ЗМІСТОВИЙ МОДУЛЬ 2
МАТЕМАТИЧНА СТАТИСТИКА

Первинне опрацювання статистичних даних

Генеральна та вибіркова сукупності

Нехай потрібно вивчити сукупність однорідних об’єктів відносно деякої якісної або кількісної ознаки, яка характеризує ці об’єкти. Наприклад, якщо досліджується партія товарів, то якісною ознакою може слугувати стандартність товару, я кількісною – наприклад, розмір товару.

Іноді проводять повне дослідження, тобто аналізується кожний об’єкт сукупності відносно ознаки, якою цікавляться. Проте на практиці, таке повне дослідження проводиться відносно рідко, і найчастіше із сукупності обирається та досліджується лише якась контрольна частина об’єктів у обмеженій кількості. Введемо відповідні означення.

Означення. Сукупність всіх можливих об’єктів даного виду, над якими проводяться спостереження, або сукупність всіх можливих спостережень, які проводяться в незмінних умовах над деякою випадковою величиною, називається *генеральною сукупністю.*

Генеральна сукупність може містити скінченну або нескінченну кількість елементів.

Означення. Відібрані з генеральної сукупності об’єкти (або результати спостережень) називаються *вибірковою сукупністю* або просто *вибіркою.*

Означення. Число N елементів *генеральної сукупності* та число n елементів *вибіркової сукупності* будемо називати відповідно *об’ємами генеральної* та *вибіркової сукупностей*.

Повторна та безповторна вибірки. Репрезентативна вибірка

При складанні вибірки можна діяти двома способами: після того, як об’єкт був відібраний та над ним було проведене спостереження, він може бути повернутий або не повернутий у генеральну сукупність. Відповідно до цього вибірки розділяють на *повторні* та *безповторні.*

Означення. *Повторною* називають *вибірку*, при якій відібраний об’єкт (перед відібранням наступного) повертають до генеральної сукупності.

Означення. *Безповторною* називають *вибірку*, при якій відібраний об’єкт не повертають до генеральної сукупності.

На практиці зазвичай користуються безповторним випадковим відбором.

Для того щоб за даними вибірки можна було достатньо впевнено судити про ознаку генеральної сукупності, яка нас цікавить, необхідно, щоб об’єкти вибірки правильно її представляли. Цю вимогу формулюють так: вибірка повинна бути *репрезентативною* (або *представницькою*). Вважається, що якщо кожний об’єкт вибірки відібраний із генеральної сукупності випадково, тобто всі об’єкти мають однакову ймовірність потрапити у вибірку, і кількість об’єктів, відібраних для спостереження, є досить великою, то вибірка буде репрезентативною. Різниця між показниками вибіркової та генеральної сукупностей становить *помилку репрезентативності.* Ці помилки виникають тому, що вибіркова сукупність неточно відображає генеральну сукупність.

Означення. Вибірка називається *репрезентативною*, якщо вона достатньо добре відтворює генеральну сукупність.

Це означення не дозоляє робити конкретні виводи, бо не вказана загальна міра відповідності між репрезентативною вибіркою та генеральною сукупністю, тому питання про репрезентативність необхідно вирішувати у конкретних задачах, спираючись на конкретні критерії відповідності.

Дискретний варіаційний ряд розподілу

Нехай із генеральної сукупності проведена вибірка об'єму n, і досліджувана випадкова величина в цій вибірці прийняла різні k значень $(x_1, x_2, \ldots, x_k)$ – ці елементи називають *варіантами*:

$$\underbrace{(\underbrace{x_1,\ldots,x_1}_{n_1\ разів},\underbrace{x_2,\ldots,x_2}_{n_2\ разів},\ldots,\underbrace{x_k,\ldots,x_k}_{n_k\ разів})}_{n\ разів}$$

Нехай у вибірці $(x_1, x_2, \ldots, x_k)$ варіанта x_1 спостерігалась n_1 разів, x_2 спостерігалась n_2 разів, і так далі відповідно до x_k, яка спостерігалась n_k разів, тобто з того, що об'єм вибірки дорівнює n, випливає:

$$n_1 + n_2 + \ldots + n_k = \sum_{i=1}^{k} n_i = n.$$

Означення. Значення $n_1, n_2, \ldots, n_k$, тобто чисельність окремої групи згрупованого ряду вибірки, називають *частотами* або *вагами варіант*. Відношення n_i до загального об'єму вибірки n, називається *відносною частотою варіанти* x_i та позначається:

$$p_i^* = \frac{n_i}{n}.$$

З означення відносної частоти випливає:

$$p_1^* + p_2^* + \ldots + p_k^* = \sum_{i=1}^{k} p_i^* = 1.$$

Означення. Якщо значення випадкової величини, яке відповідає окремій групі згрупованого ряду даних, називається *варіантою*, то змінення цього значення – *варіюванням*.

Означення. Розташування вибіркових спостережень значень випадкової величини в порядку неспадання називається *ранжуванням*.

Означення. Різниця між максимальним та мінімальним значеннями варіант, тобто інтервал варіювання, називається *розмахом вибірки* та позначається:

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

Означення. *Дискретним варіаційним рядом розподілу* (або *дискретним статистичним розподілом,* або *розподілом частот*) називається ранжована сукупність варіант x_i з відповідними частотами або відносними частотами.

У *табличній формі* він має такий вигляд:

x_i	x_1	x_2	x_3	…	x_k
n_i	n_1	n_2	n_3	…	n_k
p_i^*	p_1^*	p_2^*	p_3^*	…	p_k^*

Приклад. Проводяться спостереження над значеннями грошових виграшів у миттєвій лотереї. У результаті отримані наступні значення (у тис. грн.):

0, 1, 0, 0, 5, 0, 10, 0, 1, 0, 0, 1, 5, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 5, 0, 5, 0, 0, 1, 1, 1, 5, 10, 0, 1, 1, 0, 5, 0, 0, 0, 0, 1, 0, 1, 0, 5, 0, 0, 0, 0, 1, 0.

Складемо варіаційний ряд розподілу у табличній формі, для цього визначимо які різні значення прийняла досліджувана випадкова величина в цій вибірці, тобто визначимо варіанти, обрахуємо загальний об’єм вибірки n, відповідні частоти та відносні частоти варіант:

x_i	0	1	5	10
n_i	31	14	7	2
p_i^*	$\frac{31}{54}$	$\frac{14}{54}$	$\frac{7}{54}$	$\frac{2}{54}$

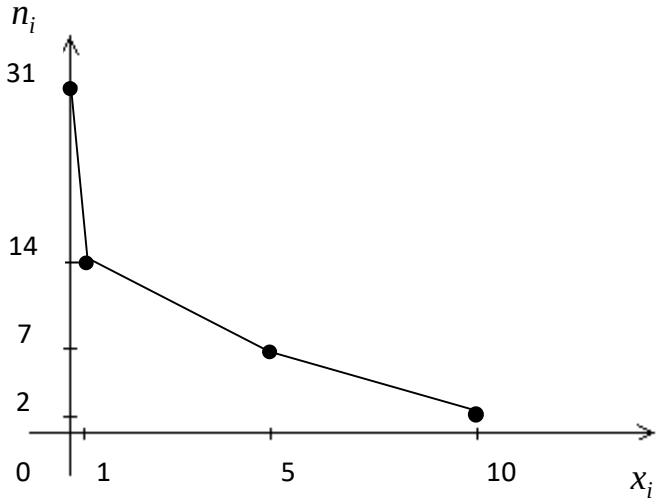
$$n = 31 + 14 + 7 + 2 = 54,$$

$$\sum_{i=1}^{k} p_i^* = 1.$$

Для графічного представлення дискретного варіаційного ряду розподілу будують *полігон (або багатокутник) частот* та *полігон (або багатокутник) відносних частот*.

Означення. Ламана лінія, відрізки якої послідовно з'єднують точки з координатами (x_i, n_i), $i = \overline{1,k}$, називається *полігоном частот*, а ламана лінія, відрізки якої послідовно з'єднують точки з координатами (x_i, p_i^*), $i = \overline{1,k}$, називається *полігоном відносних частот*.

Приклад. Побудуємо полігон частот для попереднього прикладу. Для цього відмітимо на декартовій площині точки з координатами (x_i, n_i), $i = \overline{1,k}$, які у нашому випадку будуть дорівнювати: (0,31), (1,14), (5,7), (10,2), та з'єднаємо їх послідовно відрізками прямих.



Зауважимо, що полігон, побудований за дискретним варіаційним рядом, є вибірковим аналогом багатокутника розподілу дискретної випадкової величини.

Емпірична функція розподілу

Означення. *Вибірковою (емпіричною) функцією розподілу* називається функція $F^*(x)$, яка задає для кожного значення *x* відносну частоту події $X < x$. Отже, за означенням $F^*(x) = \frac{n_x}{n}$, де n_x – число вибіркових значень величини *X*, менших за *x*, а *n* – об'єм вибірки.

Вибіркову функцію можна задати у табличному та у графічному виді. Розберемо це на прикладі.

Приклад. Побудуємо вибіркову функцію розподілу для розглянутого вище прикладу.

Об'єм вибірки $n = 54$, найменша варіанта дорівнює 0, отже $n_x = 0$ при $x \le 0$, а $F^*(x) = \frac{n_x}{n} = \frac{0}{54}$, при $x \le 0$.

При $0 < x \le 1$ нерівність $X < x$ виконується для варіанти $x_1 = 0$, і відповідно $n_x = n_1 = 31$, а $F^*(x) = \frac{n_x}{n} = \frac{31}{54} = p_1^*$.

При $1 < x \le 5$ нерівність $X < x$ виконується для варіант $x_1 = 0$ та $x_2 = 1$, і відповідно $n_x = n_1 + n_2 = 31 + 14 = 45$, а $F^*(x) = \frac{n_x}{n} = \frac{45}{54} = p_1^* + p_2^*$.

При $5 < x \le 10$ нерівність $X < x$ виконується для варіант $x_1 = 0$, $x_2 = 1$ та $x_3 = 5$, і відповідно $n_x = n_1 + n_2 + n_3 = 31 + 14 + 7 = 52$, а $F^*(x) = \frac{n_x}{n} = \frac{52}{54} = p_1^* + p_2^* + p_3^*$.

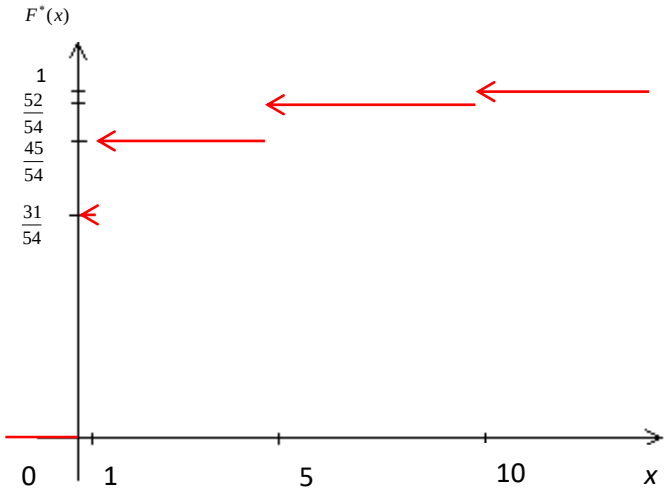
При $10 < x$ нерівність $X < x$ виконується для варіант $x_1 = 0$, $x_2 = 1$, $x_3 = 5$ та $x_4 = 10$, і відповідно $n_x = n_1 + n_2 + n_3 + n_4 = 31 + 14 + 7 + 2 = 54$, а $F^*(x) = \frac{n_x}{n} = \frac{54}{54} = 1 = p_1^* + p_2^* + p_3^* + p_4^*$.

Результати обчислень занесемо у таблицю (для зручності).

x	$F^*(x)$
$x \le 0$	0
$0 < x \le 1$	$p_1^* = \frac{31}{54}$
$1 < x \le 5$	$p_1^* + p_2^* = \frac{45}{54}$
$5 < x \le 10$	$p_1^* + p_2^* + p_3^* = \frac{52}{54}$
$10 < x$	$p_1^* + p_2^* + p_3^* + p_4^* = 1$

Можна замість таблиці записувати вибіркову функцію розподілу так само, як записували функцію розподілу дискретної випадкової величини.

Побудуємо графік вибіркової функції розподілу.



Інтервальний варіаційний ряд

Якщо досліджувана випадкова величина неперервна або дискретна величина така, що число її можливих значень достатньо велике, то для побудови варіаційного ряду використовують *інтервальний ряд розподілу*. У цьому випадку весь інтервал варіювання розбивають на декілька частинних інтервалів однакової довжини (яку називають *кроком*) та підраховують частоту попадання значень випадкової величини у кожний частинний інтервал.

Означення. *Інтервальним варіаційним рядом* (або *інтервальним розподілом частот*) називається упорядкована послідовність інтервалів варіювання випадкової величини із відповідними частотами n_i або відносними частотами p_i^* попадання у кожен з них значень випадкової величини.

У *табличній формі* він має такий вигляд:

$(x_i, x_{i+1}]$	$[x_1, x_2]$	$(x_2, x_3]$	$(x_3, x_4]$	…	$(x_k, x_{k+1}]$
n_i	n_1	n_2	n_3	…	n_k
p_i^*	p_1^*	p_2^*	p_3^*	…	p_k^*

Зауваження. Часто замість інтервалів $(x_i, x_{i+1}]$ записують $[x_i, x_{i+1})$.

Для кожної випадкової величини та для кожного об'єму вибірки необхідно визначати оптимальне число частинних інтервалів. Існують спеціальні формули для визначення цього числа. В задачах це може бути вказано у формулюванні, але часто оптимальну кількість інтервалів визначають за формулою Стерджесса:

$$k = 1 + 3{,}322 \cdot \lg n,$$

де n – обсяг вибірки, k – кількість частинних інтервалів.

Кількість інтервалів, що обчислюється за формулою Стерджесса, округляється до цілого числа. Залежно від кількості вимірювань це число

знаходиться в межах від 8 до 12. Для визначення довжини інтервалу, тобто кроку h, необхідно розмах R поділити на кількість інтервалів k:

$$h = \frac{R}{k} = \frac{x_{\max} - x_{\min}}{1 + 3{,}322 \cdot \lg n}.$$

Значення, яке отримують за цією формулою, округлюють до найбільш зручного, якщо це необхідно (часто округлюють до меншого цілого). У результаті округлення шагу нижньою границею першого інтервалу, тобто x_1, можна зсунути відносно $x_{\min}$ у бік менших значень, а верхню границю останнього інтервалу відповідно у бік більших значень, але цей зсув не повинен перевищувати $\frac{h}{2}$.

Для графічного представлення інтервального варіаційного ряду розподілу будують *гістограму частот* або *відносних частот*. Для її побудови у прямокутній системі координат на осі Ox відкладають відрізки частинних інтервалів варіювання та на цих відрізках як на основах будують прямокутники з висотами, які дорівнюють частотам або відносним частотам відповідних інтервалів.

Для інтервального варіаційного ряду розподілу, як і для дискретного, можна побудувати *полігон частот*, але для цього таблицю треба доповнити ще одним рядком, в якому записати відповідні для кожного інтервалу $(x_i, x_{i+1}]$ точки y_i, які є їх серединами:

$$y_i = \frac{x_{i+1} + x_i}{2}, \quad i = \overline{1,k}.$$

$(x_i, x_{i+1}]$	$[x_1, x_2]$	$(x_2, x_3]$	$(x_3, x_4]$	…	$(x_k, x_{k+1}]$
y_i	y_1	y_2	y_3		y_k
n_i	n_1	n_2	n_3	…	n_k
p_i^*	p_1^*	p_2^*	p_3^*	…	p_k^*

Означення. Ламана лінія, відрізки якої послідовно з'єднують точки з координатами (y_i, n_i), $i = \overline{1,k}$, називається *полігоном частот*, а ламана лінія, відрізки якої послідовно з'єднують точки з координатами (y_i, p_i^*), $i = \overline{1,k}$, називається *полігоном відносних частот.*

Означення. Якщо відносну частоту розділити на довжину інтервалу, то отримана величина буде *вибірковою оцінкою щільності ймовірності:*

$$f^*(x_i) = \frac{p_i^*}{h}.$$

Емпірична функція розподілу неперервної випадкової величини, тобто у випадку інтервального розподілу частот, будується дещо інакше, ніж для дискретної випадкової величини – використовуючи інтервальний варіаційний ряд з відносними частотами p_i^*, значення функції $F^*(x)$ обчислюються у кінцевих точках інтервалів у вигляді «накопичувальної відносної частоти». Після чого точки $(x_i, F^*(x_i))$ відмічаємо на прямокутній системі координат та з'єднуємо ці точки відрізками прямих.

Приклад. При вимірювання зросту 105 студентів (хлопців) отримана вибірка результатів:

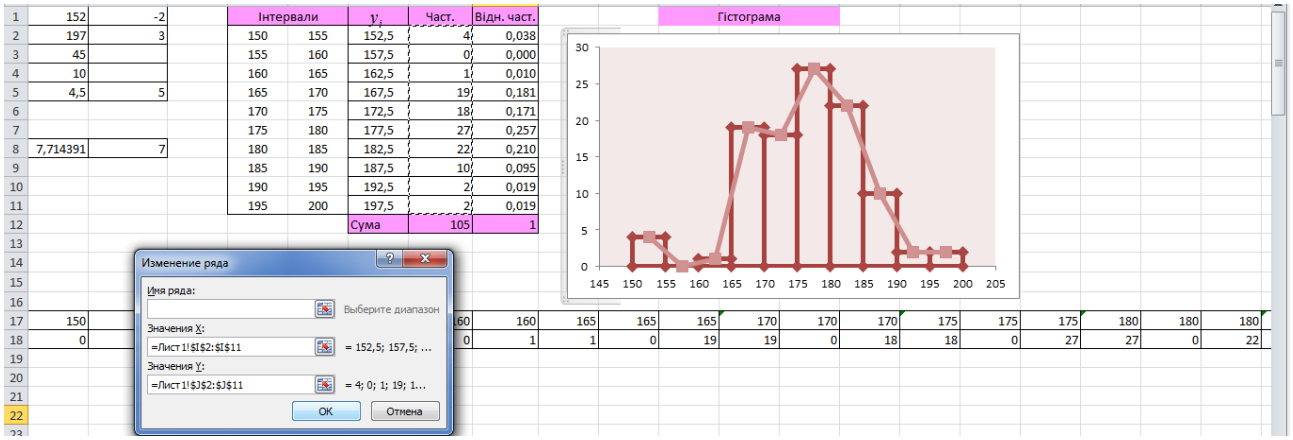
155, 170, 185, 180, 188, 152, 173, 178, 178, 168, 185,
173, 170, 183, 175, 173, 170, 183, 175, 180, 175, 193,
178, 183, 180, 197, 178, 181, 187, 168, 174, 179, 184,
183, 178, 180, 178, 163, 166, 178, 175, 182, 190, 167,
170, 178, 183, 170, 178, 181, 173, 168, 185, 175, 170,
155, 169, 186, 179, 189, 155, 174, 179, 179, 169, 186,
174, 171, 184, 175, 193, 178, 184, 180, 196, 175, 181,
188, 168, 179, 178, 183, 184, 178, 181, 177, 163, 166,

178, 175, 183, 190, 167, 170, 178, 183, 170, 178, 182,
173, 168, 186, 176, 171, 188.

Складемо інтервальний варіаційний ряд. Знайдемо мінімальне та максимальне значення випадкової величини: $x_{\min} = 152$, $x_{\max} = 197$. Тобто розмах – інтервал варіювання дорівнює $R = x_{\max} - x_{\min} = 197 - 152 = 45$. Нехай за умовами задане число частинних інтервалів, яке дорівнює 10. Зсунувши нижню границю з 152 до 150, а верхню з 197 до 200, отримаємо довжину частинного інтервалу, яка дорівнює $\frac{200-150}{10} = 5$. Складемо інтервальний варіаційний ряд (запишемо його у зручному для подальших обчислень в *MS Excel* вигляді).

Номер інтервалу i	Інтервал $(x_i, x_{i+1}]$	Середина інтервалу y_i	Частота n_i	Відносна частота p_i^*
1	(150,155]	152,5	4	0,0380952
2	(155,160]	157,5	0	0
3	(160,165]	162,5	1	0,0095238
4	(165,170]	167,5	19	0,1809524
5	(170,175]	172,5	18	0,1714286
6	(175,180]	177,5	27	0,2571429
7	(180,185]	182,5	22	0,2095238
8	(185,190]	187,5	10	0,0952381
9	(190,195]	192,5	2	0,0190476
10	(195,200]	197,5	2	0,0190476

Для графічного представлення інтервального варіаційного ряду розподілу побудуємо *гістограму частот* (проілюструємо цей приклад у подальшому на лабораторному занятті).



Числові характеристики варіаційного ряду: вибіркове середнє, вибіркова дисперсія, середньоквадратичне відхилення, мода і медіана

В результаті досліджень, пов'язаних з масовими явищами, збирається багато числових даних. На попередніх заняттях ми розв'язували задачі на обробку статистичних даних, представлення їх у вигляді таблиць, багатокутників та гістограм. Окрім розглянутого аналізу можна також зробити оцінку числових характеристик розподілу випадкової величини для більш точного описання досліджуваної випадкової величини, наприклад, характеристика положення, що фіксує деяке середнє значення випадкової величини, біля якого групуються всі можливі значення, характеристика розсіяння навколо середнього та інші.

Означення. Оцінки параметрів генеральної сукупності, отримані з аналізу вибірки, називаються *статистичними.*

Означення. *Точковою оцінкою* θ^* параметра θ називають функцію результатів спостережень над випадковою величиною X (статистику), за допомогою якої судять про значення параметра θ:

$$\theta^* = \theta^*(x_1, x_2, \ldots, x_n).$$

Точкова оцінка є випадковою величиною та залежить від закону розподілу випадкової величини та обсягу вибірки. У зв'язку з цим розглянемо вимоги, що висувають до оцінок.

Означення. Якщо математичне сподівання точкової оцінки θ^* дорівнює оцінюваному параметру θ генеральної сукупності, то така оцінка називається *незміщеною*.

Серед точкових характеристик важливу роль в статистиці грають: *вибіркове середнє, вибіркова дисперсія, мода, медіана.*

Означення. *Вибірковим (емпіричним) середнім* (або *вибірковим моментом першого порядку*) називається середнє арифметичне отриманих значень у вибірці. Якщо вибіркові дані не згруповано і n – об'єм вибірки, то:

$$\bar{x}_B = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \,.$$

У випадку, коли вже визначено *дискретний варіаційний ряд розподілу*:

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

де x_i, $i = \overline{1,k}$ – варіанта вибірки, n_i – частота варіанти, n – об'єм вибірки, p_i^* – відносна частота варіанти.

У випадку *інтервального варіаційного ряду* замість значень x_i, $i = \overline{1,k}$ беруть середини інтервалів варіаційного ряду y_i, $i = \overline{1,k}$ і формула для визначення вибіркового середнього має наступний вигляд:

$$\bar{x}_B = \frac{1}{n} \cdot \sum_{i=1}^{k} y_i \cdot n_i = \sum_{i=1}^{k} y_i \cdot p_i^* .$$

Вибіркове середнє є *незміщеною оцінкою математичного сподівання* генеральної сукупності.

Приклад 1. Знайти вибіркове середнє випадкової величини X за відомим дискретним варіаційним рядом розподілу:

x_i	2	5	7	10
n_i	5	26	10	9

Скористаємось формулою $\bar{x}_B = \frac{1}{n} \cdot \sum_{i=1}^{k} x_i \cdot n_i$, де $n = 5 + 26 + 10 + 9 = 50$:

$$\bar{x}_B = \frac{1}{n} \cdot \sum_{i=1}^{4} x_i \cdot n_i = \frac{1}{50} \cdot (2 \cdot 5 + 5 \cdot 26 + 7 \cdot 10 + 10 \cdot 9) = \frac{300}{50} = 6.$$

Приклад 2. Знайти вибіркове середнє випадкової величини X за відомим інтервальним варіаційним рядом розподілу:

$(x_i, x_{i+1}]$	[1,3]	(3,5]	(5,7]	(7,9]	(9,11]
n_i	5	10	40	35	10

Скористаємось формулою $\bar{x}_B = \frac{1}{n} \cdot \sum_{i=1}^{k} y_i \cdot n_i$, де $n = 5 + 10 + 40 + 35 + 10 = 100$. Знайдемо y_i – середини інтервалів $(x_i, x_{i+1}]$:

$(x_i, x_{i+1}]$	[1,3]	(3,5]	(5,7]	(7,9]	(9,11]
y_i	2	4	6	8	10
n_i	5	10	40	35	10

Отже вибіркове середнє дорівнює:

$$\bar{x}_B = \frac{1}{n} \cdot \sum_{i=1}^{5} y_i \cdot n_i = \frac{1}{100} \cdot (2 \cdot 5 + 4 \cdot 10 + 6 \cdot 40 + 8 \cdot 35 + 10 \cdot 10) = \frac{670}{100} = 6{,}7.$$

Означення. *Вибірковою (емпіричною) дисперсією* (або *вибірковим моментом другого порядку*) називають:

$$d_B = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x}_B)^2.$$

У випадку, коли визначено *дискретний варіаційний ряд розподілу*:

$$d_B = \frac{1}{n} \cdot \sum_{i=1}^{k} (x_i - \bar{x}_B)^2 \cdot n_i.$$

У випадку *інтервального варіаційного ряду* замість значень формула для визначення вибіркової дисперсії має наступний вигляд:

$$d_B = \sum_{i=1}^{k} (y_i - \bar{x}_B)^2 \cdot p_i^*.$$

Зауважимо, що зручніше використовувати наступну формулу для обчислення *вибіркової дисперсії*:

$$d_B = \overline{x_B^2} - (\bar{x}_B)^2,$$

де $\overline{x_B^2}$ – вибіркове середнє квадратів варіант вибірки:

$$\overline{x_B^2} = \frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2.$$

Вибіркова дисперсія є *зміщеною оцінкою* дисперсії генеральної сукупності. У якості незміщеної оцінки дисперсії потрібно брати *виправлену вибіркову дисперсію*, яка позначається як s^2 та обчислюється за формулою:

$$s^2 = \frac{n}{n-1} \cdot d_B.$$

Крім вибіркової дисперсії розглядають також *вибіркове середнє квадратичне* або *стандартне відхилення*:

$$\sigma_B = \sqrt{d_B}\,.$$

Приклад 3. Обчислимо вибіркову дисперсію, вибіркове середнє квадратичне відхилення та виправлену вибіркову дисперсію для дискретного варіаційного ряду розподілу з Прикладу 1:

x_i	2	5	7	10
n_i	5	26	10	9

Використаємо зручну формулу $d_B = \overline{x_B^2} - (\bar{x}_B)^2$, де вибіркове середнє вже обчислене та дорівнює $\bar{x}_B = 6$. Тобто необхідно ще знайти вибіркове середнє квадратів варіант вибірки:

$$\overline{x_B^2} = \frac{1}{n} \cdot \sum_{i=1}^{4} x_i \cdot n_i = \frac{1}{50} \cdot (2^2 \cdot 5 + 5^2 \cdot 26 + 7^2 \cdot 10 + 10^2 \cdot 9) = \frac{2060}{50} = 41{,}2.$$

Отже вибіркова дисперсія дорівнює:

$$d_B = \overline{x_B^2} - (\bar{x}_B)^2 = 41{,}2 - 6^2 = 5{,}2.$$

Вибіркове середнє квадратичне відхилення:

$$\sigma_B = \sqrt{d_B} = \sqrt{5{,}2} \approx 2{,}28.$$

Виправлена вибіркова дисперсія:

$$s^2 = \frac{n}{n-1} \cdot d_B = \frac{50}{50-1} \cdot 5{,}2 \approx 5{,}31.$$

Розглянуті вище статистики називають *аналітичними*. Аналітичними оцінками також є *емпіричний коефіцієнт асиметрії* та *емпіричний ексцес*. У

математичній статистиці застосовують також *порядкові* статистики. До таких статистик належать *емпірична мода*, *емпірична медіана*.

Означення. *Емпіричною* (*вибірковою*) *модою* Mo^*X називається варіанта, якій відповідає найбільша частота. Особливість моди в тому, що вона не змінюється при зміні крайніх членів ряду, тобто має певну стійкість до варіації.

Моду легко відшукати по таблиці, і ще легше на полігоні частот – це абсциса найвищої точки. Іноді таких значень декілька (з однаковою максимальною частотою), і тоді модою вважають кожне з них. Отже, для *дискретного варіаційного ряду* $Mo^*X = x_l$, якщо $n_l = \max\limits_i n_i$, тобто варіанті x_l відповідає найбільша частота.

У випадку *інтервального варіаційного ряду* з рівними інтервалами *модальний інтервал* (який містить *моду*) визначається за найбільшою частотою (при нерівних інтервалах – за найбільшою відносною частотою). Якщо маємо *інтервальний варіаційний ряд* з рівними інтервалами, то модальний інтервал легко визначити ще й за гістограмою – це інтервал з найвисоким стовпчиком.

Формула для знаходження *моди* у випадку *інтервального варіаційного ряду* має наступний вигляд:

$$Mo^*X = x_{i-1} + h \cdot \frac{n_i - n_{i-1}}{2n_i - n_{i-1} - n_{i+1}},$$

де x_{i-1} – початок модального інтервалу (з найбільшою частотою), h – довжина інтервалу, n_i – частота модального інтервалу, n_{i-1} – частота інтервалу, який передує модальному інтервалу, n_{i+1} – частота інтервалу, який є наступним за модальним інтервалом.

Означення. *Емпіричною* (*вибірковою*) *медіаною* Me^*X називають середню за розташуванням варіанту *дискретного варіаційного ряду*, якщо кількість варіант – непарна, і середнім арифметичним двох середніх за

розташуванням варіант, якщо кількість варіант – парна. Тобто медіану обчислюють наступним чином:

$$Me^*X = \begin{cases} x_{\frac{n+1}{2}}, \ при \ n - непарному; \\ \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, \ при \ n - парному. \end{cases}$$

Для визначення *медіани інтервального варіаційного ряду* необхідно визначити *медіанний інтервал*. Якщо, наприклад, на *i*-му інтервалі (x_{i-1}, x_i) значення емпіричної функції розподілу $F^*(x_{i-1}) < 0{,}5$ і $F^*(x_i) > 0{,}5$, то в середині цього інтервалу знаходиться *медіана*, яка знаходиться за формулою:

$$Me^*X = x_{i-1} + h \cdot \frac{\frac{n}{2} - \sum_{m=1}^{i-1} n_m}{n_i},$$

де x_{i-1} – початок медіанного інтервалу, тобто такого, якому відповідає перша з частот, що перевищує половину всіх спостережень, h – довжина інтервалу, n_i – частота медіанного інтервалу.

Елементи теорії кореляції

Поняття функціональної та статистичної залежностей

Економічні процеси характеризуються багатьма ознаками, які відображають ті або інші властивості. Ці признаки змінюються у часі та просторі, і часто зміни ознак взаємопов’язані. В одних випадках зв’язок (залежність) між ознаками виявляється дуже тісною (наприклад, годинне вироблення та заробітна платня), а в інших випадках зв’язок між ознаками зовсім не виявляється або виражається дуже слабо (наприклад, стать студентів та їх успішність). Чим тісніший зв’язок між ознаками, тим легше керувати системами. У кожному

конкретному зв’язку одні ознаки виступають у якості факторів, що впливають на інші та обумовлюють їх змінення, а інші ознаки – у якості результатів дії факторів впливу. Іншими словами, одні ознаки являють собою причину, інші – наслідок.

Розрізняють два типи залежності між процесами та їх ознаками: функціональну та статистичну.

Означення. *Функціональна залежність* – це зв’язок, при якому кожному значенню незалежної змінної x відповідає точно визначене значення залежної змінної y.

Для соціально-економічних процесів характерним є те, що на них впливають багато випадкових факторів, і залежність у більшості випадків є статистичною – «у загальному та середньому», при великому числі спостережень.

Означення. *Статистична залежність* – це зв’язок, при якому кожному значенню незалежної змінної x відповідає множина значень залежної змінної y, причому невідомо заздалегідь яке саме значення прийме y.

Кореляційна залежність. Коваріація, коефіцієнт кореляції

Окремим важливим випадком *статистичної залежності* є *кореляційна залежність*.

Означення. *Кореляційна залежність* – це зв’язок, при якому кожному значенню незалежної змінної x відповідає визначене математичне сподівання (середнє) залежної змінної y.

Кореляційний зв’язок є так званою «неповною» залежністю, яка проявляється не в кожному окремому випадку, а лише у середніх величинах при достатньо великій кількості випадків.

Залежність, функціональна або кореляційна, може бути прямою або зворотною:

– при *прямій залежності* зі збільшенням значень однієї ознаки зростають значення другої ознаки, а зі зменшенням значень однієї ознаки зменшуються значення другої ознаки;
– при *зворотній (оберненій) залежності* зі збільшенням значень однієї ознаки зменшуються значення другої ознаки, і навпаки – зі зменшенням значень однієї ознаки збільшуються значення другої ознаки

За аналітичним виразом залежність може бути *лінійною* та *нелінійною.*

В залежності від кількості ознак *кореляційні зв'язки* розділяють на *однофакторні* та *багатофакторні*.

Означення. *Кореляційний аналіз* – це розділ математичної статистики, який вивчає взаємозв'язки між випадковими величинам, та полягає у кількісному визначенні тісноти взаємозв'язку між ознаками.

Означення. *Кореляція* – це статистична залежність між випадковими величинами, при якій зміна однієї випадкової величини приводить до зміни математичного сподівання другої випадкової величини.

Нехай досліджується можливий взаємозв'язок двох ознак X та Y на основі отриманих відповідних спостережень – вибірок однакового об'єму n:

X	x_1	x_2	…	x_n
Y	y_1	y_2	…	y_n

Тіснота *взаємозв'язку* ознак X та Y кількісно виражається *величиною коефіцієнта кореляції*, який пов'язаний з обчисленням суми добутків відхилень значень x_i та y_i, що спостерігаються при випробуваннях, від їх середніх $\bar{x}$ та $\bar{y}$:

$$\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y}).$$

Сума таких добутків, розділена на об'єм вибірки n, називається *коваріацією*:

$$\mathrm{cov}(y,x)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{n},$$

де n – об'єм вибірки,

x_i та y_i – значення ознак X та Y, що спостерігаються при випробуваннях (дані вибірок),

$\bar{x}$ та $\bar{y}$ – вибіркові середні ознак X та Y.

На практиці часто використовують наступну формулу для обчислення *коваріації*:

$$\mathrm{cov}(y,x)=\overline{xy}-\bar{x}\cdot\bar{y},$$

де

$$\bar{x}=\frac{x_1+x_2+\ldots+x_n}{n}=\frac{1}{n}\cdot\sum_{i=1}^{n}x_i,$$

$$\bar{y}=\frac{y_1+y_2+\ldots+y_n}{n}=\frac{1}{n}\cdot\sum_{i=1}^{n}y_i,$$

$$\overline{xy}=\frac{x_1y_1+x_2y_2+\ldots+x_ny_n}{n}=\frac{1}{n}\cdot\sum_{i=1}^{n}x_iy_i.$$

Коваріація характеризує спряженість варіацій двох ознак та являє собою статистичну міру взаємодії двох випадкових величин.

1. При наявності *прямого зв'язку* більші значення x_i з'являються з більшими значеннями y_i, і, відповідно, відхилення $(x_i-\bar{x})$ та $(y_i-\bar{y})$ одночасно будуть додатними; менші значення x_i з'являються з меншими значеннями y_i, і, відповідно, відхилення $(x_i-\bar{x})$ та $(y_i-\bar{y})$ одночасно будуть від'ємними. Отже при *прямій залежності* добутки $(x_i-\bar{x})\cdot(y_i-\bar{y})$ будуть додатними величинами.

2. При наявності *зворотного зв'язку* відхилення $(x_i - \bar{x})$ та $(y_i - \bar{y})$ будуть мати різні знаки, і, відповідно, добутки $(x_i - \bar{x}) \cdot (y_i - \bar{y})$ будуть від'ємними величинами.
3. При *відсутності зв'язку* поєднання знаків будуть невпорядковані, і при сумуванні від'ємні та додатні добутки $(x_i - \bar{x}) \cdot (y_i - \bar{y})$ будуть взаємно скорочуватись, отже коваріація за значенням буде близька до нуля.

Розмірність коваріації залежить від величин значень ознак *X* та *Y*, що є недостатньо зручним при дослідженні. Для зручності використовують відносну характеристику зв'язку – для цього коваріацію ділять на максимально можливе значення, яке дорівнює добутку середньоквадратичних відхилень двох ознак σ_x та σ_y, і в результаті отримують *лінійний коефіцієнт кореляції*:

$$r_{yx} = \frac{\operatorname{cov}(y,x)}{\sigma_x \cdot \sigma_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{n \cdot \sigma_x \cdot \sigma_y},$$

де

$$\sigma_x = \sqrt{d_x} = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\overline{x^2} - \left(\bar{x}\right)^2},$$

$$\sigma_y = \sqrt{d_y} = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2} = \sqrt{\overline{y^2} - \left(\bar{y}\right)^2}.$$

На практиці часто використовують наступну формулу для обчислення *коефіцієнту кореляції*:

$$r_{yx} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - \left(\bar{x}\right)^2} \cdot \sqrt{\overline{y^2} - \left(\bar{y}\right)^2}}.$$

Коефіцієнт кореляції r_{yx} приймає значення від –1 до +1:

$$-1 \le r_{yx} \le 1.$$

Додатне значення коефіцієнта говорить про наявність прямого зв'язку, від'ємне значення – про наявність зворотного. Якщо $r_{yx} = -1$ або $r_{yx} = 1$, кореляційний зв'язок представляється лінійною функціональною залежністю, при $r_{yx} = 0$ лінійний кореляційний зв'язок відсутній.

Якісні характеристики кореляційного зв'язку

Значення r_{yx}	*Характер зв'язку*
від 0 до $\pm 0{,}3$	практично відсутній
від $\pm 0{,}3$ до $\pm 0{,}5$	слабий
від $\pm 0{,}5$ до $\pm 0{,}7$	помірний
від $\pm 0{,}7$ до ± 1	сильний

Знаючи лінійний коефіцієнт кореляції, можна визначити *парний коефіцієнт детермінації*, який дорівнює $(r_{yx})^2$ та показує, яка доля варіації змінної *y* обумовлена впливом на неї зміною *x*.

Приклад. В результаті статистичного дослідження отримані дані з середньодушового доходу населення та середньодушового обороту роздрібної торгівлі за містами регіону (місяць).

№ міста	Середньодушовий дохід населення	Середньодушовий оборот роздрібної торгівлі	№ міста	Середньодушовий дохід населення	Середньодушовий оборот роздрібної торгівлі
1	8357	5425	9	8563	5200
2	8135	5057	10	8219	4892
3	7841	4683	11	8308	5008
4	8991	5375	12	8724	5225

5	7293	4167	13	8416	4983
6	8340	4925	14	8022	5342
7	8089	4042	15	8383	5458
8	9372	5925	16	9267	5125

Обчислимо лінійний коефіцієнт кореляції за формулою

$$r_{yx} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - (\bar{x})^2} \cdot \sqrt{\overline{y^2} - (\bar{y})^2}}.$$

Знайдемо всі необхідні вибіркові середні:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \frac{1}{16} \cdot \sum_{i=1}^{16} x_i = \frac{1}{16} \cdot (8357 + 8135 + \ldots + 9267) = 8395;$$

$$\bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i = \frac{1}{16} \cdot \sum_{i=1}^{16} y_i = \frac{1}{16} \cdot (5425 + 5057 + \ldots + 5125) = 5052;$$

$$\overline{xy} = \frac{1}{16} \cdot \sum_{i=1}^{16} x_i y_i = \frac{1}{16} \cdot (8357 \cdot 5425 + 8135 \cdot 5057 + \ldots + 9267 \cdot 5125) =$$

$$= \frac{1}{16} \cdot 681173270 = 42573329{,}4;$$

$$\overline{x^2} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2 = \frac{1}{16} \cdot \sum_{i=1}^{16} x_i = \frac{1}{16} \cdot (8357^2 + 8135^2 + \ldots + 9267^2) =$$

$$= \frac{1}{16} \cdot 1131688278 = 70730517{,}4;$$

$$\overline{y^2} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i^2 = \frac{1}{16} \cdot \sum_{i=1}^{16} y_i = \frac{1}{16} \cdot (5425^2 + 5057^2 + \ldots + 5125^2) =$$

$$= \frac{1}{16} \cdot 411662886 = 25728930{,}4.$$

Отже, лінійний коефіцієнт кореляції дорівнюватиме:

$$r_{yx} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - (\bar{x})^2} \cdot \sqrt{\overline{y^2} - (\bar{y})^2}} =$$

$$= \frac{42573329{,}4 - 8395 \cdot 5052}{\sqrt{70730517{,}4 - 8395^2} \cdot \sqrt{25728930{,}4 - 5052^2}} = 0{,}7062208 \approx 0{,}71 > 0.$$

Величина коефіцієнта кореляції говорить про сильний прямий зв'язок між середньодушовий дохід населення та середньодушовий оборот роздрібної торгівлі.

Обчислимо парний коефіцієнт детермінації:

$$(r_{yx})^2 = 0{,}7062208^2 \approx 0{,}498,$$

який показує, що на 49,8% зміна обороту роздрібної торгівлі пояснюється зміною середньодушового доходу населення.