Розділ 10

Статистичні методи вивчення взаємозв'язку даних

У цьому розділі буде розглянуто:

- статистичні ряди розподілу;
- поняття кореляційного зв'язку та коефіцієнта кореляції;
- кореляційну матрицю;
- рівняння та лінії тренду;
- прогнозування даних.

Статистичні ряди розподілу

У попередньому розділі ми досліджували вибірки, дані в яких були незгруповані, тобто являли собою просто послідовності чисел. За такою послідовністю можна обчислити певні статистичні показники, але неможливо визначити *тенденцію* зміни значень досліджуваної ознаки. Наприклад, якщо є відомості про доходи 1000 осіб, можна визначити середній дохід, стандартне відхилення величини доходу, проте важко сказати, як змінюється кількість осіб, що отримують той чи інший дохід, зі зростанням його величини. Щоб дати відповідь на це питання, потрібно *згрупувати* дані, наприклад визначити кількість людей, що отримують дохід до 1000 грн, від 1000 до 2000 грн, від 2000 до 3000 грн тощо. У результаті ми отримаємо таблицю на кшталт табл. 10.1.

Таблиця 10.1. Відомості про щомісячні доходи населення

Величина доходу, грн	До 1000	1000–2000	2000–3000	3000–4000	4000–5000	5000–6000	6000–7000	7000–8000	Понад 8000
Кількість осіб	250	302	211	91	52	35	20	12	27

Побудована таблиця називається *статистичним рядом розподілу*. Загалом ряд розподілу — це два набори значень однакової довжини. В одному наборі представлені значення певної *ознаки* (у табл. 10.1 це величина доходу), а в іншому — *частоти*, тобто кількості разів, коли під час статистичного спостереження було отримано відповідне значення ознаки. Інакше кажучи, ідеться про розподіл певних об'єктів за певною ознакою. Наприклад, у табл. 10.1 наведено розподіл осіб за величиною доходу. Величина доходу — це ознака, а кількості осіб — частоти.

За рядом розподілу вже можна визначити тенденцію зміни значень досліджуваної ознаки. Так, з табл. 10.1 видно, що з ростом доходу від 0 до 2000 грн кількість осіб, які отримують цей дохід, зростає, а коли дохід перевищує 2000 грн, тенденція зворотна: що вище дохід, то менша кількість людей його отримує.

Атрибутивні та варіаційні ряди розподілу

Розрізняють атрибутивні та варіаційні ряди розподілу. Якщо за основу групування узята якісна ознака, то це *атрибутивний* ряд розподілу (розподіл за видами продукції, професіями, статтю, національною або географічною приналежністю тощо). Якщо ряд розподілу побудований за кількісною ознакою, то такий ряд є *варіаційним* (за розміром доходу, стажем роботи, числом працівників на підприємстві тощо).

Наприклад, наведений у табл. 10.1 ряд розподілу осіб за доходом є варіаційним, а ряд розподілу осіб за професіями, який наведено у табл. 10.2, — атрибутивним.

Таблиця 10.2. Приклад атрибутивного ряду розподілу

Професія	Менеджер	Медичний працівник	Військовослужбовець	Працівник освіти
Кількість осіб	151	78	92	105

Дискретні та інтервальні ряди розподілу

Варіаційні ряди розподілу, у свою чергу, поділяються на дискретні та інтервальні. У *дискретному* ряді розподілу частоти зіставляються окремим значенням ознаки, а в *інтервальному* — інтервалам таких значень. Так, ряд розподілу у табл. 10.1 є інтервальним.

У табл. 10.3 наведено приклад дискретного ряду розподілу — це розподіл кількостей випадання чисел на гральній кістці. Значення ознаки у дискретному ряді називають *варіантами*.

Таблиця 10.3. Приклад дискретного ряду розподілу

Число на гральній кістці	1	2	3	4	5	6
Кількість випадань	50	43	51	47	39	53

Інтервальний ряд розподілу можна перетворити на дискретний, взявши за значення варіант середини інтервалів. Так, у табл. 10.4 наведено дискретний ряд, який побудовано за інтервальним рядом, поданим у табл. 10.1. Зверніть увагу: хоча останній інтервал мав вигляд [8000;∞), за його середину ми взяли число 8500, припустивши, що відстань між двома останніми значеннями ознаки дорівнює відстані між передостанніми значеннями:

$$7500 - 6500 = 1000;\ 7500 + 1000 = 8500.$$

Таблиця 10.4. Дискретний ряд розподілу осіб за доходами

Величина доходу, грн	500	1500	2500	3500	4500	5500	6500	7500	8500
Кількість осіб	250	302	211	91	52	35	20	12	27

Очевидно, що атрибутивні ряди розподілу можуть бути тільки дискретними.

Абсолютні та відносні частоти

В усіх розглянутих нами рядах розподілу наведено *абсолютні частоти*, які визначають, скільки разів зустрічається певне значення ознаки. Проте часто в рядах розподілу вказують і *відносні частоти*, що дорівнюють часткам, які припадають на ту чи іншу частоту в загальному об'ємі вибірки. Приклад ряду розподілу з відносними частотами наведено в табл. 10.5.

Таблиця 10.5. Ряд розподілу з відносними частотами

Варіанти	x_1	x_2	...	x_k
Відносні частоти	$\frac{m_1}{n}$	$\frac{m_2}{n}$	$\frac{m_i}{n}$	$\frac{m_k}{n}$

Тут x_i — варіанти, m_i — абсолютні частоти, $i = 1, 2,..., k$; k — кількість різних за значенням варіант; n — об'єм вибірки.

У табл. 10.6 наведено ряд розподілу з відносними частотами, побудований на основі даних з табл. 10.1.

Таблиця 10.6. Ряд розподілу щомісячних доходів населення з відносними частотами

Величина доходу, грн	До 1000	1000–2000	2000–3000	3000–4000	4000–5000	5000–6000	6000–7000	7000–8000	Понад 8000
Кількість осіб	0,25	0,302	0,211	0,091	0,052	0,035	0,02	0,012	0,027

Побудова рядів розподілу

Припустимо, що результати статистичних спостережень необхідно згрупувати, побудувавши ряд розподілу. Ця операція виконується у кілька етапів. Насамперед необхідно визначити, який ряд розподілу будувати — інтервальний чи дискретний. Критерій такий: якщо ознака може набувати лише невелику кількість різних значень (у межах одного-двох десятків), будуйте дискретний ряд розподілу, інакше — інтервальний.

ПРИМІТКА. Не плутайте випадок, коли ознака *представлена у вибірці* невеликою кількістю значень, з випадком, коли вона *може набувати невеликої кількості значень* у генеральній сукупності. Наприклад, якщо є вибірка з відомостями про зріст семи людей, то це ще не означає, що величина «зріст» може мати лише сім значень. А якщо є вибірка днів тижня, то величина «день тижня» дійсно може набувати лише семи різних значень.

Для побудови дискретного ряду розподілу слід виписати всі можливі значення ознаки, а потім підрахувати, скільки разів кожне з них трапляється у вибірці — це будуть частоти. У Microsoft Excel для підрахунку частот слід застосувати функцію **COUNTIF** (рос. **СЧЕТЕСЛИ**), про яку йшлося в розділі 3. Розв'язання цієї задачі ми залишимо для самостійної роботи (див. завдання для самостійної роботи 1). А от принцип побудови інтервального ряду розподілу розглянемо детальніше.

Отже, для побудови за вибіркою $x_1,\ldots,x_n$ ряду розподілу, що складається з m рівних інтервалів, необхідно виконати такі кроки.

1. Визначити найбільшу та найменшу варіанти — $x_{\min}$ та $x_{\max}$.
2. Визначити величину інтервалу $h=\frac{x_{\max}-x_{\min}}{m}$.

3. Визначити межі інтервалів $[y_0;y_1]$, $[y_1;y_2]$, ..., $[y_{m-1}, y_m]$ за формулами:

$$y_0 = x_{\min};\ y_{i+1} = y_i + h,\ i = 0, ..., m - 1.$$

Тобто нижня межа першого інтервалу дорівнює найменшій варіанті, а кожна наступна межа більша за попередню на h.

4. Підрахувати, скільки варіант потрапляє у кожен інтервал — це і будуть частоти. В Excel це можна зробити за допомогою функції FREQUENCY (рос. ЧАСТОТА), яка має два аргументи:

FREQUENCY(*діапазон_вибірки;діапазон_меж_інтервалів*)

Перший аргумент — це діапазон, що містить вибірку, а другий — діапазон усіх меж інтервалів, за винятком y_0 та y_m (тобто усіх меж між інтервалами). Результатом функції буде набір частот, що відповідають кожному інтервалу. Ви вперше стикаєтеся з функцією, результатом якої є діапазон значень, а не окреме значення. Її і вводити потрібно дещо інакше, ніж інші функції. А саме, слід виділити весь діапазон, де міститимуться результати, ввести формулу функції та натиснути клавіші Ctrl+Shift+Enter.

Приклад використання функції FREQUENCY наведено на рис. 10.1, *а*.

	A	B	C	D	E	F	G
1	Виб	x_{min}= 0		Межа інтервалів		Інтервал	Частота
2	75	x_{max}= 100		20		0-20	5
3	30	h= 20		40		20-40	5
4	13			60		40-60	2
5	37			80		60-80	4
6	11					80-100	4
7	66						
8	87						
9	96						
10	36						
11	22						
12	71						
13	100						
14	27						
15	65						
16	1						
17	58						
18	53						
19	3						
20	94						
21	0						

а

=FREQUENCY(A2:A21;D2:D5)

E	F	G	H
	Інтервал	Частота	
	0-20	;D2:D5)	
	20-40		
	40-60		
	60-80		
	80-100		

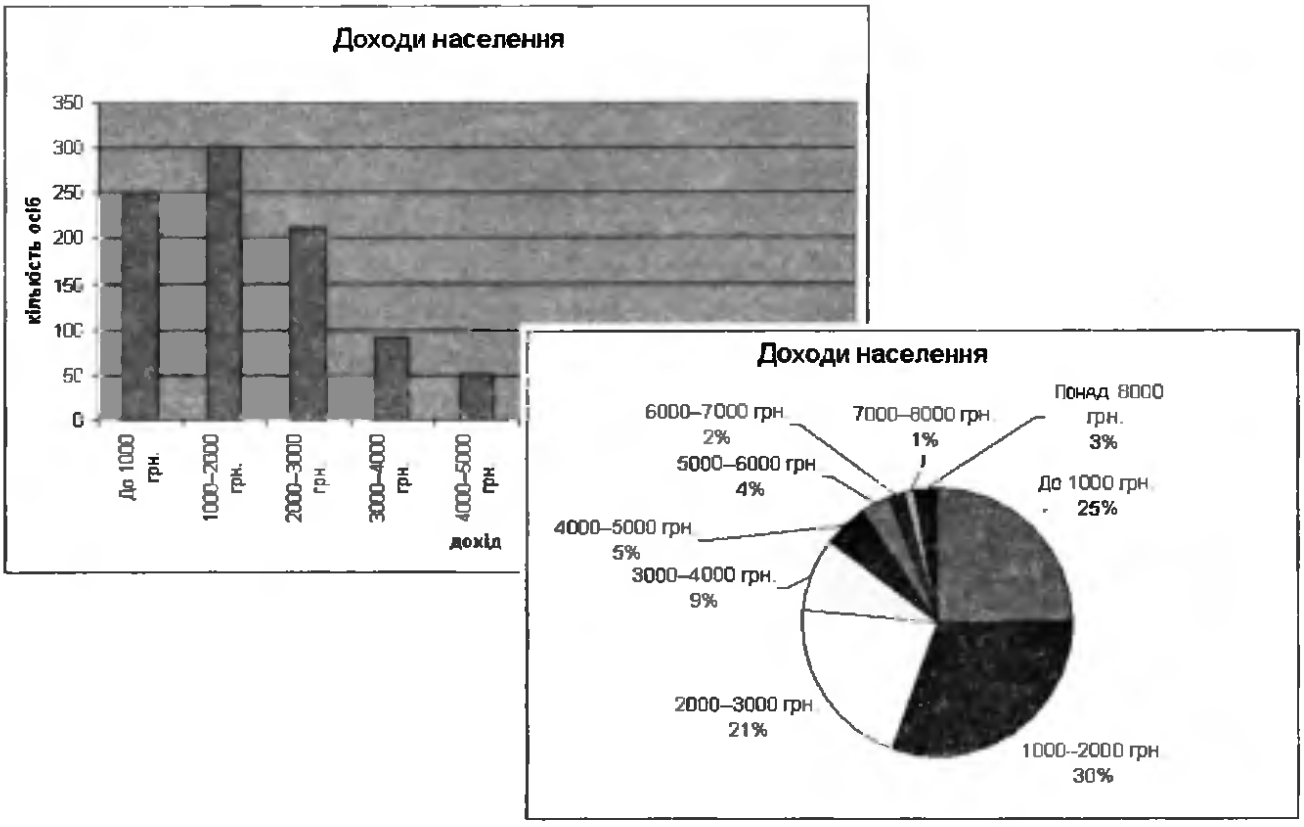
б

Рис. 10.1. Побудова інтервального ряду розподілу: *а* — результати обчислень; *б* — введення функції

Тут вибірка міститься в діапазоні A2:A21, $x_{min} = 0$, $x_{max} = 100$ і нам потрібно побудувати ряд розподілу з п'яти інтервалів. Межами між інтервалами будуть числа 20, 40, 60, 80 — вони містяться в діапазоні D2:D5. Функцію FREQUENCY введено в діапазон G2:G6, де ми бачимо результати її обчислення, тобто частоти. Процес введення функції FREQUENCY зображено на рис. 10.1, *б*.

Графічне подання рядів розподілу

Тенденції зміни частот зручно вивчати, коли ряд розподілу подано у графічному вигляді. Найчастіше для зображення рядів розподілу застосовують гістограму, а за необхідності графічно зобразити відносні частоти — кругову діаграму. На гістограмі значення ознаки відкладаються на осі x, а частоти — на осі y. Так, на рис. 10.2, *а* ряд розподілу з табл. 10.1 зображено у вигляді гістограми, а на рис. 10.2, *б* — у вигляді кругової діаграми. З гістограми відразу видно тенденцію зміни кількості осіб із ростом щомісячного доходу.



а *б*

Рис. 10.2. Графічне подання ряду розподілу: *а* — у вигляді гістограми; *б* — у вигляді кругової діаграми

Вправа 10.1. Побудова інтервального ряду розподілу

У файлі **Вправа_10_1.xls** наведено відомості про зріст учнів класу. Потрібно побудувати ряд розподілу учнів за зростом з п'ятьма рівними інтервалами, зобразити його графічно та зробити висновок щодо характеру зв'язку між зростом та кількістю учнів цього зросту.

1. Відкрийте файл **Вправа_10_1.xls**. У клітинки **C1** та **C2** уведіть формули для обчислення мінімального і максимального зросту учня: **= MIN(A2:A21)** та **=MAX(A2:A21)**. Ці значення мають бути такими: $x_{\min} = 151$, $x_{\max} = 176$.
2. У клітинці **C3** обчисліть величину інтервалу групування $h = \frac{x_{\max} - x_{\min}}{5}$. Вона повинна дорівнювати 5.
3. Обчисліть межі між інтервалами у клітинках **D2:D5**. У клітинці **D2** обчисліть значення межі $y_1 = x_{\min} + h$. У клітинку **D3** уведіть формулу **=D2+C$3**. Скопіюйте цю формулу у клітинки **D4:D5**, і ви отримаєте значення всіх інших меж. Фактично ми реалізували формулу $y_{i+1} = y_i + h$. Оскільки значення y_i змінюється, посилання **D2** є відносним. А оскільки величина h незмінна, номер рядка у посиланні **C$3** зафіксовано.
4. Уведіть межі інтервалів у клітинки **F2:F6** (рис. 10.3).

	A	B	C	D	E	F	G
1	Зріст учня, см	x_{min}=	151	Межі інтервалів		Інтервал	Частота
2	151	x_{max}=	176	156		151-156	1
3	167	h=	5	161		156-161	4
4	160			166		161-166	9
5	159			171		166-171	4
6	176					171-176	2
7	163						
8	170						
9	157						
10	165						
11	162						
12	164						
13	168						
14	173						
15	166						
16	163						
17	164						
18	160						
19	169						
20	162						
21	164						

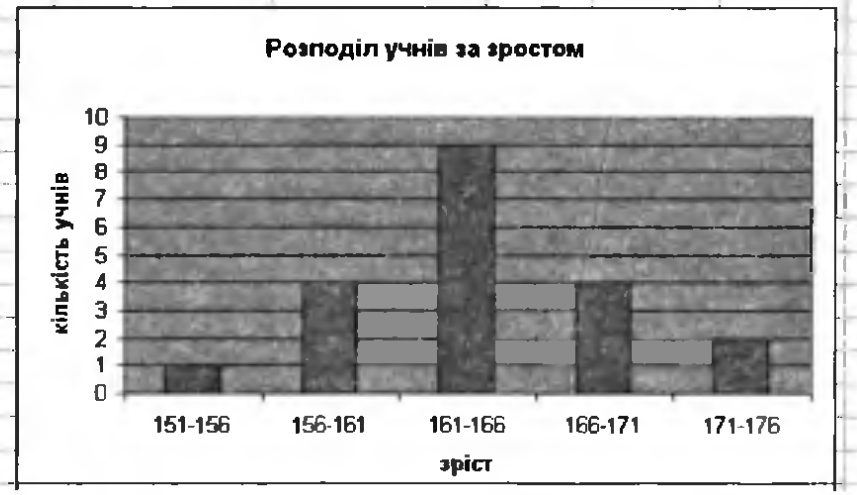


Рис. 10.3. Побудова та графічне подання ряду розподілу

5. Виділіть діапазон G2:G6 та, скориставшись кнопкою **fx** (Вставка функції), уведіть функцію FREQUENCY. Її аргументи будуть такими: діапазон вибірки — A2:A21, діапазон меж інтервалів — D2:D5. Увівши аргументи функції, не клацайте кнопку OK, а натисніть клавіші Ctrl+Shift+Enter. Частоти буде обчислено.
6. Самостійно створіть гістограму частот (див. рис. 10.3). Як будувати та форматувати діаграми, ви знаєте з розділу 4.
7* Уведіть у клітинку F2 формулу, після копіювання якої в діапазон F3:F6 у ньому буде автоматично відображено інтервали, як на рис. 10.3.

Обчислення статистичних показників варіаційних рядів розподілу

Якщо вибірку подано у вигляді варіаційного ряду розподілу, а не як набір варіант, то формули для обчислення середнього та стандартного відхилення будуть дещо складніші, ніж ті, які ми розглядали в попередньому розділі. Отже, припустимо, що є такий ряд розподілу, як показано в табл. 10.7.

Таблиця 10.7. Загальний вигляд ряду розподілу

Варіанти	x_1	x_2	...	x_k
Частоти	n_1	n_2	...	n_k

Середнє значення вибірки обчислюється за формулою середньої арифметичної зваженої:

$$\overline{x} = \sum_{i=1}^{n} x_i n_i \Big/ \sum_{i=1}^{n} n_i. \qquad (1)$$

Тут кожне значення варіанти «зважується», тобто множиться на відповідну їй частоту.

Дисперсію варіаційного ряду найлегше обчислити за такою формулою:

$$\sigma^2 = \overline{x^2} - \left(\overline{x}\right)^2 = \sum_{i=1}^{n} x_i^2 n_i \Big/ \sum_{i=1}^{n} n_i - \left(\overline{x}\right)^2. \qquad (2)$$

Ну а стандартне відхилення — це корінь із дисперсії:

$$\sigma = \sqrt{\sigma^2}. \qquad (3)$$

У Microsoft Excel спеціальних функцій для обчислення статистичних показників за формулами (1)–(3) не передбачено. Тому визначати ці величини потрібно, обчислюючи кожну суму за допомогою функції SUM (рос. СУММ) з подальшим застосуванням необхідних арифметичних перетворень.

За потреби обчислити статистичні показники для інтервального ряду розподілу його спочатку слід перетворити на дискретний, як це описано в підрозділі «Дискретні та інтервальні ряди розподілу».

Формули для визначення моди, медіани й асиметрії дискретних та інтервальних рядів розподілу ми не наводимо через їхню складність. Зазначимо також, що для атрибутивного ряду розподілу узагальнюючі статистичні показники обчислити взагалі неможливо.

Вправа 10.2. Обчислення статистичних показників

У файлі Вправа_10_2.xls показано ряд розподілу підприємств міста N за прибутком. Обчисліть середній прибуток та стандартне відхилення прибутку цих підприємств. Зробіть висновки.

1. Відкрийте файл Вправа_10_2.xls. У ньому на аркуші Аркуш1 наведено інтервальний ряд розподілу: у стовпці A вказано нижні межі інтервалів, у стовпці C — верхні, у стовпці D зазначено частоти.
2. Щоб обчислити статистичні показники, інтервальний ряд розподілу необхідно перетворити на дискретний. Для цього слід насамперед обчислити середини інтервалів. Уведіть у клітинку E2 формулу =(A2+C2)/2, скопіюйте її в діапазон E3:E11, і середини інтервалів буде відображено у стовпці E.
3. Обчисліть величини $n_i x_i$ та $n_i x_i^2$, де n_i — частоти, а x_i — середини інтервалів. Для цього введіть у клітинки F2 та G2 формули =D2*E2 та =D2*E2^2 і скопіюйте їх у діапазон F3:G11.
4. Обчисліть суми величин n_i, $n_i x_i$ та $n_i x_i^2$ у клітинках D12, F12 і G12, скориставшись функцією SUM.
5. Визначте у клітинках E15:E16 середнє значення та стандартне відхилення за формулами (1)–(3). Ви маєте отримати такі значення, як на рис. 10.4.

	A	B	C	D	E	F	G
1	Обсяг прибутку, тис. грн			Кількість підприємств, n_i	Середини інтервалів, x_i	nx_i	nx_i^2
2	0	-	50	250	25	6250	156250
3	50	-	100	177	75	13275	995625
4	100	-	150	130	125	16250	2031250
5	150	-	200	112	175	19600	3430000
6	200	-	250	86	225	19350	4353750
7	250	-	300	87	275	23925	6579375
8	300	-	350	64	325	20800	6760000
9	350	-	400	39	375	14625	5484375
10	400	-	450	32	425	13600	5780000
11	450	-	500	23	475	10925	5189375
12			Разом	1000		158600	40760000
13							
14							
15				Середнє:	158,6		
16				Стандартне відхилення:	124,924137		

Рис. 10.4. Обчислення статистичних показників ряду розподілу

6. Виходячи з отриманих значень середнього та стандартного відхилення, зробіть висновки щодо розподілу підприємств за величиною прибутку.

Основи кореляційного та регресійного аналізу

У ряді розподілу зіставляються дві послідовності значень: певної ознаки та частот. Залежність між цими послідовностями простежується не завжди. Наприклад, з поданої на рис. 10.2, *а* гістограми видно, що залежність між величиною доходу та кількістю осіб, які мають такий дохід, існує: що більший дохід, то менша кількість осіб його отримує. Якщо ж подивитися на табл. 10.3, то стає зрозуміло, що залежність між числом на гральній кістці та кількістю його випадань, скоріш за все, відсутня: значення частот зі зростанням числа на кістці не зменшуються і не збільшуються, а «стрибають», поводять себе у певному розумінні випадково.

Побудова ряду розподілу дає змогу зробити лише приблизні, «інтуїтивні» висновки щодо того, чи існує залежність між значеннями ознаки та частотами і який вона має характер. Крім того, залежності можуть існувати між довільними вибірками, а не лише між ознакою та частотами. Більш точне дослідження залежнос-

тей між двома чи більшою кількістю вибірок є завданням спеціальних розділів математичної статистики — кореляційного та регресійного аналізу. *Кореляційний аналіз* дає змогу встановити, чи існує зв'язок між явищами і наскільки цей зв'язок сильний (часто його називають *кореляційним зв'язком*). Якщо зв'язок виявився суттєвим, то доцільно скористатися методами *регресійного аналізу*, основне завдання якого полягає у визначенні характеру зв'язку і побудові його математичної моделі. На основі моделі можна передбачити ту або іншу подію, спрогнозувати, як будуть розвиватися певні процеси у разі змінення характеристик об'єкта дослідження.

Факторні та результативні ознаки

Перш ніж застосовувати кореляційний аналіз, варто визначити, які з досліджуваних ознак є *факторними* (такими, що від них залежать інші), а які — *результативними* (такими, що самі залежать від інших). Як приклад розглянемо дані про кількість хронічно хворих на астму та концентрацію чадного газу в кількох містах (табл. 10.8). Очевидно, що коли між цими ознаками існує залежність, то саме кількість хронічно хворих залежить від концентрації чадного газу, а не навпаки. Тобто концентрація чадного газу є факторною ознакою, а кількість хронічно хворих на астму — результативною.

Таблиця 10.8. Значення факторної та результативної ознак

Концентрація чадного газу, мг/м³	1,20	2,40	2,56	3,10	3,50	4,20	4,80
Кількість хронічно хворих на астму на 1000 жителів	20	35	42	48	51	59	63

Графічний аналіз кореляційного зв'язку

Як же визначити, чи існує залежність між двома ознаками? Найпростіший спосіб — побудувати *діаграму розсіювання* (рис. 10.5). У Microsoft Excel такі діаграми називають *точковими*. На осі X діаграми розсіювання розміщують значення факторної ознаки, на осі y — результативної.

	A	B	C	D	E	F	G	H
1	Захворюваність на астму							
2	Концентрація чадного газу, мг/м^3	1,2	2,4	2,56	3,1	3,5	4,2	4,8
3	Кількість хронічно хворих на астму на 1000 жителів	20	35	42	48	51	59	63

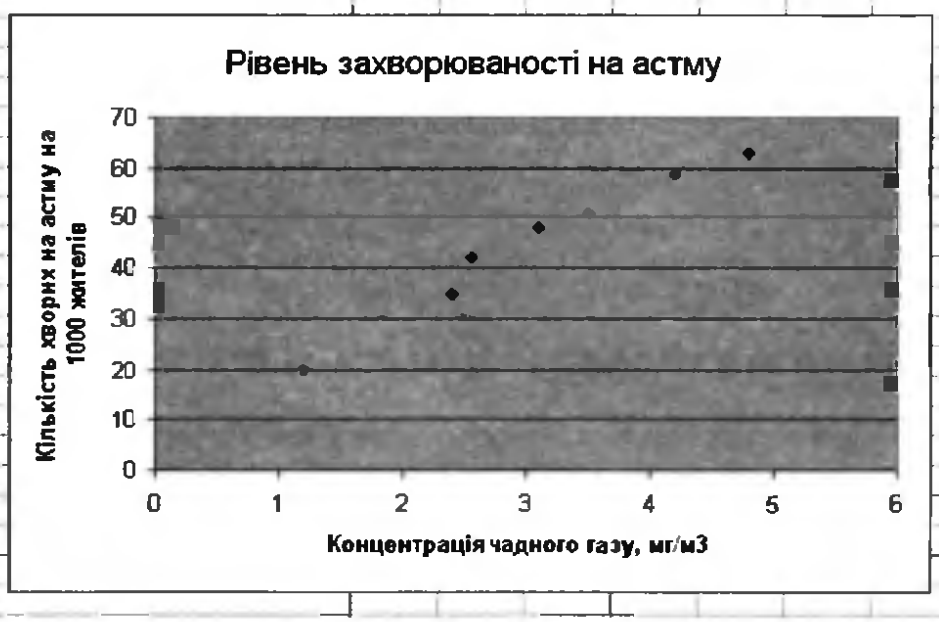
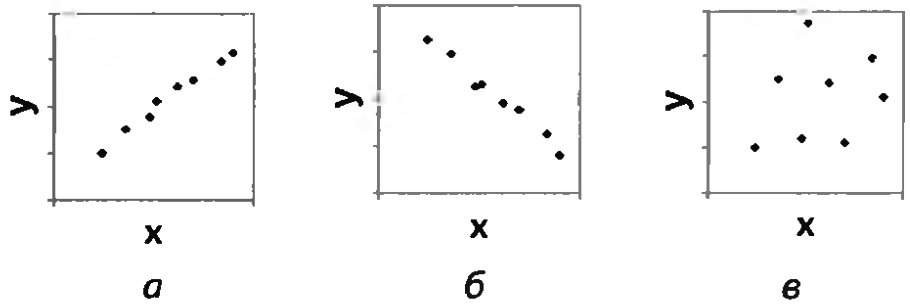


Рис. 10.5. Точкова діаграма, яка демонструє залежність між кількістю хронічно хворих на астму та рівнем концентрації чадного газу в повітрі

На цій діаграмі усі точки розташовані вздовж деякої уявної лінії, спрямованої зліва знизу вправо вверх. Називається вона *лінією тренду*. Саме через таку спрямованість лінії тренду можна говорити про наявність *прямого* кореляційного зв'язку між ознаками (рис. 10. 6, *а*): що вища концентрація чадного газу, то вищий рівень захворюваності на астму. Коли лінія тренду спрямована вправо вниз (рис. 10. 6, *б*), кореляційний зв'язок є *оберненим*, а якщо дані розсіяні хаотично і напрямок лінії тренду визначити важко (рис. 10.6, *в*), то кореляційний зв'язок взагалі відсутній.



а *б* *в*

Рис. 10.6. Кореляційний зв'язок між даними: *а* — прямий; *б* — обернений (*в* — зв'язок відсутній)

Коефіцієнт кореляції

Міцність зв'язку між двома величинами можна виразити і за допомогою *коефіцієнта кореляції*. Це число k з інтервалу $[-1, 1]$. Якщо k близьке до -1, то кореляційний зв'язок між величинами є *оберненим*, а якщо k близьке до 1 — *прямим*. Чим ближче k до нуля, тим кореляційний зв'язок слабший. Якщо говорити більш докладно, то міцність лінійного кореляційного зв'язку оцінюється так:

- $|k| \geq 0{,}8$ — сильний кореляційний зв'язок;
- $0{,}4 \leq |k| < 0{,}8$ — кореляційний зв'язок наявний;
- $|k| < 0{,}4$ — кореляційний зв'язок відсутній.

У Microsoft Excel для обчислення коефіцієнта кореляції використовується функція CORREL(*діапазон_1;діапазон_2*) (рос. КОРРЕЛ), де діапазони *діапазон_1* та *діапазон_2* містять набори значень, між якими шукається залежність. У разі визначення коефіцієнта кореляції двох вибірок, поданих на рис. 10.5, такими масивами будуть дані у діапазонах B2:H2 та B3:H3. Результатом функції CORREL у нашому випадку буде число 0,9862, що свідчить про наявність дуже сильного кореляційного зв'язку між концентрацією чадного газу в повітрі та кількістю хронічно хворих на астму.

Зазначимо, що функція CORREL визначає коефіцієнт *лінійної кореляції*, яка свідчить про наявність саме лінійного зв'язку між ознаками. Цей зв'язок буде тим сильніший, чим ближче до певної прямої розташовані точки на діаграмі розсіювання. Насправді існують й інші типи зв'язків. Наприклад, той факт, що точки на діаграмі розсіювання розташовані близько до певної параболи, свідчить про наявність між ознаками квадратичного зв'язку; щоправда, коефіцієнт лінійної кореляції при цьому може бути незначним.

Кореляційна матриця

Коли потрібно порівняти не два, а більше масивів експериментальних даних, будують *кореляційну матрицю* — таблицю, у якій коефіцієнти кореляції між ознаками розташовані на перетині відповідних рядків і стовпців. Для побудови кореляційної матриці використовують інструмент Кореляція, який запускається за допомогою команди Сервіс ▸ Аналіз даних ▸ Кореляція.

ПРИМІТКА. Якщо меню **Сервіс** не містить команди **Аналіз даних**, необхідно виконати команду **Сервіс ▸ Надбудови** та встановити прапорець **Пакет аналізу**.

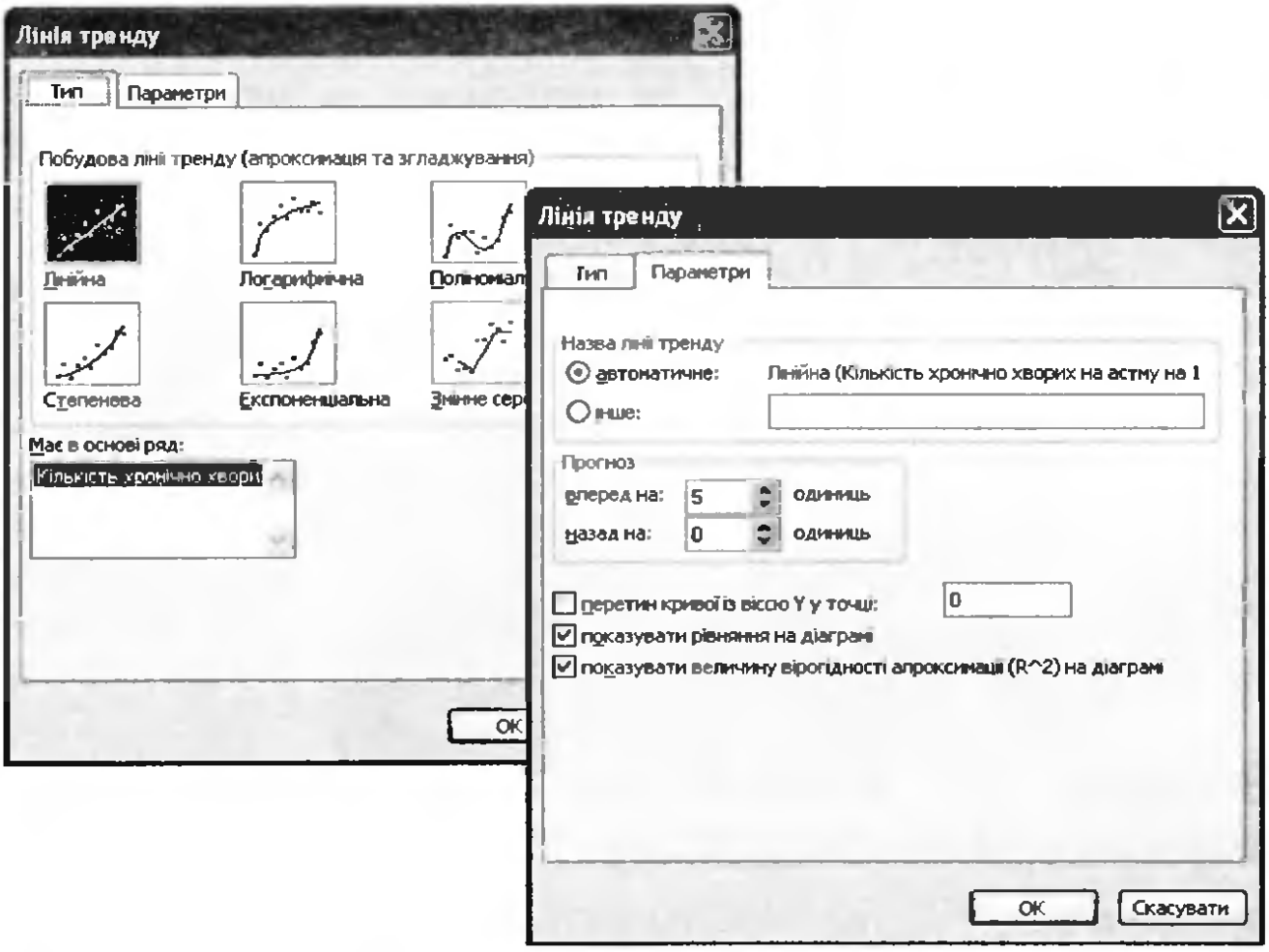
Регресійний аналіз

Як уже зазначалося, основне завдання регресійного аналізу — прогнозування. Щоб навести приклад задачі на прогнозування, повернімось до вибірок з табл. 10.8. Значення факторної ознаки (концентрації чадного газу), отримані в результаті статистичного спостереження, коливаються в межах від 1,2 до 4,8 мг/м^3. Для цих значень рівень захворюваності на астму відомий. Але задамося питанням: яким буде цей рівень, якщо концентрація чадного газу становитиме 10 мг/м^3? Тобто спробуємо спрогнозувати значення результативної ознаки у разі виходу значення факторної ознаки за межі інтервалу вибірки.

Основним методом, який використовується для прогнозування, є побудова на основі вибіркових даних *рівняння регресії* вигляду $y = f(x)$, що зв'язує факторну ознаку x і результативну ознаку y, та визначення за цим рівнянням невідомих значень результативної ознаки. Рівняння можна подати як аналітично (за допомогою формул), так і графічно. Згадана вище лінія тренду — це не що інше, як графік рівняння регресії.

У Microsoft Excel передбачена можливість автоматичної побудови лінії тренду. Для цього спочатку слід виділити діаграму розсіювання та виконати команду Діаграма ▸ Додати лінію тренду. Далі у вікні Лінія тренду на вкладці Тип (рис. 10.7, *а*) потрібно вибрати тип залежності між факторною та результативною ознаками — лінійна, поліноміальна (квадратична, кубічна тощо), логарифмічна та ін. На вкладці Параметри цього вікна (рис. 10.7, *б*) можна задати, зокрема, величину прогнозу (на скільки прогнозоване значення буде більшим за найбільше вибіркове чи меншим за найменше вибіркове). Це роблять за допомогою лічильників вперед на та назад на в області Прогноз.

На рис. 10.8 показано графік лінії тренду, доданий до точкової діаграми, зображеної на рис. 10.5. Величина прогнозу вперед для цього графіка становить 5 одиниць. З графіка видно, що за концентрації чадного газу 10 мг/м^3 рівень захворюваності на астму становитиме приблизно 120 людей на 1000 жителів міста.



а *б*

Рис. 10.7. Діалогове вікно **Лінія тренду**: *а* — вкладка **Тип**; *б* — вкладка **Параметри**

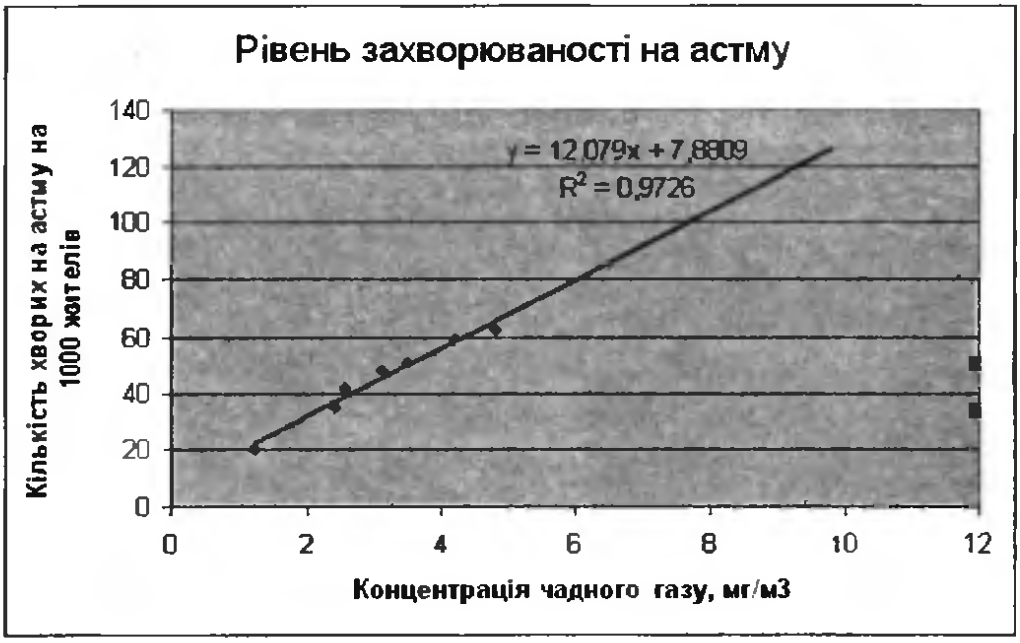


Рис. 10.8. Графік лінії тренду

Коефіцієнт детермінації

Близькість рівняння регресії та лінії тренду до вибіркових даних характеризується величиною *коефіцієнта детермінації* R^2 ($0 \le R^2 \le 1$). Рівняння регресії найбільше відповідає дійсності, коли R^2 наближається до свого максимального значення. Цей показник використовується в першу чергу для порівняння різних моделей прогнозу та вибору найкращої з них. На точковій діаграмі як значення R^2, так і саме рівняння регресії можна відобразити біля лінії тренду (див. рис. 10.8). Для цього на вкладці Параметри вікна Лінія тренду слід встановити прапорці **показувати величину вірогідності апроксимації (R^2) на діаграмі** та **показувати рівняння на діаграмі** (див. рис. 10.7, *б*). Для лінії тренду, яка наведена на рис. 10.8, $R^2 = 0{,}9726$. Це означає, що лінійне рівняння регресії добре узгоджується з вибірковими даними.

Вправа 10.3. Виявлення кореляційного зв'язку

Протягом року продовольча компанія здійснювала рекламу своєї продукції шляхом виготовлення та розповсюдження рекламних листівок у кількості від 89 000 до 345 000 шт. за місяць. Потрібно визначити, чи був цей захід ефективним та як вплине на дохід компанії виготовлення та розповсюдження протягом місяця 500 000 листівок.

1. Створіть нову електронну таблицю, введіть у неї дані, зазначені на рис. 10.9, і збережіть документ у файлі **Вправа_10_3.xls**.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Відомості про рекламну кампанію												
2	Місяць	1	2	3	4	5	6	7	8	9	10	11	12
3	Кількість рекламних листівок, тис. шт	123	234	203	155	345	231	133	205	121	89	205	307
4	Дохід компанії, млн. грн	2	4	3,5	1,3	5	3	1,2	2,3	0,2	0,5	3	3,3
5													
6	Коефіцієнт кореляції												
7	Коефіцієнт детермінації - лінійний тренд												
8	Коефіцієнт детермінації - поліноміальний тренд												
9													

Рис. 10.9. Таблиця з вихідними даними

Оскільки нас цікавить залежність доходу від кількості поширених листівок, то кількість рекламних листівок є факторною ознакою, а дохід компанії — результативною.

2. Побудуйте для створеної таблиці точкову діаграму, скориставшись кнопкою (Майстер діаграм). На осі *X* має відображатися кількість листівок, на осі *У* — дохід компанії (рис. 10.10).

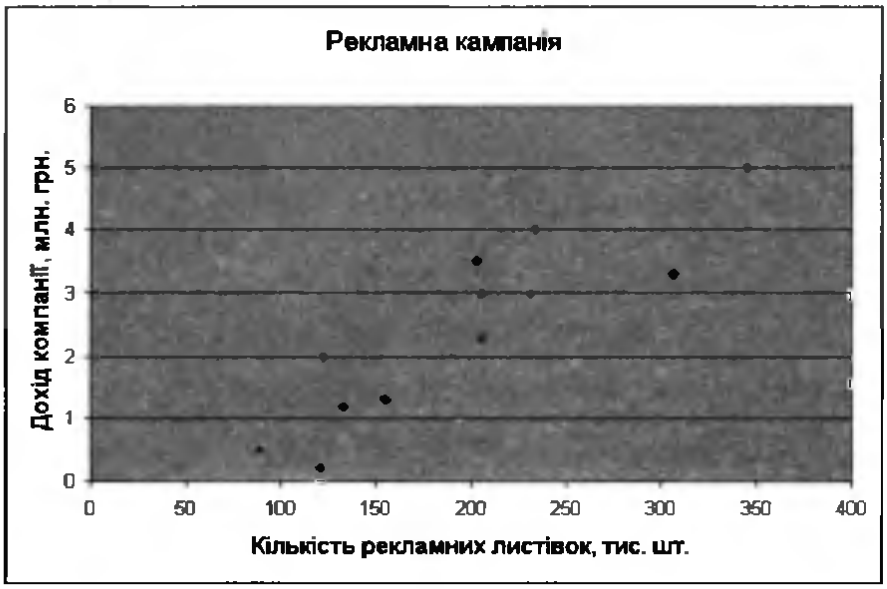


Рис. 10.10. Залежність доходу компанії від кількості проданих рекламних листівок

Як бачите, множина точок на діаграмі розсіювання витягнута зліва знизу вправо вверх. Це свідчить про існування прямого кореляційного зв'язку між кількістю розповсюджених рекламних листівок та доходом компанії.

3. Розрахуйте коефіцієнт кореляції, увівши в клітинку **B6** формулу **=CORREL(B3:M3;B4:M4)**. Отримане значення коефіцієнта кореляції (0,89) підтверджує висновок про наявність сильного прямого лінійного кореляційного зв'язку між кількістю розповсюджених рекламних листівок та доходом компанії. Отже, рекламний захід можна вважати ефективним.

4. Додайте до точкової діаграми лінію тренду лінійного типу з відображенням регресійного рівняння та значення коефіцієнта детермінації на діаграмі. Величину прогнозу вперед задайте рівною 200. Отримане значення $R^2 = 0{,}7924$ свідчить про те, що лінійна регресія достатньо добре відповідає вибірковим даним. Запишіть це значення у клітинці **B7**.

5. Перегляньте графік лінії тренду та визначте за ним, на який приблизно дохід компанії можна розраховувати в разі поширення 500 000 рекламних листівок за місяць (рис. 10.11).

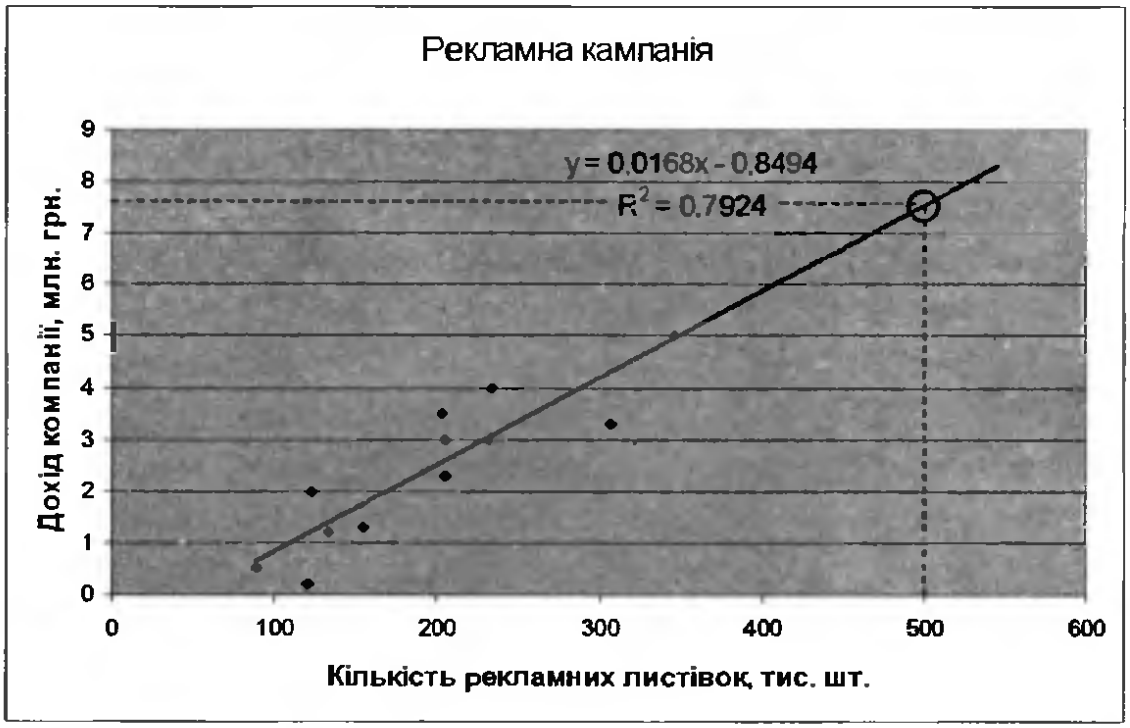


Рис. 10.11. Лінія тренду з прогнозом

6. Самостійно побудуйте поліноміальні лінії тренду другого і третього степенів. Порівняйте коефіцієнти детермінації та значення прогнозу для цих ліній тренду з відповідними значеннями для лінійного тренду. Зробіть висновки.

Практичні роботи профільного спрямування

Мета практичних робіт: закріпити уміння виявляти залежності між статистичними даними.

Практична робота 10 для групи профілів Б

Проведено статистичне спостереження за дев'ятьма респондентами, у кожного з яких виміряно масу тіла і частоту пульсу. Потрібно встановити, чи існує взаємозв'язок між цими параметрами, а також спрогнозувати, яким скоріш за все буде пульс у людини, маса тіла якої становить 100 кг.

Хід виконання

1. Створіть нову електронну книгу та введіть у неї дані, зазначені на рис. 10.12. Збережіть документ у файлі Практ_Б_10.xls.

	A	B
1	Маса, кг	Частота пульсу
2	55	65
3	89	80
4	68	66
5	70	68
6	88	69
7	53	65
8	85	80
9	70	66
10	88	69
11		
12	Коефіцієнт кореляції	

Рис. 10.12. Таблиця з вихідними даними

2. У клітинці B12 обчисліть коефіцієнт кореляції за формулою =CORREL(A2:A10;B2:B10). Він має дорівнювати 0,706. Проаналізуйте отриманий результат.
3. Для наочного відображення зв'язку між масою тіла людини та її пульсом побудуйте точкову діаграму (рис. 10.13). Логічно припустити, що саме маса тіла людини визначає частоту пульсу, а не навпаки. Тому маса тіла буде факторною ознакою, значення якої розміщуватимуться на осі X, а частота пульсу — результативною, і її значення вказуватимуться на осі Y.

ПРИМІТКА. Зверніть увагу: 50 є найменшим значенням на шкалі осі X. Щоб встановити для неї саме такий формат, відкрийте контекстне меню для елемента діаграми Вісь X(значень) та у вікні Формат осі на вкладці Шкала введіть **50** як мінімальне значення.

4. Побудуйте лінію тренду для створеної діаграми. Для цього виділіть діаграму та виконайте команду Діаграма ▸ Додати лінію тренду. На вкладці Параметри в області Прогноз у поле вперед на введіть значення 15. Задайте відображення рівняння регресії та коефіцієнта детермінації (рис. 10.13). Зробіть висновки щодо адекватності рівняння регресії вибірковим даним.
5. За лінією тренду складіть прогноз щодо частоти пульсу в людини з масою тіла 100 кг.

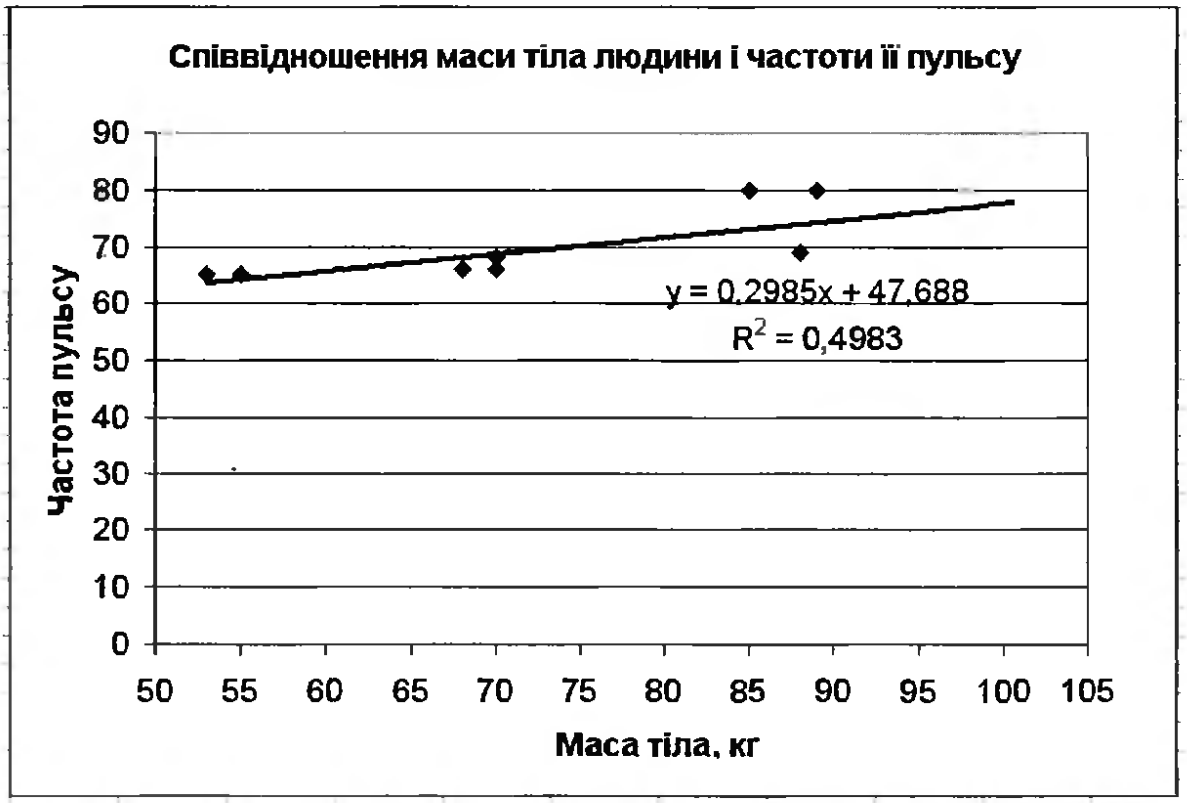


Рис. 10.13. Діаграма залежності між масою тіла людини та частотою її пульсу

Практична робота 10 для групи профілів М

Необхідно розв'язати задачу апроксимації трансцендентних функцій поліномами. Вона формулюється так: для заданої функції $f(x)$ потрібно знайти поліном $g(x) = a_n x^n + \ldots + a_1 x + a_0$, який найменше відхилявся б від функції $f(x)$ на відрізку $x \in [a;b]$. А саме, вам потрібно буде апроксимувати функцію $y = \sin x$ на відрізку $x \in [0;2\pi]$.

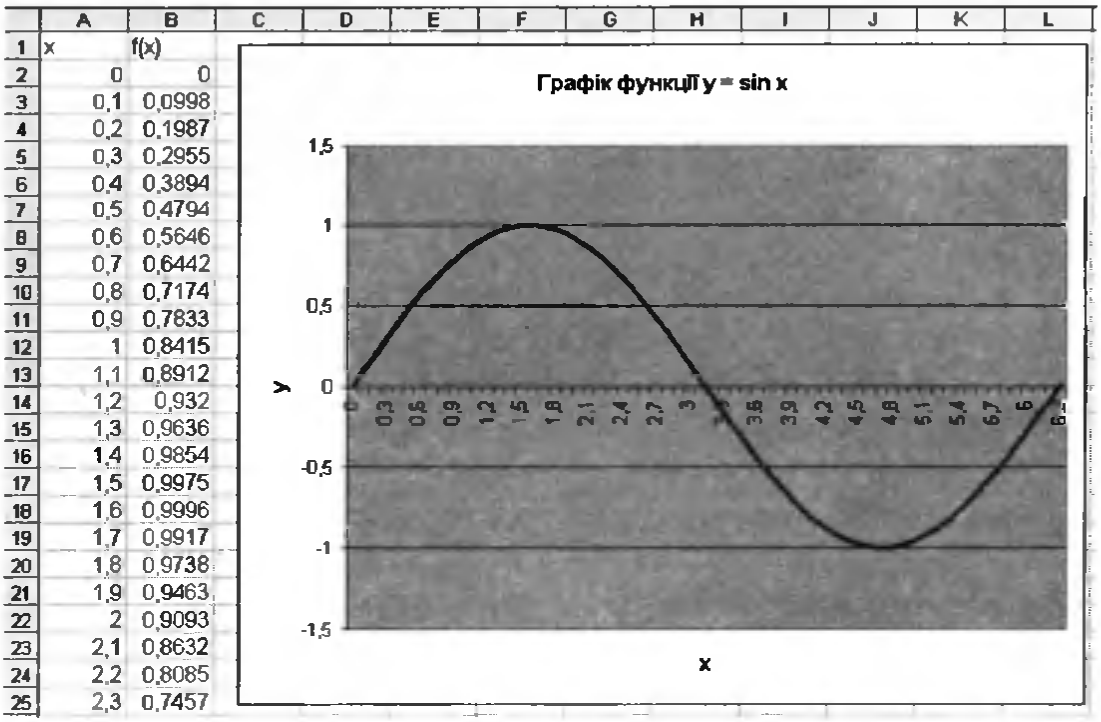
ПРИМІТКА. Апроксимація функцій поліномами широко використовується в комп'ютерних обчисленнях, зокрема в Microsoft Excel за потреби обчислити значення трансцендентної функції. Річ у тім, що алгоритм обчислення значення полінома в заданій точці очевидний, а от як саме комп'ютер має обчислювати значення трансцендентної функції, наприклад $y = \sin x$, не зрозуміло. Тому комп'ютер насправді обчислює не саму трансцендентну функцію, а поліном, який на невеликому відрізку з нею майже збігається.

Хід виконання

1. Запустіть табличний процесор Microsoft Excel і на першому аркуші електронної книги створіть таблицю значень функції $y = \sin x$ на проміжку $x \in [0;2\pi]$ (рис. 10.14). Значення x уведіть у клітинки стовпця A за допомогою прогресії з кроком

0,1. Врахуйте, що $2\pi \approx 6{,}3$. Збережіть новий документ під іменем Практ_М_10.xls.

2. За діапазоном вихідних даних A1:B65, скориставшись майстром діаграм, побудуйте графік функції $y = \sin x$ (рис. 10.14).



	A	B
1	x	f(x)
2	0	0
3	0,1	0,0998
4	0,2	0,1987
5	0,3	0,2955
6	0,4	0,3894
7	0,5	0,4794
8	0,6	0,5646
9	0,7	0,6442
10	0,8	0,7174
11	0,9	0,7833
12	1	0,8415
13	1,1	0,8912
14	1,2	0,932
15	1,3	0,9636
16	1,4	0,9854
17	1,5	0,9975
18	1,6	0,9996
19	1,7	0,9917
20	1,8	0,9738
21	1,9	0,9463
22	2	0,9093
23	2,1	0,8632
24	2,2	0,8085
25	2,3	0,7457

Рис. 10.14. Графік функції $y = \sin x$

3. Визначте степінь апроксимаційного полінома. З графіка функції на рис. 10.14 видно, що ані пряма, ані парабола не зможуть достатньо добре наблизитися до функції на всьому відрізку $x \in [0;2\pi]$. Проте функція вельми схожа на фрагмент кубічної параболи. Отже, степінь апроксимаційного полінома дорівнюватиме 3.

4. Графік апроксимаційного полінома — це не що інше, як поліноміальна лінія тренду в Excel. Отже, виділіть діаграму, виконайте команду Діаграма ▸ Додати лінію тренду, виберіть тип лінії Поліноміальна, у поле Степінь уведіть значення 3 і на вкладці Параметри задайте відображення на діаграмі рівняння регресії та коефіцієнта детермінації.

5. Клацніть кнопку ОК, і буде відображено графік апроксимаційного полінома, його рівняння $y = 0{,}00009x^3 - 0{,}009x^3 + 0{,}2066x - 0{,}3798$ та коефіцієнт детермінації $R^2 = 0{,}991$ (рис. 10.15). Близьке до 1 значення коефіцієнта детермінації свідчить про те, що поліном наближає трансцендентну функцію добре.

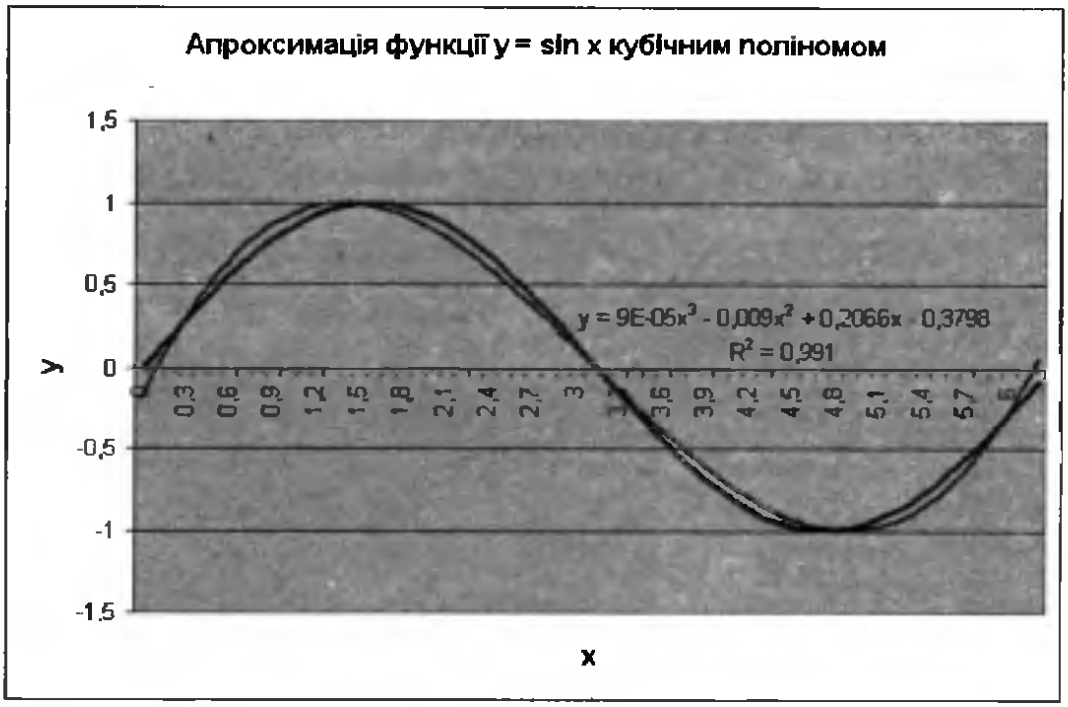


Рис. 10.15. Апроксимація трансцендентної функції поліномом

6.* Самостійно апроксимуйте поліномом функцію $y = e^x$ на відрізку $x \in [-2;2]$.

Практична робота 10 для групи профілів Е

Скориставшись відомостями про щомісячні відрахування на розвиток соціальної сфери підприємства та про рівень захворюваності його працівників протягом 12 років (рис.10.16), з'ясуйте, чи існує між цими показниками залежність, визначте її тип та спрогнозуйте, як зміниться рівень захворюваності працівників, якщо відрахування на розвиток соціальної сфери збільшити до 700 тис. грн. Визначте також, як зміниться рівень захворюваності у разі зменшення відрахувань до 100 тис. грн.

Хід виконання

1. Створіть нову електронну книгу та введіть у неї дані, зазначені на рис. 10.16. Збережіть документ у файлі Практ_Е_10.xls.
2. У клітинці C15 за формулою =CORREL(A2:A13;C2:C13) обчисліть коефіцієнт кореляції. Він має дорівнювати −0,94. Проаналізуйте отриманий результат.
3. Для наочного відображення зв'язку між відрахуваннями на розвиток соціальної сфери підприємства та рівнем захворюваності працівників побудуйте точкову діаграму (рис. 10.17).

Логічно припустити, що саме відрахування впливають на рівень захворюваності, а не навпаки. Тому обсяг відрахувань на соціальну сферу буде факторною ознакою, значення якої розміщуватимуться на осі X, а рівень захворюваності — результативною, її значення вказуватимуться на осі Y.

	A	B	C
1	Рік	Відрахування на розвиток соціальної сфери, тис. грн	Кількість пікарняних листків
2	1999	259	321
3	2000	370	260
4	2001	157	468
5	2002	259	318
6	2003	201	399
7	2004	201	420
8	2005	459	201
9	2006	257	350
10	2007	587	196
11	2008	129	452
12	2009	253	389
13	2010	257	370
14			
15		Коефіцієнт кореляції	

Рис. 10.16. Таблиця з вихідними даними

4. Побудуйте пряму лінію тренду для створеної діаграми. Для цього виділіть діаграму та виконайте команду **Діаграма ▸ Додати лінію тренду**. На вкладці **Параметри** в області **Прогноз** в обидва поля, **вперед на** і **назад на**, введіть значення 100. Задайте відображення рівняння регресії та коефіцієнта детермінації.
5. Коефіцієнт детермінації дорівнює 0,88, що свідчить про непогану відповідність лінії регресії вибірковим даним. Проте резерв для покращення значення R^2 теж є. Множина точок на діаграмі розсіювання дещо вигнута дугою вниз, і це свідчить про те, що, можливо, парабола відповідатиме їй краще, ніж пряма. Тож додайте до діаграми ще одну лінію тренду, на цей раз поліноміальну степеня 2, і задайте для неї відображення коефіцієнта детермінації. Він дорівнює 0,94 — отже, поліноміальна лінія тренду відповідає вибірковим даним краще.
6. Виділіть пряму лінію тренду та видаліть її, натиснувши клавішу **Del**. Клацніть поліноміальну лінію тренду правою кнопкою миші, виберіть з її контекстного меню команду **Формат лінії тренду** і в однойменному вікні на вкладці **Параметри** задайте відображення прогнозу вперед на 120 одиниць і назад

на 100 одиниць. Лінія тренду має набути такого вигляду, як на рис. 10.17.

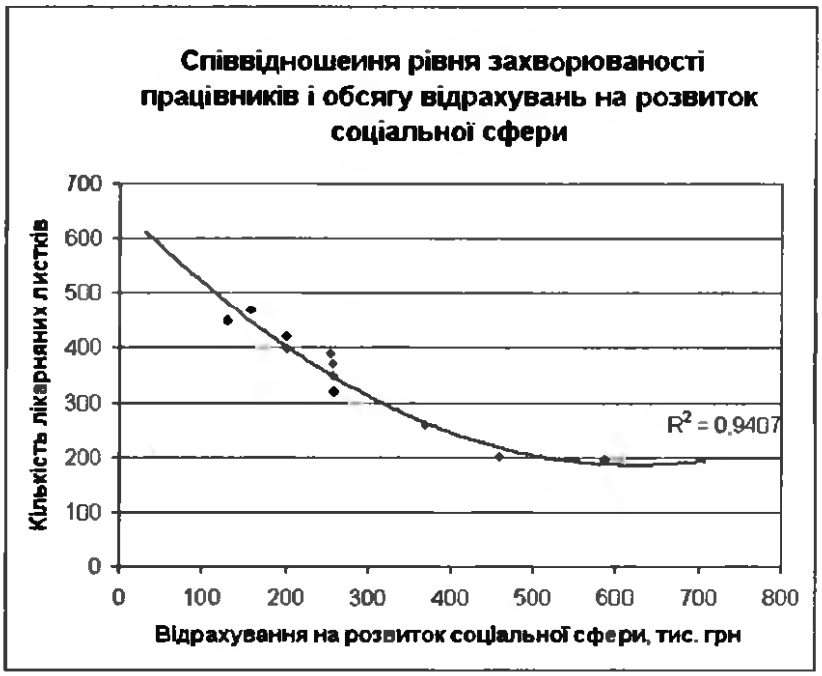


Рис. 10.17. Діаграма залежності між обсягом відрахувань на соціальну сферу та рівнем захворюваності

7. За лінією тренду складіть прогноз щодо рівня захворюваності у разі, якщо відрахування на соціальну сферу становитимуть 100 тис. і 700 тис. грн.

Самостійна робота

1. У файлі Самостійна_10_1.xls, у діапазоні A2:A101, наведено дані опитування 100 жителів міста щодо вподобань на виборах міського голови. У діапазоні C1:D10 побудуйте атрибутивний ряд розподілу кількості виборців, готових віддати голос за кожного з кандидатів. Прізвища кандидатів вже наведено у діапазоні C2:C10. Ви маєте ввести у клітинку D2 таку формулу, щоб, скопіювавши її в діапазон D3:D10, отримати потрібний результат. Використайте у формулі функцію COUNTIF (рос. СЧЕТЕСЛИ).
2. Знайдіть поліном, що апроксимує функцію $y = tgx$ на відрізку $x \in [-1{,}5;1{,}5]$.