МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
СХІДНОЄВРОПЕЙСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ЛЕСІ УКРАЇНКИ

Математичний факультет
Кафедра математичного аналізу

О.Г.Ханін

Статистичні методи в менеджменті та бізнесі

Методична розробка

.

Луцьк – 2014

Зміст

«Все, що не можна виразити у цифрах, -
не наука, а просто думка»
(Р.Хайнлайн, американський письменник)

ВСТУП

1. Статистика – наука і мистецтво. Генеральна сукупність і вибірка.

Перед менеджером будь-якого напрямку і рівня стоїть задача прийняття правильних управлінських рішень. Звичайно саме поняття «правильних» залежить від контексту задачі. Але, у будь-якому випадку, для прийняття правильних рішень потрібна інформація (*informare* з латині — «навчати»).

Приклад 1.1

Нехай нам відомі дані про щоденний товарообіг магазину на протязі місяця (тис.грн.):

55,42	51,43	65,05
46,02	36,93	56,82
41,90	46,37	59,25
56,32	53,93	53,36
39,75	54,12	56,11
57,27	42,71	45,82
48,64	57,85	46,58
48,20	44,38	29,18
51,77	65,63	50,47
45,02	53,76	45,06

Що дають нам ці дані? Спробуйте відповісти, дивлячись на ці дані, погано чи добре йдуть справи у магазині? Який рівень товарообігу очікувати завтра? Післязавтра? Тобто нам потрібно навчитись аналізувати дані і отримувати з них інформацію. Математичні методи, зокрема статистика, як раз і дозволяють з великої купи даних отримати корисну інформацію.

Таким чином, статистика – це наука та мистецтво збирання та аналізу даних. Треба лише зауважити, що статистика не вміє працювати з унікальними даними, а лише з багатократними, масовими даними. Ця наука гнучка, в певному розумінні близька до мистецтва, бо часто вона не пропонує жорстких рішень, а служить доповненням до вашої інтуїції та досвіду. Вона полегшує процес прийняття рішень, але не замінює цей процес.

Самі статистичні дані називаються вибірками бо їх, так би мовити, беруть з реального світу. А сам цей реальний світ, тобто всі можливі дані, називається генеральною сукупністю. Дані у вибірці називають її елементами. Кількість елементів у вибірці називається її об’ємом. Тобто у нашому прикладі ми мали вибірку об'єму 30. Зауважимо, що, як правило, елементи вибірки незалежні один від одного. Так знання сьогоднішнього товарообігу у магазині не визначає його точного значення завтра.

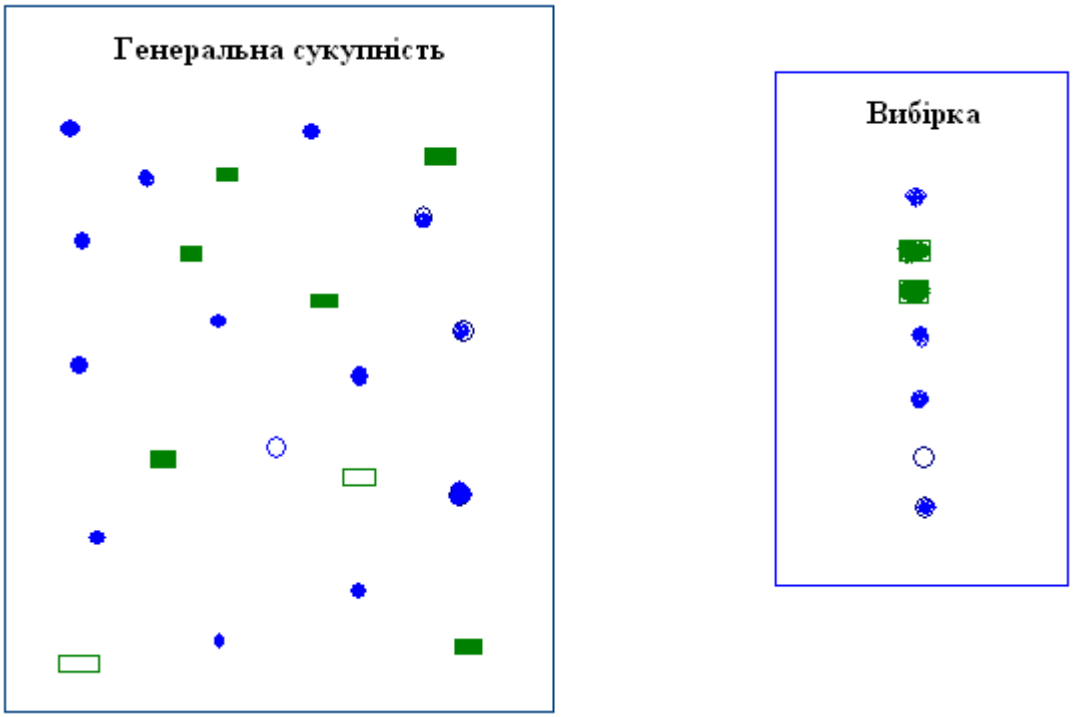
Перед тим як аналізувати дані, необхідно їх вміло зібрати. Припустимо, що ми хочемо скласти портрет покупця печива фірми «Роги і копита» для того, щоб визначити політику просування. Ми прийшли під стіни Волинського університету і опитали кожного десятого. І що отримали? Типовий користувач нашого печива – дівчина 17-22 років, незаміжня і голодна від напруженого навчання. Стосовно упаковки, то краще за все на ній зобразити симпатичного хлопця.

А як же бабусі зі своїми онуками? Та і самі онуки – школярі? А голодні чоловіки?

Статистики кажуть, що вибірка повинна бути репрезентативною, тобто адекватно відображати реальний світ – генеральну сукупність. **Адекватно – значить всі елементи генеральної сукупності повинні мати однакову ймовірність потрапити до вибірки.** Це призведе до того, що однакові елементи будуть зустрічатись у вибірці так само часто, як і у природі. Наприклад, якщо у місті – 10% одружених чоловіків віком від 25 до 35 років, то

приблизно такий відсоток вони мають складати серед опитаних. Це дуже непроста задача, в певному сенсі мистецтво. Галузь статистики, яка вивчає методи правильного складання вибірок, називається плануванням експерименту. Зауважимо, що іноді репрезентативної вибірки може не існувати, оскільки кожен об'єкт може бути унікальним. Тому на практиці намагаються отримати «достатньо» чи «практично» репрезентативну вибірку.

Рис.1.1



Вибірка на рис.1.1 є достатньо репрезентативною, але не повністю, т.я. до неї не увійшов жоден незамальований прямокутник.

Репрезентативність залежить від контексту задачі, так як цей контекст визначає саму генеральну сукупність. Повернемося до прикладу 1.1. Якщо ми збирали дані по місту Луцьку, то наша вибірка буде репрезентативна лише у межах міста, а не у межах усієї країни. Бо уся країна – це інша генеральна сукупність, і наша вибірка, яка була репрезентативною для однієї генеральної сукупності, не буде репрезентативною для іншої

Приклад 1.2.

Генеральна сукупність: 826 ящиків з різним комп'ютерним обладнанням, які щойно поступили на склад. Ви хотіли би перевірити на місці вміст окремих ящиків, щоб переконатися, наскільки вміст відповідає накладній. Розглянемо декілька способів.

Зручний спосіб полягає у тім, щоб узяти 10 найближчих ящиків і перевірити їх вміст. Але така вибірка навряд чи буде репрезентативною. До того ж, якщо постачальники раптом довідаються про вашу методику, то навряд чи ви зможете отримати користь з такої вибірки.

Ви можете узяти для перевірки також 2 великих ящики, три середніх та три маленьких. Але така вибірка буде нерепрезентативною, якщо більшість ящиків, скажімо, має великі розміри.

Можна взяти накладну і випадково відібрати певну кількість ящиків для перевірки. Потім необхідно знайти і відкрити усі відібрані ящики. Це буде найбільш правильна вибірка, оскільки гарантує рівну ймовірність потрапити до вибірки кожного ящика на складі, тобто кожного елементу вашої генеральної сукупності.

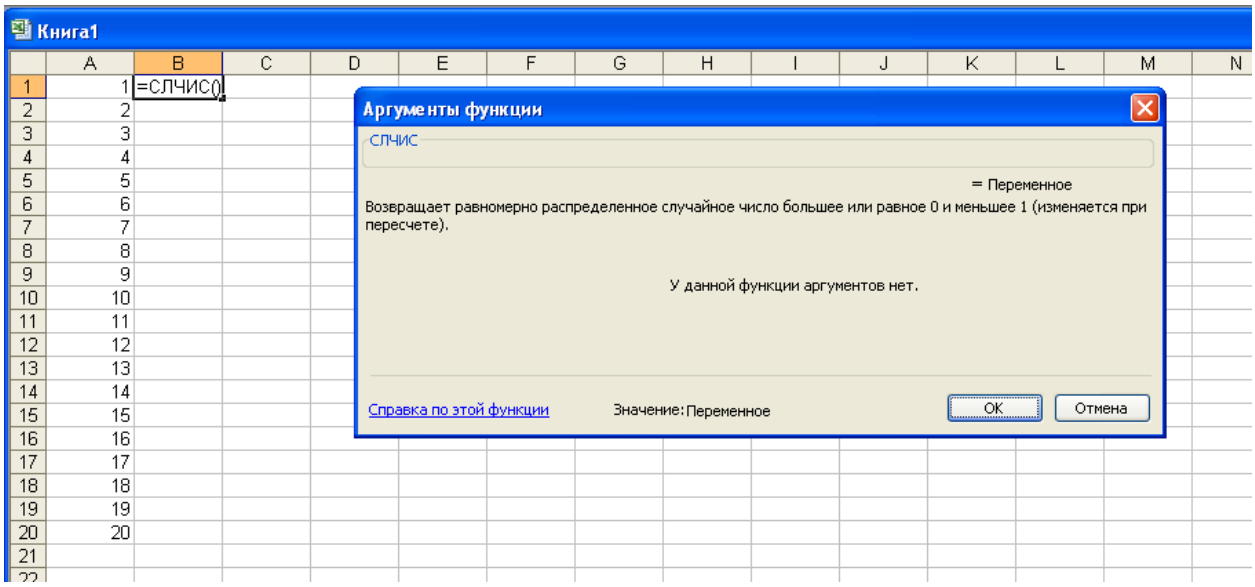
«Є три види брехні:
просто брехня, нахабна брехня і статистика»
(Марк Твен)

2. Методи побудови репрезентативних вибірок

Складність формування репрезентативної вибірки призводить до отримання помилкової інформації і хибних висновків. Крім того, як ми побачимо далі, отримана статистична інформація потребує змістовної інтерпретації. Все це разом і призводить до того, що у невмілих руках статистика стає особливим видом брехні.

Для формування репрезентативної вибірки на практиці можна поступити наступним чином. Нехай ми маємо генеральну сукупні об'єму ***N.*** Необхідно отримати репрезентативну вибірку об'єму n. Розташуємо в одному з стовпчиків аркуша Excel числа від 1 до ***N***. Далі заповнимо ***N*** рядочків сусіднього стовпчика рівномірно розподіленими випадковими числами. Випадкові числа – це значення деякої випадкової величини. Рівномірно розподіленими – значить, що ця випадкова величина може набувати кожне з своїх значень з однаковою ймовірністю. Для отримання таких випадкових значень скористаємось математичною функцією СЛЧИС(). Ця функція без аргументів при кожному її виклику повертає нове випадкове число від 0 до 1. Вставимо випадкове число у комірку B1 (рис. 2.1) і потягнемо виділення до комірки B20. Отримаємо ***N*** випадкових чисел (в нашому прикладі – 20) (рис. 2..2).

Рис. 2.1



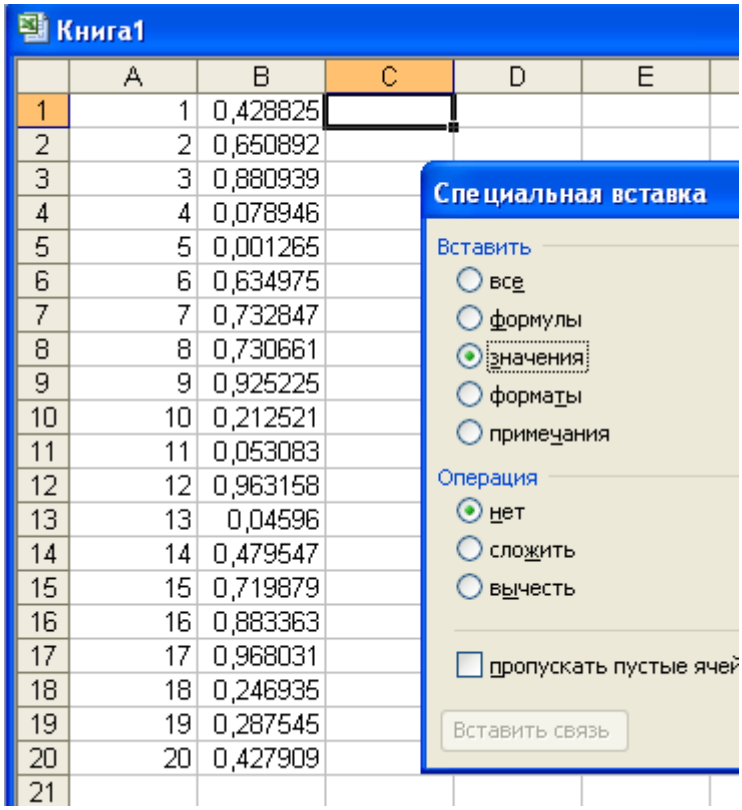
Виділимо тепер комірки B1:B20, скопіюємо і за допомогою спеціальної вставки вставимо у стовпчик C значення (рис. 2.3)

Рис. 2.2

Книга1

	A	B
1	1	0,428825
2	2	0,650892
3	3	0,880939
4	4	0,078946
5	5	0,001265
6	6	0,634975
7	7	0,732847
8	8	0,730661
9	9	0,925225
10	10	0,212521
11	11	0,053083
12	12	0,963158
13	13	0,04596
14	14	0,479547
15	15	0,719879
16	16	0,883363
17	17	0,968031
18	18	0,246935
19	19	0,287545
20	20	0,427909
21		

Рис.2.3



Відсортуємо стовпчики A, B і C так, щоб значення у третьому стовпчику були впорядковані за зростанням. Тоді елементи генеральної сукупності (числа у першому стовпчику) будуть перемішані випадковим чином (рис. 2.4).

Рис. 2.4

Книга1

	A	B	C
1	5	0,767719	0,001265
2	13	0,657822	0,04596
3	11	0,443486	0,053083
4	4	0,062645	0,078946
5	10	0,961634	0,212521
6	18	0,046203	0,246935
7	19	0,531485	0,287545
8	20	0,188983	0,427909
9	1	0,557305	0,428825
10	14	0,251856	0,479547
11	6	0,605534	0,634975
12	2	0,675216	0,650892
13	15	0,684209	0,719879
14	8	0,315982	0,730661
15	7	0,08622	0,732847
16	3	0,506211	0,880939
17	16	0,170773	0,883363
18	9	0,276585	0,925225
19	12	0,418713	0,963158
20	17	0,359942	0,968031
21			

Щоб сформувати репрезентативну вибірку об'єму ***n*** візьмемо з генеральної сукупності елементи з номерами, рівними першім n числам стовпчика А. Скажімо, якщо вибірка має об'єм 5, то до неї увійдуть елементи генеральної сукупності з номерами 5, 13, 11, 4, 10.

Іноді генеральна сукупність містить ясні, легко ідентифіковані групи, які розрізняються між собою, але однорідні всередині. Такі групи називаються стратами. Коли будується випадкова вибірка, може статися так, що певні страти будуть представлені недостатньо або навпаки надлишково. Тому доцільно здійснювати випадкові вибірки окремо у кожній страті генеральної сукупності, а потім об'єднати їх у єдину вибірку. Розміри вибірок у кожній страті можуть бути різними, але пропорційними тій частці, яку становить дана страта у генеральній сукупності.

Приклад 2.1.

Для розробки маркетингової стратегії просування високотехнологічної аудіо- та відеопродукції необхідна відповідна інформація про потенційних покупців. В залежності від обізнаності про дану технологію покупців можна розділити на 2 категорії: обізнані - які хочуть знати технічні подробиці продукції, та необізнані - яких влаштовує лише базова інформація загального характеру.

Щоб спрогнозувати, яку суму грошей у цьому році планує витратити на вашу продукцію типовий покупець, доцільно сформувати стратифіковану вибірку, оскільки розумно очікувати, що група обізнаних покупців планує більш крупні витрати, тобто страти суттєво розрізняються.

Ключові слова: генеральна сукупність, вибірка, об'єм вибірки, репрезентативна вибірка, стратифікована вибірка.

«Якщо вам зрозуміла мова чисел, то ви читаєте вже не числа,
подібно тому як у книзі ви читаєте не слова. Ви відразу читаєте зміст»
(Гарольд Дженін, американський менеджер)

Точкове оцінювання

3. Поняття про статистичну оцінку

Припустимо ми зарибили водойму і через рік вирішили дослідити, наскільки успішно просувається наш проект, тобто наскільки виросли наші мальки. Таким чином нас цікавить параметр генеральної сукупності - середня довжина риби у водоймі. Як визначити цей параметр? Зрозуміло, що точне його значення можна встановити лише вимірявши усіх риб, але це неможливо і недоцільно. В такому разі формують випадкову вибірку, тобто виловлюють певну кількість риб і вимірюють їх. Середня довжина риб, отримана по даних вибірки, може не збігатися точно з середньою довжиною по всій генеральній сукупності. Вона, як кажуть статистики, є оцінкою параметра генеральної сукупності, яким в нашому прикладі є середня довжина риби у водоймі. Із зростанням об'єму вибірки точність оцінки збільшується.

Даний приклад ілюструє застосування вибіркового методу і поняття статистичної оцінки, яка має наближений характер. Ця оцінка лягає в основу певних статистичних висновків. Якість статистичної оцінки визначають по тому, наскільки статистичні висновки, що ґрунтуються на ній, відображають реальну практику. Зауважимо, що рішення про адекватність чи неадекватність оцінки визначається дослідником в залежності від конкретної ситуації. В одному випадку похибка, скажімо у 1,5%, є незначною, у інших критичною. Якщо точність статистичної оцінки недостатня, необхідно збільшити об'єм вибірки.

4. Вибіркове середнє.

Припустимо, що в місті А - 4 супермаркети «Край», а в місті В – 8. Відомі місячні обсяги реалізації цукерок «Смак життя» по супермаркетах (тис. грн.):

Місто А		**Місто В**	
Супермаркет 1	1500	Супермаркет 1	1000
Супермаркет 2	1000	Супермаркет 2	2800
Супермаркет 3	2500	Супермаркет 3	4000
Супермаркет 4	3200	Супермаркет 4	1200
		Супермаркет 5	1900
		Супермаркет 6	2500
		Супермаркет 7	2000
		Супермаркет 8	1000

Необхідно вирішити, у якому місті краще купуються цукерки даного сорту.

Проаналізуємо ситуацію. Зрозуміло, що об'єм реалізації по місту В більший, але і супермаркетів там більше. Чи можна сказати, що при збільшенні кількості супермаркетів у місті А до 8 обсяг реалізації цукерок буде більшим, ніж у місті А?

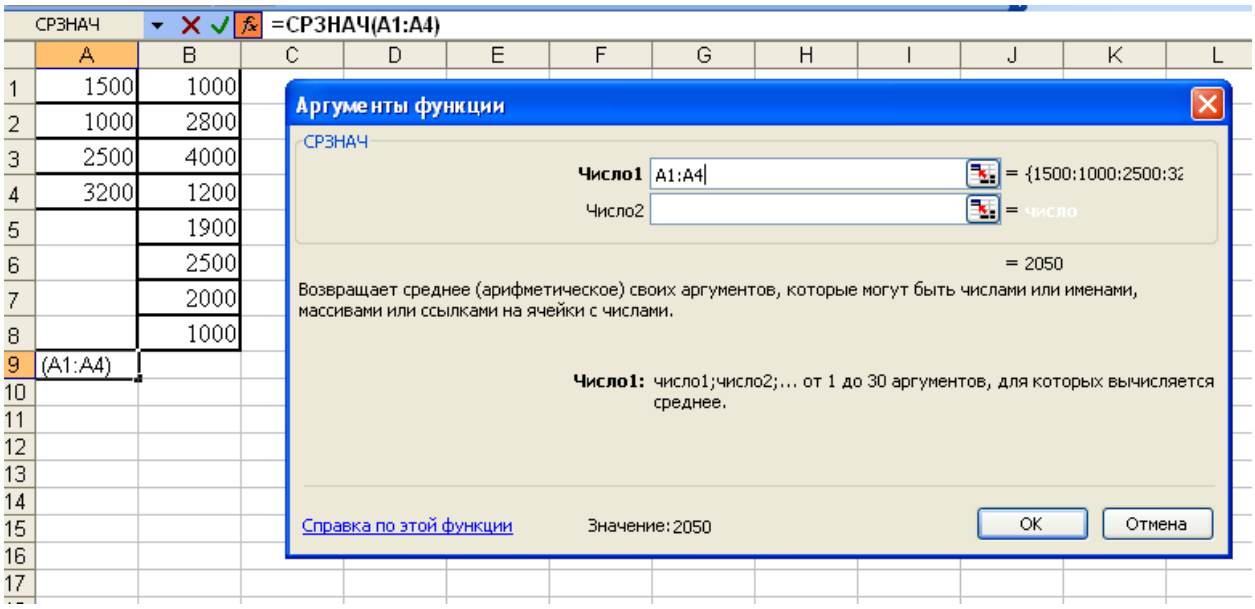
Одним з можливих шляхів розв'язання цього питання є обчислення вибіркового середнього.

Нехай ми маємо вибірку об'єму n з елементами $x_1, x_2, \dots, x_n$. Тоді вибіркове середнє

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Фактично, вибіркове середнє, яке є оцінкою середнього по всій генеральній сукупності, є просто середнім арифметичним. У Excel є статистична функція для обрахування вибіркового середнього: СРЗНАЧ. Обрахуємо вибіркове середнє по містам А і В.

Рис.4.1



Для цього застосуємо функцію СРЗНАЧ до комірок А1:А4 та, окремо, В1:В8.
Виявляється, що по обох містах середній об'єм реалізації цукерок на 1 супермаркет однаковий – 2050 грн.

Приклад 4.1

Фірма, яка планує продавати медичні товари, для того, щоб оцінити ринок у Луцьку поставила перед маркетологами питання: «Скільки в цілому витрачають за місяць на медичні товари мешканці міста?» Була побудована репрезентативна вибірка мешканців міста об'ємом 50 і проведено опитування, результати якого (грн.) зведені у таблицю.

51,69	55,42	51,43	65,05	51,89
44,77	46,02	36,93	56,82	47,11
48,99	41,90	46,37	59,25	43,63
42,86	56,32	53,93	53,36	58,18
69,19	39,75	54,12	56,11	38,26
54,56	57,27	42,71	45,82	45,46
58,70	48,64	57,85	46,58	44,20
62,79	48,20	44,38	29,18	48,87
57,30	51,77	65,63	50,47	52,74
57,80	45,02	53,76	45,06	54,01

Вважаючи, що населення міста складає 250 000, визначити об'єм ринку.

Скопіюємо ці значення у Excel. Обрахуємо за допомогою функції СРЗНАЧ вибіркове середнє, яке є оцінкою середніх витрат на медичні товари по всьому місту (по генеральній сукупності). Це середнє становить 50,76 грн. Тобто в середньому кожен мешканець міста витрачає на медичні товари 50,76 грн. на місяць. Значить всі мешканці міста витрачають за місяць 250 000 * 50,76 = 12 690 652,60 грн. Звичайно – це є оцінка, тобто наближене значення.

5. Вибіркова дисперсія та стандартне (середньоквадратичне) відхилення

Приклад 5.1.

Уявимо собі, що при дослідженні витрат на медичні товари по містам А і В ми отримали 2 вибірки (грн.). По місту А: 10, 90, 20, 80, 15, 85. По місту В: 49,5, 50,5, 48, 52, 49, 51. Обрахуємо вибіркове середнє за допомогою функції СРЗНАЧ. Виявляється, що середні значення однакові і становлять 50. Але чи можна при оцінці ринку міста А вважати, що кожен мешканець витрачає 50 грн.? А для міста В?

Ми бачимо, що у першому випадку значення елементів вибірки далекі від середнього, а у другому – навпаки близькі. Тобто для міста В ми напевно могли б для розрахунків користуватись середнім значенням, а для міста А отримали б велику похибку. Значить для визначення якості середнього необхідна певна характеристика – міра розсіяння вибірки навколо її середнього. Такими характеристиками є дисперсія вибірки та стандартне відхилення.

Дисперсією вибірки називається середній квадрат відстані від значень вибірки до вибіркового середнього:

$$s = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}.$$

(n-1 замість n у знаменнику пов'язано із властивістю незміщеності статистичних оцінок).

Для обрахування дисперсіє користуються функцією Excel ДИСП.

Але оскільки дисперсія – це квадрат відстані, то частіше користуються коренем квадратним з дисперсії, який називається **стандартним або середньоквадратичним відхиленням (СКВ).** Для обрахування стандартного відхилення користуються функцією Excel СТАНДОТКЛОН. Порахуємо дисперсію та стандартне відхилення для вибірок з прикладу 5.1.

	Місто А		**Місто В**
	10		49,5
	90		50,5
	20		48
	80		52
	15		49
	85		51
Дисперсія	1490,00		2,10
СКВ	38,60		1,45

Ми бачимо, що дані по місту А значно більше розсіяні навколо середнього 50, ніж дані по місту В. Але, чи є вичерпними дисперсія або стандартне відхилення. Дійсно, порахуємо середнє та стандартне відхилення наступних вибірок:

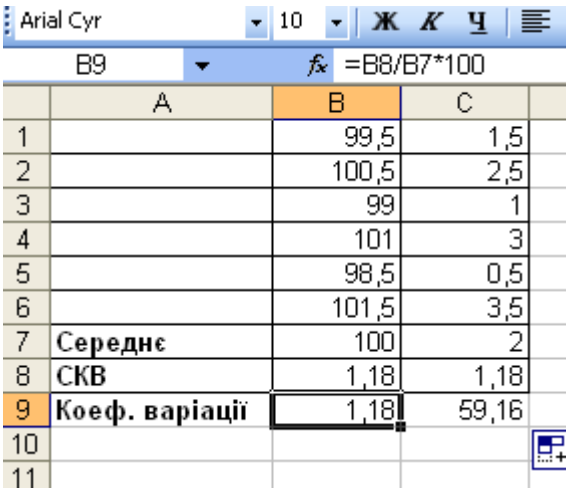
	99,5	1,5
	100,5	2,5
	99	1
	101	3
	98,5	0,5
	101,5	3,5
Середнє	100	2
СКВ	1,18	1,18

Обидві вибірки мають однакові значення стандартного відхилення: 1,18. Тобто середньоквадратична відстань від даних до середнього однакова. Дивлячись на першу вибірку можна сказати. Що дані щільно розташовані навколо середнього 100. А дивлячись на другу вибірку? Чи щільно розташовані дані навколо середнього 2, якщо стандартне відхилення 1,18? Зрозуміло, ні! Таким чином, для визначення ступеня розсіяння даних навколо середнього необхідно порівнювати це розсіяння (СКВ) із самим середнім.

Відношення стандартного відхилення до середнього значення вибірки називається коефіцієнтом варіації. Його часто виражають у відсотках.

Для обчислення цього коефіцієнту Excel не має стандартної функції, тому порахуємо його самостійно.

Рис. 5.1



	A	B	C
1		99,5	1,5
2		100,5	2,5
3		99	1
4		101	3
5		98,5	0,5
6		101,5	3,5
7	**Середнє**	100	2
8	**СКВ**	1,18	1,18
9	**Коеф. варіації**	1,18	59,16
10			
11			

Таким чином, для першої вибірки коефіцієнт варіації складає лише 1,18%, а для другої – 59,16% !

6. Кількісні та якісні дані

До сих пір ми розглядали кількісні дані: витрати, товарообіг тощо. Але існують дані іншої природи - якісні дані. Розглянемо кілька прикладів.

Приклад 6. 1.
Припустимо, що області України розташовані за алфавітом і пронумеровані:

1	Вінницька
2	Волинська
3	Дніпропетровська
…	…

Необхідно визначити середню область з тих областей, звідки родом працівники вашої компанії. Звичайно, можна знайти формальне середнє значення номерів цих областей, але це середнє не буде мати змістовного навантаження. Навряд чи результат у 2,75, тобто між Волинською і Дніпропетровською областями, буде представляти для вас інтерес. В даному випадку ми маємо справу із якісними даними, а нумерація – це умовність, бо можна пронумерувати області і зовсім інакше.

Приклад 6.2.
Нехай ми маємо результати опитування користувачів стосовно освіти:
1,2,2,1,2,3,4,1,1,1,2,1,3,4,3,1,1,1,2,1,2,2,1,1,3,
де 1 – вища освіта, 2 – середня спеціальна, 3 – загальна середня, 4 – неповна середня.

Що каже нам вибіркове середнє 2,04? Звичайно, ми також маємо справу з якісними даними, бо числова нумерація не має ніякого змістовного навантаження, вона могла би бути виконана абсолютно інакше.

Якісні дані вказують до якої якісної категорії, класу відноситься той чи інший об'єкт.

Існують два типу якісних даних: **порядкові** (ординальні), для яких існує порядок, який має змістовне навантаження та **номінальні**, для яких не існує порядку, що мав би змістовну інтерпретацію.

Питання: з якими якісними даними ми мали справу у прикладах 6.1 і 6.2 ?

Набір якісних даних є порядковим, якщо можна вести мову про перший (кращий) об'єкт, другий, третій і т.д. Можна ранжувати дані у відповідності до цього порядку і використовувати це ранжування для аналізу.

Приклад 6.3. Порядкові дані.

- Посади керівників компанії: президент, віце-президент, начальник відділу, заступник начальника відділу і т.д.
- Кредитний рейтинг типу AA+, AA, AA-,A+, A, A-, B+,B, B-.
- Результати опитування про освіту з прикладу 7.

7. Медіана вибірки

Ми побачили, що обрахування вибіркового середнього для якісних даних не доцільно. Аналогом середнього значення для якісних порядкових даних виступають медіана та мода вибірки.

Медіана – це значення, яке розбиває вибірку на 2 рівні частини: кількість даних, менших за медіану, дорівнює кількості даних, більших за медіану.

Тобто медіана знаходиться у центрі вибірки і в цьому розумінні є аналогом середнього значення. Зауважимо, що, якщо вибірка не має єдиного центрального значення (об'єм вибірки парний), то в якості медіани беруть середнє арифметичне двох сусідніх елементів вибірки, які знаходяться посередині.

Приклад 7.1.

Обрахувати медіану вибірки з прикладу 6.2:
1,2,2,1,2,3,4,1,1,1,2,1,3,4,3,1,1,1,2,1,2,2,4,1,3.

Перед тим, щоб обчислювати медіану, впорядкуємо вибірку за зростанням. Така **впорядкована вибірка називається варіаційним рядом,** а її елементи - **варіантами**.

1,1,1,1,1,1,1,1,1,1,1,**2,**2,2,2,2,2,2,3,3,3,3,4,4,

Оскільки у варіаційному ряді 25 елементів, то, очевидно, двійка, яка стоїть на 13-му місці (виділена) і буде медіаною. Тобто медіана дорівнює 2.

Розглянемо тепер той самий варіаційний ряд, тільки без останнього елементу:

1,1,1,1,1,1,1,1,1,1,1,**1,2,**2,2,2,2,2,2,3,3,3,3,4.

Тепер варіаційний ряд складається з парної кількості елементів. Тоді медіана – це середнє арифметичне варіант, що стоять на 12-му та 13-му місцях: (1+2)/2=1,5.

Якщо вибірка впорядкована за зростанням, тобто побудовано варіаційний ряд, то **номери варіант називаються їх рангами.**

Медіана, в такому випадку, це варіанта з рангом (n+1)/2, якщо n - непарне, або середнє арифметичне варіант рангів n/2 і n/2+1, якщо n – парне.

Медіану можна обчислити за допомогою функції Excel МЕДИАНА.

Зауважимо, що медіану можна обраховувати не тільки для якісних даних, але й для кількісних.

8. Мода вибірки

Модою називається елемент вибірки, який зустрічається у вибірці найчастіше.
Тобто мода – це типове значення вибірки. Моду можна обраховувати як для якісних, так і кількісних даних.

Приклад 8.1.
З метою аналізу причин відмови комп'ютерної техніки була побудована вибірка об'єму 250 і вивчені причини несправностей.

Причина проблеми	Кількість випадків
Пайка з'єднань	37
Пластмасовий корпус	86
Блок живлення	194
Бруд	8
Удар (при падінні)	1

Причини відмови комп'ютерів – це якісні, причому номінальні дані. Модою нашої вибірки є. очевидно, «блок живлення», т.я. саме ця причина зустрічається найчастіше. Для обрахування моди існує функція Excel МОДА.

Дуже корисно користуватися модою при обробці результатів опитування.

Приклад 8.2.
Нехай опитування щодо освіти респондентів дали наступні результати (приклад 6.2):
1,2,2,1,2,3,4,1,1,1,2,1,3,4,3,1,1,1,2,1,2,2,4,1,3.
Очевидно, дуже важливо визначити, яке із значень вибірки зустрічається найчастіше, тобто який варіант відповідей був обраний найбільшою долею респондентів. Це є мода даної вибірки. В нашому випадку вона дорівнює 1, тобто більшість опитаних мали вищу освіту.

Вибірка може мати одну моду (унімодальна), дві моди (бімодальна), або більше.
Так у вибірці 1,2,2,1,2,3,4,1,1,1,2,1,3,4,3,1,2,1,1,2,2,1,2,2,4,2,2,1,3 - дві моди 1 і 2, бо обидва елементи зустрічаються по 11 разів, в той час, як інші елементи зустрічаються рідше.

Зауваження: вибіркове середнє, медіана та мода можуть мати суттєво різні значення. Іноді із змісту задачі зрозуміло, якою характеристикою краще користуватися. Для неперервних кількісних даних (об'єм реалізації тощо) користуються середнім або медіаною, для якісних порядкових даних – медіаною або модою, для якісних номінальних даних – модою.

Ключові слова: якісні дані, порядкові та номінальні дані, вибіркове середнє, дисперсія, стандартне відхилення, коефіцієнт варіації, медіана, мода, варіаційний ряд, варіанта, ранг.

Непараметричне оцінювання

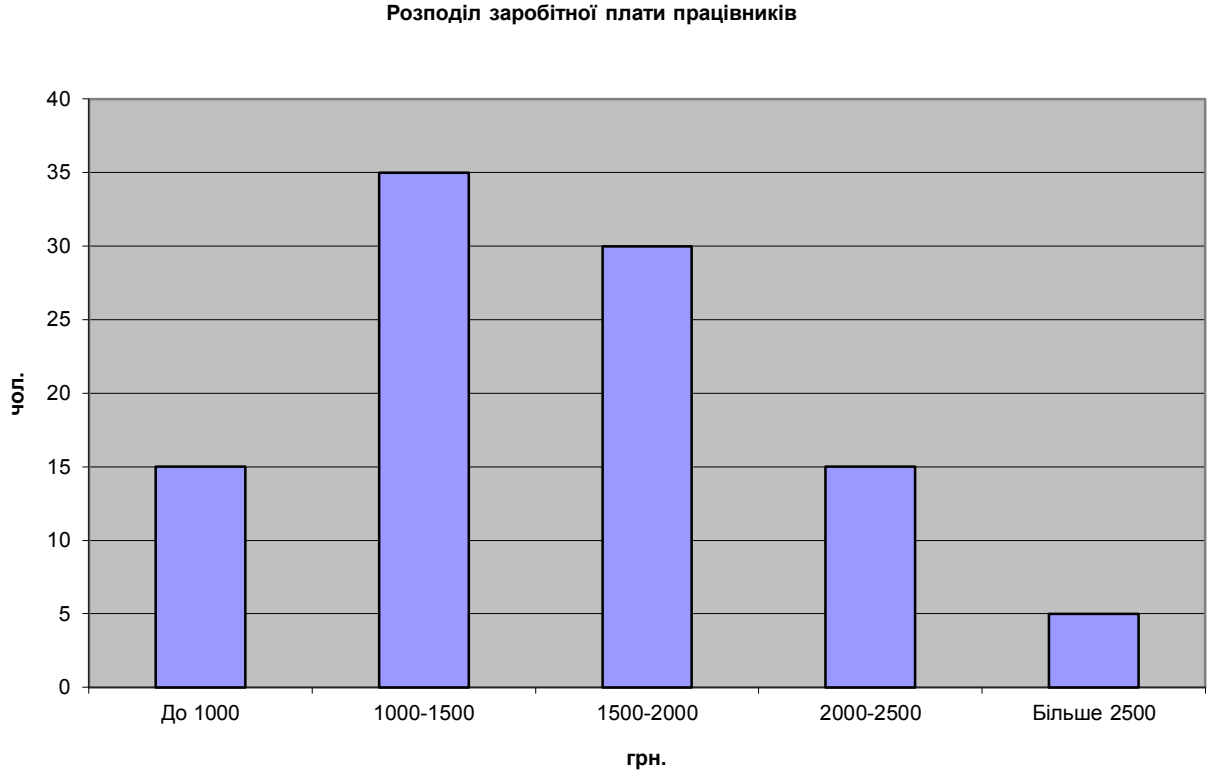
9. Поняття про гістограму: погляд на розподіл даних

Повернемося до кількісних даних.

Приклад 9.1. Розподіл заробітної плати у компанії (грн.)
:

Зарплата	Кількість працівників
До 1000	15
1000-1500	35
1500-2000	30
2000-2500	15
Більше 2500	5

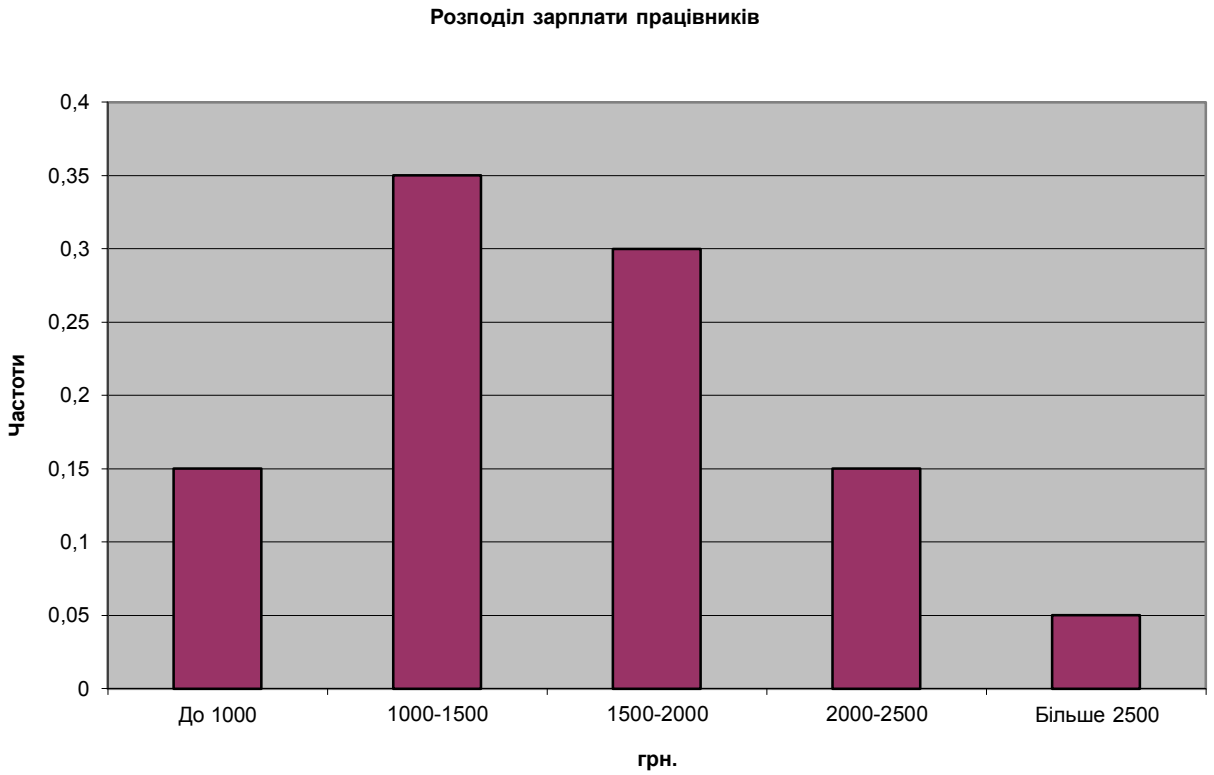
Зобразимо розподіл працівників в залежності від рівня зарплати у вигляді стовпчикової діаграми. Для цього скористаємося майстром діаграм у Excel:



Така діаграма називається гістограмою. **Гістограма – це стовпчикова діаграма частот.**
«Але де ж тут частоти?»- запитаєте ви. Для того, щоб відобразити частоти, треба було поділити кількість працівників, що попадають у певний діапазон на їх загальну кількість (100 чоловік). Тобто таблиця частот мала б вигляд:

Зарплата	Частота
До 1000	0,1
1000-1500	0,35
1500-2000	0,30
2000-2500	0,15
Більше 2500	0,05

Побудуємо по цій таблиці гістограму:



Порівняйте з попередньою діаграмою. Ми бачимо, що вони ідентичні, бо ділення кожного значення у правому стовпчику на одне те саме число (100) – це просто зміна масштабу, і на форму гістограму це не впливає. Єдина відмінність – одиниці виміру по осі OY.

Якщо у попередньому прикладі була вже зроблена попередня обробка даних – вони були підсумовані в межах кожного інтервалу зарплат і зведені у таблицю, то на практиці ситуація виглядає дещо інакше.

Приклад 9.2. Розглянемо дані з прикладу 4.1. (місячні витрати на медичні товари, грн.)

51,69	55,42	51,43	65,05	51,89
44,77	46,02	36,93	56,82	47,11
48,99	41,90	46,37	59,25	43,63
42,86	56,32	53,93	53,36	58,18
69,19	39,75	54,12	56,11	38,26
54,56	57,27	42,71	45,82	45,46
58,70	48,64	57,85	46,58	44,20
62,79	48,20	44,38	29,18	48,87
57,30	51,77	65,63	50,47	52,74
57,80	45,02	53,76	45,06	54,01

Проаналізувати цю інформацію дуже важко. А якби кількість елементів вибірки складала би не 50, а 1000 ? Для аналізу інформації необхідно по цих даних побудувати розподіл частот (гістограму) для того, щоб наочно побачити у яких межах знаходяться витрати більшості мешканців міста. Зауважимо, що гістограма будується не по даних, а по частотах. Тобто перед тим, як будувати гістограму, необхідно визначити певні інтервали витрат і порахувати частки людей, витрати яких попадають у ці інтервали (частоти).

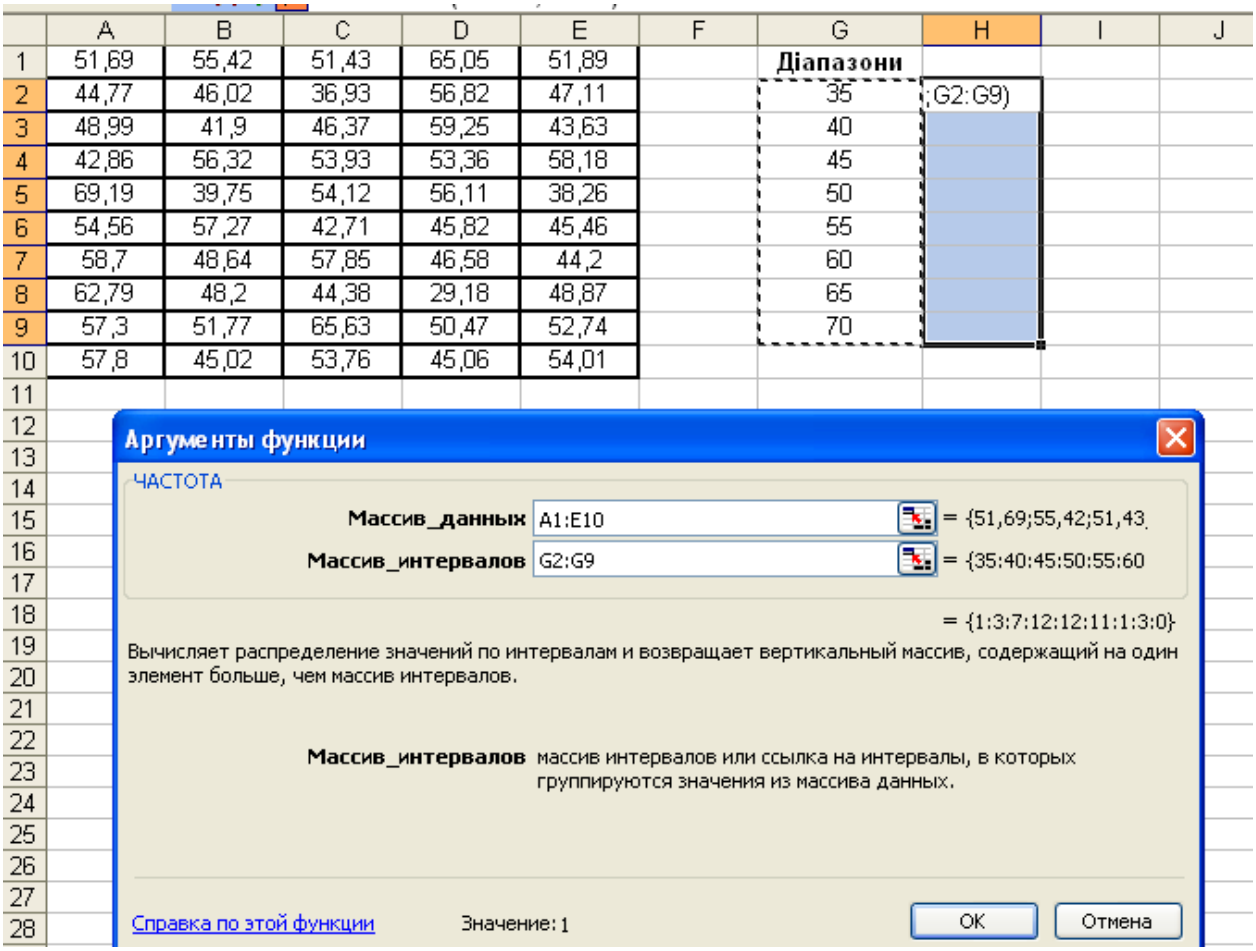
Зробимо для цього наступні кроки. Знайдемо найменше і найбільше значення вибірки за допомогою функцій МИН і МАКС. Це 29,18 і 69,19 відповідно. Цей діапазон розділимо на декілька інтервалів. (Це, в певному розумінні, мистецтво. Розіб'ємо на занадто велику кількість інтервалів, у деякі не буде попадати жодного значення, розіб'ємо на занадто малу – аналіз буде не достатньо деталізованим.)
69,19-29,18 = 40,01. Розіб'ємо діапазон від 29, 18 до 69,19 на 8 інтервалів довжиною ≈ 5 наступним чином:

До 35
35 - 40
40- 45
45 -50
50 - 55
55 - 60
60 - 65
65- 70

На аркуші Excel це виглядатиме наступним чином:

	A	B	C	D	E	F	G
1	51,69	55,42	51,43	65,05	51,89		**Діапазони**
2	44,77	46,02	36,93	56,82	47,11		35
3	48,99	41,9	46,37	59,25	43,63		40
4	42,86	56,32	53,93	53,36	58,18		45
5	69,19	39,75	54,12	56,11	38,26		50
6	54,56	57,27	42,71	45,82	45,46		55
7	58,7	48,64	57,85	46,58	44,2		60
8	62,79	48,2	44,38	29,18	48,87		65
9	57,3	51,77	65,63	50,47	52,74		70
10	57,8	45,02	53,76	45,06	54,01		
11							

Далі, щоб порахувати кількість даних, що попадають у кожний з інтервалів, скористаємося статистичною функцією ЧАСТОТА. Інтервалів 8, значить ця функція поверне нам 8 значень відразу. Такі функції, які повертають не одне, а кілька значень, називаються функціями масиву. Перед тим, як викликати таку функцію, необхідно виділити в аркуші Excel стільки комірок, скільки значень поверне функція. В нашому випадку, поряд із стовпчиком «**Діапазони**» виділимо стовпчик, довжиною у 8 комірок і викличемо функцію ЧАСТОТА.



Щоб ввести масив даних, виділимо весь діапазон даних: комірки A1:E10. Щоб ввести масив інтервалів, виділимо всі інтервали: комірки G2:G9.

Не можна закінчувати роботу з функцією масиву натисканням кнопки «ОК» або клавіші «ENTER». Для закінчення натискають комбінацію клавіш: SHIFT CTRL ENTER. Після цього у виділеному стовпчику з'являються шукані частоти (кількість даних, що потрапляє у відповідний інтервал).

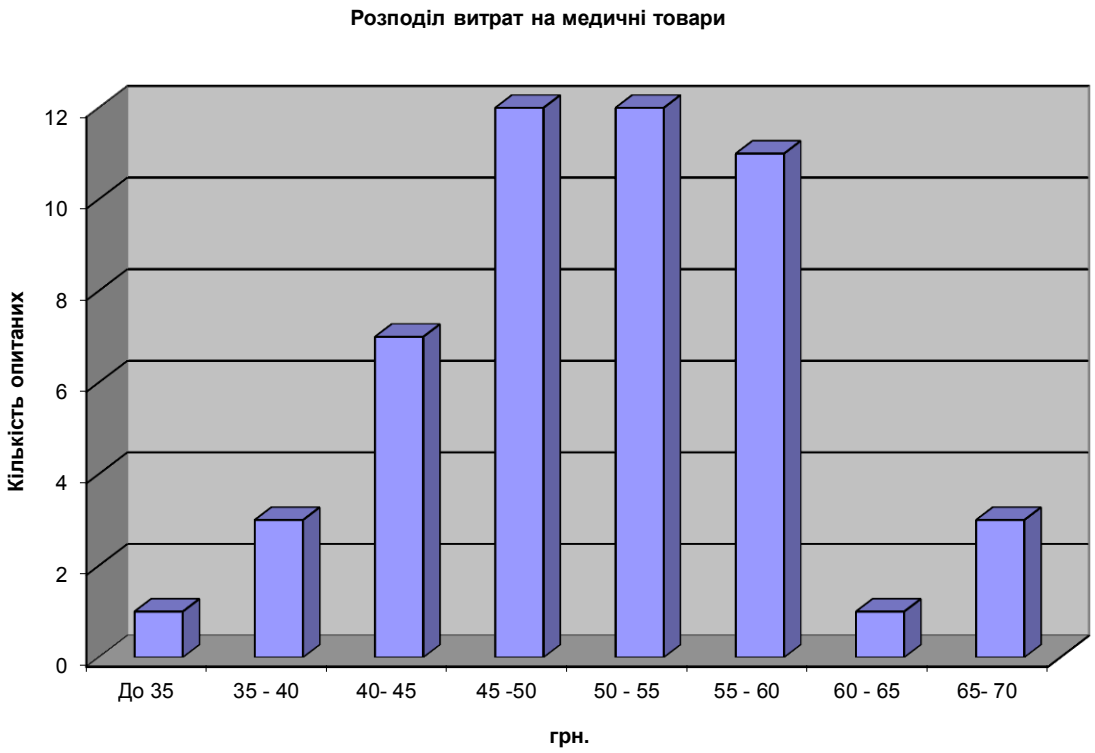
K20

	A	B	C	D	E	F	G	H
1	51,69	55,42	51,43	65,05	51,89		**Діапазони**	**Частоти**
2	44,77	46,02	36,93	56,82	47,11		35	1
3	48,99	41,9	46,37	59,25	43,63		40	3
4	42,86	56,32	53,93	53,36	58,18		45	7
5	69,19	39,75	54,12	56,11	38,26		50	12
6	54,56	57,27	42,71	45,82	45,46		55	12
7	58,7	48,64	57,85	46,58	44,2		60	11
8	62,79	48,2	44,38	29,18	48,87		65	1
9	57,3	51,77	65,63	50,47	52,74		70	3
10	57,8	45,02	53,76	45,06	54,01			
11								

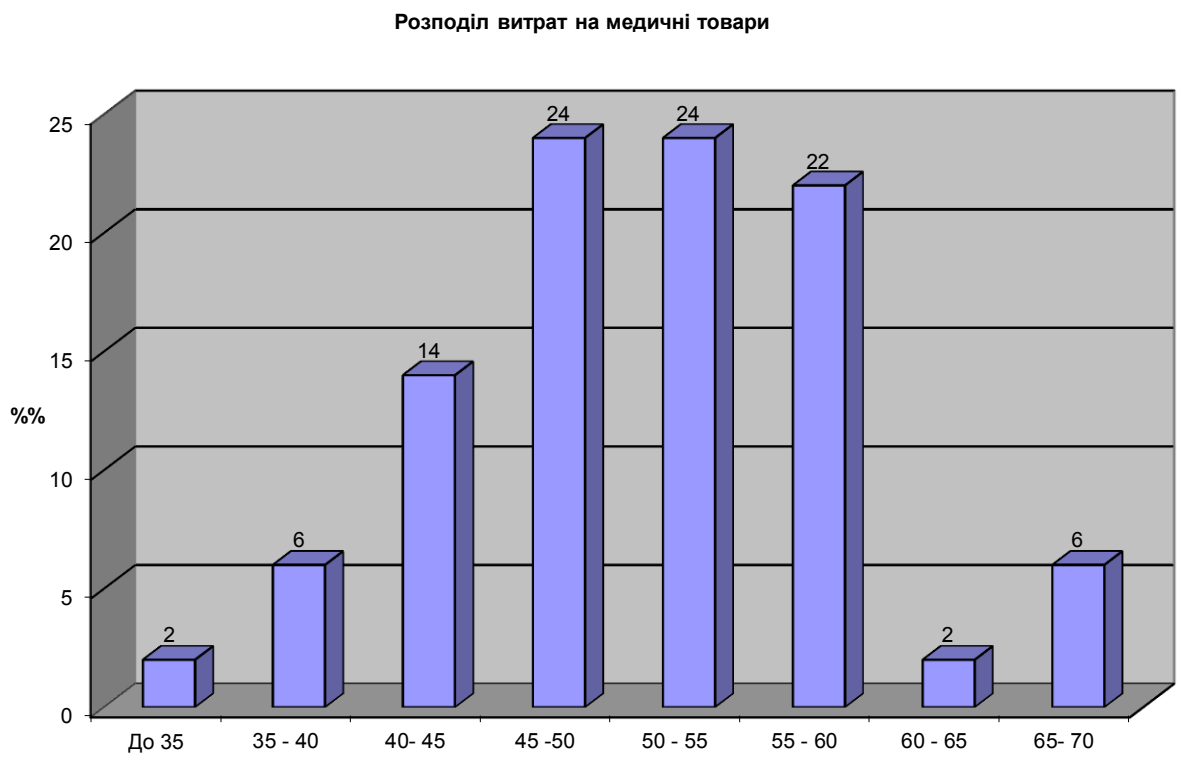
Ці частоти представляють собою розподіл витрат на медичні товари по інтервалах. Побудуємо по них гістограму. Запишемо попередньо в один з стовпчиків аркуша Excel інтервали у такому вигляді.

До 35
35 - 40
40- 45
45 -50
50 - 55
55 - 60
60 - 65
65- 70

Гістограма буде мати вигляд:



Якщо ж значення по осі OY поділити на кількість елементів вибірки (50) и помножити на 100%, то отримаємо розподіл витрат у відсотках:



З гістограми ми бачимо, що більшість мешканців міста (70%) витрачає на медичні товари 45-60 грн. на місяць. Всього 8% витрачає більше 60 грн. і т.д. Тобто гістограма дає нам можливість отримати з даних певну інформацію - знання про витрати населення на медичні товари і, відповідно, дає можливість оцінити обсяг ринку, розробити маркетингові заходи тощо.

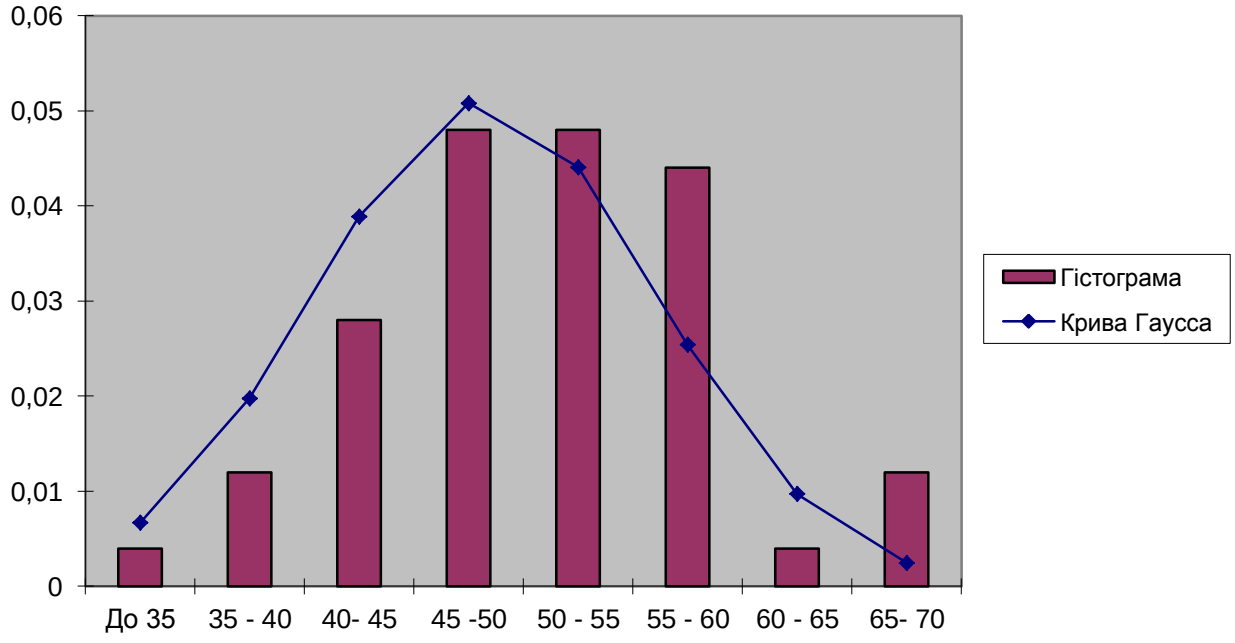
Гістограма будується за певною вибірко., тобто є вибірковою оцінкою такої характеристики генеральної сукупності, як густина розподілу. Із зростанням об’єму вибірки гістограма буде наближатися до густини розподілу генеральної сукупності.

10. Нормальний розподіл та деякі інші види розподілів

У багатьох практичних ситуаціях генеральна сукупність має, так званий, нормальний розподіл (розподіл Гауса). Густина цього розподілу залежить від двох параметрів: середнього значення ***a*** та стандартного відхилення ***σ***. Її графік називається кривою Гауса (диференціальною) і має «дзвоноподібну» форму:

.

Якщо накласти на гістограму розподілу витрат на медичні товари криву Гауса, то побачимо, що крива Гауса наближає (апроксимує) гістограму:



Крива Гауса – це густина розподілу генеральної сукупності, а гістограма – її вибіркова оцінка. Нормальній розподіл (розподіл Гауса) має надзвичайну практичну цінність, яка є наслідком центральної граничної теореми.

Центральна гранична теорема (ЦГТ) каже, що сума великої кількості маленьких за абсолютною величиною незалежних випадкових величин (впливів) має розподіл, близький до нормального. Причому, чим більше таких величин у сумі і менші їх абсолютні значення, тим ближче цей розподіл до нормального.

На практиці у величезній кількості випадків ми маємо справу як раз з великою кількістю незалежних незначних впливів. Нормальний розподіл має цікаві особливості і ширше, ніж будь-який інший, застосовується у статистиці. Зокрема, якщо ми спостерігаємо за значенням деякої випадкової величини, тобто маємо певну вибірку, розподіл якої близький до нормального, то у 95% випадків елементи вибірки будуть знаходитись в межах двох стандартних відхилень від середнього значення.

Ключові слова: розподіл даних, гістограма, густина нормального розподілу (диференціальна крива Гауса), центральна гранична теорема, функція масиву, функція ЧАСТОТА.

Інтервальне оцінювання та елементи перевірки статистичних гіпотез

11. Поняття про довірчій інтервал

Повернемося до прикладу 4.1 – даних про витрати на медичні товари. Ми вже зрозуміли, що для оцінки об'єму ринку можна скористатися середніми витратами кожного мешканця міста. Але необхідно розрізняти середнє значення витрат, узяте по усьому місту, тобто середнє значення генеральної сукупності, і вибіркове середнє, побудоване по певній вибірці з генеральної сукупності. Вибіркове середнє є лише оцінкою генерального середнього. А яка ж похибка такої оцінки? Чи можна вказати з певною ймовірністю такий інтервал, у який буде входити невідоме середнє значення усієї генеральної сукупності? Виявляється можна, правда коли наші дані розподілені за нормальним законом, або близьким до нього.

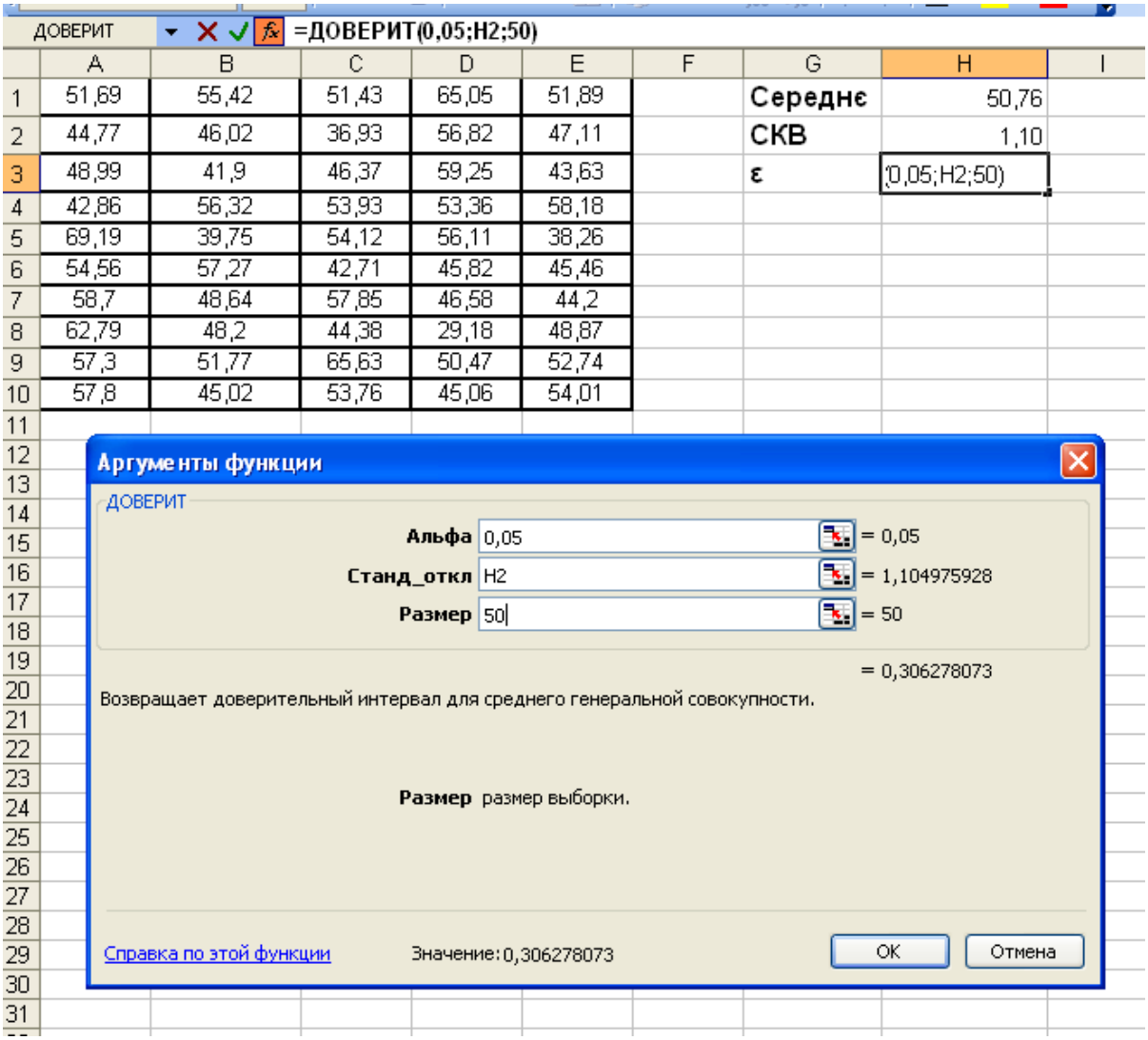
Довірчим інтервалом, називають такий обчислений за даними вибірки інтервал, який з наперед відомою ймовірністю містить цікавий для нас невідомий параметр генеральної сукупності. Ця наперед задана ймовірність ***p*** називається надійністю інтервалу або рівнем довіри. Як правило користуються надійністю 0,95 або 0,99. Величина, рівна ***1-p*** (0.05 або 0,01) називається рівнем значущості і представляє собою ймовірність помилки при побудові довірчого інтервалу. Тобто рівень значущості - це ймовірність того, цікавий для нас параметр генеральної сукупності не буде входити в побудований нами довірчий інтервал. **Якщо об'єм вибірки більший за 40, то для обчислення довірчого інтервалу для середнього значення генеральної сукупності можна скористатися функцією ДОВЕРИТ.** При використанні цієї функції необхідно ввести об'єм вибірки, задати рівень значущості і стандартне відхилення вибіркового середнього, **яке дорівнює стандартному відхиленню елементів вибірки, діленому на** $\sqrt{n}$ **(**де ***n*** **-** об'єм вибірки).

Приклад 11.1
Обчислити для середнього значення витрат не медичні товари мешканцями міста довірчий інтервал надійності 0,95.

Для розв'язання цієї задачі нам потрібно знати об'єм нашої вибірки, вибіркове середнє, стандартне відхилення та рівень значущості. Об'єм вибірки (розмір) дорівнює 50, рівень значущості (альфа) дорівнює 1-0,95=0,05. Вибіркове середнє та стандартне відхилення (СКВ) вибірки порахуємо за допомогою функцій СРЗНАЧ та СТАНДОТКЛОН. Отримаємо 50,76 та 7,81 відповідно. **Поділимо 7,81 на** $\sqrt{50}$ **і отримаємо 1,10**. Далі скористаємося функцією ДОВЕРИТ , значення якої позначимо ***ε***. Отримаємо ***ε***= 0,30.(см. рис. 11.1). Тоді довірчий інтервал для генерального середнього становить: (50,76-0,30; 50,76+0,30), тобто (50,46; 51,06). Це означає, що із ймовірністю 0,95 середні витрати на медичні товари по усьому місту знаходяться у межах від 50,46 до 51,06 грн. А це вже результат!

При оцінці об'єму ринку будемо обережні і скористаємося лівою межею – 50,46. Тобто з ймовірністю не меншою за 0,95 об'єм ринку не менший за 50,46 * 250 000=12 615 000 грн.

Рис. 11.1



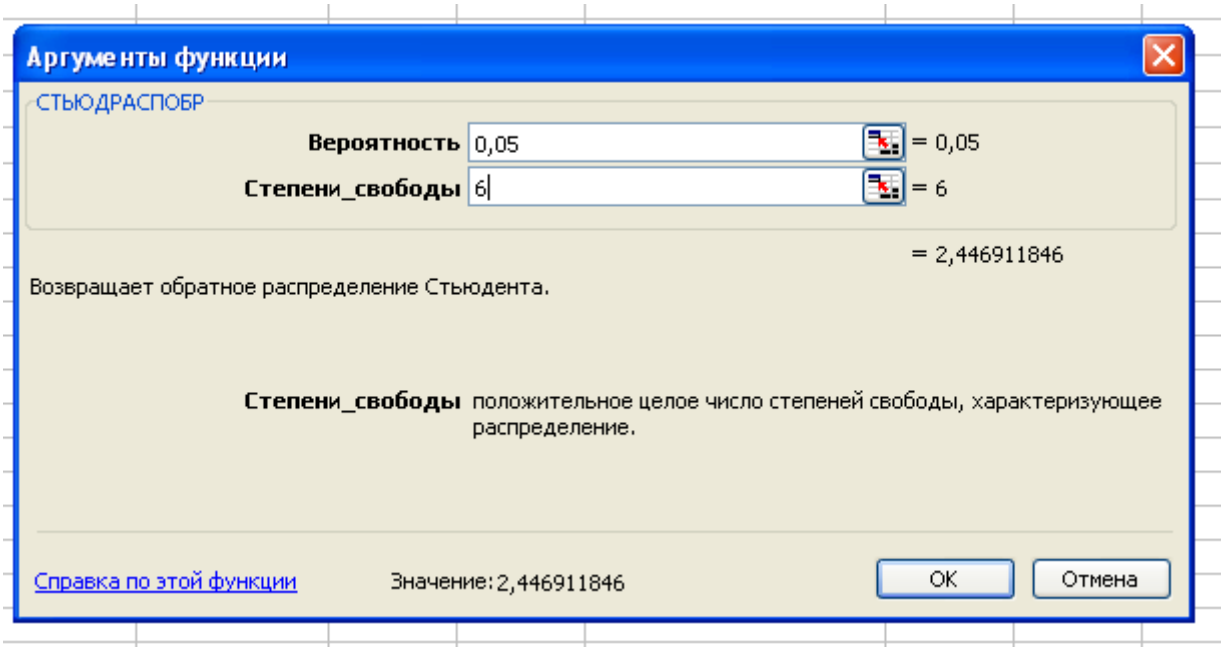
Якщо об'єм вибірки, менший за 40, то для отримання більш точного результату при побудові довірчого інтервалу користуються функцією СТЬЮДРАСПОБР. Це функція, обернена до функції розподілу Стьюдента (псевдонім відомого англійського статистика В.Госсета). Щоб скористатися цією функцією необхідно задати кількість ступенів вільності (на 1 менше за об'єм вибірки) та рівень значущості.

Приклад 11.2
Нехай ви випробуєте домішок до сировини, який має збільшувати обсяг продукції. Домішок випробовували 7 діб, тобто об'єм вибірки дорівнює 7.

Середній добовий обсяг продукції за минулий тиждень при використанні добавки (тонн)	$\bar{x}$	39,6
Стандартне відхилення вибіркового середнього	S	4,2
Розмір вибірки	N	7

Генеральне середнє в цьому випадку – це середній добовий обсяг продукції, порахований за довгий проміжок часу до застосування нового домішку. Побудуємо довірчий інтервал для генерального середнього надійності 0,95. Скористаємося функцією СТЬЮДРАСПОБР з 6 ступенями вільності і рівнем значущості 0,05 і обчислимо її значення ***t*** (рис. 11.2):

Рис. 11.2



Тобто t=2,45. Тоді шуканий довірчий інтервал надійності 0,95 для генерального середнього дорівнює $(\bar{x} - t * 4.2; \bar{x} + t * 4.2)$ =(29,31;49,89).

Таким чином, в загальному вигляді, довірчий інтервал надійності ***p*** для генерального середнього вибірки об'єму ***n*** має вигляд:

$$(\bar{x} - t * S_{\text{ñä}}; \bar{x} + t * S_{\text{ñä}}),$$

де ***t*** називається ***p***-квантілем розподілу Стьюдента і обраховується так само, як і в наведеному вище прикладі, а $S_{ср.}$ – стандартнее відхилення вибіркового середнього, яке отримують діленням стандартного відхилення вибірки ***S*** на $\sqrt{n}$ **.**

Зауваження: перед використанням довірчих інтервалів необхідно побудувати гістограму вибірки і переконатися, що форма густини розподілу наближається до форми кривої Гауса.

Вище ми будували довірчий інтервал для вибіркового середнього. Покажемо, як будується так званий інтервал передбачення. **Інтервал передбачення - це такий інтервал, побудований за даними вибірки, у який з певною ймовірністю попаде наступний елемент вибірки.**
Для побудови інтервалу передбачення не можна користуватися функцією ДОВЕРИТ. Для нього так само обраховують значення функції СТЬЮДРАСПОБР за заданим рівнем значущості та кількістю ступенів вільності. Отримане значення ***t*** множать на стандартне відхилення елементів вибірки ***S*** (а не вибіркового середнього). Отриманий добуток додатково. множать на величину $\sqrt{1+\frac{1}{n}}$, де n – об'єм вибірки. Шуканий інтервал має вигляд $(\bar{x} - t * S * \sqrt{1+\frac{1}{n}}; \bar{x} + t * S * \sqrt{1+\frac{1}{n}})$.

Побудуємо інтервал передбачення для витрат на медичні товари: n=50, рівень значущості 0,05.

Скористаємося функцією СТЬЮДРАСПОБР і обрахуємо величину t (рис. 11.3). t=2,01. Стандартне відхилення S= 7,81, $\sqrt{1+\frac{1}{50}} \approx 1{,}01$. Вибіркове середнє дорівнює 50,76. Тобто інтервал передбачення має вигляд: (50,76-2,01*7,81*1,01; 50,76+2,01*7,81*1,01)=(34,91; 66,92).
Таким чином, якщо ми опитаємо нового мешканця міста, то з ймовірністю 0,95 його витрати на медичні товари будуть у межах від 34,91 до 66,92 грн.

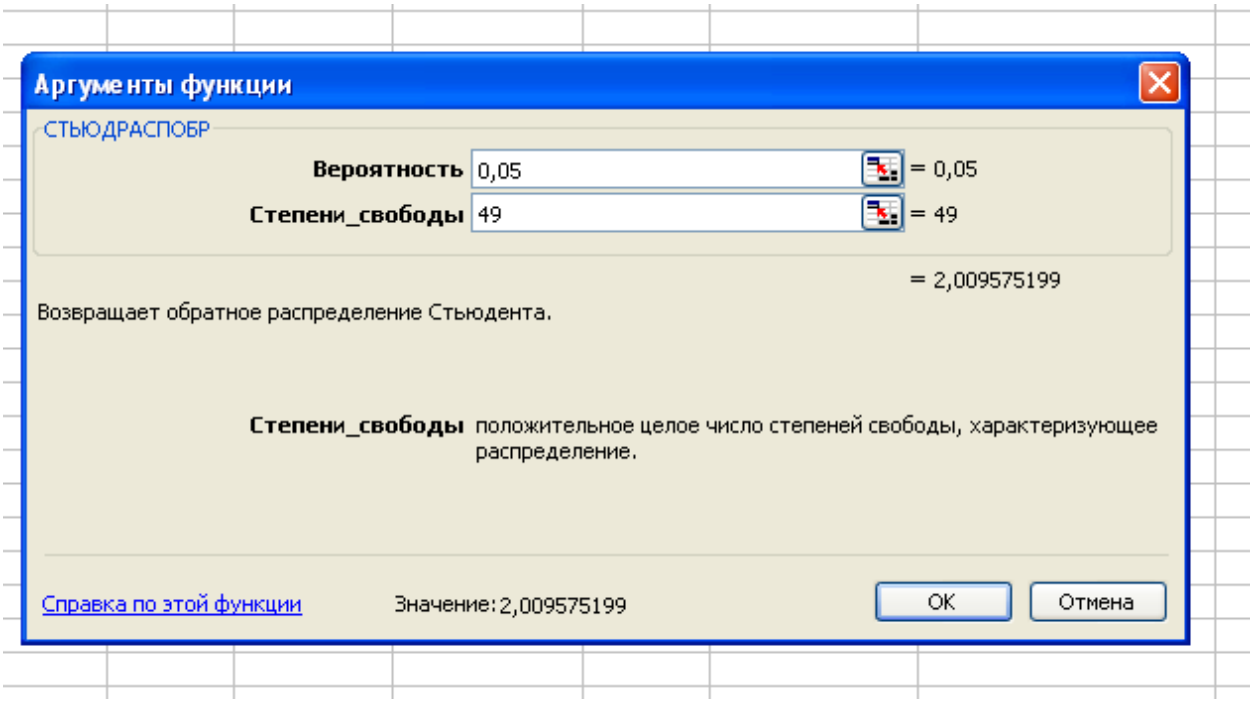


Рис. 11.3

Виникає слушне запитання: а що буде, якщо надійність довірчого інтервалу збільшити до 0,99 (99%) або навіть 1 (100%) ? В цьому випадку ймовірність влучення того чи іншого показника у довірчий інтервал збільшиться, але сам інтервал розшириться. І якщо надійність 0,99 іноді застосовується на практиці, то ймовірність 1 означає, що всі можливі елементи вибірки мають попадати у такий інтервал. А чи багато користі в тому, що ми знатимемо, що з ймовірністю 1 витрати на медичні товари мешканцями міста знаходяться у межах між 0 і 100 000 грн.?

12. Поняття про перевірку статистичних гіпотез

Розглянемо таку ситуацію. Наполегливий комерційний агент запропонував вам домішок для збільшення продуктивності вашого нафтопереробного заводу. Ви випробували його на протязі тижня і ніби переконалися у ефективності. В котре ви чуєте у телефонну трубку: «Об'єм продукції виріс? А що ж я вам казав?» Але ви сумніваєтесь, бо об'єм продукції і так не є сталим. Він може збільшуватися, може зменшуватися, бо залежить від багатьох факторів. Чи є збільшення обсягів виробництва випадковістю чи закономірністю?

Подивимося на цю ситуацію з точки зору комерційного агента. Припустимо, що цей домішок не впливає на обсяги виробництва. Він був запропонований для випробування 100 підприємствам. Тоді, внаслідок випадкових впливів, за тиждень випробувань приблизно у половини з них обсяги виробництва зменшаться і з ними не варто мати справу. Але у інших 50 підприємств обсяг виробництва збільшиться і з ними можна надалі працювати!

Таким чином, нам необхідно зробити вибір: чи збільшення обсягів є випадковістю, чи відображає реальну корисність домішку. Це класична задача перевірки статистичних гіпотез.

Перевірка статистичних гіпотез дозволяє на основі вибіркових даних зробити вибір між двома припущеннями (гіпотезами) щодо генеральної сукупності. Перевірка статистичних гіпотез – є одним з компонентів прийняття рішень.

Нульова гіпотеза H_0 представляє собою таке твердження, яке приймається тоді, коли немає переконливих аргументів для прийняття альтернативної гіпотези. **Дослідницька (альтернативна) гіпотеза H_1** приймається лише тоді, коли є переконливі статистичні аргументи, що призводять до її прийняття. Нульова і альтернативна гіпотези взаємно виключають одна одну. На відміну від нульової гіпотези, прийняття альтернативної потребує переконливих аргументів. Тому, при перевірці статистичних гіпотез, прийняття нульової гіпотези, означає лише відсутність підстав прийняти альтернативну гіпотезу, а не істинність нульової гіпотези.

Часто дослідницька гіпотеза містить скриті плани дослідника, а нульова гіпотеза висувається лише для того, щоб бути відхиленою. Прийняття тої чи іншої гіпотези має ймовірнісний зміст і не є гарантією її абсолютної істинності..

Зупинимося на застосуванні довірчих інтервалів до перевірки гіпотез. Повернемося до нашого прикладу про домішок, що «збільшує обсяг продукції». Нехай

Середній добовий обсяг продукції за минулий тиждень при використанні домішку (тонн)	$\bar{x}$	39,6
Стандартне відхилення	S	4,2
Розмір вибірки	N	7
Середнє значення добового обсягу продукції за довгий період до використання домішку (тонн)	μ_0	32,1

Дивлячись на ці дані, виникає припущення (дослідницька гіпотеза), що додавання домішку суттєво впливає на обсяг виробництва. Нульова гіпотеза полягає у тому, що домішок не впливає на результативність виробництва (середнє суттєво не змінилося при додаванні домішку). Прийняття альтернативної (дослідницької) гіпотези розумно, якщо значення середньодобового обсягу виробництва μ_0 за довгий період до застосування домішку (генеральне середнє) , вийде за межі довірчого інтервалу надійності 0,95. Пригадаємо, що цей інтервал дорівнює (29,31;49,89). Тобто генеральне середнє μ_0 не виходить за межі інтервалу і слушних підстав прийняти дослідницьку гіпотезу немає.

Робимо висновок, що підстав вважати домішок корисним **немає.** Дійсно**, з** точки зору здорового глузду, середнє значення добового обсягу виробництва без усіляких домішок з ймовірністю 0,95 знаходиться у межах (29,31;49,89). Після застосування домішку середньодобовий обсяг залишився у тому ж інтервалі!

Прийняття нульової гіпотези - більш слабке рішення, ніж прийняття альтернативної, бо для цього не потрібно переконливих аргументів «за», а потрібна лише відсутність переконливих аргументів «проти». Тому у нашому випадку правильніше казати, що немає достатніх підстав вважати домішок ані корисним, ані некорисним.

Якщо б середнє значення добового обсягу продукції за довгий період до використання домішку становило б 29 тонн, то тоді ми могли б вважати, що домішок збільшує обсяги виробництва, а якщо б 50 тонн, то зменшує, тобто альтернативна гіпотеза була би прийнята!

Ключові слова: довірчий інтервал, довірчий інтервал для генерального середнього, інтервал передбачення, надійність інтервалу, рівень значущості, стандартне відхилення вибіркового середнього, функція ДОВЕРИТ, функція СТЬЮДРАСПОБР, перевірка статистичних гіпотез, нульова гіпотеза, альтернативна (дослідницька) гіпотеза.

Кореляція: міра взаємозв'язку

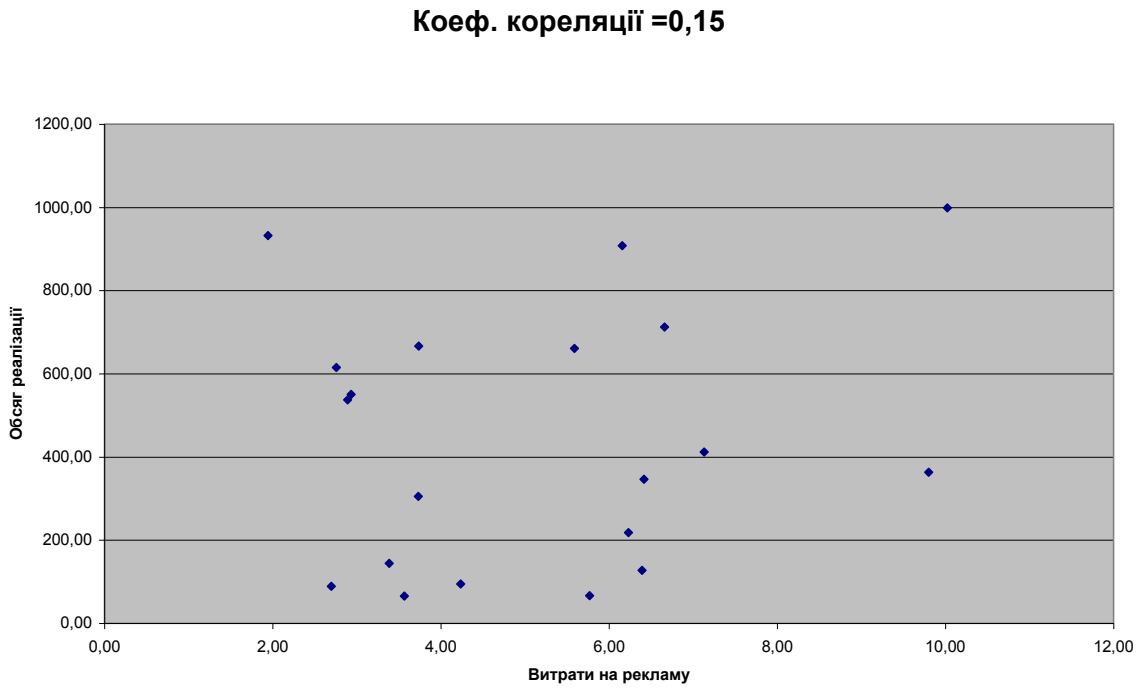
13. Діаграми розсіювання

Приклад 13.1
Нехай ми маємо дані по витратах на рекламу та по обсягах реалізації певної послуги за 20 місяців.

Витрати на рекламу (тис. грн.)	Обсяг реалізації (тис. грн.)
3,56	65,45
5,77	66,31
2,70	88,67
4,24	94,13
6,39	127,75
3,39	143,88
6,23	217,63
3,73	305,35
6,42	346,51
9,80	363,62
7,13	412,01
2,89	537,75
2,93	550,59
2,76	615,28
5,59	660,58
3,74	666,07
6,66	712,50
6,16	908,04
1,95	932,37
10,02	998,57

Якщо побудувати точкову діаграму, абсциси точок, якої - це витрати, а ординати - значення обсягів реалізації, то це і буде діаграма розсіювання.

Рис. 13.1



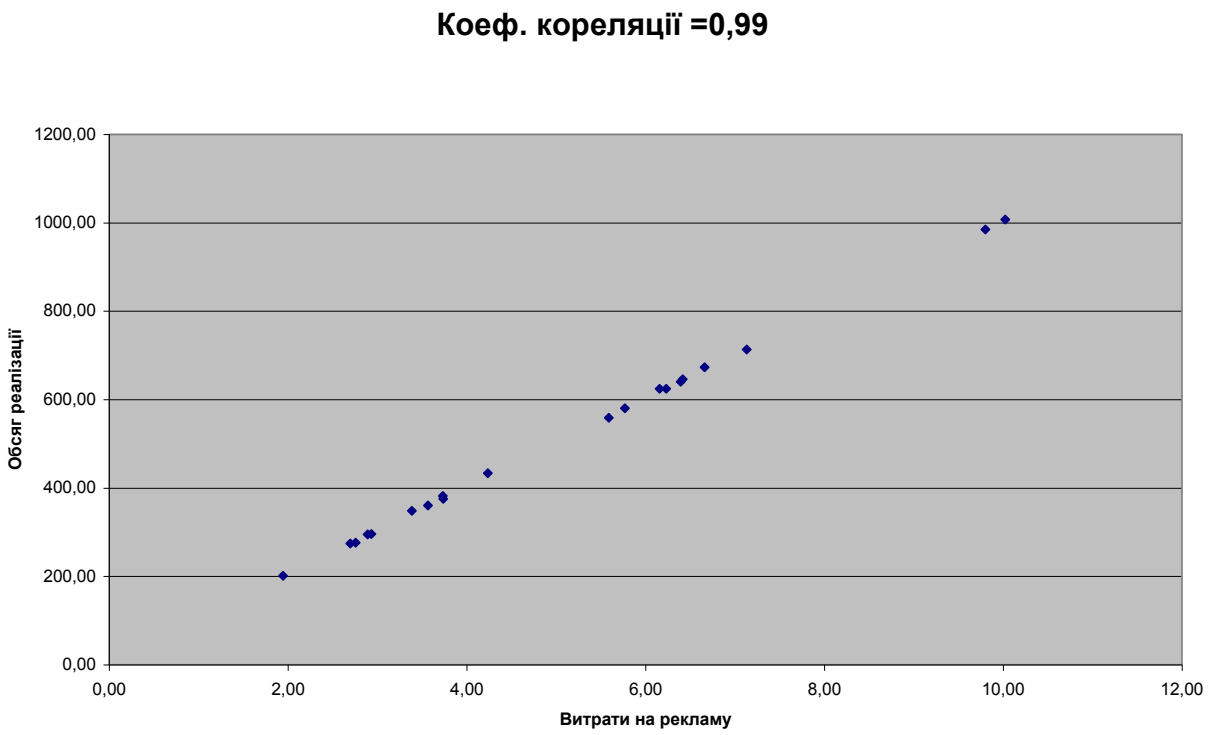
Діаграма розсіювання відображає структуру двовимірних даних (витрати – обсяг реалізації), дозволяє побачити певні особливості тощо. Якщо одне з вимірювань (в нашому випадку – витрати на рекламу) розглядати як причину, а інше (обсяги реалізації) як наслідок, то діаграма розсіювання дозволяє встановити якісний ступінь взаємозв'язку між ними.

Приклад 13.2
Нехай ми знову маємо дані по витратах на рекламу та по обсягах реалізації за 20 місяців

3,56	360,21
5,77	580,02
2,70	274,17
4,24	433,52
6,39	639,82
3,39	348,48
6,23	624,53
3,73	381,45
6,42	646,33
9,80	985,04
7,13	713,58
2,89	294,93
2,93	295,50
2,76	276,03
5,59	559,21
3,74	375,35
6,66	673,43
6,16	624,02
1,95	201,57
10,02	1007,50

Побудуємо по цих даних діаграму розсіювання. Вона матиме вигляд

Рис. 13.2

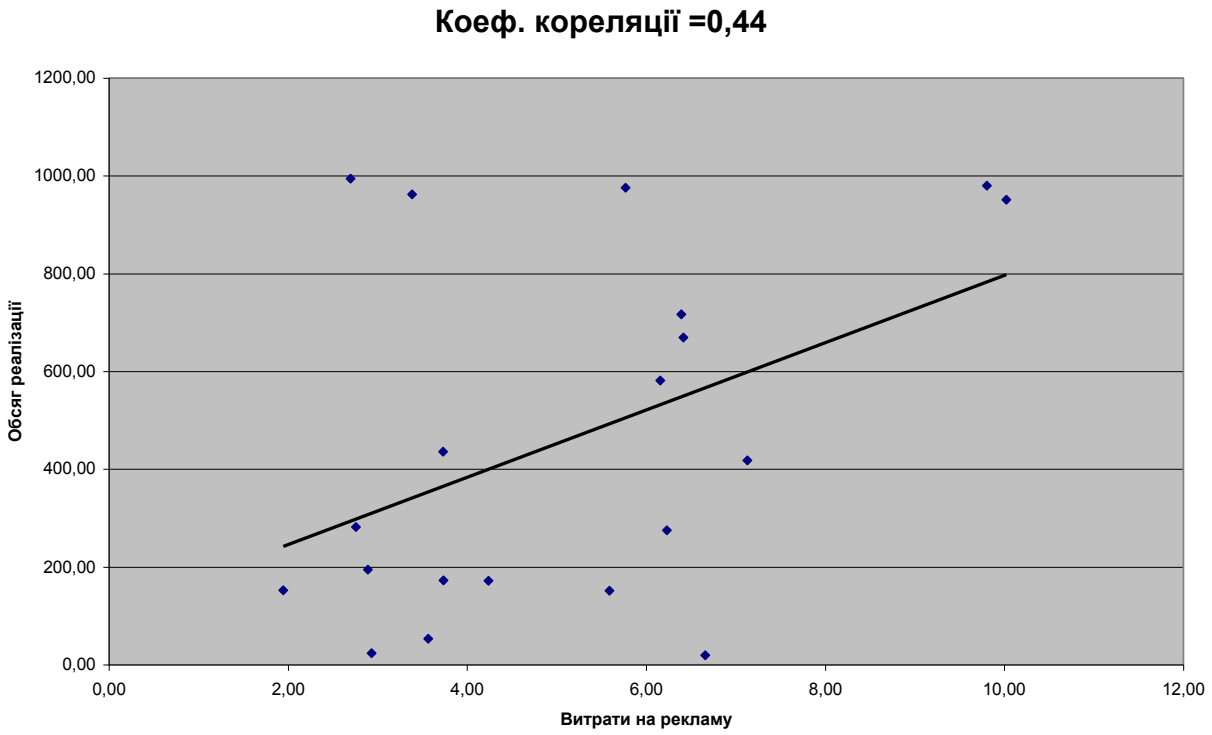


Ми бачимо, що на рис. 13.1 точки діаграми утворюють хмару, якоїсь помітної тенденції зміни обсягів реалізації при зростанні витрат на рекламу не спостерігається. Таким чином взаємозв'язок між цими параметрами майже відсутній.

На рис. 13.2 точки діаграми розташовані практично вздовж прямої. Це каже по тісний, практично лінійний зв'язок між параметрами. Причому нахил прямої каже про додатній зв'язок: тобто збільшення витрат на рекламу веде до збільшення обсягів реалізації. Якщо б пряма утворювала з віссю абсцис тупий кут, то ми б казали про зворотній зв'язок: збільшення одного з показників приводило б до зменшення іншого.

Звичайно, на практиці бувають не настільки однозначні ситуації, наприклад діаграма розсіювання може мати наступний вигляд:

Рис. 13.3



Тобто наявне помітне розсіювання даних, але в цілому вони групуються навколо прямої, хоча і не так щільно, як у попередньому прикладі. Ми бачимо, що зростання витрат на рекламу призведе до зростання обсягів реалізації, але залежність тут не буде цілком лінійною.

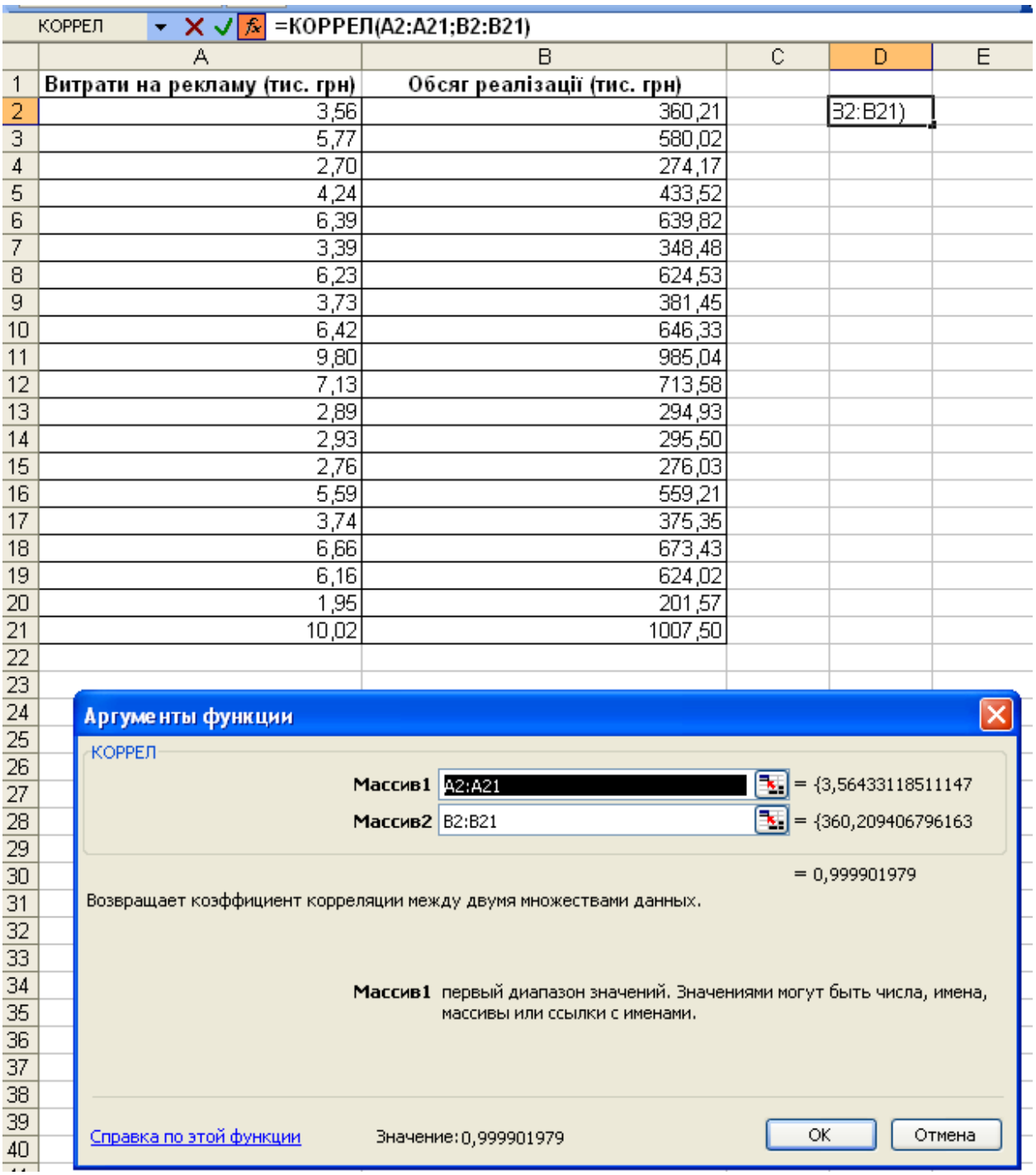
14. Коефіцієнт кореляції Пірсона: числова оцінка ступеня взаємозв'язку числових даних

Виникає питання як же оцінити ступінь взаємозв'язку більш точніше, у числовому вигляді. Взагалі – це достатньо складна задача. Але якщо зупинитися на вимірюванні ступеня лінійного взаємозв'язку числових даних (наскільки щільно дані знаходяться біля певної прямої), то для цього застосовують коефіцієнт кореляції Пірсона.

Коефіцієнт кореляції – це числова міра лінійного зв'язку між вибірками. Його значення знаходиться у межах між -1 і 1. Якщо значення коефіцієнта кореляції близьке до 1, то це каже про майже лінійний зв'язок між вибірками, тобто збільшення одного показника, веде до відповідного збільшення другого. Якщо коефіцієнт кореляції близький до -1, то це каже також про практично лінійний зв'язок, але обернений, тобто зростання одного показника веде до спадання іншого. Звичайна інтерпретація проміжних значень у діапазоні між -1 і 1 полягає у тім, що абсолютна величина коефіцієнту кореляції вказує на «силу» взаємозв'язку, а знак (додатній або від'ємний) – на напрямок такого зв'язку (прямий або обернений). Якщо є коефіцієнт кореляції близький до 0, то кажуть про практичну відсутність взаємозв'язку.

Зауваження: до такої інтерпретації проміжних значень коефіцієнту кореляції слід відноситись з обережністю, т.я. нелінійність взаємозв'язку та викиди даних можуть суттєво викривлювати картину. У цьому випадку доцільно оцінити загальну картину за допомогою діаграм розсіювання.

Для практичного обчислення коефіцієнту кореляції застосовують функцію EXCEL КОРРЕЛ. Застосуємо цю функцію до даних з прикладу 13.2.



	A	B	C	D	E
1	**Витрати на рекламу (тис. грн)**	**Обсяг реалізації (тис. грн)**			
2	3,56	360,21		32:B21)	
3	5,77	580,02			
4	2,70	274,17			
5	4,24	433,52			
6	6,39	639,82			
7	3,39	348,48			
8	6,23	624,53			
9	3,73	381,45			
10	6,42	646,33			
11	9,80	985,04			
12	7,13	713,58			
13	2,89	294,93			
14	2,93	295,50			
15	2,76	276,03			
16	5,59	559,21			
17	3,74	375,35			
18	6,66	673,43			
19	6,16	624,02			
20	1,95	201,57			
21	10,02	1007,50			

Отримане значення майже дорівнює 1.

Звернемо увагу на значення коефіцієнту кореляції на усіх діаграмах розсіювання. Можна побачити взаємозв’язок між виглядом діаграми і значенням коефіцієнту кореляції.

15. Коефіцієнт кореляції Спірмена: числова оцінка ступеня взаємозв'язку якісних (порядкових) даних

Приклад 15.1.
В результаті маркетингового опитування отримали наступні дані:

Стать	Місячні витрати на послугу (грн.)
1	2
1	4
2	2
1	1
2	1
2	3
2	2
1	4
1	4
1	4
1	3
2	3
1	1
2	3
2	3
1	3
2	4
2	2
2	4
1	3

де «Стать»: 1 – чоловіча, 2 – жіноча,
«Витрати»: 1- до 50 грн., 2- 50 – 100 грн., 3 - 100- 150 грн., 4 – більше 150 грн.
Необхідно визначити ступінь взаємозв'язку між статтю і витратами на послугу.

Ми не можемо застосувати коефіцієнт кореляції Пірсона, т.я. наші дані не є числовими (їх числові значення – умовність). В таких випадках англійський психолог Чарльз Спірмен запропонував знаходити коефіцієнт кореляції між рангами елементів вибірок, а ранг, як було зазначено вище, це порядковий номер елемента у впорядкованій вибірці. Впорядкуємо обидві вибірки і визначимо ранги їх елементів.

Витрати	Ранг
1	1
1	1
1	1
2	4
2	4
2	4
2	4
3	8
3	8
3	8
3	8
3	8
3	8
3	8
4	15
4	15
4	15
4	15
4	15
4	15

Стать	Ранг
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
2	11
2	11
2	11
2	11
2	11
2	11
2	11
2	11
2	11
2	11

Дійсно, розглянемо впорядковану вибірку витрат: «1» має ранг 1, бо є першим елементом вибірки, «2» вже має ранг 4, бо перед двійкою стоїть 3 одиниці, «3» має ранг – 8, бо перед трійкою стоять 3 одиниці і 4 двійки і т.д. Обчислити ранги можна за допомогою функції РАНГ. Ранговий коефіцієнт кореляції Спірмена обчислюється за формулою:

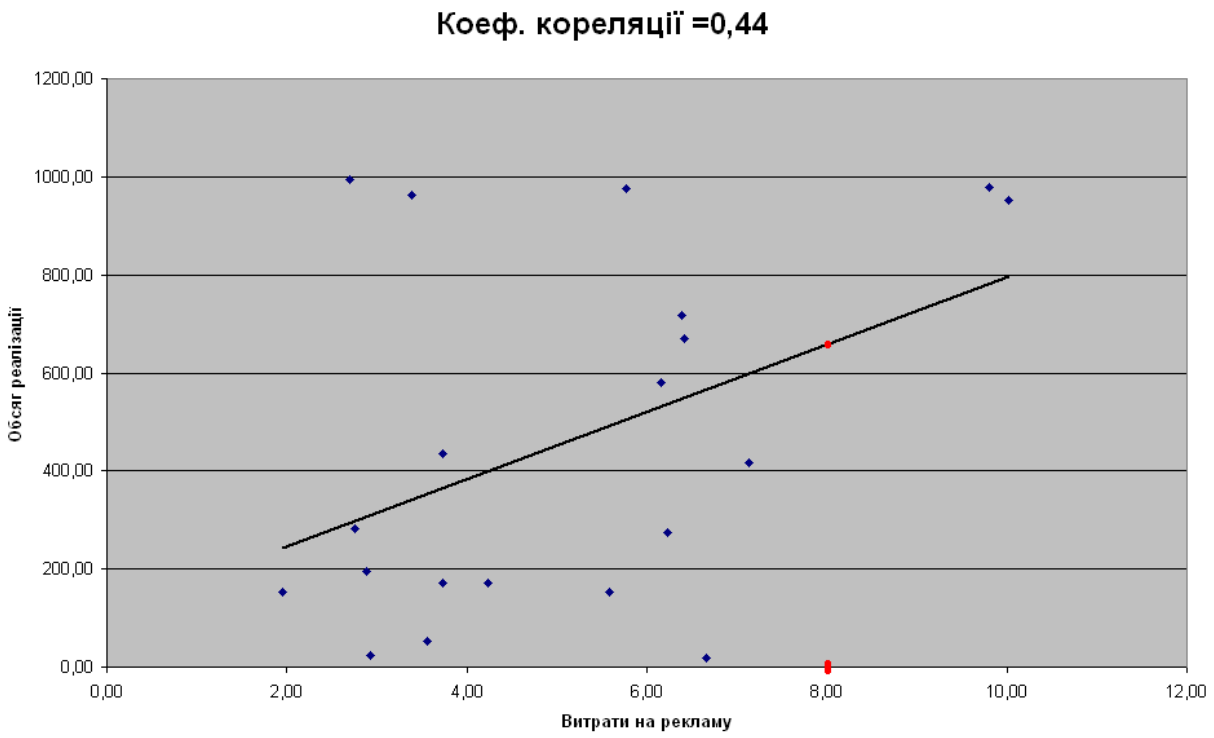
$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

де $d_i = x_i - y_i$ - різниця між рангами кожної відповідної пари спостережень над змінними ***x*** та ***y***. Для наших виборок отримаємо значення 0,79. Таким чином, між статтю і витратами на послугу існує достатньо тісний додатній зв'язок: більшому значенню показника статі відповідає більша сума витрат. Тобто жінки на дану послугу витрачають більше, ніж чоловіки.

16. Регресія: передбачення одного фактора по іншому

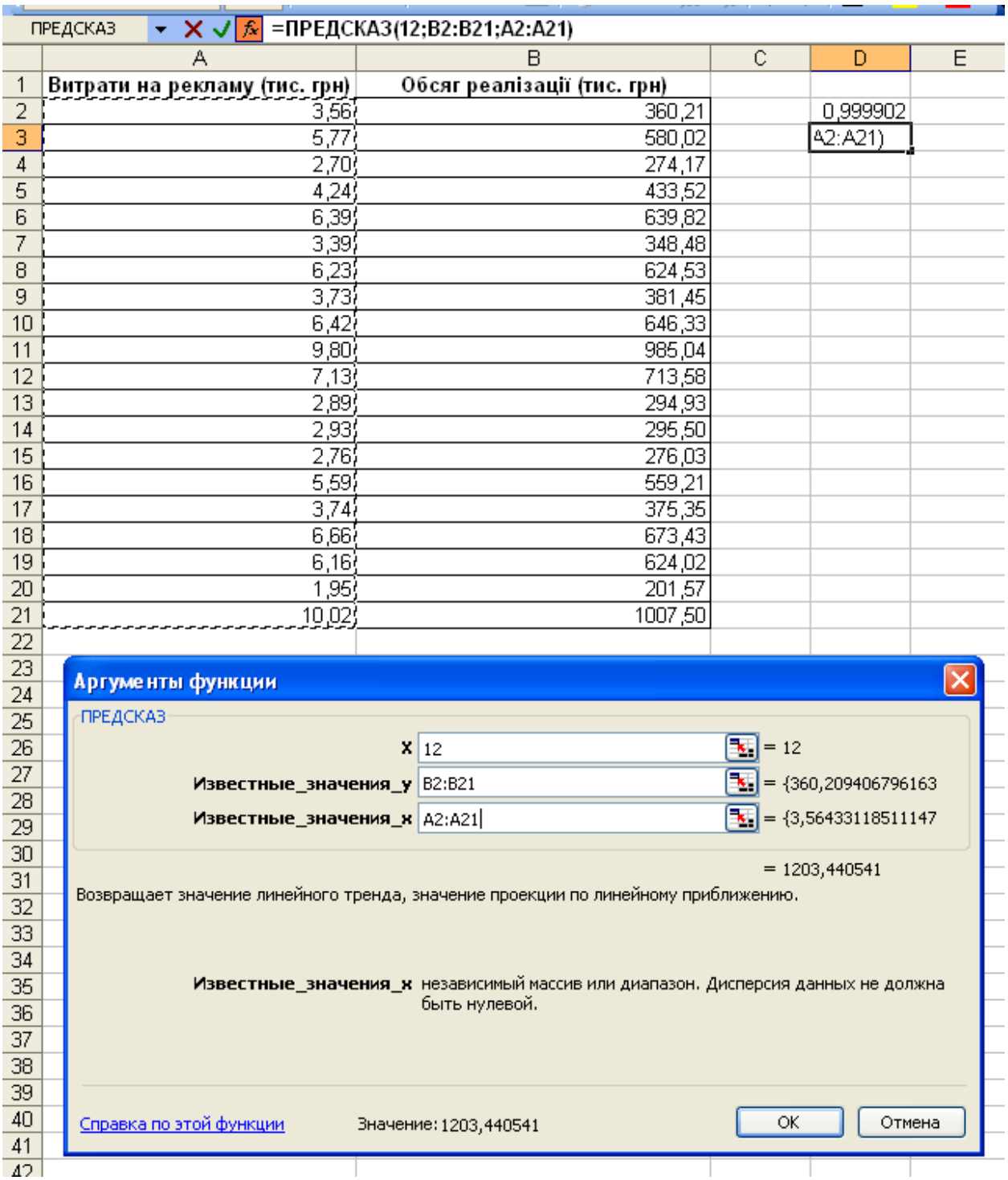
Задамося питанням, а чи можна знаючи заплановані витрати на рекламу на наступний місяць спрогнозувати обсяг реалізації. Чи навпаки, якщо задатися бажаним обсягом реалізації, то якими мають бути витрати на рекламу? Відповідь на це питання дає регресійний аналіз. Він дозволяє передбачати одну змінну на основі іншої з використанням прямої лінії, що характеризує взаємозв'язок між змінними. Ця лінія називається лінією регресії. Вона будується так, щоб сума квадратів відстаней до неї від точок діаграми розсіювання була найменшою. Тобто, в певному сенсі, вона є найближчою до усіх точок. Така лінія зображена на малюнку 13.3. Тоді, якщо відомо значення показника ***x***, то прогнозоване значення показника ***y*** – це ордината точки на прямій

регресії, яка відповідатиме абсцисі ***x***. Тобто, якщо задатися витратами на рекламу, рівними 8 тис. грн., то відповідний обсяг реалізації буде ≈650 тис. грн. (червона точка на прямій):



Для отримання точного числового значення прогнозу, побудованого за допомогою лінії регресії, користуються функцією EXCEL ПРЕДСКАЗ.

Спрогнозуємо за допомогою цієї функції по даних прикладу 13.2 обсяг реалізації, що відповідатиме витратам на рекламу у сумі 12 тис. грн.:



Тобто прогнозоване значення становить 1203,44 тис. грн.

Зауважимо, що точність прогнозу зростає із наближенням коефіцієнту кореляції між показниками за абсолютною величиною до 1. Тобто. Чим щільніші дані знаходяться біля прямої регресії, тим точніший буде прогноз. Для перевірки точності прогнозування, можна спрогнозувати по попередніх даних елемент вибірки, реальне значення якого вже відомо (ретроспективний прогноз), а потім порівняти прогнозоване і реальне значення.

Ключові слова: діаграма розсіювання, коефіцієнт кореляції Пірсона, коефіцієнт кореляції Спірмена, лінія регресії, функція РАНГ, функція ПРЕДСКАЗ.

«Бізнес у більшому ступені, ніж будь-яка інша справа

щоденно має справу з майбутнім;
це неперервний розрахунок ймовірностей,
інстинктивна вправа у передбаченні»
(Г.Люс, американський видавець)

Динамічний ряд

17. Лінійне прогнозування та виявлення трендів

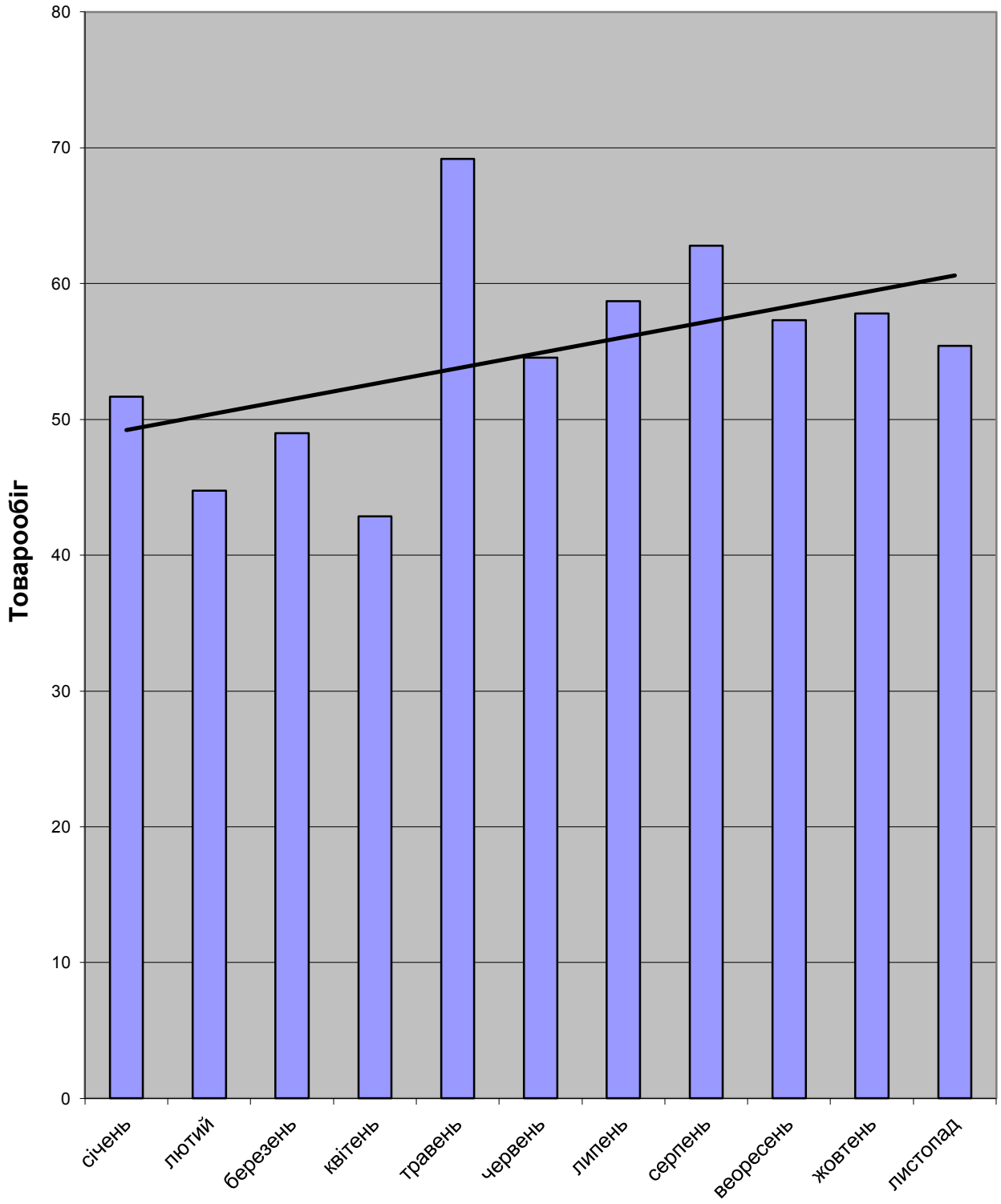
На відміну від звичайної вибірки **динамічний (часовий) ряд представляє собою послідовність даних, розташованих у порядку зростання моментів часу, у які ці дані спостерігалися.** Тобто порядок даних динамічного ряду суттєвий, дані у ньому розташовані послідовно. Наприклад, на протязі 11 місяців року спостерігається місячний товарообіг торгівельної точки:

Місяць	Товарообіг (тис. грн.)
січень	51,69
лютий	44,77
березень	48,99
квітень	42,86
травень	69,19
червень	54,56
липень	58,7
серпень	62,79
вересень	57,3
жовтень	57,8
листопад	55,42

По суті ми маємо двовимірні дані: час і товарообіг. Для таких даних можна будувати діаграму розсіювання, лінію регресії та доцільно поставити питання прогнозування.

Зобразимо діаграму розсіювання не у вигляді точкової, а у вигляді стовпчикової діаграми і додамо лінію регресії, яка для динамічних рядів має назву лінії тренду (див. нижче).

Лінія тренду показує генеральний напрямок зміни даних, тенденцію. Ми бачимо, що обсяги реалізації мають додатній тренд, тобто їх значення мають тенденцію до збільшення у часі.



Так, як і вище на основі лінії регресії, на основі лінії тренду можна робити прогнози. Щоб зробити такий прогноз на 1 наступний часовий період достатньо виділити у Excel стовпчик з вихідними даними і потягнути його на один рядок донизу (за крапочку справа внизу виділеного стовпчика). Якщо бажаємо прогнозувати на 2 часових періоди, то треба потягнути на 2 рядочки униз і т.д.

B2 fx 51,69

	A	B
	Місяць	**Товарообіг (тис. грн)**
2	січень	51,69
3	лютий	44,77
4	березень	48,99
5	квітень	42,86
6	травень	69,19
7	червень	54,56
8	липень	58,7
9	серпень	62,79
10	веоресень	57,3
11	жовтень	57,8
12	листопад	55,42
13	грудень	
14		

Таким чином можна спрогнозувати значення товарообігу на грудень.

. Для перевірки якості прогнозу можна так само користуватися ретроспективним прогнозом.

18. Врахування сезонної компоненти динамічного ряду

Наведений вище метод прогнозування базується на побудові лінії тренду динамічного ряду. Але лінія тренду показує лише загальну довготермінову тенденцію зміни даних з часом. Якщо ж у даних присутня сезонна компонента, тобто елементам генеральної сукупність кожного року властиві сезонні коливання, то прогноз за допомогою лінії тренду буде неточним (рис. нижче).

Рис. 18.1

Крім задачі прогнозування часто виникає потреба порівняти значення певних показників у різні періоди часу.

Приклад 18.1. Ми знаємо, що обсяги реалізації у грудні, більші, ніж у листопаді, т.я. перед новим роком збільшується попит. У листопаді нами була проведена рекламна кампанія. Нас цікавить, чи вплинула ця кампанія на збільшення обсягів реалізації, чи збільшення обсягів викликано лише сезонними коливаннями попиту?

Приклад 18.2. Як зрозуміти, чи збільшення попиту на нерухомість у другому кварталі порівняно із першим продиктовано лише сезонними коливаннями, чи відображає певну закономірність і вимагає відповідних дій?

Таким чином, нам необхідно навчитися враховувати і аналізувати сезонну компоненту динамічного ряду.

Насправді, будь-який динамічний ряд містить у собі 4 базові компоненти: довготерміновий тренд, сезонну компоненту, циклічну компоненту (яка складається із послідовних понижень та підвищень, що не повторюються кожного року і тому не входять до сезонної компоненти) та випадкової (нерегулярної) компоненти (яка присутня об'єктивно, але яку неможливо пояснити чи передбачити. У цій компоненті проявляється вплив тих однократних, унікальних подій, які відбуваються випадково, а не систематично). Нижче ми проаналізуємо лише перші дві компоненти, вплив яких, як правило, найбільший.

Щоб позбавитися сезонної компоненти скористаємося **ковзаючим середнім**. Ковзаючим середнім називається динамічний ряд, який отримують усередненням сусідніх спостережень за наступною схемою: до значення показника за поточний квартал додають значення за попередній та наступний квартали та половини значень за квартал, що передує попередньому, та квартал, що іде слідом за наступним. Отриману суму ділять
на 4. Так, якщо дані за 5 кварталів мають вигляд,

1 кв. 2001 р.	2 кв. 2001 р.	3 кв. 2001 р.	4 кв. 2001 р.	1 кв. 2002 р.
17115	19833	17205	17898	20636

То ковзаюче середнє для 3-го кварталу 2001 року має значення (17115/2+19833+17205+17898+20636/2)/4=18453. Ділення даних за перші квартали призводить до того, що ці дані враховуються у ковзаючому середньому точно також, як і дані за інші квартали (бо перші квартали входять двічі).

Для обрахування ковзаючого середнього для помісячних даних, до даних за поточний місяць додають значення за 6 попередніх та 6 наступних, причому крайні значення (за самий першій та останній місяці з обраних) ділять навпіл. Отриману суму ділять на 12.

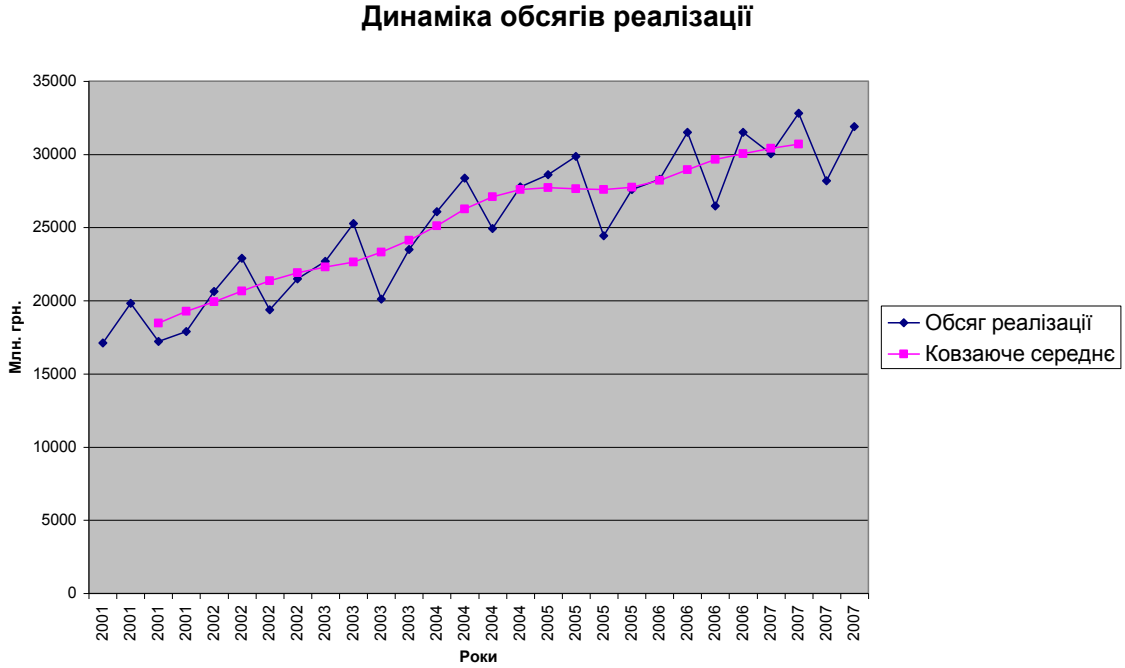
Для обрахування ковзаючого середнього неважко створити відповідну формулу вручну. У пакеті «Аналіз даних» програми Excel (Сервіс: Аналіз даних) є можливість швидкого обчислення ковзаючого середнього, але за методикою, яка дещо відрізняється від наведеної (рис. 18.2).

Рис. 18.2

D4 =(0,5*C2+C3+C4+C5+0,5*C6)/4

	A	B	C	D
1	Рік	Квартал	Обсяг реалізації (млн. грн)	Ковзаюче середнє
2	2001	1	17115	
3	2001	2	19833	
4	2001	3	17205	18453
5	2001	4	17898	19277
6	2002	1	20636	19931
7	2002	2	22903	20652
8	2002	3	19370	21358
9	2002	4	21498	21909
10	2003	1	22686	22297
11	2003	2	25264	22640
12	2003	3	20107	23315
13	2003	4	23511	24127
14	2004	1	26070	25118
15	2004	2	28375	26252
16	2004	3	24926	27101
17	2004	4	27766	27603
18	2005	1	28601	27727
19	2005	2	29861	27645
20	2005	3	24437	27586
21	2005	4	27597	27754
22	2006	1	28297	28212
23	2006	2	31505	28953
24	2006	3	26459	29659
25	2006	4	31505	30039
26	2007	1	30037	30419
27	2007	2	32805	30685
28	2007	3	28196	
29	2007	4	31897	

Якщо побудувати на одному графіку динамічний ряд і ряд, складений з ковзаючих середніх, то видно, що ковзаючі середні вже не місять сезонної компоненти:

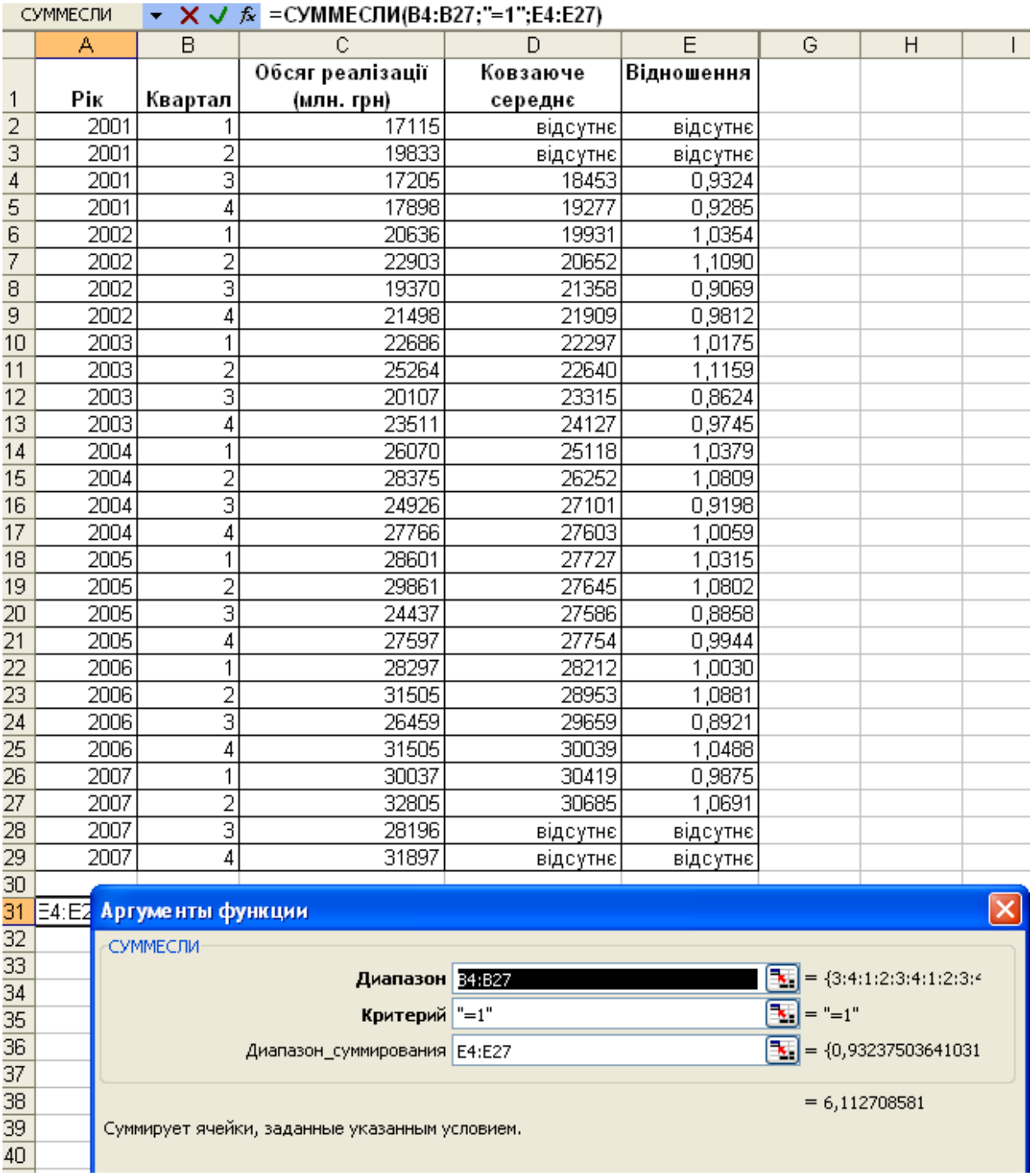


Щоб виділити суто сезонну компоненту з динамічного ряду, необхідно знайти відношення елементів динамічного ряду до відповідних ковзаючих середніх (там, де вони визначені). Обчислимо ці відношення за допомогою Excel (див. стовпчик «Відношення»):

Рік	Квартал	Обсяг реалізації (млн. грн)	Ковзаюче середнє	Відношення
2001	1	17115	відсутнє	відсутнє
2001	2	19833	відсутнє	відсутнє
2001	3	17205	18453	0,9324
2001	4	17898	19277	0,9285
2002	1	20636	19931	1,0354
2002	2	22903	20652	1,1090
2002	3	19370	21358	0,9069
2002	4	21498	21909	0,9812
2003	1	22686	22297	1,0175
2003	2	25264	22640	1,1159
2003	3	20107	23315	0,8624
2003	4	23511	24127	0,9745
2004	1	26070	25118	1,0379
2004	2	28375	26252	1,0809
2004	3	24926	27101	0,9198
2004	4	27766	27603	1,0059
2005	1	28601	27727	1,0315
2005	2	29861	27645	1,0802
2005	3	24437	27586	0,8858
2005	4	27597	27754	0,9944
2006	1	28297	28212	1,0030
2006	2	31505	28953	1,0881
2006	3	26459	29659	0,8921
2006	4	31505	30039	1,0488
2007	1	30037	30419	0,9875
2007	2	32805	30685	1,0691
2007	3	28196	відсутнє	відсутнє
2007	4	31897	відсутнє	відсутнє

Значення відношень, усереднені за один той самий період (місяць, квартал), називаються **індексами сезонності**. Тобто індекс сезонності для 1-го кварталу – це середнє арифметичне відношень по усіх перших кварталах, для 2-го кварталу – по усіх других кварталах і т.д.

Для знаходження сум відношень за певний квартал можна скористатися математичною функцією СУММЕСЛИ. При цьому діапазон – це ряд кварталів, для яких визначені відношення, критерій – це той квартал, за який знаходиться сума, а діапазон сумування – це ряд з відношень. Отриману суму необхідно поділити на 6 – кількість врахованих у сумі кварталів.



	A	B	C	D	E	G	H	I
1	Рік	Квартал	Обсяг реалізації (млн. грн)	Ковзаюче середнє	Відношення			
2	2001	1	17115	відсутнє	відсутнє			
3	2001	2	19833	відсутнє	відсутнє			
4	2001	3	17205	18453	0,9324			
5	2001	4	17898	19277	0,9285			
6	2002	1	20636	19931	1,0354			
7	2002	2	22903	20652	1,1090			
8	2002	3	19370	21358	0,9069			
9	2002	4	21498	21909	0,9812			
10	2003	1	22686	22297	1,0175			
11	2003	2	25264	22640	1,1159			
12	2003	3	20107	23315	0,8624			
13	2003	4	23511	24127	0,9745			
14	2004	1	26070	25118	1,0379			
15	2004	2	28375	26252	1,0809			
16	2004	3	24926	27101	0,9198			
17	2004	4	27766	27603	1,0059			
18	2005	1	28601	27727	1,0315			
19	2005	2	29861	27645	1,0802			
20	2005	3	24437	27586	0,8858			
21	2005	4	27597	27754	0,9944			
22	2006	1	28297	28212	1,0030			
23	2006	2	31505	28953	1,0881			
24	2006	3	26459	29659	0,8921			
25	2006	4	31505	30039	1,0488			
26	2007	1	30037	30419	0,9875			
27	2007	2	32805	30685	1,0691			
28	2007	3	28196	відсутнє	відсутнє			
29	2007	4	31897	відсутнє	відсутнє			
30								
31	E4:E2							
32								
33								
34								
35								
36								
37								
38								
39								
40								

Значення індексів сезонності, порахованих за допомогоюExcel, зібрані у таблиці:

Квартал	Індекс сезонності
1	1,018785
2	1,090524
3	0,899902
4	0,988875

Індекс сезонності показує наскільки більшим (чи меншим) є показник, що розглядається, у певний момент часу, порівняно із типовим періодом року. Наприклад, значення 1,018 означає, що у першому кварталі обсяги реалізації на 1,8% більші, ніж у типовому кварталі. Дані з поправкою на сезонність наведені у таблиці:

Порівняємо тепер обсяги реалізації продукції за 4-ий квартал 2006 року та 2-ий квартал 2007 року. Ці дані становлять 31505 та 32805 млн. грн. відповідно. Але для порівняння необхідно внести поправку на сезонні коливання, тобто виключити з даних вплив сезонності. Для цього дані ділять на відповідний сезонний індекс. Для 4-го кварталу значення з поправкою на сезон дорівнює 31859, для 2-го - 30082. Тобто, по суті, якщо відкинути вплив сезонності, то обсяг реалізації у 2-му кварталі 2007 року не збільшився, а зменшився!

Таким чином, значення з поправкою на сезонність – це дані динамічного ряду, поділені на відповідні індекси сезонності. Саме такі значення використовують для порівняння даних за різні періоди часу (різні місяці, квартали тощо).

Отримаємо тепер з вихідного динамічного ряду значення з поправкою на сезонність. Для цього можна скористатися логічною функцією ЕСЛИ.

Arial Cyr | 10 | Ж К Ч

G4 | =ЕСЛИ(B4=1;C4/B$31;ЕСЛИ(B4=2;C4/B$32;ЕСЛИ(B4=3;C4/B$33;C4/B$34)))

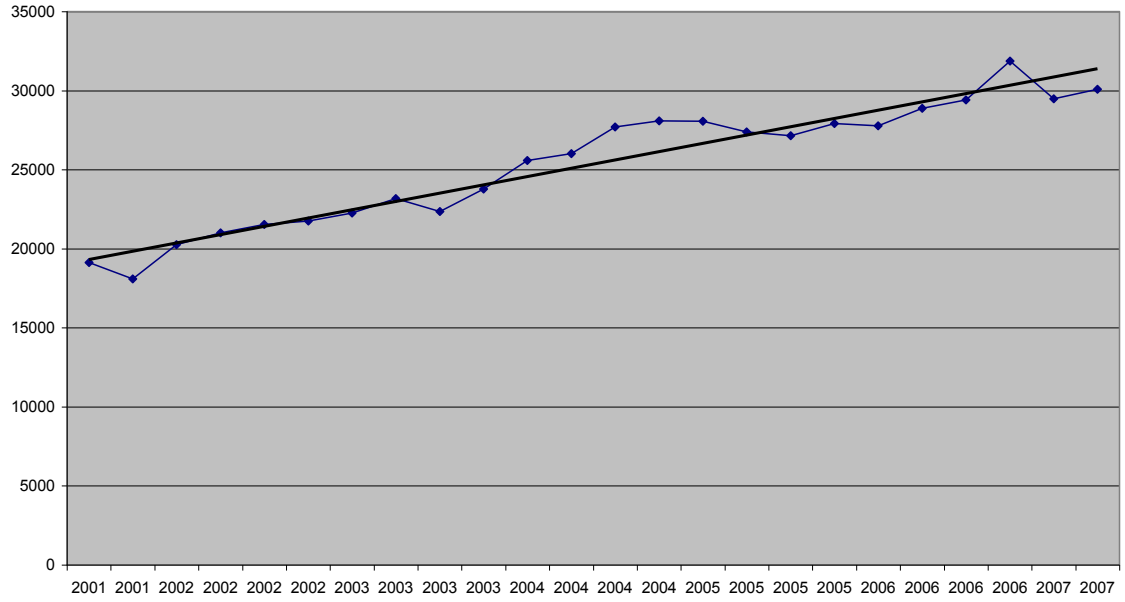
	A	B	C	D	E	G	H	I
1	**Рік**	**Квартал**	**Обсяг реалізації (млн. грн)**	**Ковзаюче середнє**	**Відношення**			
2	2001	1	17115	відсутнє	відсутнє	відсутнє		
3	2001	2	19833	відсутнє	відсутнє	відсутнє		
4	2001	3	17205	18453	0,9324	19119		
5	2001	4	17898	19277	0,9285	18099		
6	2002	1	20636	19931	1,0354	20256		
7	2002	2	22903	20652	1,1090	21002		
8	2002	3	19370	21358	0,9069	21525		
9	2002	4	21498	21909	0,9812	21740		
10	2003	1	22686	22297	1,0175	22268		
11	2003	2	25264	22640	1,1159	23167		
12	2003	3	20107	23315	0,8624	22344		
13	2003	4	23511	24127	0,9745	23775		
14	2004	1	26070	25118	1,0379	25589		
15	2004	2	28375	26252	1,0809	26020		
16	2004	3	24926	27101	0,9198	27699		
17	2004	4	27766	27603	1,0059	28078		
18	2005	1	28601	27727	1,0315	28074		
19	2005	2	29861	27645	1,0802	27382		
20	2005	3	24437	27586	0,8858	27155		
21	2005	4	27597	27754	0,9944	27907		
22	2006	1	28297	28212	1,0030	27775		
23	2006	2	31505	28953	1,0881	28890		
24	2006	3	26459	29659	0,8921	29402		
25	2006	4	31505	30039	1,0488	31859		
26	2007	1	30037	30419	0,9875	29483		
27	2007	2	32805	30685	1,0691	30082		
28	2007	3	28196	відсутнє	відсутнє	відсутнє		
29	2007	4	31897	відсутнє	відсутнє	відсутнє		
30		**Індекс сезонності**						
31	6,112709	1,018785						
32	6,543142	1,090524						
33	5,399414	0,899902						
34	5,933252	0,988875						

Нижче наведена таблиця, що містить початковий динамічний ряд та ряд з поправками на сезонність.

Рік	Квартал	Обсяг реалізації (млн. грн)	Обсяг реалізації з урахуванням сезонності (млн. грн.)
2001	1	17115	відсутнє
2001	2	19833	відсутнє
2001	3	17205	19119
2001	4	17898	18099
2002	1	20636	20256
2002	2	22903	21002
2002	3	19370	21525
2002	4	21498	21740
2003	1	22686	22268
2003	2	25264	23167
2003	3	20107	22344
2003	4	23511	23775
2004	1	26070	25589
2004	2	28375	26020
2004	3	24926	27699
2004	4	27766	28078
2005	1	28601	28074
2005	2	29861	27382
2005	3	24437	27155
2005	4	27597	27907
2006	1	28297	27775
2006	2	31505	28890
2006	3	26459	29402
2006	4	31505	31859
2007	1	30037	29483
2007	2	32805	30082
2007	3	28196	31332
2007	4	31897	32256

Зобразимо графічно ряд з поправкою на сезонність і відповідну йому лінію тренду.

Обсяги реалізації із поправкою на сезонність



Ми бачимо, що дані з поправкою на сезонність знаходять набагато щільніше до лінії тренду, ніж первинні данні (рис. 18.1), тому і лінійний прогноз, отриманий за допомогою лінії тренду, буде набагато якіснішим.

Прогнозовані значення з поправкою на сезонність на 3-й та 4-й квартали 2007 року дорівнюють 31332 та 32256. Для того, щоб тепер повернути у прогнозовані дані сезонність, тобто врахувати її реальний вплив, їх множать на індекси сезонності, які відповідають 3-му та 4-му кварталам відповідно. Отримаємо 28196 та 31897 млн. грн.

Таким чином, для прогнозування динамічного ряду, який має виражену сезонність, його позбавляють впливу сезонності, поділивши його значення на відповідні індекси сезонності, і формують ряд з поправкою на сезонність. Для цього ряду здійснюють лінійний прогноз, і прогнозовані дані множать на відповідні індекси сезонності, щоб знову врахувати її реальний вплив.

Література

1. Э.Ф.Сигел. Практическая бизнес-статистика.- М.: Вильямс, 2002, 1056 с.
2. Кеннет Берк, Патрик Кэйри. Анализ данных с помощью Microsoft Excel
3. П.Корнелл. Анализ данных в Excel. Просто как дважды два.
4. Липпе П. фон дер. Экономическая статистика. -- Штудгарт: Йена, 1995. -- 629 с.
5. Статистика: Навч.-метод. посіб. для самост. вивч. дисц. / За ред. А. М. Єріної та Р. М. Моторина. -- К.: КНЕУ, 2002.
6. К.Калберг. Бизнес-анализ с помощью Excel 2000. М.: Вильямс, - 480 с.