розподілу, найчастіше гістограми.

Оцінки параметра передбачуваного закону розподілу можна отримати: методом моментів, найменших квадратів чи максимальної правдоподібності. Перевірка відповідності передбачуваного закону розподілу експериментальним даним виконується виходячи з критеріїв згоди, як-от критерій Пірсона чи Колмогорова.

Критерій Пірсона або χ^2 ("хі")-квадрат

В якості міри розбіжності між передбачуваним теоретичним і статистичним розподілами приймається величина, що дорівнює зваженій сумі квадратів відхилень між теоретичною ймовірністю, розрахованою за передбачуваним законом (p_i), і статистичною ймовірністю (p_i^*).

Алгоритм застосування критерію Пірсона

1. Виходячи з відомих значень емпіричних частот попадання в i -тий інтервал (частіше за все за видом гістограми розподілу), висувається нульова гіпотеза про передбачуваний закон розподілу випадкової величини X і знаходяться параметри цього закону за одним з описаних вище методів.

2. Визначають теоретичні частоти np_i , відповідні частотам, отриманим з експерименту m_i :

$$p_i = P(x_i < X < x_{i+1}) = F(x_{i+1}) - F(x_i),$$

тут p_i – теоретична ймовірність попадання в i -ий інтервал. Якщо серед експериментальних частот є нечисленні, то їх необхідно об'єднати з сусідніми. Число інтервалів після об'єднання має бути не менше 5.

1. Визначається міра розбіжності між теоретичними і емпіричними частотами:

$$\chi^2 = \sum_{i=1}^k \frac{(m_i - np_i)^2}{np_i}.$$

4. Визначають число ступенів свободи $r = k - s - 1$, де k – кількість інтервалів, s – число параметрів передбачуваного теоретичного розподілу.

5. На заданому рівні значущості α і з розрахованим числом ступенів свободи r за таблицею розподілу χ^2 знаходять критичне значення: $\chi_{кр}^2 = \chi^2(1 - \alpha, r)$.

6. Якщо фактично знайдене за вибіркою значення χ^{2*} : $\chi^{2*} < \chi_{кр}^2$ – приймається нульова гіпотеза (теоретичний закон розподілу не суперечить емпіричним даним).

Якщо фактично спостерігаємо значення $\chi^2: \chi^2 \geq \chi_{кр}^2$ – нульову гіпотезу відкидають.

Зауваження. Статистика $\chi^2 = \sum_{i=1}^k \frac{(m_i - np_i)^2}{np_i}$ – має розподіл “хі”-квадрат лише при $n \rightarrow \infty$, в даному випадку, необхідно, щоб у кожному інтервальному угрупованні було достатньо спостережень (≥ 5), інакше сусідні інтервали об’єднують чи заново стоять угруповання. Це стосується як емпіричних m_i , так теоретичних частот np_i .

Критерій узгодження Колмогорова

Для критерію Колмогорова в якості міри розбіжності приймається величина, пропорційна максимуму абсолютної величини відхилень функцій розподілу передбачуваного теоретичного закону і емпіричної функції розподілу:

$$D = \max |F(x) - F^*(x)|,$$

де $F^*(x)$ – емпірична функція розподілу, $F(x)$ – теоретична функція розподілу.

Алгоритм застосування критерію Колмогорова

1. Виходячи з відомих значень емпіричних частот попадання в i -тий інтервал, висувають нульову гіпотезу про передбачуваний закон розподілу випадкової величини X і знаходять його параметри.

2. В результаті n незалежних спостережень будується $F^*(x)$ – емпірична функція розподілу неперервної випадкової величини X .

За розрахованими параметрами будується передбачувана теоретична функція розподілу $F(x)$.

Визначається міра розбіжності між теоретичними і емпіричними значеннями функції розподілу:

$$\lambda^* = \sqrt{n} \cdot D = \sqrt{n} \cdot \max |F(x) - F^*(x)|.$$

2. На заданому рівні значущості α за таблицею розподілу критичних значень для критерію Колмогорова знаходять критичне значення λ_α .

5. Якщо $\lambda^* < \lambda_\alpha$ – приймається нульова гіпотеза (теоретичний закон розподілу не суперечить емпіричним даним). Якщо $\lambda^* \geq \lambda_\alpha$ – нульову гіпотезу відкидають.

Таблиця 5.1 – Критичні значення критерію Колмогорова

α	0.1	0.05	0.025	0.01	0.005	0.001
λ_α	1.22	1.36	1.48	1.63	1.73	1.95

Завдання до лабораторної роботи

1. Для досліджень скористатися одновимірною вибіркою та її обробкою у лабораторних роботах №1-2.
2. На вигляд гістограми висунути припущення про закон розподілу СВ.
3. Методами моментів та максимальної правдоподібності визначити оцінки параметрів передбачуваного теоретичного розподілу.
4. Перевірити гіпотези про відповідність теоретичного передбачуваного закону розподілу емпіричних даних за критерієм Пірсона та Колмогорова.

Контрольні питання

1. Загальна характеристика критеріїв узгодження.
2. Як визначити вид передбачуваного теоретичного закону розподілу?
3. Методи визначення оцінок параметрів передбачуваного теоретичного закону розподілу.
4. Що таке рівень значущості і як він вибирається?
5. Критерій Пірсона. Міра розбіжності або статистика критерію. Число ступенів свободи. Нерівність для критичної галузі. Обмеження на теоретичні та емпіричні частоти.
6. Критерій Колмогорова. Міра розбіжності або статистика критерію. Нерівність для критичної галузі.

Література [3-4, 6-7]

Рекомендована література

1. Жлуктенко В. І. Теорія ймовірностей і математична статистика: навч.-метод. посібник : у 2-х ч. Ч. І : Теорія ймовірностей / В.І. Жлуктенко, С.І. Наконечний. – К.: КНЕУ. – 2000. – 304 с.
2. Кармелюк Г.І. Теорія ймовірностей і математична статистика. Посібник з розв’язування задач : навч. посібник / Г.І. Кармелюк. – К.: Центр учбової літератури. – 2007. – 576 с.
3. Кремер Н. Ш. Теория вероятностей и математическая статистика: учебник и практикум для вузов / Н.Ш. Кремер. – 5-е изд., перераб. и доп. – М.: Издательство Юрайт, 2019. – 538 с.
4. Вентцель Е.С. Теория вероятностей. – М.: ЮСТИЦИЯ, 2018. – 658 с.
5. Малярець Л. М. Теорія ймовірностей та математична статистика: навчальний посібник / Л.М. Малярець, І.Л. Лебедева, Е.Ю. Железнякова, З.Г. Попова. – Харків: Вид. ХНЕУ. – 2010. – 404 с.
6. Методичні вказівки і завдання до лабораторних і практичних робіт з курсу «Емпіричні методи програмної інженерії» (Для студентів, що навчаються за спеціальністю 121 «Інженерія програмного забезпечення» всіх форм навчання) / укладач: проф. Дмитрієва О.А. - Покровськ: ДонНТУ, 2017 р. – 83 с.
7. Методические указания и задания к лабораторным работам по курсу “Статистика. Часть1” для студентов области знаний 0305 “Экономика и предпринимательство”/ разраб.: Назарова И.А., Яценко Е.А. – Донецк: ДонНТУ, 2009 – 45 с.