ЛЕКЦІЯ 16

ВИБІРКОВИЙ МЕТОД

16.1. Генеральна сукупність та вибірка. Варіаційний ряд. Статистичний розподіл вибірки

Математична статистика — наука про методи збору, систематизації, обробки і використання для наукових і практичних висновків статистичних даних, здобутих у результаті випробувань із випадковими наслідками. Ці дані є можливими значеннями деякої випадкової величини (досліджуваної ознаки) X, наприклад похибки вимірювань фізичної величини, відхилення розмірів виробу від стандарту, процента зайнятості крісел на авіарейсах, прибутку підприємства тощо.

При цьому досліджувані ознаки можуть бути дискретними (набувати окремих ізольованих значень) або неперервними (набувати будь-яких можливих значень з деякого інтервалу).

При дослідженні ознаки X зібрати та обробити всі її можливі значення здебільшого буває неможливо через їх занадто великий об'єм, а також у випадках, коли, наприклад, аналізуючи відповідність стандарту геометричних розмірів виробів, доводиться виконувати масові вимірювання або, аналізуючи міцність виробів, руйнувати їх. Тому з усієї множини N даних про ознаку X, яка називається *генеральною сукупністю* об'ємом n, випадковим чином відбирається досить представницька (репрезентативна) її частина, що містить n даних і називається *вибірковою сукупністю*, або *вибіркою* об'єму n. При цьому об'єм вибірки n має бути суттєво меншим за об'єм генеральної сукупності N, але дані вибірки мають достатньо повно відбивати властивості ознаки X генеральної сукупності.

Різні можливі значення x_i ознаки X, які потрапили до вибірки, називаються *варіантами*, а система варіант $x_1, x_2, \dots, x_k$, розміщена в зростаючому порядку, називається *варіаційним рядом*.

Якщо варіанти $x_1, x_2, \dots, x_k$ спостерігаються у вибірці відповідно $n_1, n_2, \dots, n_k$ раз, то значення $n_1, n_2, \dots, n_k$ називаються *частотами варіант*, і для них виконується умова $\sum_{i=1}^{k} n_i = n, \left(i = \overline{1,k}\right)$, а відношення

$$\omega_1 = \frac{n_1}{n}, \omega_2 = \frac{n_2}{n}, \dots, \omega_k = \frac{n_k}{n}$$

називаються *відносними частотами* варіант, і для них виконується умова $\sum_{i=1}^{k} \omega_i = 1$,

Таблиця, в першому рядку якої наведено впорядковані варіанти x_i (варіаційний ряд), а в другому — відповідні їм частоти n_i (або відносні частоти ω_i), називається дискретним статистичним розподілом вибірки (табл. 16.1 або табл. 16.2).

Таблиця 16.1. *Варіативний ряд*

x_i	x_1	x_2	…	x_{k-1}	x_k
n_i	n_1	n_2	…	n_{k-1}	n_k

Таблиця 16.2. *Варіативний ряд у відносних частотах*

x_i	x_1	x_2	…	x_{k-1}	x_k
ω_i	ω_1	ω_2	…	ω_{k-1}	ω_k

Якщо кількість варіант у вибірці надто велика або досліджувана ознака X є неперервною випадковою величиною, то будується інтервальний статистичний розподіл вибірки, аналогічний дискретному, у першому рядку якого замість дискретного варіаційного ряду записується *інтервальний ряд*

$$[x_1; x_2), [x_2; x_3), \ldots, [x_l; x_{l+1}).$$

Кількість інтервалів m для вибірки обсягом n орієнтовно вважається такою, що дорівнює $\sqrt{n}$ (із округленням до цілого значення), а розмір (крок) h інтервалу обчислюється за формулою (16.1):

$$h = \frac{x_k - x_1}{l}, \qquad (16.1)$$

де x_1 і x_k — відповідно найменша і найбільша варіанти у вибірці, l – кількість інтервалів.

Частоти n_i^* варіант, які потрапили в кожний частинний інтервал $[x_i; x_{i+1})$, $(i = \overline{1, l})$ записуються в другий рядок табл. 16.3.

Таблиця 16.3. *Інтервальний статистичний ряд*

$[x_i; x_{i+1})$	$[x_1; x_2)$	$[x_2; x_3)$	…	$[x_l; x_{l+1})$
n_i	n_1	n_2	…	n_l

Якщо межа двох частинних інтервалів збігається зі значенням варіанти, то частота цієї варіанти або ділиться порівну між цими двома інтервалами, або в усіх таких випадках цілком заноситься у правий (або лівий) від варіанти інтервал.

Правило Стерджеса — емпіричне правило визначення оптимальної кількості інтервалів, на які розбивається діапазон зміни випадкової величини, що спостерігається, при побудові гістограми щільності її розподілу. Названо на ім'я американського статистика Герберта Стерджеса (Herbert Arthur Sturges, 1882-1958).

Кількість інтервалів визначається як:

$$l = 1 + log_2(n), \qquad (16.2)$$

де n – загальна кількість спостережень величини, $log_2(n)$ – логарифм за основою 2.

Часто зустрічається записаним через десятковий логарифм:

$$l = 1 + 3{,}322 \cdot lg(n), \qquad (16.3)$$

де n – загальна кількість спостережень величини, $lg(n)$ – логарифм за основою 10.

Також зустрічається записаним через натуральний логарифм:

$$l = 1 + 3{,}322 \cdot ln(n). \qquad (16.4)$$

де n – загальна кількість спостережень величини, $ln(n)$ – натуральний логарифм.

Приклад 16.1. *У результаті вибіркового аналізу добової виручки авіакомпанії дістали вибірку обсягом* $n = 40$ (*млн грн*):

0,87 0,94 0,98 0,90 0,90 0,87 0,85 0,87
0,90 0,94 0,87 0,87 0,82 0,90 0,94 0,90
0,85 0,85 0,87 0,94 0,81 0,82 0,87 0,97
0,90 0,94 0,85 0,81 0,87 0,85 0,90 0,82
0,98 0,90 0,94 0,82 0,97 0,81 0,85 0,87

Скласти: ***а)*** *варіаційний ряд*; ***б)*** *інтервальний розподіл.*

Розв'язання. а) Випишемо різні значення варіант, які потрапили у вибірку:

0,87; 0,94; 0,98; 0,90; 0,85; 0,82; 0,81; 0,97.

Розмістивши їх у порядку зростання, дістанемо дискретний варіаційний ряд:

0,81; 0,82; 0,85; 0,87; 0,90; 0,94; 0,97; 0,98.

Обчислимо частоту кожної варіанти із варіаційного ряду і складемо таблицю, відповідну табл. 16.1. Дістанемо дискретний статистичний розподіл вибірки (табл. 16.4).

Таблиця 16.4. *Варіативний ряд*

x_i	0,81	0,82	0,85	0,87	0,90	0,94	0,97	0,98
n_i	3	4	6	9	8	6	2	2

б) За обсягом вибірки $n = 40$ визначаємо за формулою (16.2) орієнтовну кількість $m = 6$ частинних інтервалів в інтервальному статистичному розподілі, а за формулою (16.1) обчислюємо крок інтервалу

$$h = \frac{0{,}98 - 0{,}81}{6} = 0{,}03$$

Тоді інтервальний варіаційний ряд запишеться у вигляді: $[0{,}81; 0{,}84)$, $[0{,}84; 0{,}87)$, $[0{,}87; 0{,}90)$, $[0{,}90; 0{,}93)$, $[0{,}93; 0{,}96)$, $[0{,}96; 0{,}99)$.

Тоді інтервальний статистичний розподіл вибірки набере вигляду (табл. 16.5):

Таблиця 16.5. *Інтервальний варіативний ряд*

$[x_i; x_{i+1})$	$[0{,}81; 0{,}84)$	$[0{,}84; 0{,}87)$	$[0{,}87; 0{,}90)$	$[0{,}90; 0{,}93)$	$[0{,}93; 0{,}96)$	$[0{,}96; 0{,}99)$
n_i^*	7	6	9	8	6	4

Якщо частоти варіант, які співпали з межами інтервалів 0,87 та 0,90 розподілити порівну між суміжними інтервалами (табл. 16.6), то:

Таблиця 16.6. *Інтервальний варіативний*

$[x_i; x_{i+1})$	$[0{,}81; 0{,}84)$	$[0{,}84; 0{,}87)$	$[0{,}87; 0{,}90)$	$[0{,}90; 0{,}93)$	$[0{,}93; 0{,}96)$	$[0{,}96; 0{,}99)$
n_i^*	7	10	9	4	6	4

Відповідь. а) табл. 16.4; б) табл. 16.5.

16.2. Полігон та гістограма. Емпірична функція розподілу

Для створення наочного уявлення про статистичні розподіли застосовуються полігон та гістограма частот.

Полігоном частот (або *відносних частот*) називається ламана лінія, відрізки якої сполучають на площині точки з координатами (x_i, n_i) (або (x_i, ω_i)). Полігон застосовується також для графічного зображення інтервальних статистичних розподілів вибірки. У цьому випадку за абсциси x_i точок беруться центри частинних інтервалів.

Гістограма застосовується для графічного зображення інтервальних статистичних розподілів. Для побудови гістограми частот (або відносних частот) на осі абсцис відкладаються відрізки, що дорівнюють довжині (кроку) h частинних інтервалів, і на цих відрізках як на основах будуються прямокутники з висотами $\frac{n_i^*}{h}$ (або $\frac{\omega_i^*}{h} = \frac{n_i^*}{n \cdot h}$).

Оскільки величини $\frac{n_i^*}{h}\left(\frac{\omega_i^*}{h}\right)$ є щільностями частот (відносних частот) на відповідних інтервалах, то при достатньо великих обсягах вибірки n (і відповідно малих h) гістограма може бути достатньо близьким статистичним аналогом щільності ймовірності досліджуваної ознаки X у генеральній сукупності.

Приклад 16.2. *Побудувати полігон відносних частот для дискретного статистичного розподілу* (*табл.* 16.2) *та гістограму відносних частот для інтервального розподілу* (*табл.* 16.3).

Розв'язання. Обчислимо відносні частоти $\omega_i^* = \frac{n_i}{n}$ для дискретного розподілу і щільності відносних частот $\frac{\omega_i^*}{h} = \frac{n_i^*}{n \cdot h}$ для інтегрального розподілу при $n = 40$, $h = 0{,}03$. Дістанемо відповідно розподіли (табл. 16.7 і 16.8).

Таблиця 16.7. *Варіативний ряд у відносних частотах*

x_i	0,81	0,82	0,85	0,87	0,90	0,94	0,97	0,98
ω_i^*	0,075	0,10	0,15	0,225	0,20	0,15	0,05	0,05

Таблиця 16.8. *Інтервальний варіативний ряд*

$[x_i; x_{i+1})$	[0,81; 0,84)	[0,84; 0; 87)	[0,87; 0,90)	[0,90; 0,93)	[0,93; 0,96)	[0,96; 0,99)
$\frac{\omega_i^*}{h}$	5,83	8,33	7,5	3,33	5	3,33

Полігон та гістограму відносних частот побудовано відповідно на рис. 16.1 і 16.2.

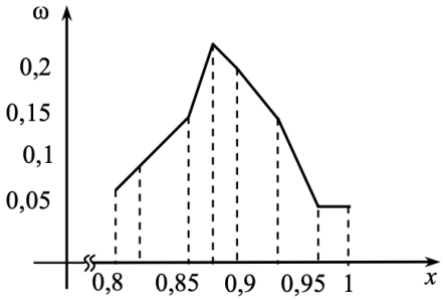

Рис. 16.1. Полігон

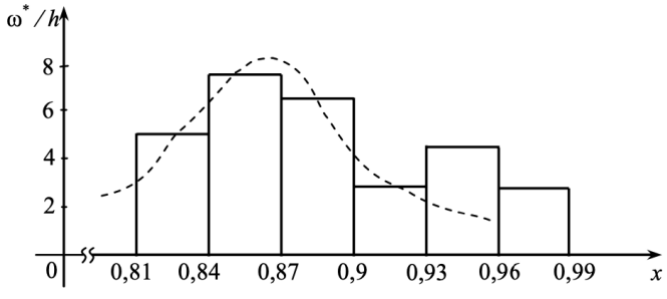

Рис. 16.2. Гістограма

Відповідь. Рис. 16.1 і 16.2.

За загальним виглядом гістограми можна скласти певне уявлення про графік щільності розподілу ознаки X у генеральній сукупності, наприклад припустити, що він має вигляд, зображений на рис. 16.2 пунктирною лінією (при згладженні деяких сплесків частот, викликаних, можливо, малим обсягом вибірки).

Розподіл ознаки у варіаційному ряду за накопиченими частотами (частинами) зображується за допомогою кумуляти. Кумулята або кумулятивна крива, на відміну від полігону, будується за накопиченими частотами або частотами. У цьому на осі абсцис поміщають значення ознаки, але в осі ординат — накопичені частоти чи частоти.

Кумулята — графічне порівняння двох або більше варіаційних розподілів з рівними чи нерівними інтервалами. Вона будується за кумулятивним розподілом накопичених частот, при цьому використовуються праві кінці інтервалів (рис. 16.3).

Огіва — це різновид кумулятивного розподілу. Вона є дзеркальним відображенням кумуляти. На осі ординат відкладаємо межі інтервалів, по осі абсцис — накопичені частоти (рис. 16.4).

Приклад 16.3. *Для наступних статистичних даних* (*табл.* 16.9) *побудувати кумуляту та огіву.*

Таблиця 16.9. *Вхідні дані*

Вибірка	Частота	Накопичувана частота
0,1	0	0
0,3	5	5
0,5	6	11
0,7	18	29
0,9	16	45
1,1	12	57
1,3	10	67
1,5	1	68
1,7	2	70

Розв'язання. На рис. 16.3 та 16.4 наведено відповідно графіки кумуляти та огіви.

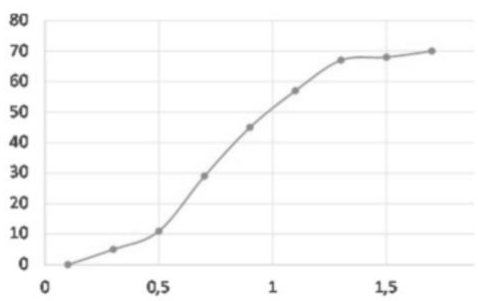

Рис. 16.3. Кумулята

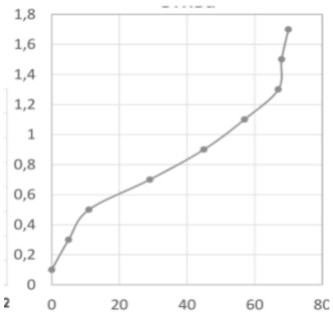

Рис. 16.4. Огіва

Відповідь. Рис. 16.3 і 16.4.

Перейдемо до розгляду емпіричної функції розподілу, яка утворюється за даними вибірки і є наближеною оцінкою реальної або так званої теоретичної функції розподілу $F(x)$ ознаки X генеральної сукупності.

Емпіричною функцією статистичного розподілу називається функція $F^*(x)$ яка для кожного значення x дорівнює відносній частоті події $X < x$, тобто (16.5):

$$F^*(x) = \frac{n_{x_i}}{n}, \tag{16.5}$$

де n_{x_i}— сумарна частота всіх варіант x_i , менших за x.

Для дискретного статистичного розподілу (табл. 16.1) функція $F^*(x)$ набуває таких значень:

$F^*(x) = 0$, при $x \le x_1$, оскільки в цьому випадку немає варіант, менших за x;

$F^*(x) = \frac{n_1}{n}$, при $x_1 < x \le x_2$;

$F^*(x) = \frac{n_1+n_2}{n}$, при $x_2 < x \le x_3$;

.....

$F^*(x) = \frac{n_1+n_2+\cdots+n_{k-1}}{n}$, при $x_{k-1} < x \le x_k$;

$F^*(x) = \frac{n_1+n_2+\cdots+n_k}{n} = 1$, при $x > x_k$, оскільки в цьому випадку додаються всі варіанти.

Графік функції $F^*(x)$ будується аналогічно графіку функції розподілу випадкової величини. Для неперервної ознаки X генеральної сукупності емпірична функція будується за інтервальним статистичним розподілом. Із цією метою обчислюються значення функції $F^*(x)$ у кількох точках, за які найчастіше беруть межі інтервалів. Знайдені точки $(x_i, F^*(x_i))$ наносять на графік і сполучають прямолінійними відрізками.

Якщо x_1 — ліва межа першого інтервалу, а x_{m+1} — права межа останнього інтервалу в інтервальному варіаційному ряду, то $F^*(x) = 0$ при $x \le x_1$ і $F^*(x) = 1$ при $x > x_{m+1}$.

Приклад 16.4. *Побудувати емпіричну функцію розподілу $F^*(x)$ за дискретним* (*табл.* 16.2) *та інтервальним* (*табл.* 16.3) *статистичними розподілами вибірки.*

Розв'язання. Для обчислення значень емпіричної функції скористаємось дискретним статистичним розподілом із відносними частотами ω_i (табл. 16.5), який знайдено у прикладі 16.2. Значення функції $F^*(x)$ обчислюються простим нагромадженням відносних частот:

$$F^*(x)=\begin{cases}0, \text{при } x\le 0{,}81,\\ 0+0{,}075=0{,}075, \text{при } 0{,}81<x\le 0{,}82,\\ 0{,}075+0{,}1=0{,}175, \text{при } 0{,}82<x\le 0{,}85,\\ 0{,}175+0{,}15=0{,}325, \text{при } 0{,}85<x\le 0{,}87,\\ 0{,}325+0{,}225=0{,}55, \text{при } 0{,}87<x\le 0{,}90,\\ 0{,}55+0{,}2=0{,}75, \text{при } 0{,}90<x\le 0{,}94,\\ 0{,}75+0{,}15=0{,}9, \text{при } 0{,}94<x\le 0{,}97,\\ 0{,}9+0{,}05=0{,}95, 0{,}97<x\le 0{,}99,\\ 0{,}95+0{,}05=1, x>0{,}99.\end{cases}$$

Графік функції $F^*(x)$ подано на рис. 16.5 (східчаста розривна лінія).

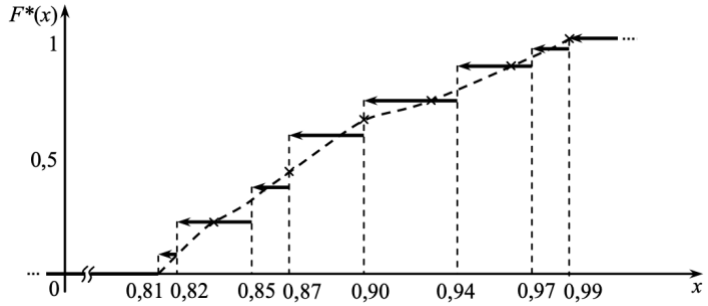

Рис.16.5. Емпірична функція розподілу

Для побудови функції $F^*(x)$ за інтервальним статистичним розподілом (табл. 16.3) обчислимо значення функції на межах інтервалів:

$x=0{,}81, F^*(x)=0$, оскільки варіант, менших за 0,81, у вибірці немає;

$x=0{,}84, F^*(x)=\frac{7}{40}=0{,}175$, оскільки варіанти, менші за 0,84, містяться в першому інтервалі, і їх кількість 7;

$x=0{,}87, F^*(x)=\frac{7+10}{40}=0{,}425$;

$x=0{,}90, F^*(x)=\frac{7+10+9}{40}=0{,}65$;

$x=0{,}93, F^*(x)=\frac{7+10+9+4}{40}=0{,}75$;

$x=0{,}96, F^*(x)=\frac{7+10+9+4+6}{40}=0{,}9$;

$x=0{,}99, F^*(x)=\frac{7+10+9+4+6+4}{40}=1$.

Нанесемо знайдені точки на графік рис. 16.5 і сполучимо їх відрізками прямої. Дістанемо графік функції $F^*(x)$, зображений на рис. 16.5 пунктирною лінією.

Відповідь. Рис. 16.5.

16.3. Мода та медіана статистичної вибірки

Більш докладно про моду та медіану розповідалося в лекції 6 паралельно з модою та медіаною випадкової величини.

Мода статистичної вибірки — це значення, яке найчастіше зустрічається у вибірці даних.

Особливості моди:

Мода визначається як значення або категорія з найбільшою частотою.

Вибірка може мати:
- Одну моду (унімодальний розподіл),
- Дві моди (бімодальний розподіл),
- Більше двох мод (мультимодальний розподіл).

Приклад 16.5. *Обчислити моду вибірки* X: 2,4,4,6,7,4,8,9.

Розв'язання. Значення 4 зустрічається тричі, тоді як інші значення з'являються лише раз.

$$Mo(X) = 4.$$

Відповідь. Мода вибірки дорівнює 4.

Переваги використання моди:
- Можна застосовувати як для кількісних, так і для якісних даних.
- Є стійкою до викидів (аномально великих або малих значень).

Недоліки моди:
- Мода може бути неунікальною (багато мод).
- Вона не завжди відображає центр розподілу, як це роблять середнє значення чи медіана.

Застосування моди:
- Аналіз категоріальних даних (наприклад, найбільш популярний товар).
- Визначення найбільш часто зустрічаючогося значення в наборах кількісних даних.

Мода є простою, але ефективною характеристикою центральної тенденції, особливо в практичних застосуваннях.

Медіана статистичної вибірки — це значення, яке ділить упорядковану вибірку навпіл: 50% спостережень менші або рівні медіані, а 50% — більші або рівні медіані.

Властивості медіани:
- Стійкість до викидів: на відміну від середнього арифметичного, медіана менш чутлива до аномально великих або малих значень.
- Використовується як міра центральної тенденції для кількісних даних.

Спосіб обчислення медіани:

1. Впорядкувати значення у вибірці у зростаючому порядку.
2. Залежно від кількості значень у вибірці:
- Непарна кількість значень: медіана — це значення, що стоїть посередині.
- Парна кількість значень: медіана — це середнє арифметичне двох центральних значень.

Приклад 16.6. *Обчислити медіану вибірок* X *та* Y: 3,5,7,9,11 *та* 2,4,6,8 *відповідно*.

Розв'язання. Вибірка X має непарну кількість значень, отже, центральне значення дорівнює 7.

$$Me(X) = 7.$$

Вибірка Y має парну кількість значень, отже, центральні значення дорівнюють 4 та 6.

$$Me(Y) = \frac{4+6}{2} = 5.$$

Відповідь. Мода вибірки X дорівнює 7; мода вибірки Y дорівнює 5.

Переваги медіани:

- Використовується, коли дані містять викиди або асиметричний розподіл.
- Зручна для рангових і порядкових даних.

Недоліки медіани:

- Може бути менш точною, якщо розподіл даних симетричний (у такому випадку середнє значення краще відображає центр).
- Не враховує всі значення у вибірці.

Застосування медіани:

- Економіка: наприклад, медіанний дохід показує типову заробітну плату, ігноруючи викиди з дуже великими значеннями.
- Медицина: аналіз тривалості життя пацієнтів.
- Соціальні дослідження: оцінка середніх показників на основі неповних або асиметричних даних.

Медіана — це зручна і наочна характеристика розподілу даних, особливо у випадках, коли середнє значення є занадто чутливим до аномалій.

16.4. Часові ряди

Часові ряди — це набір даних, що відображають значення певної змінної через регулярні проміжки часу. Важливою особливістю часових рядів є те, що порядок спостережень має значення, оскільки він відображає зміну показника у часі (рис. 16.6, 16.7).

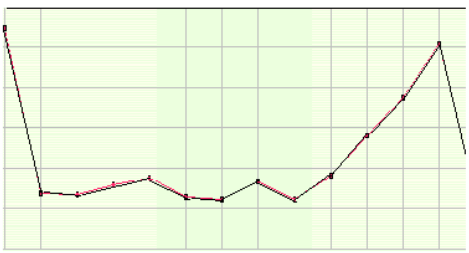

Рис. 16.6. Фрагмент часового ряду за сезонний період

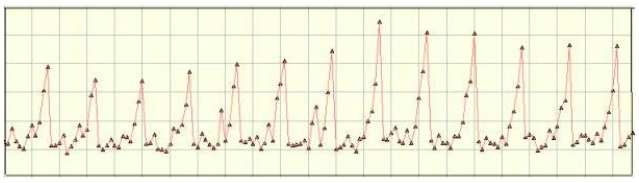

Рис. 16.7. Фрагмент часового ряду за 12 місяців

Основні поняття та компоненти часових рядів

Тренд (*Trend*) – це довгострокова тенденція, яка відображає загальний напрямок зміни даних (зростання, спад або стабільність) протягом тривалого часу. Наприклад, довгострокове зростання ВВП або поступове збільшення населення.

Сезонність (*Seasonality*) – це регулярні коливання, що повторюються з певною періодичністю (рік, квартал, місяць, тиждень) і пов’язані з певними

факторами. Наприклад, збільшення продажів під час свят або сезонні коливання температури.

Циклічність (*Cyclical Component*) – це довгострокові коливання, які можуть тривати кілька років і пов'язані з економічними циклами, наприклад, економічні підйоми та спади. Циклічні коливання часто важко передбачити і вони можуть мати непостійний період.

Шум (*Noise або Random Component*) – це випадкові коливання в часовому ряді, які не піддаються прогнозуванню і відображають вплив непередбачуваних факторів. Вони додають «шуму» до загальної картини і ускладнюють аналіз.

Аналіз часових рядів

Аналіз часового ряду здійснюється з метою:

- визначення природи ряду;
- прогнозування майбутніх значень ряду.

У процесі визначення структури і закономірностей часового ряду передбачається виявлення: шумів і викидів, тренду, сезонної компоненти, циклічної компоненти. Визначення природи тимчасового ряду може бути використано як своєрідна розвідка даних. Знання аналітика про наявність сезонної компоненти необхідно, наприклад, для визначення кількості записів вибірки, яке повинно брати участь в побудові прогнозу.

Шуми і викиди будуть детально обговорюватися нижче. Вони ускладнюють аналіз часового ряду. Існують різні методи визначення та фільтрації викидів, що дають можливість виключити їх з метою більш якісного Data Mining.

Основою для прогнозування (однієї з основних задач Data Mining) служить історична інформація, що зберігається в базі даних у вигляді часових рядів.

$$y_t = tr_t + s_t + c_t + r_t \text{ або } y_t = tr_t + s_t + r_t, \qquad (16.6)$$

де tr_t – тренд, s_t – сезонна складова, c_t – циклічна складова, r_t – випадкова складова (шум).

Приклади часових рядів

Економічні дані: Валовий внутрішній продукт (ВВП) за квартал, рівень безробіття щомісяця, індекс цін на акції щодня.

Фінансові дані: Щоденні ціни на акції, обсяги торгівлі, валютні курси.

Метеорологічні дані: Щоденні температури, місячні опади, швидкість вітру.

Соціальні дані: Щорічний приріст населення, рівень злочинності за місяць, кількість користувачів соціальних мереж за рік.

16.5. Перехід від часових рядів до статистичних вибірок

Для контролю якості прогнозів (зокрема курсу криптовалюти), необхідно сформувати датасет з наявної статистичної інформації (табл. 16.10).

Таблиця 16.10. *Значення курсу обраної криптовалюти за вказаний проміжок часу*

Момент часу	Реальне значення курсу	Прогнозоване значення курсу
…	…	…
t_1	x_1	y_1
t_2	x_2	y_2
…	…	…
t_k	x_k	y_k
…	…	…

В таблиці 16.10 наведено реальні курси обраної криптовалюти $x_1, x_2, \ldots, x_k$ в обраний проміжок часу $[t_1;t_k]$, які можна взяти з сайту крипто біржі Binance, прогнозовані курси криптовалюти $y_1, y_2, \ldots, y_k$ отримано за допомогою обраного алгоритму прогнозування.

Таким чином, реальні та прогнозовані курси криптовалюти представляють собою часові ряди $X = (x_1, x_2, \ldots, x_k)$ та $Y = (y_1, y_2, \ldots, y_k)$ відповідно за період часу $[t_1;t_k]$.

Найчастіше часовий ряд є послідовністю, взятою на рівновіддалених точках в часі, які йдуть одна за одною. Таким чином, він є послідовністю даних дискретного часу. Тобто, для аналізу часового ряду параметр часу виступає одним із визначальних чинників. Це відрізняє часовий ряд від звичайної випадкової вибірки, де індекси вводять лише для зручності ідентифікації. Принциповою відмінністю часового ряду від простих статистичних сукупностей є:

- по-перше, рівні часового ряду є залежними. Інакше кажучи, якщо майбутні значення змінної можна визначити, то вони є функцією від минулих значень цієї змінної;
- по-друге, рівні часового ряду неоднаково розподілені. Закон розподілу ймовірностей цих випадкових величин і, зокрема, їхні математичні сподівання та дисперсії можуть залежати від часу.

З іншого боку, якщо абстрагуватися від того, що означають величини x_i та y_i, $i = \overline{1;k}$, але при цьому продовжувати враховувати порядок їх слідування, то можна перейти до статистичних вибірок $X = (x_1, x_2, \ldots, x_k)$ та $Y = (y_1, y_2, \ldots, y_k)$, де k – об'єм вибірок, частоти елементів вибірок дорівнюють 1.

Варто підкреслити, що принциповим є той факт, що обидві вибірки було отримано з реальних та прогнозованих курсів криптовалюти за однаковий період часу і в ті самі моменти часу. Це дає змогу обґрунтувати правомірність переходу від часових рядів до статистичних вибірок і, таким чином, коректно обчислювати середню відносну похибку прогнозування $MAPE$.

$$MAPE = \frac{1}{k} \cdot \sum_{i=1}^{k} \left(\frac{|x_i - y_i|}{x_i} \right) \cdot 100\%, \qquad (16.7)$$

де x_i – елементи вибірки X, y_i – елементи вибірки Y, $i = \overline{1;k}$, k – об'єм вибірок X та Y.

Також слід зазначити, що $MAPE$ для добре побудованого прогнозу має наближатися до нуля, але аналітик сам має змогу встановити його допустиму верхню межу згідно наступної шкали (табл. 16.11).

Таблиця 16.11. *Рівень адекватності моделі прогнозування*

$MAPE$	Точність прогнозу
менше 10%	Висока
10%– 20%	Добра
20%– 40%	Задовільна
40%– 50%	Погана
більше 50%	Незадовільна

Загалом, процес перевірки якості прогнозу складається з наступних кроків:

КРОК 1: Встановлення верхньої допустимої межі $MAPE$.

КРОК 2: Створення статистичних вибірок $X = (x_1, x_2, \dots, x_k)$ та $Y = (y_1, y_2, \dots, y_k)$ з реальних та прогнозованих курсів криптовалюти відповідно за період часу $[t_1; t_k]$.

КРОК 3: Обчислення $MAPE$ для даних вибірок.

КРОК 4: Прийняття рішення щодо точності прогнозу. Якщо значення $MAPE$ менше верхньої допустимої межі, то прогноз вважається якісним. В протилежному випадку потрібно збільшити об'єм вибірок або змінити алгоритм прогнозування за яким було отримано вибірку Y.

Перераховані кроки представлено на рис. 16.8:

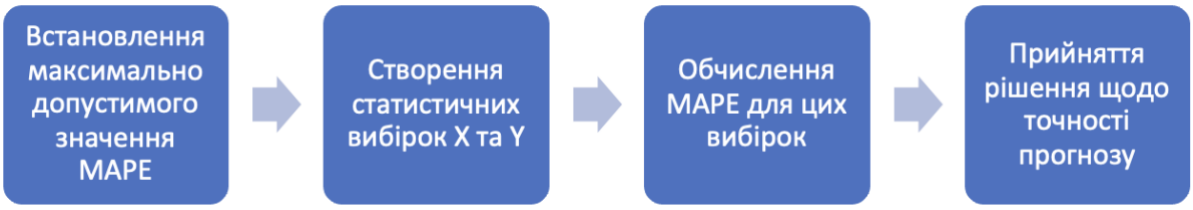

Рис. 16.8. Схема перевірки якості прогнозу

Окрім вищесказаного, перехід до статистичного підходу дає змогу застосовувати для оцінки якості отриманого прогнозу класичні методи статистичного аналізу:

- обчислення основних точкових характеристик вибірок, таких як вибіркове середнє, вибіркове середньоквадратичне відхилення,
- обчислення коефіцієнти кореляції (зокрема коефіцієнти рангової кореляції Спірмена або Кендала, оскільки потрібно враховувати порядок слідування елементів виборок),
- застосування статистичних критеріїв, зокрема t-критерію, що дає змогу перевірити гіпотезу про значущість кореляційного зв'язку між виборками, а також критерію $\mathcal{X}^2$ –Пірсона для перевірки гіпотези про тип розподілу генеральної сукупності (див. лекцію 18).

Усі перераховані методи статистичного аналізу можуть бути використані не лише для контролю якості прогнозів, а й для побудови неперервних функцій, які є щільностями відповідного типу розподілу. За допомогою цих функцій також можна отримати потрібні прогнози.