ЛЕКЦІЯ 13
3. ЕЛЕМЕНТИ МАТЕМАТИЧНОЇ СТАТИСТИКИ

3.1. Основні поняття математичної статистики

Математична статистика – наука про методи систематизації, обробки і використання для наукових і практичних висновків *статистичних даних*, одержаних при проведенні випробувань з випадковими наслідками.

При розв'язуванні багатьох наукових і виробничих задач як початкова інформація використовуються так звані статистичні дані про ті чи інші події або процеси випадкового характеру. Такими, наприклад, є задачі аналізу і прогнозування пасажирських потоків, визначення надійності авіаційної техніки, розрахунку потреби в запчастинах тощо.

В подібних задачах виникає необхідність оцінювати числові характеристики випадкових величин, визначати закони їх розподілу, функції розподілу або щільності ймовірностей. Всі ці характеристики розглядались нами раніше в припущенні, що для випадкових величин вони відомі або можуть бути визначені з умов розв'язуваних задач.

В практичній же діяльності при розв'язуванні задач, пов'язаних з випадковими явищами, даних для визначення таких характеристик, як правило, немає. Є лише можливість проводити різноманітні випробування над випадковими явищами і в результаті одержувати той статистичний матеріал, який дає змогу принаймні наближено оцінити ці характеристики. Тому *статистичні дані* можна розглядати як сукупність значень, одержаних внаслідок проведення певної кількості випробувань над випадковою величиною або системою випадкових величин.

Статистичні дані, одержані в результаті випробувань, як правило, невпорядковані і мають безсистемний характер, що не дає можливості скласти бодай наближене уявлення про поведінку випадкової величини і про характер її розподілу. Крім того, всяке випробування пов'язане з похибками спостережень і вимірювань, отже, в результаті випробування можна одержати лише наближені характеристики випадкової величини. В зв'язку з цими обставинами виникає ряд задач, розв'язуванням яких займається математична статистика:

– упорядкування статистичного матеріалу, представлення його в найбільш зручному для огляду та аналізу вигляді;

– формування на основі теоретичних міркувань гіпотез про якісний характер (закон) розподілу досліджуваної випадкової величини та перевірка правильності цих гіпотез на одержаних статистичних даних;

– оцінка на основі статистичних даних кількісних (числових) характеристик досліджуваної випадкової величини і визначення точності цих оцінок для проведеної кількості випробувань.

Розв'язування цих задач складає основу предмета математичної статистики, але далеко не вичерпує її важливі проблеми. В багатьох своїх розділах математична статистика ґрунтується на методах теорії ймовірностей, які дозволяють оцінити надійність і точність висновків, зроблених на основі обмеженого статистичного матеріалу.

3.1.2. Генеральна та вибіркова сукупності. Поняття про вибірковий метод

В математичній статистиці вивчення множини схожих об'єктів, явищ, процесів може проводитись як по якісних, так і по кількісних ознаках. Наприклад, при обстеженні відділом технічного контролю партії виробів якісною ознакою може бути міцність виробів, а кількісною — відповідність виробу встановленим геометричним розмірам.

При цьому обстеження всієї партії виробів не завжди можливе, зокрема, якщо при перевірці якісної ознаки відбувається руйнування виробу (випробування на розрив) або при перевірці кількісної ознаки доводиться виконувати вимірювання занадто великої кількості виробів.

В таких випадках з усієї сукупності обстежуваних однорідних об'єктів, яка називається *генеральною сукупністю*, випадковим чином відбирається достатньо представницька її частина, яка називається *вибірковою сукупністю* або *вибіркою*. Результати обстеження об'єктів, які потрапили у вибірку, дозволяють зробити висновок про стан досліджуваної ознаки, яка називається ще *варійовною ознакою*, в усій генеральній сукупності.

Кількість об'єктів сукупності (генеральної – N або вибіркової – n) називається *об'ємом* цієї сукупності. Наприклад, якщо із тисячі виробів вибрано для контролю 100, то об'єм генеральної сукупності $N=1000$, а об'єм вибірки $n=100$.

Зроблена вибірка повинна достатньо повно відбивати досліджувану ознаку об'єктів генеральної сукупності, тобто бути представницькою або репрезентативною. Вона буде такою за умов:

– якщо кожний об'єкт генеральної сукупності має однакову ймовірність потрапити у вибірку;

– якщо всі об'єкти, що потрапили у вибірку, відібрані випадково.

Сформована одним із вказаних способів вибірка являє собою початкову статистичну сукупність, в якій досліджувана ознака ще ніяким чином не впорядкована. Для подальшого вивчення початкову статистичну сукупність потрібно перш за все впорядкувати, тобто надати їй характер системності.

Статистична обробка може здійснюватися і з усією генеральною сукупністю, якщо її об'єм не надто великий. Тоді можна вважати, що вибірка співпадає з генеральною сукупністю.

3.1.3. Варіаційний ряд. Дискретний та інтервальний статистичні розподіли вибірки

Нехай з генеральної сукупності для дослідження деякої її варійовної ознаки X одним із розглянутих способів проведена вибірка об'єму n. Такою ознакою може бути, наприклад, щоденний обсяг продукції підприємства, рівень щомісячного заробітку його працівників, відсоток зайнятості крісел на рейсах авіакомпанії, щоденний прибуток авіакомпанії тощо.

Одержана вибірка являє собою невпорядкований статистичний матеріал і складає *первинну* статистичну сукупність. Першим етапом її обробки є її впорядкування, тобто розташування в певному (як правило, неспадному) порядку або одержання так званої *впорядкованої* статистичної сукупності. По ній можна вже зробити певні висновки про досліджувану ознаку X, зокрема, які різні можливі значення приймає ознака X і як часто кожне значення зустрічається у вибірці.

Другим етапом обробки вибіркових даних є складання *дискретного і інтервального варіаційних рядів*.

Нехай у вибірці об'єму n варійовна ознака X прийняла в порядку зростання значення x_1 n_1 раз, значення x_2 - n_2 раз і т.д. і, нарешті, значення x_k - n_k раз $(n_1 + n_2 + ... + n_k = n)$. Всі різні значення x_i $(\mathrm{i} = 1,2,...,k)$ ознаки X називаються *варіантами*, а кількості n_i появи кожної з варіант у вибірці – їх *частотами*.

Означення 3.1. *Дискретним* або *перервним варіаційним рядом* називаються розташовані в порядку зростання варіанти

$$x_1, x_2, ..., x_k, \tag{3.1}$$

які являють собою окремі ізольовані одне від іншого значення варійовної ознаки X.

Якщо кількість варіант у дискретному ряду (3.1) занадто велика або ознака X генеральної сукупності є неперервною випадковою величиною, то виникає потреба побудови інтервального варіаційного ряду.

Означення 3.2. *Інтервальним (неперервним) варіаційним рядом* називається ряд, в якому значення варіант задані у вигляді інтервалів, тобто значення ознаки X можуть відрізнятись одне від одного на як завгодно малу величину.

Для побудови інтервального варіаційного ряду потрібно визначити кількість інтервалів, на які розбивається множина варіант дискретного ряду (3.1), і знайти величину (крок) інтервалу.

Орієнтовно кількість інтервалів $k_{інт}$ для вибірки об'єму n обчислюється за формулою

$$k_{інт} = \sqrt{n}. \tag{3.2}$$

Більш точною вважається формула, запропонована Стерджессом

$$k_{інт} = 1 + 3.322 \ln n. \tag{3.3}$$

Застосовується також формула

$$k_{інт} = 5 \ln n. \tag{3.4}$$

Для вибірок об'єму $n < 1000$ формули (3.2)-(3.4) дають порівнянні результати, зокрема, при $n = 500$ для $k_{інт}$ формули (3.2), (3.3) і (3.4) дають відповідно значення 22, 21 і 31, в той час, як при $n = 10000$ результат за формулою (3.2) суттєво переважає результати двох інших формул (відповідно 100, 32 і 46). Тому для вибірок об'єму $n > 1000$ слід віддавати перевагу формулі (3.3) або (3.4).

Величина (крок) h інтервалу обчислюється за формулою

$$h = \frac{x_k - x_1}{k_{інт}}, \tag{3.5}$$

а інтервальний варіаційний ряд являє собою сукупність інтервалів

$$(x_1; x_1 + h),\ (x_1 + h; x_1 + 2h), ...,\ (x_1 + (l-1)h; x_m), \tag{3.6}$$

де $l = k_{інт}$.

Зауваження. Оскільки $k_{інт}$ за формулами (3.2)-(3.4) обчислюється орієнтовно, слід цю величину вибирати так, щоб вона була, з одного боку, близькою до обчисленої за формулою, а, з другого, - давала по можливості „округле" значення h за формулою (3.5).

Третім етапом обробки вибіркових даних є побудова дискретного та інтервального статистичного розподілу вибірки.

Означення 3.3. *Дискретним статистичним розподілом* вибірки називається відповідність між варіантами x_i варіаційного ряду (3.1) та їх частотами n_i (або відносними частотами $\omega_i = \frac{n_i}{n}$).

Дискретний статистичний розподіл подається у вигляді табл.3.1

Таблиця 3.1

Варіанти, x_i	x_1	x_2	…	x_i	…	x_k	
Частоти, n_i	n_1	n_2	…	n_i	…	n_k	$\sum_i n_i = n$
Відносні частоти, ω_i	$\frac{n_1}{n}$	$\frac{n_2}{n}$	...	$\frac{n_i}{n}$	…	$\frac{n_k}{n}$	$\sum_i \omega_i = 1$

По дискретному статистичному розподілу будується інтервальний, який також являє собою таблицю, в першому рядку якої представлені інтервали, з яких складається варіаційний ряд (3.6), а в другому і третьому рядках записуються сумарні частоти n_i і, відповідно, сумарні відносні частоти ω_i варіант, які групуються в кожному інтервалі. При цьому, якщо границя між двома сусідніми інтервалами в точності співпадає із значенням варіанти, частота останньої розподіляється нарівно між цими інтервалами або приписується повністю в усіх таких випадках одному з них (лівому або правому).

Інтервальний статистичний розподіл вибірки у загальному вигляді поданий табл.3.2.

Таблиця 3.2

Інтервали	$(x_1; x_1 + h)$	$(x_1 + h; x_1 + 2h)$	…	$(x_1 + (l-1)h; x_k)$
Частоти, n_i^*	n_1^*	n_2^*	…	n_l^*
Відносні частоти, ω_i^*	$\frac{n_1^*}{n}$	$\frac{n_2^*}{n}$	…	$\frac{n_l^*}{n}$

Приклад 3.1. Для аналізу денної виручки авіафірми від реалізації пасажирських авіабілетів із статистичних даних за рік зроблено вибірку об'єму $n = 40$, до складу якої потрапили такі значення денної виручки X (тис.грв.):

110; 105; 116; 108; 95; 103; 119; 108; 110; 105;
108; 105; 92; 112; 105; 110; 103; 116; 99; 116;
119; 99; 110; 103; 99; 112; 105; 105; 103; 108;
105; 99; 108; 105; 108; 103; 112; 95; 108; 110.

Скласти дискретний та інтервальний статистичні розподіли досліджуваної ознаки X.

Розв'язання. Запишемо вибіркові дані у неспадному порядку, тобто складемо впорядковану статистичну сукупність

92; 95; 95; 99; 99; 99; 99; 103; 103; 103;

103; 103; 105; 105; 105; 105; 105; 105; 105; 105;
108; 108; 108; 108; 108; 108; 108; 110; 110; 110;
110; 110; 112; 112; 112; 116; 116; 116; 119; 119.

З цієї впорядкованої сукупності визначаємо кількість варіант в дискретному варіаційному ряду $(k=10)$, частоту кожної варіанти і складаємо дискретний статистичний розподіл ознаки X (табл.3.3):

Таблиця 3.3

Варіанти, x_i	92	95	99	103	105	108	110	112	116	119	Σ
Частоти, n_i	1	2	4	5	8	7	5	3	3	2	40
Відносні частоти, ω_i	0,025	0,05	0,1	0,125	0,2	0,175	0125	0,075	0,075	0,05	1

Для побудови інтервального статистичного розподілу обчислимо кількість інтервалів за спрощеною формулою (3.2)

$$k_{інт}=\sqrt{40}=6$$

і крок інтервалу – за формулою (3.5)

$$h=\frac{119-92}{6}=4.5.$$

Тоді інтервальний варіаційний ряд матиме вигляд

(92; 96.5), (96.5; 101), (101; 105.5), (105.5; 110), (110; 114.5), (114.5; 119).

Підрахуємо сумарні частоти варіант, які групуються в кожному інтервалі. Оскільки границя четвертого і п’ятого інтервалів співпадає з варіантою 110, розподілимо її частоту між цими інтервалами, додавши 2 одиниці до четвертого і 3 одиниці до п’ятого інтервалів.

Одержимо інтервальний статистичний розподіл ознаки X (табл.3.4).

Таблиця 3.4

Інтервали	(92; 96,5)	(96,5; 101)	(101; 105,5)	(105,5; 110)	(110; 114,5)	(114,5; 119)	Σ
Частоти, n_i^*	3	4	13	9	6	5	40
Відносні частоти, ω_i^*	0,075	0,1	0,325	0,225	0,15	0,125	1

3.1.4. Емпірична функція розподілу. Полігон частот та гістограма

Дискретний статистичний розподіл вибірки (табл.3.1) є аналогом ряду розподілу (2.1) дискретної випадкової величини, лише замість ймовірностей p_i в ньому представлені відносні частоти ω_i варіант. Тому за статистичним розподілом можна побудувати функцію розподілу $F^*(x)$, яка називається *емпіричною* або *функцією розподілу вибірки* і є наближеним виразом теоретичної функції розподілу $F_T(x)$ ознаки X генеральної сукупності.

Означення 3.4. *Емпіричною функцією розподілу* називається функція $F^*(x)$, яка для кожного значення x дорівнює відносній частоті варіант, менших x:

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

де n_{x_i} — кількість варіант вибірки, менших x.

Для дискретного розподілу, наведеного в табл.3.1, функція $F^*(x)$ має такий аналітичний вираз:

$$F^*(x) = \begin{cases} 0 & \text{при} \quad x \le x_1; \\ \dfrac{n_1}{n} & \text{при} \quad x_1 < x \le x_2; \\ \dfrac{n_1 + n_2}{n} & \text{при} \quad x_2 < x \le x_3; \\ \dots\dots\dots\dots\dots\dots\dots\dots & \\ \dfrac{n_1 + n_2 + \dots + n_{k-1}}{n} & \text{при} \quad x_{k-1} < x \le x_k; \\ 1 & \text{при} \quad x > x_k. \end{cases} \qquad (3.7)$$

(див. (2.4)) і графік, відповідний представленому на рис.2.2.

Для побудови емпіричної функції $F^*(x)$ за інтервальним статистичним розподілом значення функції $F^*(x)$ обчислюються в кількох точках, за які, як правило, вибираються границі інтервалів x_i. Точки $(x_i, F^*(x))$ наносяться на графік і послідовно сполучаються відрізками прямих.

Якщо x_1 — ліва границя першого інтервалу, а x_k — права границя останнього в інтервальному розподілі вибірки, то $F^*(x) = 0$ при $x < x_1$ і $F^*(x) = 1$ при $x > x_k$.

Для більшої наочності при виробленні висновків про вид розподілу ознаки X генеральної сукупності по вибірковим даним застосовується графічне зображення статистичних розподілів вибірки, а саме *полігон* розподілу та *гістограма*.

Полігон розподілу будується як для дискретних, так і для інтервальних статистичних розподілів вибірки. Для дискретного розподілу (табл.3.1) в прямокутній системі координат наносяться точки з координатами (x_i, n_i) або (x_i, ω_i), які сполучаються відрізками прямих. Одержана ламана лінія є полігоном частот. Для інтервального статистичного розподілу (табл.3.2) полігон будується аналогічно, лише за абсциси точок приймаються середини інтервалів.

Гістограма застосовується для графічного зображення інтервальних розподілів. Для її побудови на осі абсцис відкладаються відрізки, рівні кроку інтервалу h, і на цих відрізках, як на основах, будуються прямокутники з висотами $\dfrac{\omega_i}{h}$. Оскільки величини $\dfrac{\omega_i}{h}$ є щільностями відносних частот на відповідних інтервалах, то гістограма дає наближений вигляд щільності розподілу ознаки X в генеральній сукупності. Якщо з'єднати середини верхніх основ прямокутників плавною кривою, то цю криву можна прийняти за статистичний аналог щільності розподілу ознаки X.

Побудова по вибірці статистичних розподілів, емпіричної функції розподілу, полігону частот і гістограми складає первинну обробку статистичної сукупності.

Приклад 3.2. За статистичними розподілами вибірки, одержаними в прикладі 3.1 для денної виручки авіафірми, побудувати емпіричну функцію розподілу, полігони частот та гістограму.

Розв'язання. 1) Значення емпіричної функції розподілу одержимо за формулою (3.7), складаючи послідовно відносні частоти десяти варіант дискретного статистичного розподілу:

Для інтервального статистичного розподілу значення емпіричної функції $F^*(x)$ на границях інтервалів також одержуються додаванням відносних частот, які містяться в відповідних інтервалах:

$$F^*(x)=\begin{cases}0 & \text{при} \quad x\le 92;\\ 0+0.025=0.025 & \text{при} \quad 92<x\le 95;\\ 0.025+0.05=0.075 & \text{при} \quad 95<x\le 99;\\ 0.075+0.1=0.175 & \text{при} \quad 99<x\le 103;\\ 0.175+0.125=0.3 & \text{при} \quad 103<x\le 105;\\ 0.3+0.2=0.5 & \text{при} \quad 105<x\le 108;\\ 0.5+0.175=0.675 & \text{при} \quad 108<x\le 110;\\ 0.675+0.125=0.8 & \text{при} \quad 110<x\le 112;\\ 0.8+0.075=0.875 & \text{при} \quad 112<x\le 116;\\ 0.875+0.075=0.95 & \text{при} \quad 116<x\le 119;\\ 0.95+0.05=1 & \text{при} \quad x>119.\end{cases}$$

$$F^*(x)=\begin{cases}0 & \text{при} \quad x=92;\\ 0+0.075=0.075 & \text{при} \quad x=96.5;\\ 0.075+0.1=0.175 & \text{при} \quad x=101;\\ 0.175+0.325=0.5 & \text{при} \quad x=105.5;\\ 0.5+0.225=0.725 & \text{при} \quad x=110;\\ 0.725+0.15=0.875 & \text{при} \quad x=114.5;\\ 0.875+0.125=1 & \text{при} \quad x=119.\end{cases}$$

Побудуємо графіки функції $F^*(x)$ для обох випадків (рис.3.1). Для дискретного статистичного розподілу графік – розривна східчаста лінія, а для інтервального – неперервна ламана лінія.

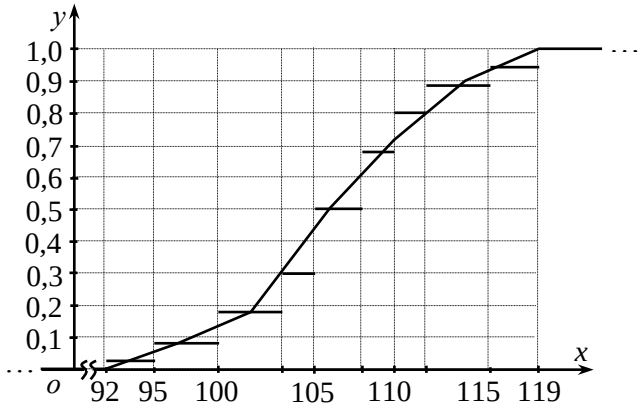

Рис.3.1

Одержані графіки дають загальне уявлення про функцію розподілу досліджуваної ознаки X генеральної сукупності. Зрозуміло, що інша вибірка по денній виручці авіафірми дала б дещо іншу картину, хоча загальна тенденція при цьому б збереглась. При збільшенні об'єму вибірки довжини сходинок і висоти стрибків у точках розриву зменшуються і розривна східчаста лінія, побудована для дискретного статистичного розподілу, наближається до деякої плавної кривої, яку можна вважати графіком теоретичної функції розподілу $F_T(x)$ ознаки X генеральної сукупності. Те ж саме відбувається і з неперервною ламаною лінією – графіком функції $F^*(x)$ для інтервального розподілу, оскільки при збільшенні об'єму вибірки скорочуються довжини інтервалів, а разом з ними і довжини прямолінійних ланок ламаної лінії.

2) Полігон відносних частот для дискретного розподілу вибірки, зображений на рис.3.2 суцільною ламаною лінією, ланки якої з'єднують послідовні точки з координатами (x_i, ω_i) з дискретного розподілу. Полігон відносних частот для інтервального розподілу поданий на тому ж рис.3.2 пунктирною ламаною лінією, яка з'єднує точки, абсцисами яких є центри інтервалів: 94,25; 98,75; 103,25; 107,75; 112,25; 116,75, а ординатами – відносні частоти ω_i^* на відповідних інтервалах.

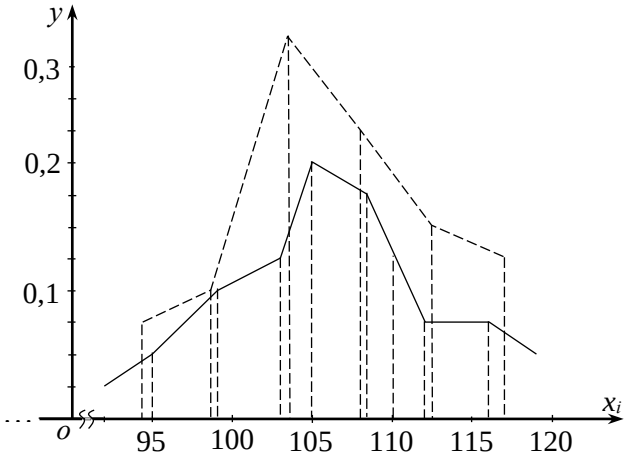

Рис.3.2

3) Для побудови гістограми обчислимо щільності відносних частот $\frac{\omega_i^*}{h}$ для кожного інтервалу інтервального статистичного розподілу

Інтервали	(92; 96,5)	(96,5; 101)	(101; 105,5)	(105,5; 110)	(110; 114,5)	(114,5; 119)
$\frac{\omega_i^*}{h}$	0,017	0,022	0,072	0,05	0,033	0,028

Гістограма, побудована за цими даними, представлена на рис.3.3.

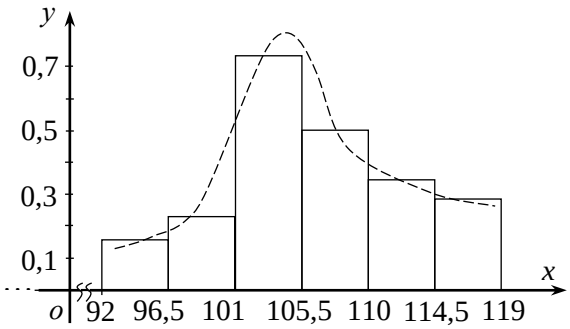

Рис.3.3

Одержана гістограма і плавна лінія, яка проходить через середини верхніх основ прямокутників, вже складають певне уявлення про закон розподілу денної виручки авіафірми в генеральній сукупності, зокрема, вони не виключають її нормального розподілу, хоча й відзначаються певною асиметрією, яка може бути наслідком малого об’єму або недосконалості вибірки.

Обґрунтування попередніх висновків про закон розподілу ознаки X генеральної сукупності проводиться на подальших етапах аналізу та дослідження вибіркової сукупності.