Янцевич А. А., Дьячкова О. В. Математична статистика

Харківський національний університет імені В. Н. Каразіна

А. А. Янцевич
О. В. Дьячкова

ТЕОРІЯ ЙМОВІРНОСТЕЙ і МАТЕМАТИЧНА СТАТИСТИКА

Частина II

Математична статистика

Навчальний посібник

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
імені В. Н. КАРАЗІНА

А. А. Янцевич, О. В. Дьячкова

ТЕОРІЯ ЙМОВІРНОСТЕЙ І МАТЕМАТИЧНА СТАТИСТИКА

Навчальний посібник у двох частинах

Частина II

Математична статистика

Харків – 2018

УДК 519.2 (075.8)
Я 99

Рецензенти:

В. О. Золотарьов – доктор фіз.-мат. наук, професор, провідний науковий співробітник ФТІНТ імені Б. І. Вєркіна НАН України;

О. О. Аршава – канд. фіз.-мат. наук, доцент, зав. кафедри вищої математики ХНУБА;

Є. В. Свищова – канд. фіз.-мат. наук, доцент кафедри інформаційних технологій і математики ХГУ «НУА».

Затверджено до друку рішенням Вченої ради
Харківського національного університету імені В. Н. Каразіна
(протокол № 10 від 29.10.2018 року)

Янцевич А. А.

Я 99 Теорія ймовірностей і математична статистика : навч. посібник : у 2-х ч. Ч. 2. Математична статистика / А. А. Янцевич, О. В. Дьячкова. — Х. : ХНУ імені В. Н. Каразіна, 2018. — 152 с.

ISBN 978-966-285-551-7

Видання призначено для студентів соціально-економічних і управлінських спеціальностей (усіх форм навічання), які вивчають базовий курс теорії ймовірностей і математичної статистики. Другу частину присвячено основним статистичним поняттям і методам. Зокрема, розглянуто найважливіші завдання математичної статистики – оцінювання параметрів, перевірка гіпотез, а також елементи кореляційного і регресійного аналізу.

Викладення супроводжується прикладами з розв'язаннями, питаннями для самоконтролю, насичено багатим ілюстративним рядом – графіками, схемами, діаграмами. Відмінною рисою посібника є наявність широкого довідкового апарату: основних формул математичної статистики, статистичних таблиць і комп'ютерних функцій, предметного покажчика та перекладного словника з математичної статистики. Це дозволяє використовувати посібник і як довідник студентам, аспірантам, викладачам, науковим співробітникам та всім бажаючим, які опановують імовірнісні та статистичні методи.

УДК 519.2 (075.8)

ISBN 978-966-285-551-7

КОРОТКИЙ ЗМІСТ

ЗМІСТ

Вступ

Масові випадкові явища відбуваються в результаті дії сукупності неконтрольованих чинників, і часто неможливо виявити вплив кожного з них.

У таких випадках для вивчення явищ застосовують *статистичні методи* – тобто досліджують результати спостережень (статистичні дані) методами теорії ймовірності, щоб виявити закономірності явищ і їх взаємозв'язку.

Математична статистика займається як систематизацією результатів вимірювань або спостережень, так і побудовою і перевіркою адекватності відповідних імовірнісних моделей.

Серії спостережень за одних і тих же умов будуть різними, але ці відмінності при великому числі вимірювань фактично зникають. У такому разі йдеться про *статистичну стійкість*.

Окрім вказівки способів систематизації (збору, угрупування) даних, математична статистика вирішує такі основні задачі:

– визначення закону розподілу випадкової величини;
– оцінка невідомих параметрів розподілу;
– оцінка залежності випадкової величини від однієї або декількох інших випадкових величин;
– перевірка правдоподібності гіпотез про закон розподілу, про значення оцінюваного параметра, про форму зв'язку між випадковими величинами та ін.

Методи математичної статистики широко використовують при вирішенні багатьох задач економіки, соціології, медицини, біології, фізики тощо.

Автори висловлюють глибоку вдячність Євгенії Віталіївні Свищовій, доцентові кафедри інформаційних технологій і математики ХГУ «НУА», за ретельний і професійний аналіз рукопису і зроблені зауваження.

Глава 1

Основи статистичних методів

Основні питання:

- ♣ *генеральна сукупність і вибірка*
- ♣ *дискретний і інтервальний варіаційні ряди*
- ♣ *емпірична функція розподілу та її властивості*
- ♣ *числові характеристики вибірки: вибіркове середнє, вибіркова дисперсія, вибіркове середнє квадратичне відхилення*
- ♣ *операція «риса»*

1.1. Генеральна і вибіркова сукупності

Статистичні дослідження застосовують до сукупності однотипних об'єктів для вивчення однієї або декількох ознак. При цьому припускають, що інші ознаки, що характеризують об'єкти, рівноправні – тобто множина об'єктів **однорідна**.

Множина однорідних об'єктів, які вивчають з точки зору їх розподілу за деякою ознакою (ознаками), називають **генеральною сукупністю**.

У більшості випадків досліджують *кількісні* ознаки об'єктів, а для некількісних (*якісних*) ознак існують способи числового їхнього подання.

Приклад 1.1. Приклади генеральної сукупності :
– партія деталей, досліджуваних за їхніми вагою і розміром;
– товари в магазині, що вивчаються за кількістю їхніх продажів за день;
– поля зернової культури – за їхньою врожайністю;
– населення держави – за середньодушовим доходом;
– новинні інтернет-сайти – за їхньою відвідуваністю;
– підприємства галузі – за величиною їхніх активів і пасивів або прибутку / збитків;
– працівники підприємства – за їхніми добовою продуктивністю, стажем роботи;
– студенти групи або факультету – за оцінками за останню сесію і т. ін.

Як бачимо, досліджувані кількісні ознаки можуть бути як дискретними, так і неперервними.

Повне обстеження усіх об'єктів генеральної сукупності можливе далеко не завжди. Наприклад, це дуже важко або неможливо, якщо число об'єктів велике або нескінченне, вони важкодоступні, унаслідок дослідження об'єкти можуть бути знищені, дослідження трудомісткі або дорогі тощо. У таких випадках обстежують лише частину генеральної сукупності – *вибірку*.

Вибірковою сукупністю або **вибіркою** називають сукупність об'єктів, випадково відібраних із генеральної сукупності.

Об'єкти вибірки зазвичай позначають $x_1, x_2, \ldots, x_n$.

Число елементів у вибірковій сукупності називають **обсягом вибірки**.

Обсяг генеральної сукупності позначимо N (N може бути скінченним або нескінченним), вибірки – n. Часто значення N невідоме. Значення n відоме завжди і, як правило, воно значно менше N: $n \ll N$.

Вибірку можна використовувати для вивчення ознаки (ознак) генеральної сукупності лише тоді, коли ці вибірки правильно відображають цю ознаку – тобто вибірка має бути *репрезентативною*. Для цього вибірку формують випадковим чином, дотримуючись також певних вимог до її обсягу. Тоді згідно з законом великих чисел можна буде стверджувати, що вибірка репрезентативна.

Способи відбору об'єктів у вибірку можна розділити на два класи:

- простий випадковий вибір (із поверненням відібраного об'єкту назад у генеральну сукупність або без повернення): при цьому кожен об'єкт має однакову ймовірність потрапити у вибірку;
- розбиття генеральної сукупності на частини, а потім відбір об'єктів з цих частин. Можливе комбінування різних способів відбору.

Приклад 1.2. Вибіркова сукупність може, наприклад, включати:
а) 15 % випадково вибраних транспортних фірм регіону (для аудиторської перевірки сплати податків усіма його транспортними фірмами);
б) кожного чотирьохсотого відвідувача веб-сайту (для опитування з метою аналізу усієї аудиторії за віком і статтю);
в) кожен п'ятий пристрій із вироблених підприємством А і кожен десятий пристрій – підприємством В (для перевірки якості пристроїв, що випускаються на обох підприємствах, причому на підприємстві А в удвічі більшому обсязі);
г) усіх студентів двадцяти випадково обраних студентських груп університету (для виявлення студентської думки загалом) і так далі.

Ці вибірки можуть бути або не бути репрезентативними. Наприклад:
вибірка (а) не буде репрезентативна, якщо аналізується сплата податків автопідприємствами різних форм власності (що впливає на сплату) – тоді цей чинник потрібно враховувати і при відборі;
вибірка (г) не показова, якщо ставлення до досліджуваної теми може бути різним у студентів різних факультетів (отже, у вибірці факультети мають бути подані пропорційно своїй чисельності) і так далі.

Оскільки у об'єктів генеральної сукупності досліджуються кількісні ознаки, то можна замість розподілу цих об'єктів за ознакою розглядати розподіл деякої випадкової величини X. При цьому випробуванням буде випадковий вибір об'єкта із сукупності, а значення досліджуваної ознаки і буде значенням випадкової величини.

Оскільки на результат вимірювання в загальному випадку кожного разу впливають неконтрольовані чинники, то вибірку $x_1, x_2, \ldots, x_n$ можна розглядати як сукупність випадкових величин.

Якщо усі x_k незалежні і однаково розподілені, то вибірка називається стандартною. У цьому випадку усі ймовірнісні характеристики x_k однакові – зокрема, у них співпадають функції розподілу. Тому існує гіпотетична випадкова величина X із цією функцією розподілу. У такому разі прийнято говорити про вибірку із заданого закону розподілу (нормального, показового і т. ін.). Надалі, якщо не обумовлене інше, розглядаються лише стандартні вибірки.

1.2. Варіаційні ряди та їх графічні характеристики

Нехай отримана деяка вибірка – тобто внаслідок проведення n випробувань отримані числові дані $x_1, x_2, \ldots, x_n$, що характеризують досліджувану ознаку X.

Спостережувані (числові) значення $x_1, x_2, \ldots, x_n$ (елементи вибірки) називають **варіантами**. Послідовність варіантів, записаних у зростаючому порядку, називається простим **варіаційним рядом**.

Оскільки для дискретної випадкової величини (ВВ) деякі значення можуть повторюватися, то для кожного варіанта розраховують число його повторень.

Частотою n_k деякого варіанта називають кількість його повторень у вибірці. Відношення частоти n_k до обсягу вибірки n називають **відносною частотою** (позначають n_k^*): $n_k^* = \frac{n_k}{n}$.

Вочевидь, якщо у вибірці m різних варіантів, то:

$$\sum_{k=1}^{m} n_k = n \quad \text{и} \quad \sum_{k=1}^{m} n_k^* = 1.$$

Сукупність впорядкованих значень x_k і відповідних їм частот n_k, яку заносять у вигляді таблиці (матриці), найчастіше називають **дискретним варіаційним рядом**:

значення X	x_1	x_2	…	x_m
частоти n_k	n_1	n_2	…	n_m

Якщо обсяг вибірки n великий, а кожен варіант зустрічається рідко, то увесь діапазон значень вибірки розбивають на інтервали, що не перетинаються. Так само роблять, якщо випадкова величина ξ неперервна. Тоді одержують інтервальний (неперервний) варіаційний ряд.

Інтервальний варіаційний ряд – таблиця, де кожному інтервалу значень вибірки $[\alpha_{k-1}, \alpha_k)$ ставиться у відповідність кількість n_k значень, що потрапили до нього:

інтервали $[\alpha_{k-1}, \alpha_k)$	$[\alpha_0, \alpha_1)$	$[\alpha_1, \alpha_2)$	…	$[\alpha_{m-1}, \alpha_m)$
частоти n_k	n_1	n_2	…	n_m

Зазвичай інтервали мають однакову довжину.

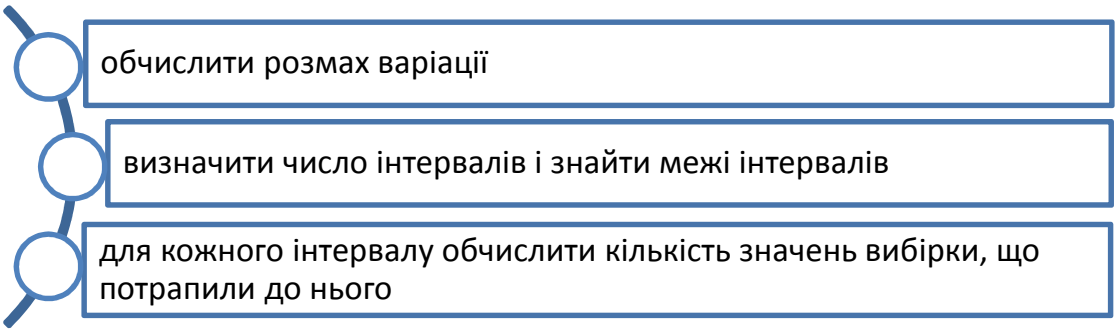
Щоб побудувати інтервальний варіаційний ряд за вибіркою, часто поступають таким чином:

1) визначають **розмах варіації** – різницю між максимальним і мінімальним значеннями вибірки: $R = x_{max} - x_{min}$;
2) задають число інтервалів – так, щоб ряд не був занадто громіздким, але і дозволяв виявити характерні властивості ВВ. Дослідним шляхом встановлено, що можна узяти значення $m = 1 + \log_2 n = 1 + 3{,}322 \lg n$. (Зауважимо, що m залишається скінченним і тоді, коли обсяг вибірки прагне до нескінченності). Довжина інтервалів дорівнює $h = \frac{R}{m}$;
3) за початок першого інтервалу приймають величину $\alpha_0 = x_{min} - \frac{h}{2}$, наступних – $\alpha_k = \alpha_{k-1} + h$ $(k = \overline{1,m})$;

4) для кожного інтервалу підраховують кількість значень вибірки, що потрапили у нього. Якщо значення потрапило на межу двох інтервалів, його відносять до попереднього інтервалу.

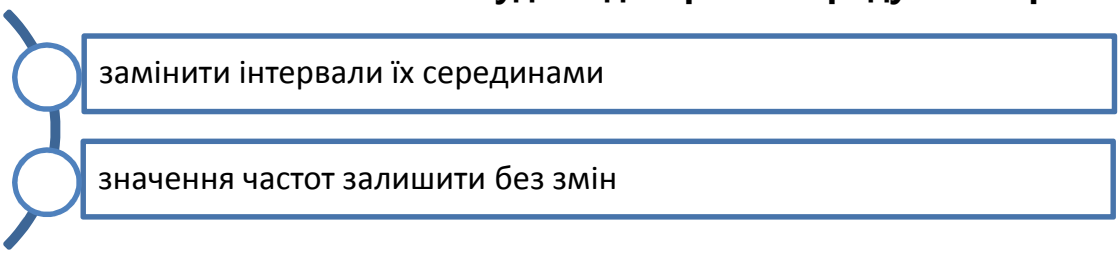
Таким чином, алгоритм побудови інтервального ряду за вибіркою зводиться до наступного.

Побудова інтервального ряду за вибіркою



Якщо вже є інтервальний варіаційний ряд, то перейти до дискретного не завдає труднощів: потрібно замінити інтервали числами – серединами цих інтервалів $x_k=\frac{\alpha_{k-1}+\alpha_k}{2}$ (а значення частот залишити без змін).

Побудова дискретного ряду за інтервальним



Приклад 1.3. Для аналізу кількості щоденних замовлень деякого товару була отримана вибірка: 37, 43, 43, 39, 37, 40, 43, 39, 34, 39, 40, 37, 40, 39, 40, 39, 37, 39, 40, 39. Побудувати дискретний варіаційний ряд.

Розв'язання.

Наведемо приклад побудови статистичного розподілу а) частот і б) відносних частот.

Обсяг вибірки $n=20$. Розташуємо усі значення за збільшенням: 34, 37, 37, 37, 37, 39, 39, 39, 39, 39, 39, 39, 40, 40, 40, 40, 40, 43, 43, 43. Залишимо лише варіанти, що не повторюються: 34, 37, 39, 40, 43.

а) обчислимо частоти тих варіант, що не повторюються. Наприклад, для значення $x_1=34$ $n_1=1$, для $x_2=37$ $n_2=4$ і т. д.

Тоді дискретний варіаційний ряд (як статистичний розподіл частот):

X	34	37	39	40	43
n	1	4	7	5	3

Контроль обчислень: 1+4+7+5+3 = 20.

б) Знайдемо відносні частоти, розділивши частоти на $n=20$: Наприклад, $n_1^* = \frac{1}{20} = 0{,}05$, $n_2^* = \frac{4}{20} = 0{,}2$ тощо.

Тоді можна побудувати дискретний варіаційний ряд як статистичний розподіл відносних частот:

X	34	37	39	40	43
n*	0,05	0,2	0,35	0,25	0,15

Контроль обчислень: 0,05 + 0,2 + 0,35 + 0,25 + 0,15 = 1.

Таким чином, вихідними для подальшого статистичного аналізу є дискретний або інтервальний варіаційний ряд. Мистецтво дослідника полягає в тому, щоб отримати максимум інформації, аналізуючи ці дані.

Зазвичай вибірку спочатку систематизують, групують, відкидають аномальні значення, перевіряють на однорідність тощо, а потім розпочинають аналіз даних за вибіркою. У цьому посібнику розглядаються такі задачі аналізу :

1) оцінка невідомих параметрів заданого закону розподілу (точкова або інтервальна);
2) перевірка правдоподібності гіпотез про закон розподілу;
3) оцінка статистичного зв'язку між випадковими величинами та ін.

Зазвичай аналіз починають з побудови полігону (багатокутника) частот або гістограми.

Полігон частот (*багатокутник частот*) – ламана, сполучаюча точки (x_1, n_1), (x_2, n_2), …, (x_m, n_m).

Гістограма – ступінчаста фігура, складена із прямокутників, основою яких є інтервали $[a_{k-1}, a_k)$, а висоти дорівнюють відповідним частотам n_k.

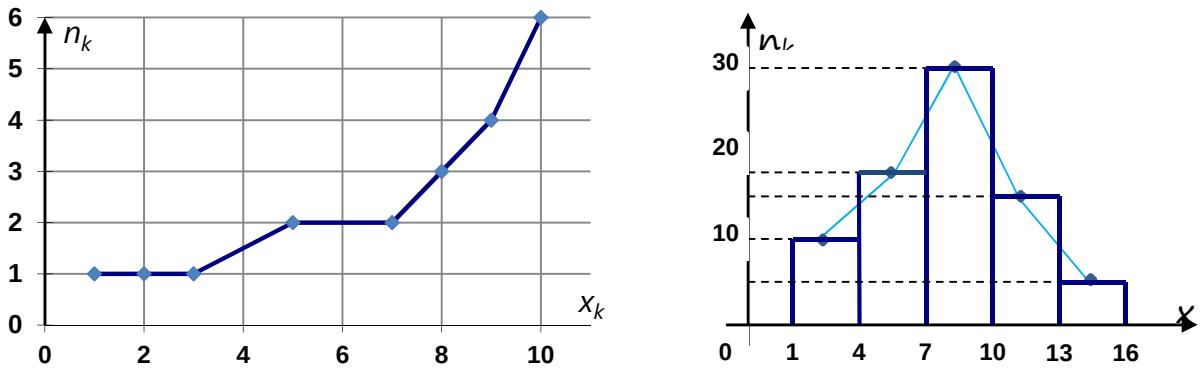


Рис. 1.1. а) полігон дискретного варіаційного ряду;
б) полігон і гістограма інтервального варіаційного ряду

Гістограму застосовують для відображення інтервального ряду. Полігон – і для дискретного, і для інтервального варіаційних рядів (для інтервального ряду лініями сполучають точки, значення $x_1, x_2, \ldots, x_n$ яких – середини інтервалів).

Графічне зображення варіаційного ряду дозволяє уявити приблизно закон розподілу випадкової величини.

1.3. Емпірична функція розподілу

Для вибіркової сукупності визначають емпіричну функцію розподілу.

Емпіричною функцією розподілу[1] називають функцію $F_n(x)$, що показує для кожного x відносну частоту тих спостережень, у яких значення вибірки менше x.

Таким чином, $F_n(x)=\frac{n_x^{(н)}}{n}$, де n – обсяг вибірки,

$n_x^{(н)}=\sum_{x_k<x} n_k$ – накопичена частота (сума частот значень вибірки, менших ніж x).

Наприклад, для дискретного варіаційного ряду

X	x_1	x_2	…	x_m
n	n_1	n_2	…	n_m

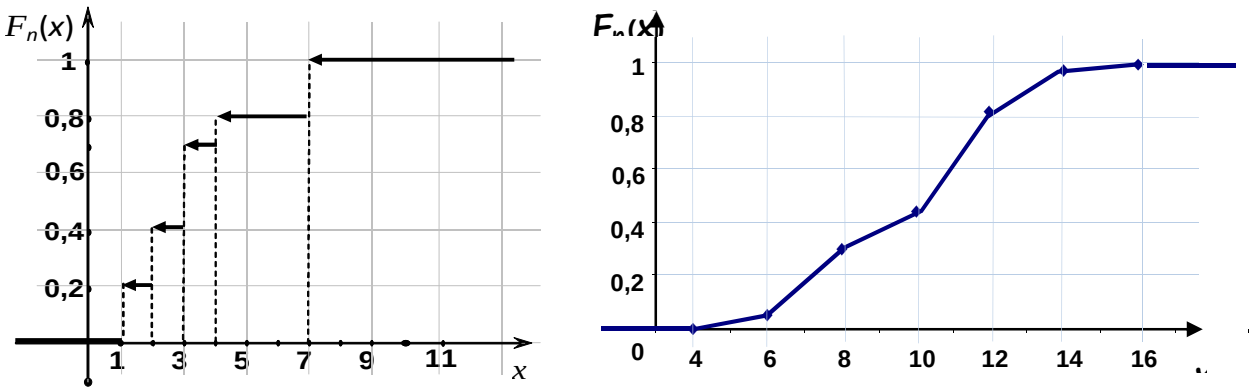
емпірична функція розподілу матиме вигляд:

$$F_n(x)=\begin{cases} 0 & \text{при } x\le x_1 \\ \frac{n_1}{n} & \text{при } x_1<x\le x_2 \\ \frac{n_1+n_2}{n} & \text{при } x_2<x\le x_3 \\ \frac{n_1+n_2+n_3}{n} & \text{при } x_3<x\le x_4 \\ \ldots\ldots & \\ \frac{n_1+n_2+\ldots+n_m}{n}=1 & \text{при } x>x_m \end{cases}$$

Для інтервального ряду емпіричну функцію розподілу можна побудувати лише приблизно. Для цього зазвичай обчислюють значення $F_n(x)$ на кінцях інтервалів. Щоб побудувати графік, відображають ці точки (накопичені частоти відповідають верхнім межам інтервалів, для нижньої межі першого інтервалу накопичена частота дорівнює нулю) і з'єднують їх прямими (іноді кривими) лініями.

---

[1] У зарубіжній літературі зазвичай позначають EDF, e.d.f. (*англ.* Empirical Distribution Function).



а) б)

Рис. 1.2. Емпірична функція розподілу: а) дискретного, б) інтервального ряду

Функцію розподілу $F(x)$ генеральної сукупності, що визначає ймовірність події $\xi<x$, називають **теоретичною функцією розподілу** .

На відміну від неї, емпірична функція розподілу $F_n(x)$ визначає відносну частоту події.

За теоремою Бернуллі (наслідок закону великих чисел), при достатньо великому числі незалежних випробувань відносна частота події скільки завгодно мало відрізняється від її ймовірності. Отже, можна використовувати емпіричну функцію розподілу $F_n(x)$ стандартної вибірки для наближеного подання теоретичної функції $F(x)$.

До того ж *властивості емпіричної функції* $F_n(x)$ співпадають із властивостями теоретичної функції розподілу $F(x)$:

1. Значення $F_n(x)$ належать інтервалу $[0; 1]$, тобто: $0 \le F_n(x) \le 1$.
2. $F_n(x)$ – неспадна функція, тобто з $x_1 \le x_2$ випливає, що $F_n(x_1) < F_n(x_2)$.
3. Якщо x_1 – найменший варіант, то $F_n(x) = 0$ при $x \le x_1$.
   Якщо x_m – найбільший варіант, то $F_n(x) = 1$ при $x > x_m$.
4. $F_n(x)$ неперервна зліва: $\lim\limits_{x \to x_0 - 0} F_n(x) = F_n(x_0)$.

1.4. Числові характеристики вибірки

Часто замість вивчення усіх елементів вибірки достатньо проаналізувати деякі їх числові характеристики. До основних із них відносять вибіркове середнє, вибіркову дисперсію, вибіркове середнє квадратичне відхилення.

Вибірковим середнім $\bar{x}$ називають середнє арифметичне варіантів вибіркової сукупності з урахуванням їхніх частот.

Якщо дані подано у вигляді дискретного варіаційного ряду $x_1, x_2, \ldots, x_n$ і усі значення різні, то

$$\bar{x} = \frac{x_1+x_2+\ldots+x_n}{n} = \frac{1}{n}\sum_{k=1}^{n} x_k \qquad [1.1]$$

Якщо значення $x_1, x_2, \ldots x_m$ мають частоти $n_1, n_2, \ldots, n_m$ (при цьому $\sum_{k=1}^{m} n_k = n$), то

$$\bar{x} = \frac{x_1 n_1+x_2 n_2+\ldots+x_m n_m}{n} = \frac{1}{n}\sum_{k=1}^{m} x_k n_k$$

Вибіркове середнє – основна характеристика розподілу вибірки, і надалі усереднювання величин буде використано багаторазово. Тому запровадимо для цієї операції назву – «риса» – і розглянемо її властивості (див. нижче п. 1.5 Операція «риса»).

Якщо дані подано у вигляді інтервального варіаційного ряду, то для обчислення вибіркової середньої або дисперсії необхідно спочатку перейти до дискретного варіаційного ряду.

Вибірковою дисперсією (дисперсією варіаційного ряду) називають середнє арифметичне квадратів відхилень варіантів від їхньої вибіркової середньої $\bar{x}$.

$$S_x^2 = \frac{1}{n}\sum_{k=1}^{n} (x_k - \bar{x})^2 \qquad [1.2]$$

Якщо значення $x_1, x_2, \ldots x_m$ мають частоти $n_1, n_2, \ldots, n_m$, ця формула набирає вигляду:

$$S_x^2 = \frac{1}{n}\sum_{k=1}^{m} (x_k - \bar{x})^2 n_k$$

На практиці для обчислення S_x^2 використовують таку властивість вибіркової дисперсії: вона дорівнює вибіркової середньої квадратів варіантів без квадрата їхньої вибіркової середньої:

$$S_x^2 = \frac{1}{n}\sum_{k=1}^{n} x_k^2 - \left(\frac{1}{n}\sum_{k=1}^{n} x_k\right)^2 = \overline{x^2} - (\bar{x})^2. \qquad [1.3]$$

Ця властивість S_x^2 випливає з її визначення:

$$S_x^2 = \frac{1}{n}\sum_{k=1}^{n} (x_k - \bar{x})^2 =$$

розкриємо квадрат двочлена, а потім розіб'ємо на три суми:

$$= \frac{1}{n}\sum_{k=1}^{n}\left(x_k^2 - 2x_k\bar{x} + (\bar{x})^2\right) = \frac{1}{n}\sum_{k=1}^{n}x_k^2 - 2\bar{x}\,\frac{1}{n}\sum_{k=1}^{n}x_k + (\bar{x})^2\frac{1}{n}n =$$

скористаємося для другого доданку визначенням вибіркового середнього:

$$= \frac{1}{n}\sum_{k=1}^{n}x_k^2 - 2\bar{x}\,\bar{x} + (\bar{x})^2 = \overline{x^2} - (\bar{x})^2$$

Таким чином, ми записали вираз для вибіркової дисперсії за допомогою тільки операції «риса».

Вибірковим середнім квадратичним відхиленням S_x називають арифметичне значення кореня квадратного з вибіркової дисперсії:

$$S_x = \sqrt{S_x^2}$$

Вибіркові середнє, дисперсія, середнє квадратичне відхилення залежать від значень вибірки, тому вони самі є випадковими величинами.

Приклад 1.4. Нижче наведено результати виміру середньоденної температури у Харкові в 2017 році. Знайти вибіркове середнє і вибіркову дисперсію температури.

Температура	[-20, -15)	[-15, -10)	[-10, -5)	[-5, 0)	[0, 5)	[5, 10)	[10, 15)	[15, 20)	[20, 25)	[25, 30]
К-ть днів	4	6	15	41	65	69	37	60	43	25

Розв'язання.

Перейдемо до дискретного варіаційного ряду, розрахувавши середини інтервалів:

Температура	[-20, -15)	[-15, -10)	[-10, -5)	[-5, 0)	[0, 5)	[5, 10)	[10, 15)	[15, 20)	[20, 25)	[25, 30]
Середина інтервалу	-17,5	-12,5	-7,5	-2,5	2,5	7,5	12,5	17,5	22,5	27,5
К-ть днів	4	6	15	41	65	69	37	60	43	25

За формулою вибіркового середнього:

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{m}x_k n_k = \frac{1}{365}((-17{,}5)\cdot 4 + (-12{,}5)\cdot 6 + (-7{,}5)\cdot 15 + (-2{,}5)\cdot 41 + 2{,}5\cdot 65 +$$

$$+ 7{,}5\cdot 69 + 12{,}5\cdot 37 + 17{,}5\cdot 60 + 22{,}5\cdot 43 + 27{,}5\cdot 25) \approx 9{,}56.$$

Скористаємося властивістю вибіркової дисперсії: $S_x^2 = \overline{x^2} - (\bar{x})^2 =$

$$= \frac{1}{365}((-15)^2\cdot 4 + (-10)^2\cdot 6 + (-5)^2\cdot 15 + 0^2\cdot 41 + + 5^2\cdot 65 + 10^2\cdot 69 + 15^2\cdot 37 +$$

$$+ 20^2\cdot 60 + 25^2\cdot 43 + 30^2\cdot 25) - 9{,}56^2 = 107{,}01.$$

Відповідь: вибіркове середнє 9,56; вибіркова дисперсія 107,01.

1.5. Операція «риса»

Уведемо операцію «риса»:

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

і розглянемо її властивості. Ясно, що

$$\overline{x+y} = \frac{1}{n}\sum_{k=1}^{n}(x_k+y_k) \qquad \overline{x^2} = \frac{1}{n}\sum_{k=1}^{n} x_k{}^2$$

$$\overline{xy} = \frac{1}{n}\sum_{k=1}^{n} x_k\, y_k \qquad \overline{(x-\bar{x})^2} = \frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x})^2$$

Властивості операції «риса».

1) $\overline{C} = C$, де C – стала

2) $\overline{Cx} = C\bar{x}$

3) $\bar{\bar{x}} = \bar{x}$

4) $\overline{x+y} = \bar{x}+\bar{y}$

5) $\left|\overline{xy} - \bar{x}\,\bar{y}\right| \le \sqrt{\overline{(x-\bar{x})^2}\cdot\overline{(y-\bar{y})^2}}$

Доказати властивостей операції «риса».

▲ Перші 4 властивості очевидні. Доведемо 5-у властивість.

Розглянемо $z = (x-\bar{x})+\lambda(y-\bar{y})$.

Через властивості (2) и (4) $\bar{z} = 0$

Обчислимо z^2:

$$z^2 = (x-\bar{x})^2 + 2\lambda(x-\bar{x})(y-\bar{y}) + \lambda^2(y-\bar{y})^2$$

Застосуємо операцію «риса»:

$$\overline{z^2} = \overline{(x-\bar{x})^2} + 2\lambda\overline{(x-\bar{x})(y-\bar{y})} + \lambda^2\overline{(y-\bar{y})^2}.$$

Очевидно, що $\overline{z^2} \ge 0$ для будь-якого $\lambda \in R$.

Розглянемо $\overline{z^2}$ як квадратичний тричлен відносно λ. Оскільки його значення ненегативні для будь-якого λ і його коефіцієнт при λ^2 теж, то дискримінант тричлена має бути $D \le 0$. Запишемо дискримінант:

$$D = 4\cdot(\overline{xy} - \bar{x}\,\bar{y})^2 - 4\cdot\overline{(x-\bar{x})^2}\cdot\overline{(y-\bar{y})^2} \le 0$$

Отже,

$$\left|\overline{xy} - \bar{x}\,\bar{y}\right| \le \sqrt{\overline{(x-\bar{x})^2}\cdot\overline{(y-\bar{y})^2}}$$ ■

Зауваження. Для величини $\overline{(x-\bar{x})^2}$ легко отримати подання

$$\overline{(x-\bar{x})^2} = \overline{x^2} - \bar{x}^2$$

Дійсно, $\overline{(x-\bar{x})^2} = \overline{(x^2 - 2x\bar{x} + \bar{x}^2)} = \overline{x^2} - 2\bar{x}^2 + \bar{x}^2 = \overline{x^2} - \bar{x}^2$.

Перевірте, чи засвоїли ви такі **ключові поняття:**

- ♣ генеральна сукупність
- ♣ вибіркова сукупність (вибірка)
- ♣ обсяг выбірки
- ♣ варіаційний ряд
  - – дискретний
  - – інтервальний
- ♣ полігон частот
- ♣ гістограма
- ♣ емпірична функція розподілу
- ♣ теоретична функція розподілу
- ♣ вибіркове середнє
- ♣ вибіркова дисперсія
- ♣ вибіркове середнє квадратичне відхилення
- ♣ операція «риса»

Питання для самоконтролю

1. Які вимоги до об'єктів, що становлять генеральну сукупність? Які ознаки цих об'єктів можуть вивчатися в математичній статистиці – кількісні, якісні, дискретні, неперервні?
2. Які вимоги до формування вибірки?
3. У чому відмінності дискретного і інтервального варіаційних рядів? Чи можна отримати який-небудь один з іншого? Яким чином?
4. Назвіть і опишіть графічні характеристики варіаційних рядів.
5. У чому відмінність між теоретичною і емпіричною функціями розподілу?
6. Чому емпіричну функцію розподілу можна використовувати для оцінки теоретичної?
7. Які з перерахованих характеристик є сталими, а які – випадковими величинами: дисперсія, вибіркова дисперсія, вибіркове середнє, мат. очікування?
8. Для кожного з законів розподілу вкажіть його параметри і числові характеристики, з якими ці параметри пов'язані: нормальний, показовий, рівномірний, Бернуллі, Пуассона.

Глава 2

Оцінки параметрів розподілу

Основні питання:

- *оцінки точкові та інтервальні*
- *оцінки незміщені, обґрунтовані, ефективні*
- *емпірична функція розподілу та її властивості*
- *точкові оцінки характеристик ВВ*
- *метод максимальної правдоподібності*
- *метод моментів*
- *інтервальні оцінки характеристик ВВ*

При розв'язанні практичних задач у дослідника, як правило, наявна лише сукупність результатів спостережень (вибірка). У загальному випадку зазвичай невідомий ані точний закон розподілу ймовірностей спостережуваних величин, ані його параметри. Дослідникові належить зробити висновки про них на підставі лише вибіркових результатів.

Розглянемо спочатку випадок, коли з яких-небудь міркувань вигляд розподілу відомий. Наприклад, при вивченні явища, на яке впливає безліч слабо залежних малозначимих чинників, можна на підставі центральної граничної теореми припускати нормальний закон розподілу. Щоб конкретизувати цей закон, потрібно знати його параметри. Так, для нормального закону розподілу це параметри a і σ, для показового – λ тощо.

Таким чином, належить розв'язати задачу: на підставі вибіркової сукупності отримати наближені значення параметрів відомого закону розподілу.

Оскільки для багатьох законів розподілу випадкової величини їхні параметри виражаються через числові характеристики цієї ВВ (математичне очікування, дисперсія, середнє квадратичне відхилення та ін.), то проводять оцінку цих числових характеристик. Наприклад, визначити параметри нормального закону розподілу – означає отримати оцінку відповідно математичного очікування і середнього квадратичного відхилення.

2.1. Поняття про оцінку параметрів. Види оцінок

Нехай маємо стандартну вибірку $x_1, x_2, \ldots x_n$. Значення вибірки можна розглядати як незалежні однаково розподілені випадкові величини $x_1, x_2, \ldots x_n$, і отже, їх розподіл збігається з розподілом генеральної сукупності $F(x)$.

Нехай для функції $F(x)$ відомий її вигляд, але невідомі параметри $\Theta_1, \Theta_2, \ldots \Theta_m$, що її визначають: $F(x, \Theta_1, \Theta_2, \ldots \Theta_m)$. Тоді виникає задача оцінці цих невідомих параметрів за вибіркою.

Оцінкою $\hat{\Theta}$ параметра Θ називають його наближене значення, залежне від вибіркових даних $x_1, x_2, \ldots, x_n$.

Розрізняють два види статистичних оцінок параметрів – точкові та інтервальні.

Точковою називають оцінку, яка визначається одним числом $\hat{\Theta}$.

Оскільки значенням $x_1, x_2, \ldots x_n$ повинне відповідати одне значення $\hat{\Theta}$, це означає, що точкова оцінка може бути подана як функція від цих випадкових значень: $\hat{\Theta}=\hat{\Theta}_n(x_1, x_2, \ldots x_n)$. Такі функції, залежні від результатів спостережень, називають статистичними оцінками, або *статистиками*. Вочевидь, $\hat{\Theta}$ залежить від обсягу вибірки n.

Знайти оцінку – означає знайти функцію від $x_1, x_2, \ldots, x_n$, яка при підстановці до неї конкретних результатів спостереження і дає наближене значення Θ.

Оскільки оцінка $\hat{\Theta}_n$ залежить від випадкових значень, то вона є випадковою величиною.

Якщо вибірка малого обсягу, точкова оцінка може суттєво відрізнятися від оцінюваного параметра. У цьому випадку доцільно використовувати інтервальні оцінки.

Інтервальною називають оцінку, яка визначається двома числами Θ_1 та Θ_2 – кінцями інтервалу, що містить із заданою ймовірністю γ оцінюваний параметр Θ. Межі інтервалу – функції від вибіркових даних $x_1, x_2, \ldots x_n$.

 Інтервал (Θ_1,Θ_2), що містить справжнє значення параметра Θ з заданою ймовірністю γ, називають **довірчим інтервалом** для параметра Θ: $P(\Theta_1<\Theta<\Theta_2)=\gamma$.

Імовірність γ називають **довірчою ймовірністю** (або *надійністю оцінки, рівнем довіри*), а число $\alpha=1-\gamma$ – **рівнем значущості** .

Оскільки обидві межі довірчого інтервалу Θ_1, Θ_2 визначають за наслідками спостережень, то вони є випадковими величинами. Тому говорять, що довірчий інтервал містить оцінюваний параметр Θ з імовірністю γ. Зазвичай довірчу ймовірність γ задають заздалегідь як число, близьке до одиниці, – найчастіше використовують значення $0,90$; $0,95$; $0,99$; $0,999$.

2.2. Точкові оцінки параметрів

Часто існує декілька статистик для оцінки одного і того ж параметра. Задача – вибрати таку функцію $\hat{\Theta}_n$, щоб отримувані оцінки були б найбільш близькі до значення Θ, а зі збільшенням обсягу вибірки точність оцінки зростала. Тому для вибору найкращої оцінки висувають низку вимог до оцінок.

Щоб точкова оцінка $\hat{\Theta}$ була близька до значення параметра Θ, бажано, щоб вона мала властивості незміщеності, обґрунтованості й ефективності. Це основні властивості, які характеризують якість оцінок і дозволяють вибрати якнайкращу.

 Оцінку $\hat{\Theta}$ називають **незміщеною**, якщо її математичне очікування дорівнює оцінюваному параметру Θ при будь-якому обсязі вибірки:

$$M\hat{\Theta}=\Theta \qquad [2.1]$$

Якщо рівність не виконується, оцінку називають *зміщеною*, а різницю $M\hat{\Theta}-\Theta$ називають *зміщенням* або *систематичною похибкою*.[2]

Незміщеність оцінки означає, що вона не накопичує односторонніх помилок – у бік збільшення або зменшення. Тому намагаються використовувати саме такі оцінки параметрів (якщо вони існують). Незміщених оцінок може бути декілька, а може не існувати..

[2] Слабкіша умова вимагає, щоб із ростом обсягу вибірки $n\to\infty$ математичне очікування оцінки збігалося до справжнього значення параметра – такі оцінки називають *асимптотично незміщеними*.

Оцінку $\hat{\Theta}$ називають **обґрунтованою**[3], якщо при необмеженому збільшенні числа спостережень n вона наближається до оцінюваного параметра Θ з імовірністю, скільки завгодно близькою до одиниці:

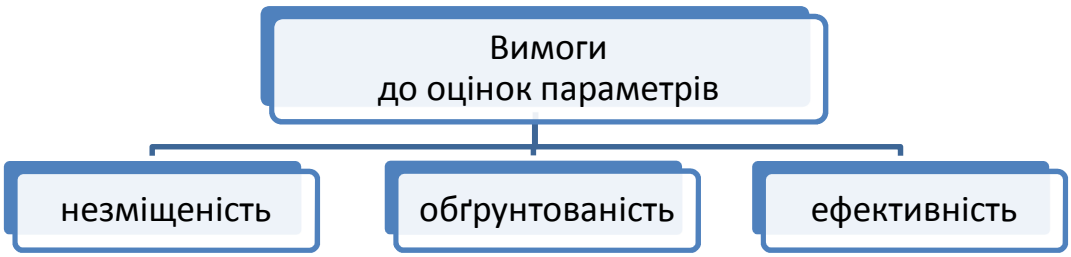
$$\lim_{n\to\infty} P\left(\left|\hat{\Theta}-\Theta\right|<\varepsilon\right)=1 \qquad \forall \varepsilon>0 \qquad [2.2]$$

Обґрунтованість оцінки означає, що чим більше обсяг вибірки, тим ближче оцінка $\hat{\Theta}$ наближається до оцінюваного параметра Θ. У такому разі говорять: «оцінка $\hat{\Theta}$ **збігається за ймовірністю** до Θ». Обґрунтованих оцінок також може бути декілька.

Часто може бути знайдено декілька обґрунтованих, а іноді і незміщених оцінок одного і того ж параметра. Для того, щоб порівняти їх між собою, вибирають деяку функцію ризику, за допомогою якої порівнюють між собою відхилення оцінок від справжнього значення. Кращою вважається та оцінка, для якої функція ризику набуває найменшого значення. Для незміщених оцінок в якості функції ризику зазвичай розглядають їхню дисперсію – для неї існує мінімальне значення.

Оцінку $\hat{\Theta}$ називають **ефективною**, якщо вона має найменшу дисперсію серед усіх інших незміщених оцінок параметра Θ, обчислених за вибіркою того ж обсягу n.

Іншими словами, ефективна оцінка – найточніша серед названих. Проте така існує не завжди[4] – у цьому випадку зазвичай можна вибрати ефективнішу (тобто з меншою дисперсією).



При здобутті оцінок параметрів прагнуть до того, щоб вони були незміщеними, обґрунтованими й ефективними. Але на практиці не завжди вдається задовольнити усі ці три умови одночасно.

Властивість обґрунтованості є обов'язковою: необґрунтовані оцінки практично не використовуються.

---

[3] Обґрунтована, конзистентна, слушна оцінка (*укр.*).

[4] Слабкіша умова вимагає, щоб із ростом обсягу вибірки $n \to \infty$ дисперсія незміщеної оцінки збігалась до мінімально можливого значення – такі оцінки називаються *асимптотично ефективними*.