ЛЕКЦІЯ.
ПЕРВИННІ ФОРМИ ВІЗУАЛІЗАЦІЇ ДАНИХ

Ваш партнер вже півгодини розглядає величезну таблицю витрат споживачів на покупку виробів ваших конкурентів, сподіваючись дізнатися якомога більше з чисел в колонках і навіть частково досягнувши успіху в цьому (про це свідчать періодичні вигуки типу «Більшість витрачає від 10 до 15 доларів!», «Практично ніхто не витрачає більше 35 доларів!» та «О-о! Один витратив 58 доларів!»). Ви розумієте, що слід порадити партнеру використовувати замість цієї таблиці якийсь графік, наприклад гістограму, оскільки це заощадить час і дасть більш повну картину. Єдина проблема – чисто психологічна: як пояснити це партнерові, не зачепивши його самолюбства.

У цій лекції ви дізнаєтеся, як надати сенс колонці чисел. Гістограма – це графічне зображення даних, яке дає візуальне уявлення про основні властивостях набору даних в цілому і дозволяє відповісти на наступні питання:

Перше. Які значення типові для цього набору даних?

Друге. Як розрізняються між собою значення?

Третє. Навколо якого значення сконцентровані дані?

Четверте. Який характер має ця концентрація даних? Зокрема, чи однаковий характер «загасання» для малих і великих значень даних?

П'яте. Чи є в цьому наборі такі значення, які настільки сильно відрізняються від інших, що вимагають спеціальної обробки?

Шосте. Чи можна сказати, що це в цілому однорідний набір або чітко спостерігається наявність груп, які необхідно аналізувати окремо?

Багато стандартних методів статистичного аналізу вимагають, щоб набір даних був приблизно нормально розподіленим. Ви дізнаєтеся, як розпізнати цю, схожу на дзвін, форму і як перетворити дані, якщо вони не задовольняють цій вимозі.

2.1. Послідовність даних

Набір даних найпростішого виду – це послідовність чисел, що представляють деяку властивість (єдина статистична змінна), виміряна для кожного з аналізованих об'єктів (для кожної елементарної одиниці). Послідовність чисел можна представити в кількох формах, які, на перший погляд, здаються такими, що сильно різняться між собою. Допомогти відрізнити результати вимірювань (значень) від частот може відповідь на питання: «Що є елементарною одиницею, для яких проводилися вимірювання?».

Приклад. Діяльність регіональних менеджерів з продажу

Розглянемо приклад дуже короткої послідовності (тільки три спостереження), де змінною є «обсяг продажів останнього кварталу», а елементарними одиницями – «регіональні менеджери з продажу».

Ім'я	Обсяг продажів (десятки тисяч)
Білл	28
Дженіфер	32
Генрі	18

Цей набір даних на додаток до трьох чисел обсягу продажів містить інформацію для інтерпретації (тобто ім'я менеджера з продажу, яке позначає кожну елементарну одиницю набору даних). Іноді така перша колонка опускається, і значення змінної записуються безпосередньо в першу колонку.

Приклад. Розмір домогосподарства

Іноді послідовність чисел має вигляд таблиці частот, як у наведеному нижче прикладі даних про кількість членів сім'ї у вибірці з 17 домогосподарств.

Розмір домогосподарства (кількість осіб)	Кількість домогосподарств (частота)
1	3
2	5
3	6
4	2
5	0
6	1

При інтерпретації такої таблиці необхідно враховувати, що вона являє собою таку послідовність чисел, в якій кожне число з лівої колонки (розмір домогосподарства) повторюється таку кількість разів, яку зазначено у відповідному рядку в правій колонці (частота такого спостереження, кількість домогосподарств такого розміру). У таблиці представлений наступний перелік чисел, що відображає кількість людей в кожному домогосподарстві:

1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 4, 4, 6

Число 1 повторюється в цьому списку тричі (як показано в першому рядку таблиці), число 2 – 5 разів (що випливає з другого рядка) і т.д.

Таблиця частот особливо корисна для подання довгих переліків чисел з відносно невеликою кількістю різних значень. Тому для вибірки великого розміру розміри домогосподарств можна було б узагальнити наступним чином.

Розмір домогосподарства (кількість осіб)	**Кількість домогосподарств (частота)**
1	342
2	581
3	847
4	265
5	23
6	11
7	2

У цій таблиці представлено багато даних! Відповідний перелік чисел починається послідовністю з 342 одиниць, потім йде 581 двійки і т.д. Таблиця містить розміри всіх 2 071 домогосподарств з цієї великої вибірки[1].

Числова вісь

Щоб наочно уявити значення послідовності, ми розташуємо числа вздовж прямої. Числова вісь являє собою пряму лінію з нанесеною на ній шкалою числових значень.

[1] Число 2071 – це фінальна частота, тобто сума всіх чисел правої колонки.

Важливо розташувати числа на осі рівномірно і без пропусків[2]. Щоб показати місце кожного з чисел послідовності, можна зробити позначки у відповідних місцях на осі. Наприклад, три цифри, що відповідають продажам: 28, 32, 18, можна зобразити на числової осі таким чином.

Ця діаграма дає наочне уявлення про те, як ці значення співвідносяться між собою. Зокрема, відразу ж видно, що два значення достатньо близькі один до одного за величиною і набагато більше третього значення.

Використання такого роду або інших графіків більш інформативне для аналізу, ніж просто розглядання послідовностей чисел. Хоча числа добре підходять для реєстрації даних, вони не мають наочності в поданні цілого ряду властивостей даних. Наприклад, послідовність

0 1 2 3 4 5 6 7 8 9

не дає ніякої конкретної візуальної вказівки про послідовне збільшення значень; при русі за переліком чисел зліва направо числа не стають більше за розміром, не стають темнішими і т.п. У той же час числова вісь явно показує цю важлива властивість.

2.2. Використання гістограм для відображення частот

Гістограма демонструє частоти у вигляді діаграми з стовпчиків, які розташовані над числовою віссю і показують, наскільки часто різні значення зустрічаються в наборі даних. По горизонтальній осі відкладають виміряні значення з набору даних (виражені в доларах, кількості людей, милях на галон і інших одиницях виміру), по вертикальній – частоту появи цих значень. Висоти прямокутників відповідають частотам значень, найвищий стовпчик

[2] Якщо необхідно розірвати числову вісь, наприклад, щоб пропустити значення, що не цікавлять нас, то слід явно показати розрив на числової осі, що дозволить не створювати помилкове враження звичайної безперервної лінії.

відповідає значенню, що найбільш часто зустрічається в наборі даних, а найнижчий – значенню, що зустрічається рідше всіх.

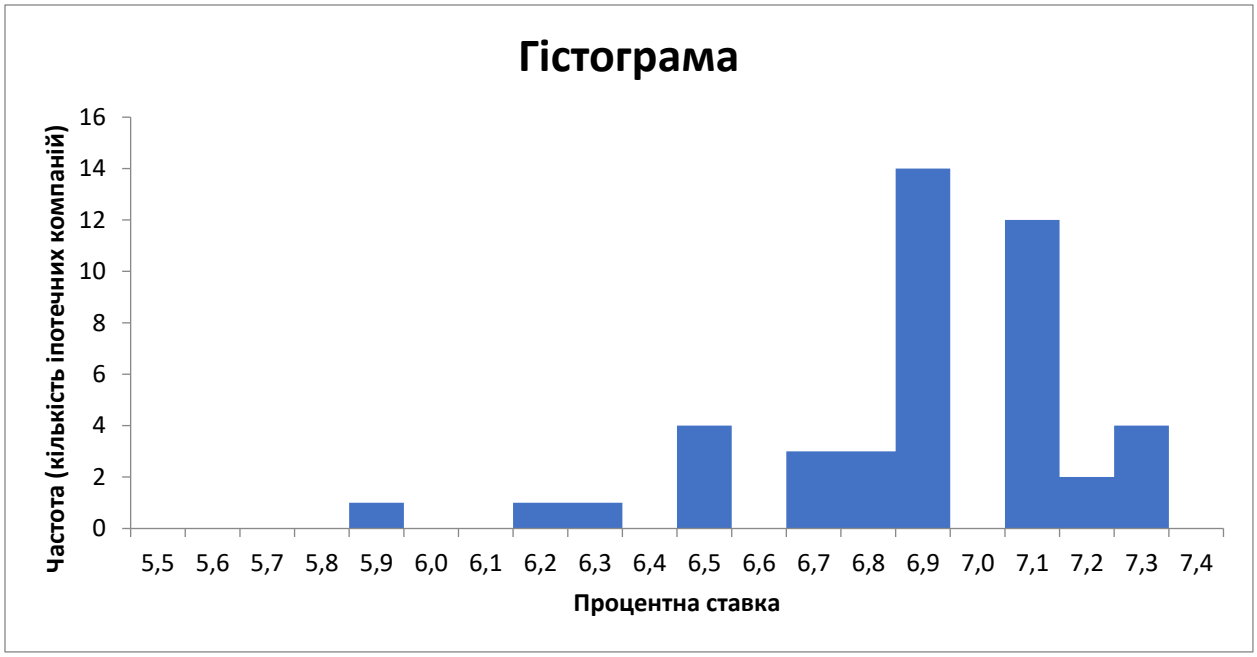
Приклад. Процентні ставки позики під заставу нерухомості

У табл. 2.2.1 представлені розміри фіксованої процентної ставки позик під заставу нерухомості, що надаються на 30 років іпотечними компаніями Сіетла (США).

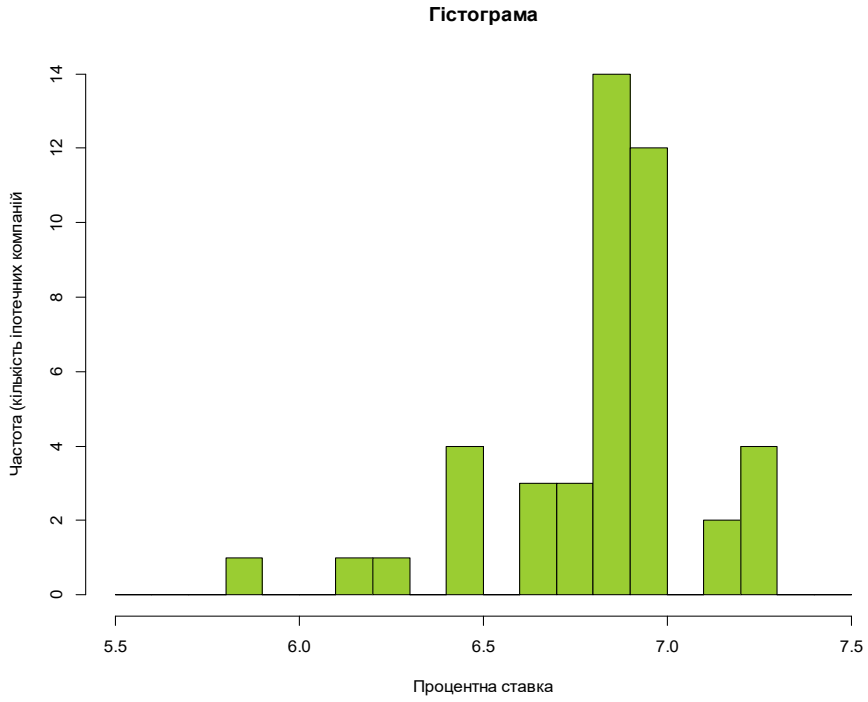
Таблиця 2.2.1. Ставки на позику під заставу нерухомості

Кредитор		**Процентна ставка**
Accubanc Mortgage Corp.	1.	7,000
Alpine Mortgage Services	2.	6,875
American Investment Mrtg.	3.	6,875
Bay Mortgage	4.	6,750
Capital Mortgage Corp.	5.	6,875
Castle Mortgage Corp.	6.	7,250
Choice Mortgage	7.	6,875
Citizen's Mortgage Inc.	8.	7,000
City Mortgage	9.	6,875
Community National Mrtg.	10.	7,000
Countrywide Home Loans	11.	7,250
Edmonds Mortgage Inc.	12.	7,000
Equity Northwest, Inc.	13.	7,000
Evergreen Pacific Services	14.	6,125
First American Mortgage	15.	6,750
First Mark Mortgage	16.	7,125
First National Home Mrtg.	17.	7,125
Goldmark Financial Corp.	18.	7,000
Group One Mortgage, Inc.	19.	7,000
Guaranty Mortgage Co.	20.	7,000
Home Loans Online	21.	6,875
Home Mortgage Corp.	22.	6,875
Integral Real Estate & Mrtg	23.	6,500
Intercontinental Mtg.	24.	6,500
Lincoln Federal Mortgage	25.	6,500
Merrill Lynch Credit	26.	7,250
Millennium Mortgage	27.	6,750
Mortgage Broker Services	28.	6,875
Mortgage Network, Inc.	29.	6,875
Mortgage Solutions	30.	6,875
Nu-West Mortgage	31.	6,875
Pacific Mountain Mortgage	32.	6,500
Park Place	33.	6,875
Performance Mortgage	34.	7,000
Portia Financial Services	35.	6,875
Prime Port Mortgage	36.	7,000

Producer $ Mortgage Service	37.	7,250
Raintree Financial Network	38.	7,000
Redmond Mortgage Co.	39.	6,625
Residential Mtg. Brokers Inc,	40.	6,875
Select Mortgage	41.	6,625
US Discount Mortgage Co.	42.	6,625
Wash. Women's Mrtg. Crp.	43.	6,250
Western Heritage Mrtg. Svgs.	44.	5,875
Wohletz Mortgage	45.	7,000



Мал. 2.2.1. Гістограма процентних ставок позик під заставу нерухомості (створена за допомогою MS Offcie Excel)



Мал. 2.2.1. Гістограма процентних ставок позик під заставу нерухомості (створена за допомогою Rstudio)

Тепер опишемо загальний підхід до інтерпретації гістограм і одночасно з'ясуємо, про що говорить нам цей конкретний графік розглянутих процентних ставок.

Числа на горизонтальній осі в нижній частині малюнка вказують на значення процентних ставок, виражені в процентах. Числа на вертикальній осі показують частоту появи кожної процентної ставки. Наприклад, передостанній стовпчик праворуч (розташований по горизонталі між процентними ставками, рівними 7,1% і 7,2%) має частоту (висоту), що дорівнює 2, що означає, що 2 фінансові організації пропонують ставку між 7,1% і 7,2%[3]. Таким чином, ви маєте графічне зображення характеру зміни процентних ставок, яке показує, які значення зустрічаються найбільш часто, які – найменш часто, а які ставки взагалі не пропонуються.

Що можна дізнатися про процентні ставки з цієї гістограми?

[3] Прийнято відносити всі значення даних, які потрапляють точно між кордонами двох стовпчиків гістограми, до стовпчика, розташованому праворуч. В даному конкретному випадку стовпчик між числами 7,1 і 7,2 на горизонтальній осі включає всі компанії, чиї ставки дорівнюють або перевищують ліве значення (7,1%), але менше правого (7,2%). Організація, що пропонує 7,2%, увійде в наступний стовпчик, розташований праворуч від значення 7,2.

1. Розмах (діапазон) значень. Розмах процентних ставок становить більше однієї процентної одиниці від найменшого значення (близько 5,8%) до максимального значення (близько 7,3%) – це відповідно ліва і права межі гістограми.

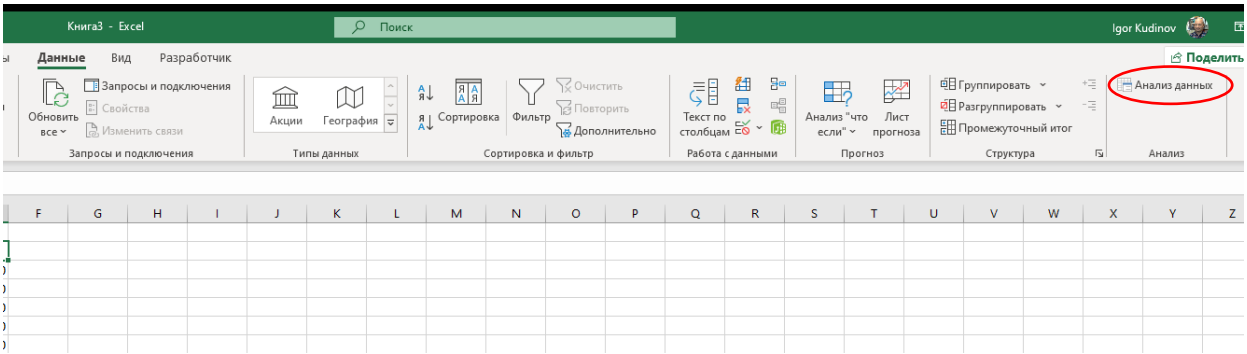
2. Типові значення. Ставки розміром від 6,8% до 7,1% зустрічаються найчастіше (зверніть увагу на високі стовпчики в цій частині діаграми).

3. Розсіювання. Типова різниця ставок для різних фінансових організацій становить приблизно 0,5% (помірно високі стовпчики стоять один від одного по горизонтальній осі приблизно на 0,5 процентних одиниць).

4. Загальна конфігурація даних. Більшість організацій зосереджені правіше середини діапазону (тут найвищі стовпчики), і незначна кількість організацій пропонують або дуже низькі, або дуже високі ставки (короткі стовпчики праворуч і ліворуч).

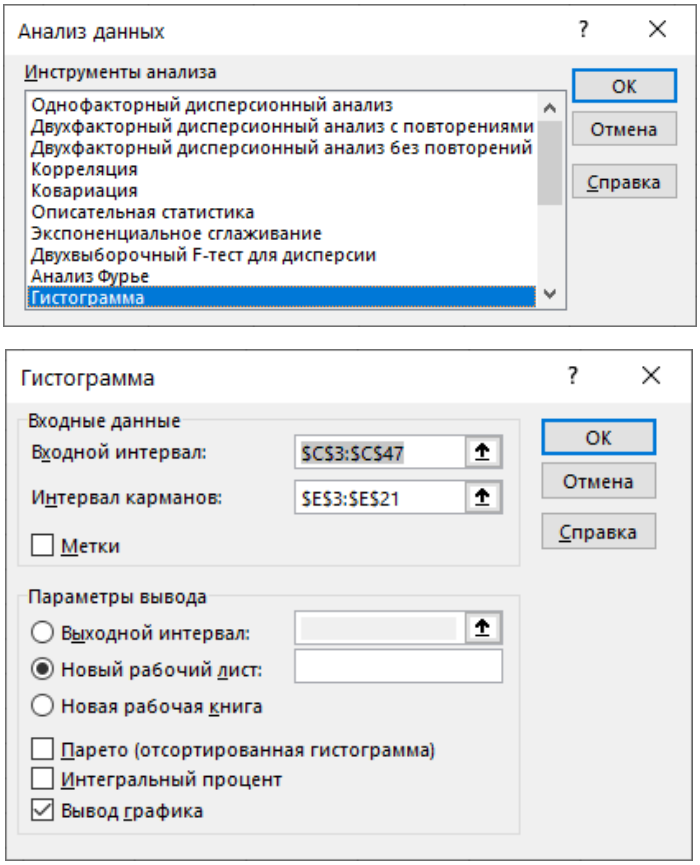
5. Характерні особливості. Ймовірно, ви помітили, що на гістограмі в цьому прикладі пропущена область від 6,9 до 7,0. Мабуть, жодна компанія не пропонує ставку в інтервалі від 6,9% до 7,0%. Це обумовлено тим, що, як правило, вказують ставки, кратні 1/8 відсотка (наприклад, 6,5%; 6,625%; 6,75%, 6,875% і 7%).

Незважаючи на те, що Microsoft Excel дозволяє будувати гістограми, часто краще використовувати або інші додатки, або окремі статистичні пакети програм. Щоб побудувати гістограму за допомогою Excel, виберіть у меню Дані пункт Аналіз даних, а потім пункт Гістограма.

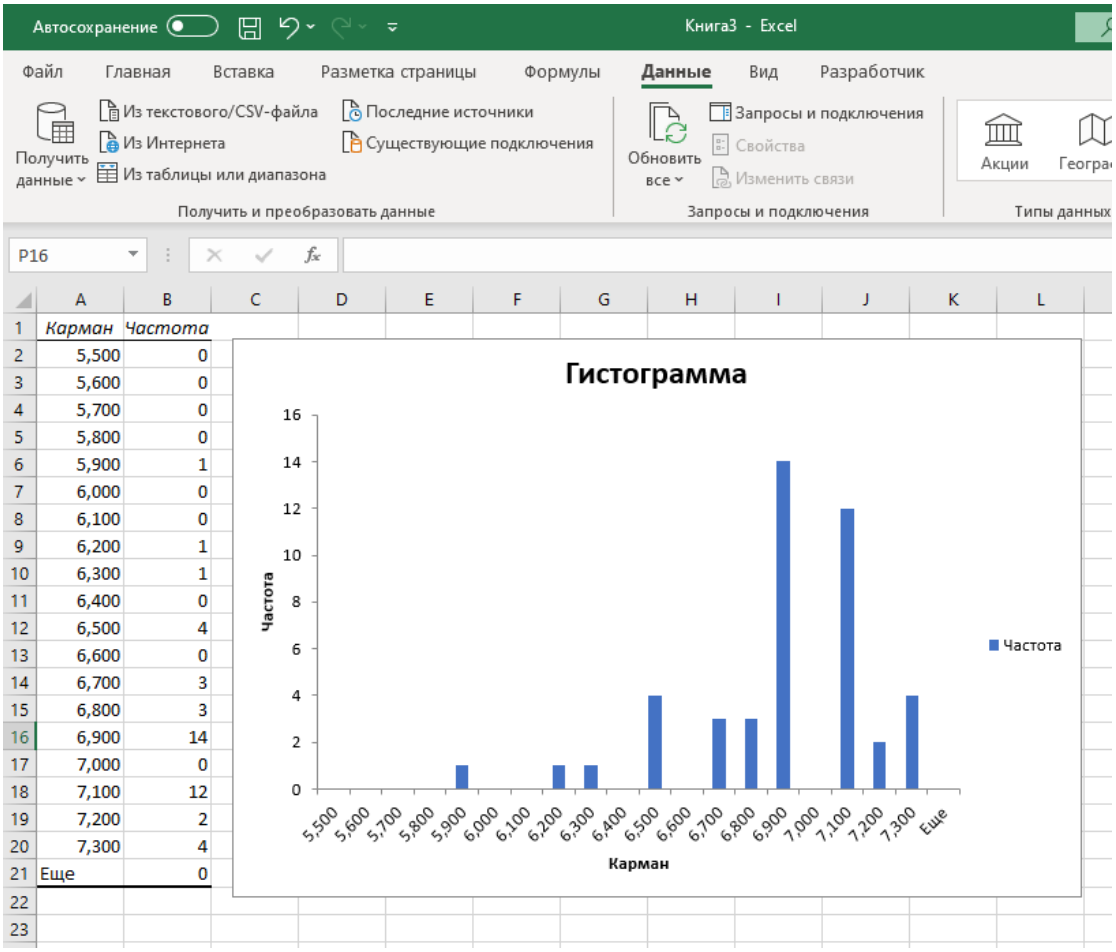


У діалоговому вікні вкажіть дані (виділяючи дані за допомогою миші у вікні або, якщо дані названі, вводячи відповідну назву), поставте позначку біля

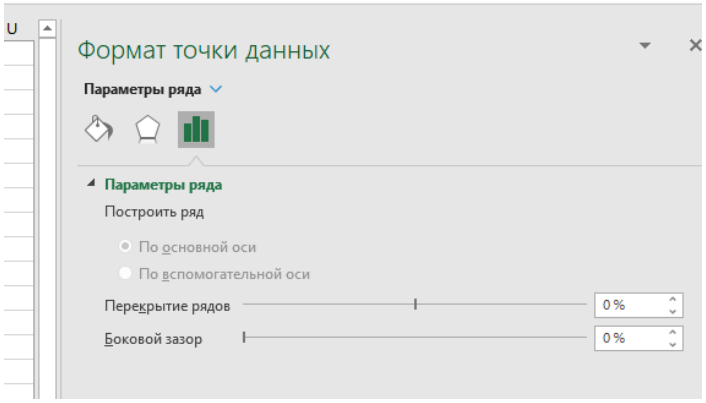
«Вивід графіка» і вкажіть, куди необхідно вивести результат («Новий робочий лист».



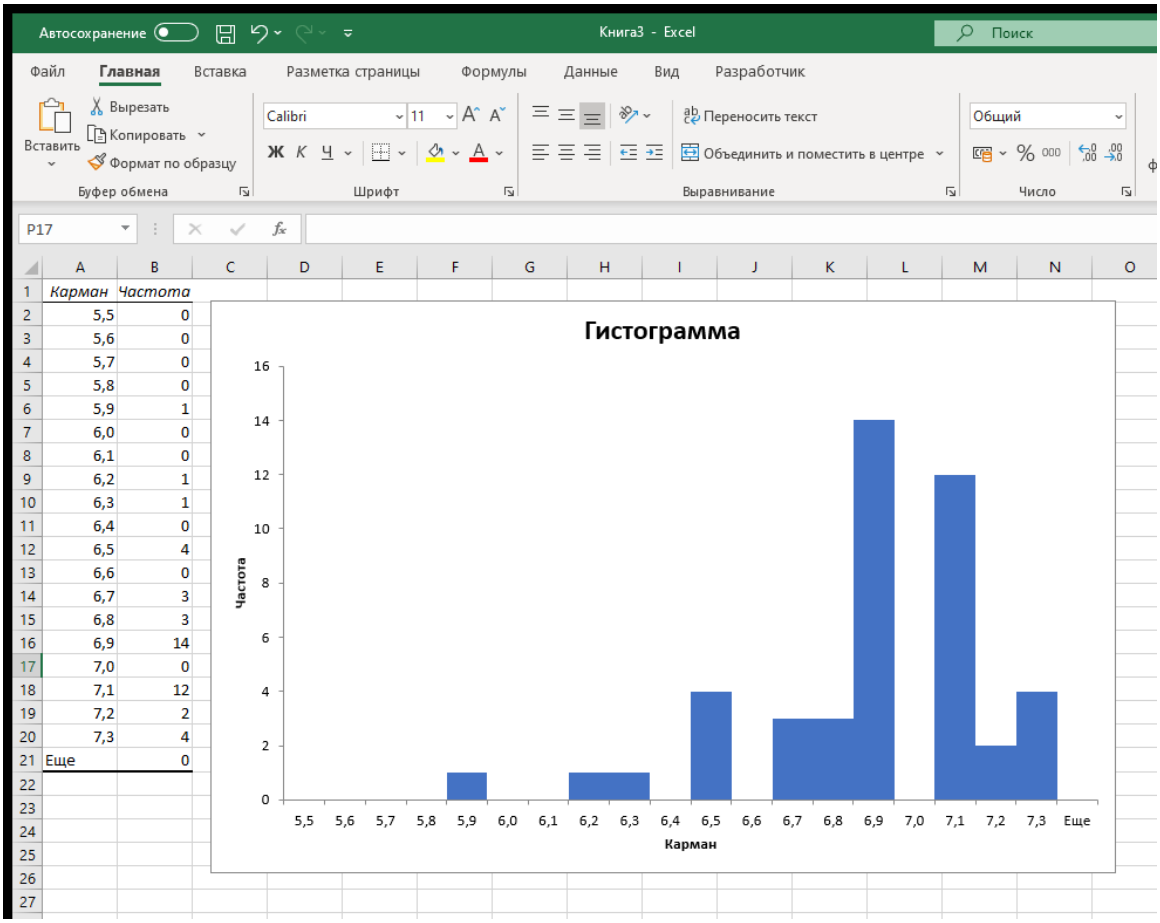
Побудовану гістограму можна збільшити (клацнути мишею біля краю і потім розтягнути картинку), щоб більш чітко побачити деталі.



	A	B
1	*Карман*	*Частота*
2	5,500	0
3	5,600	0
4	5,700	0
5	5,800	0
6	5,900	1
7	6,000	0
8	6,100	0
9	6,200	1
10	6,300	1
11	6,400	0
12	6,500	4
13	6,600	0
14	6,700	3
15	6,800	3
16	6,900	14
17	7,000	0
18	7,100	12
19	7,200	2
20	7,300	4
21	Еще	0
22		
23		



Однак стовпчики на цьому малюнку занадто вузькі, щоб це була дійсно гістограма. Щоб поправити це, двічі клацніть на зображенні стовпчика, виберіть у вікні вкладку «Параметри», встановіть нульове значення для параметра Боковий зазор і клацніть на кнопці «ОК». В результаті отримаємо таку гістограму.



Ви бачите, що побудувати гістограму в Excel непросто, особливо якщо ви хочете побудувати нестандартну гістограму зі стовпчиками певної ширини (вказуючи значення для параметра «Інтервал кишень» в діалоговому вікні). Тому в якості альтернативи можна використовувати статистичний пакет R або інший програмний продукт.

Гістограми і стовпчикові діаграми

Гістограма – це стовпчикова діаграма частот, а не даних. Висота кожного стовпчика на гістограмі показує, як часто вказане на горизонтальній осі значення зустрічається в наборі даних. Це дає візуальне уявлення про місця підвищеної і зниженої концентрації даних. Кожен стовпчик на гістограмі може представляти багато значень даних (фактично висота стовпчика точно відображає кількість значень набору даних, які стосуються відповідного діапазону). Це відрізняє гістограму від стовпчикової діаграми фактичних значень даних, де кожному значенню відповідає свій стовпчик. Також зверніть

увагу, що у гістограми числа на горизонтальній осі завжди мають змістовну інтерпретацію, а у стовпчиковій діаграмі – не обов'язково.

Приклад. Стартова заробітна плата випускників ФСУ з дипломами магістра

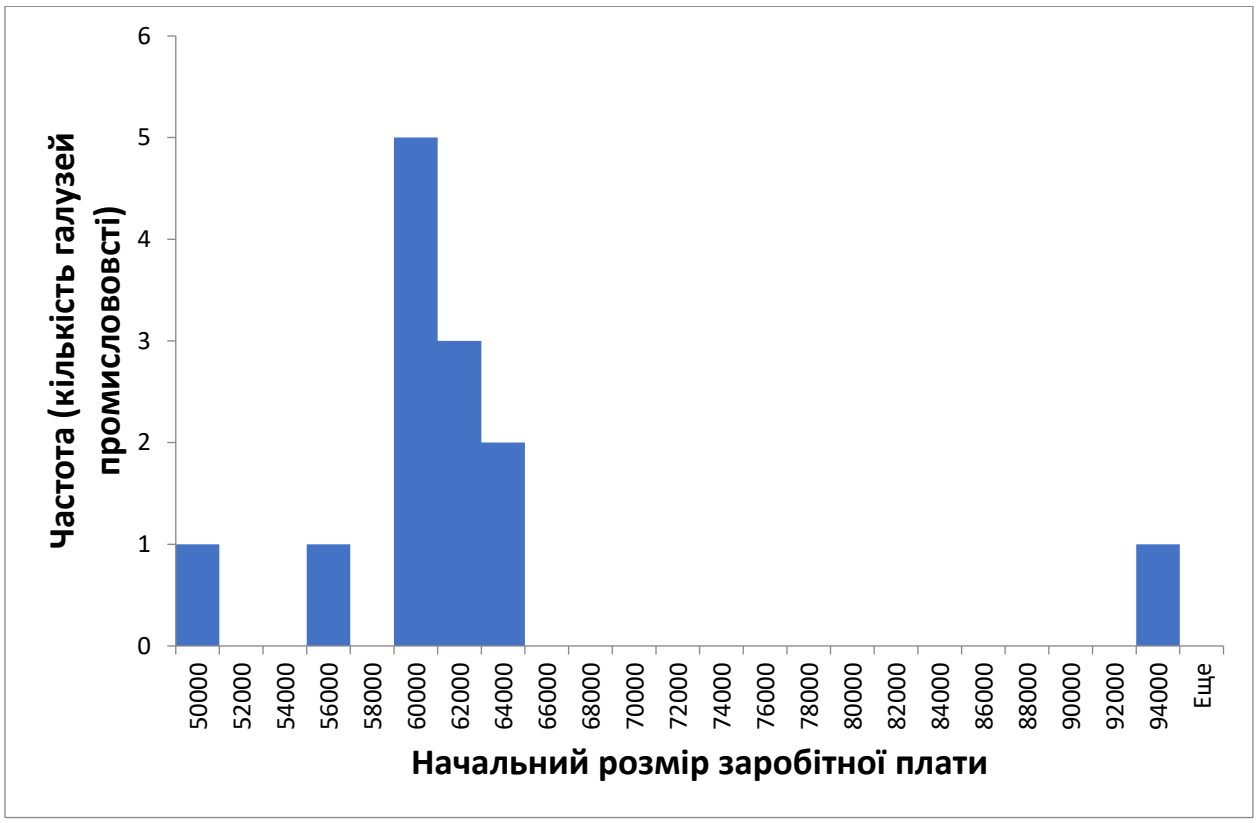
Розглянемо розмір типовою початкової заробітної плати в різних областях промисловості випускників Факультету соціології та управління (ФСУ), які отримали в 2016 році ступінь магістра. Відповідні дані наведені в табл. 2.2.2.

Порівняйте гістограму значень даних (рис. 2.2.2) і стовпчикову діаграму, наведену на рис. 2.2.3. Зверніть увагу, що стовпчики на гістограмі показують кількість галузей в кожному з діапазонів заробітної плати, а стовпчики на стовпчиковій діаграмі – фактичне значення заробітної плати в конкретній галузі.

Корисні обидва графічних зображення. Стовпчикову діаграм краще використовувати, коли бажано ідентифікувати всі значення з набору даних, за умови, що набір даних досить невеликий. Однак для отримання загального уявлення про набір даних більше підходить гістограма, особливо при великих наборах даних з безліччю чисел.

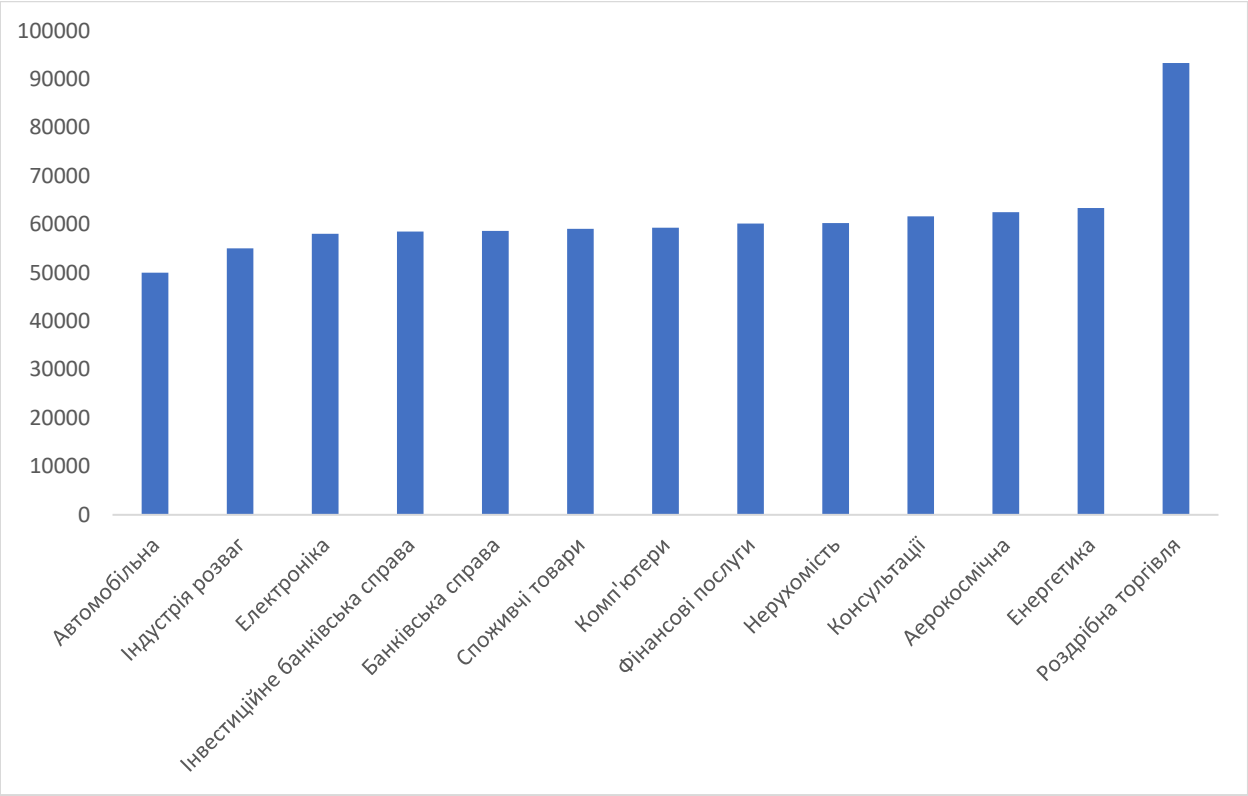
Таблиця 2.2.2. Стартова заробітна плата випускників ФСУ з дипломами магістра

Галузь	**Заробітна плата, грн.**
Аерокосмічна	62 500
Автомобільна	50 000
Банківська справа	58 611
Комп'ютери	59 280
Консультації	61 625
Споживчі товари	59 062
Електроніка	58 016
Енергетика	63 333
Індустрія розваг	55 000
Фінансові послуги	60 125
Інвестиційне банківська справа	58 500
Нерухомість	60 250
Роздрібна торгівля	93 300



Мал. 2.2.2. Гістограма значень початкового розміру заробітної плати

Зверніть увагу, що кожен стовпчик може представляти більше однієї галузі (див. Кількість на вертикальній осі зліва). Стовпчики показують, які діапазони заробітної плати частіше, а які рідше зустрічаються в цьому наборі даних

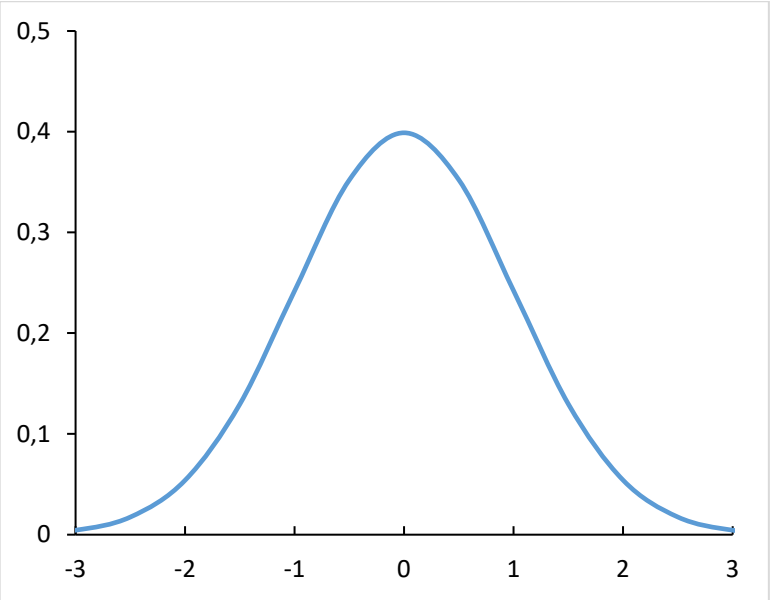


Мал. 2.2.3. Стовпчикова діаграма значень початкового розміру заробітної плати

Зверніть увагу, що кожен стовпчик представляє одну галузь промисловості

2.3. Нормальний розподіл

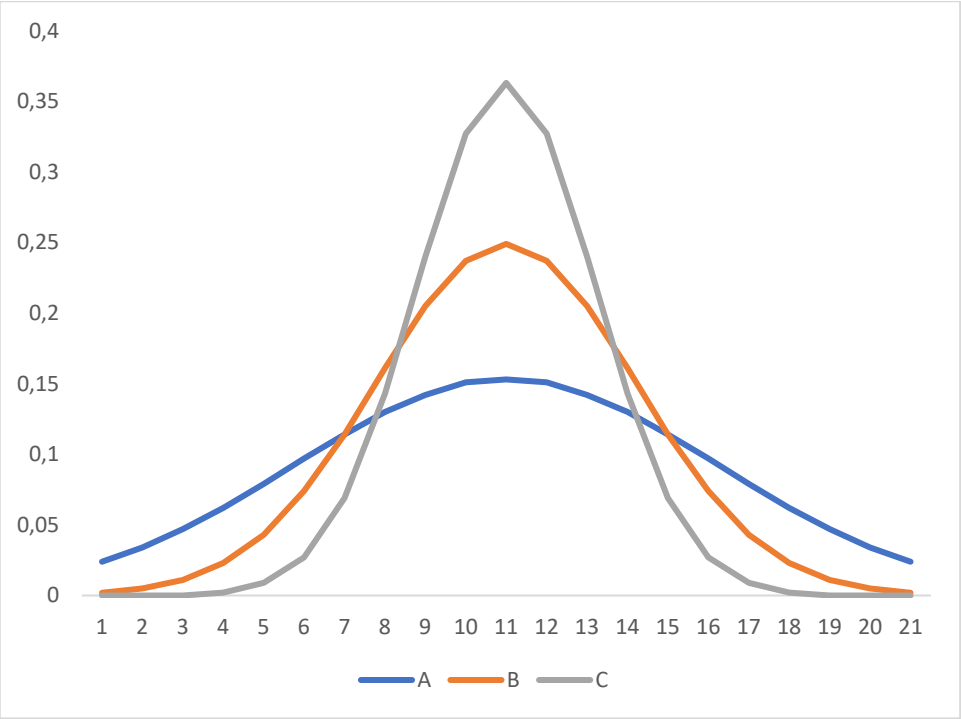
Нормальний розподіл – це теоретична гладка гістограма в формі дзвона без випадкових відхилень. Така крива представляє ідеальний набір даних, в якому більшість чисел сконцентровано в середній частині діапазону значень, а решта значень із загасанням симетрично розташовані по обидва боки від вершини дзвона. Такий ступінь гладкості не притаманний реальним даним. На рис. 2.3.1 приведена крива нормального розподілу.



Мал. 2.3.1. Ідеальна (теоретична) крива нормального розподілу

Реальні нормально розподілені набори даних мають деякі випадкові відхилення від цієї ідеально гладкої кривої.

Фактично існує багато різних кривих нормального розподілу, форма яких нагадує симетричний дзвін. Вони відрізняються розташуванням центру і масштабом (шириною дзвону). Щоб побудувати конкретну криву нормального розподілу, слід взяти базову криву у формі дзвону, перемістити її по горизонталі в точку, де передбачається розмістити центр, а потім розтягнути (або стиснути). На рис. 2.3.2 наведено кілька кривих нормального розподілу.

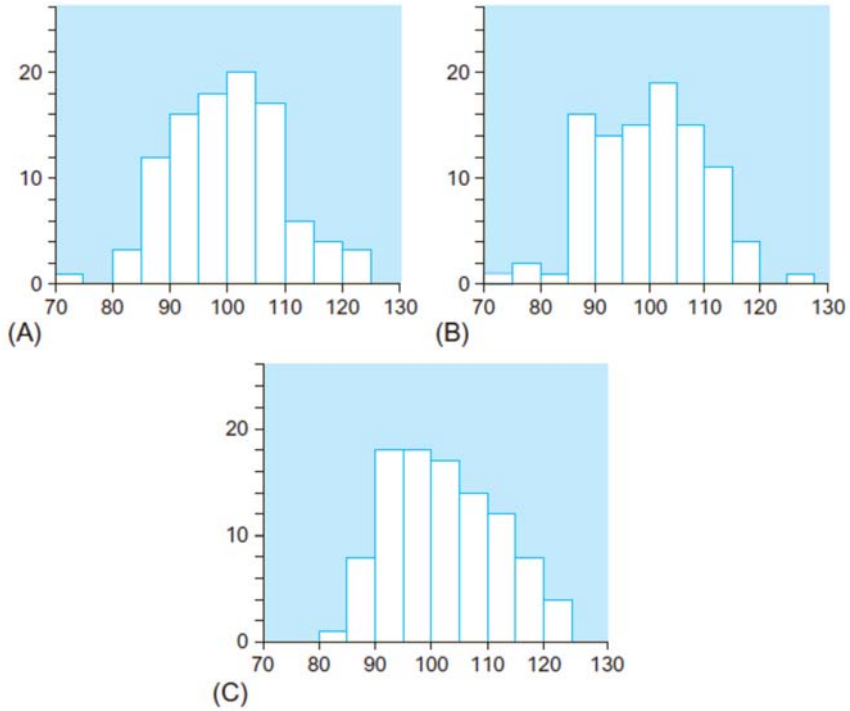


Мал. 2.3.2. Криві нормального розподілу з різними центрами і по-різному розтягнуті (в різних масштабах)

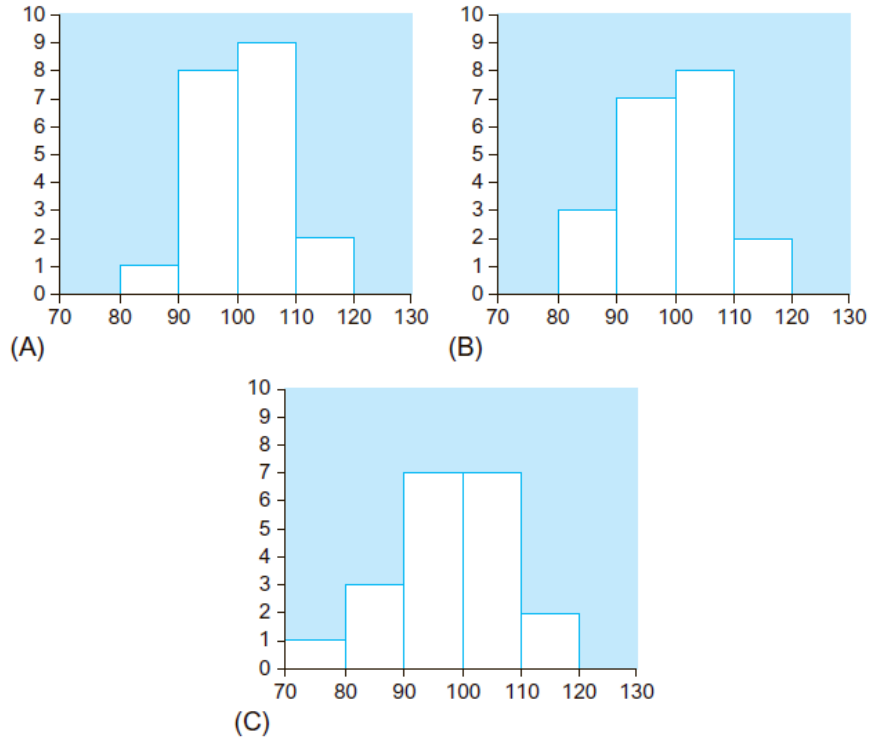
Чому нормальний розподіл відіграє таку важливу роль в статистиці? Зазвичай в статистиці припускають, що розподіл даних приблизно відповідає нормальному. Фахівці-статистики знають властивості нормального розподілу і використовують їх всякий раз, коли гістограма схожа на криву нормального розподілу.

В якому випадку можна сказати, що набір даних підпорядковується нормальному розподілу? Хороший спосіб полягає в тому, щоб уважно вивчити гістограму. На рис. 2.3.3 представлені гістограми для різних вибірок обсягом 100 значень кожна з нормально розподіленим набором даних. Цей малюнок демонструє, наскільки випадковою може бути форма розподілу при обмеженому розмірі вибірки.

Зменшення кількості даних призводить до збільшення випадковості, оскільки немає достатньо інформації для подання повної картини розподілу. Це наочно показано на рис. 2.3.4, де наведені гістограми для вибірок обсягом 20 значень кожна з нормально розподіленим набором даних.



Мал. 2.3.3. Гістограми для даних, витягнутих з нормально розподіленого набору. Обсяг кожної вибірки дорівнює 100. Порівняння цих трьох гістограм демонструє, який ступінь випадковості можна очікувати

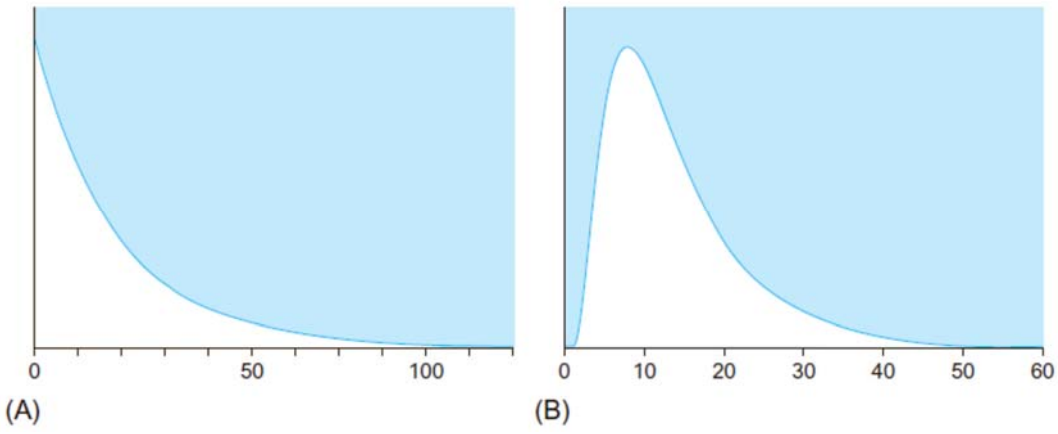


Мал. 2.3.4. Гістограми для даних, витягнутих з нормально розподіленого набору. Обсяг кожної вибірки дорівнює 20. Порівняння цих трьох гістограм демонструє, який ступінь випадковості можна очікувати

Чи справді в реальному житті все набори даних підкоряються нормальному розподілу? Звичайно, ні. Використовуючи гістограму, важливо визначити, чи є дані нормально розподіленими. Це особливо важливо, якщо подальший аналіз передбачає використання стандартних статистичних процедур, які вимагають нормального розподілу даних. У наступному розділі ми розглянемо один вид відмінності даних від нормального розподілу і запропонуємо спосіб впоратися з цією проблемою.

2.4. Несиметричні розподіли і перетворення даних

Несиметричний (скошений) розподіл не є ні симетричним, ні нормальним, оскільки значення даних на одній стороні кривої загасають швидше, ніж на інший. У реальному житті часто можна зустріти асиметрію в наборах даних, які відображають величини, виражені позитивними числами (наприклад, обсяги продажів або розміри активів). Це пов'язано з тим, що такі дані не можуть приймати негативні значення (наявність кордону з одного боку) і значення не обмежені зверху. В результаті на гістограмі багато значень даних сконцентровано близько нуля, і кількість значень стає все менше і менше при русі по горизонтальній осі гістограми вправо. На рис. 2.4.1 міститься кілька прикладів теоретичних кривих несиметричних розподілів.



Мал. 4.4.1. Приклади згладжених ідеалізованих кривих несиметричних розподілів. Реальні асиметрично розподілені набори даних мають деякі випадкові відхилення від таких ідеально гладких кривих

Приклад. Активи комерційних банків

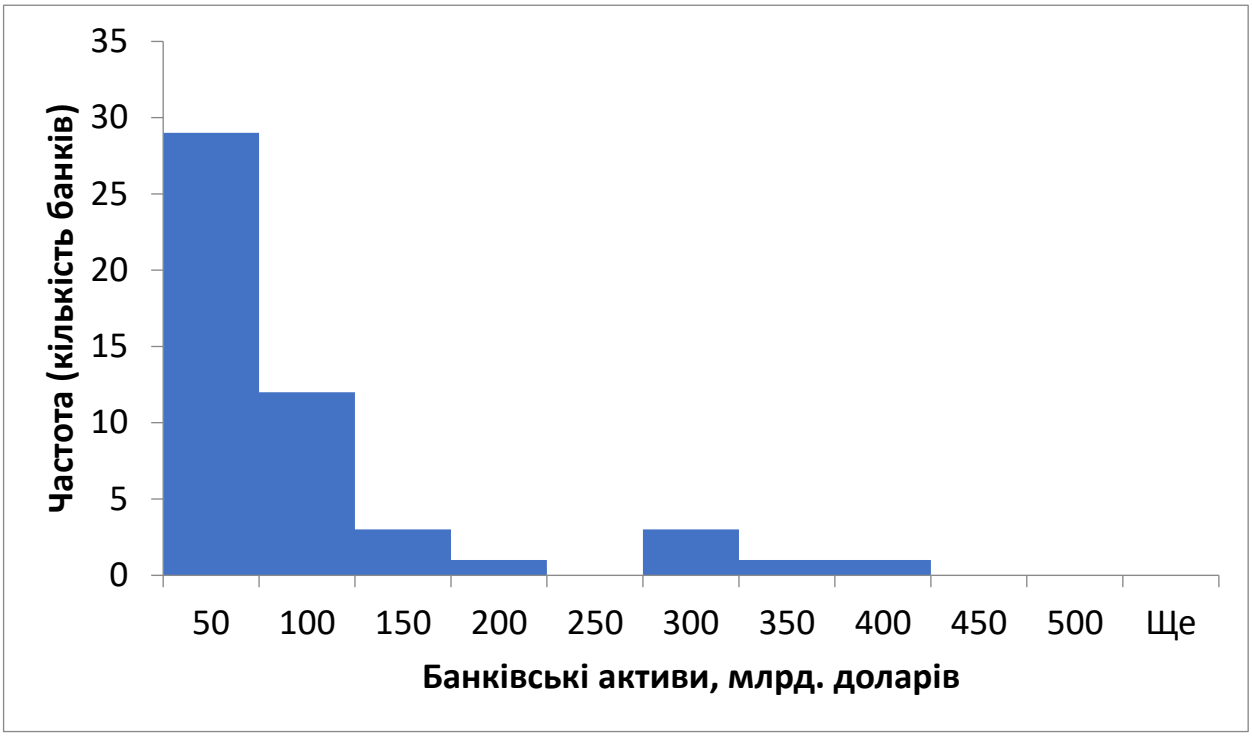
У табл. 2.4.1 містяться дані про активи комерційних банків зі списку Fortune 1000. Ці дані представляють хороший приклад сильно несиметричного (сильно скошеного) розподілу.

Таблиця 2.4.1. Активи комерційних банків з Fortune 1000

Банк	**Активи, млрд дол.**
Chase Manhattan Corp.	366
Citicorp	311
NationsBank Corp.	265
J. P. Morgan & Co	262
BankAmerica Corp.	260
First Union Corp.	157
Bankers Trust New York Corp.	140
Bank One Corp.	116
First Chicago NBD Corp.	114
Wells Fargo & Co.	97
Norwest Corp.	89
Fleet Financial Group	86
PNC Bank Corp.	75
KeyCorp	74
U. S. Bancorp	71
BankBoston Corp.	69
Wachovia Corp.	65
Bank of New York Co.	60
Sun Trust Banks	58
Republic New York Corp.	56
National City Corp.	55
CoreStates Financial Corp.	48
Barnett Banks	47
Mellon Bank Corp.	45
State Street Corp.	38
Comerica	36
South Trust Corp.	31
Summit Bancorp	30
Mercantile Bancorp.	30
BB & T Corp.	29
Huntington Bancshares	27
Northern Trust Corp.	25
Crestar Financial Corp.	23
Regions Financial	23
Fifth Third Bancorp	21
MBNA	21
First of America Bank Corp.	21
Firstar Corp.	20
Marshall & llsley Corp.	19
Popular	19

Amsouth Planters Corp.	18
First Security Corp.	17
Pacific Century Financial	15
First Tennessee National Corp.	14
First Empire State Corp.	14
Old Kent Financial Corp.	14
Compass Bancshares	13
Synovus Financial Corp.	9
First National of Nebraska	7
Providian Financial	4

На рис. 2.4.2 приведена гістограма цього набору даних. Через асиметрію розподіл даних не можна віднести до нормального.

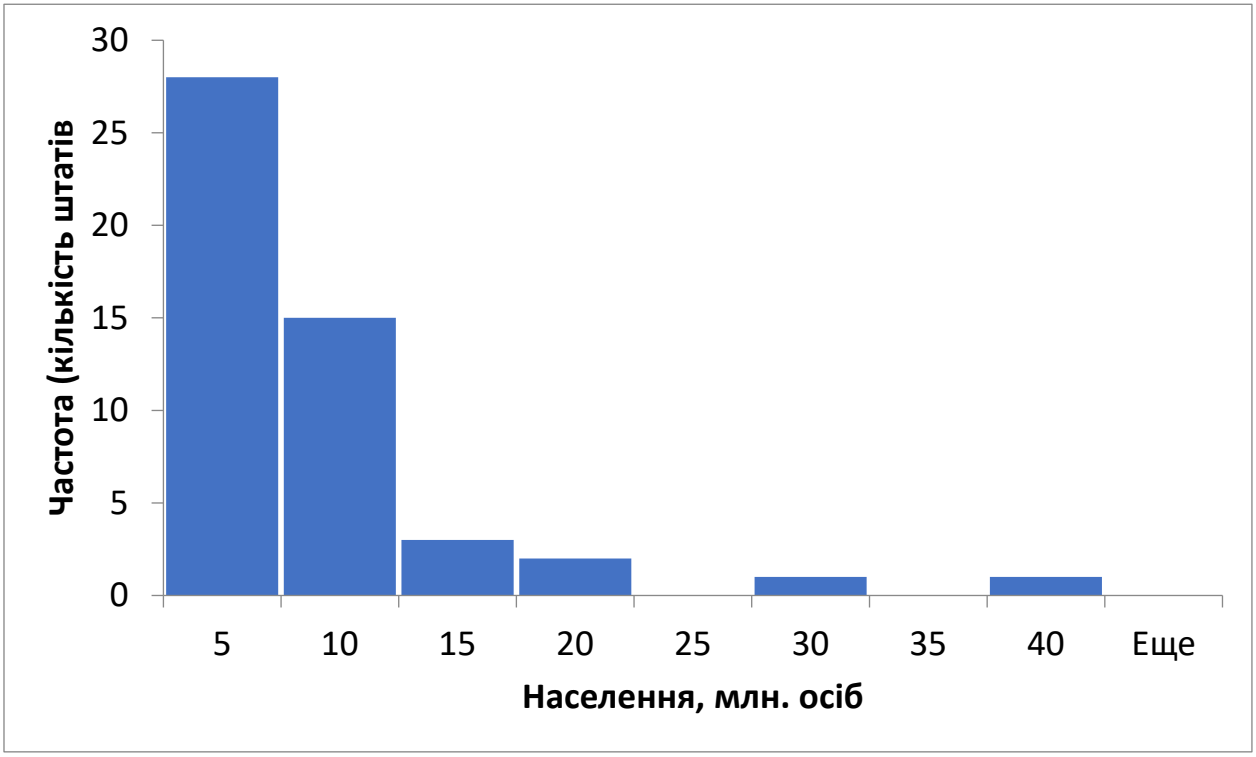


Мал. 2.4.2. Гістограма розмірів активів (в мільярдах доларів) комерційних банків за даними Fortune 1000. Це асиметричний (скошений), а не нормальний розподіл

Дуже високий стовпчик зліва на гістограмі представляє більшість з цих банків, які мають активи менше 50 мільярдів доларів. Кілька стовпчиків, розташованих правіше, представляють відносно невелику кількість більших банків. Дуже невеликий стовпчик праворуч на гістограмі представляє один банк – Chase Manhattan Corp, що має активи 366 мільярдів доларів.

Приклад. Чисельність населення штатів

Інший приклад несиметричного розподілу дають дані про чисельність населення окремих штатів в США, представлені у вигляді списку чисел. Асиметрія відображає той факт, що є багато штатів з невеликим або середнім розміром населення і всього кілька штатів з великим розміром населення (три найбільших штати: Каліфорнія, Техас і Нью-Йорк). Гістограма приведена на рис. 2.4.3.



Мал. 2.4.3. Гістограма чисельності населення штатів у 2020 році: несиметричний розподіл

Проблема з асиметрією

Одна з проблем, пов'язаних з асиметрією даних, полягає в тому, що більшість з найбільш поширених статистичних методів вимагають, щоб дані були, принаймні, приблизно нормально розподіленими. Якщо ці методи застосовують до несиметричних даних, то отриманий результат може бути неточним або просто невірним. І навіть якщо результати виходять в основному коректними, буде певна втрата ефективності аналізу, оскільки не

забезпечується найкраще використання всієї інформації, що міститься в наборі даних.

Вихід за допомогою перетворення

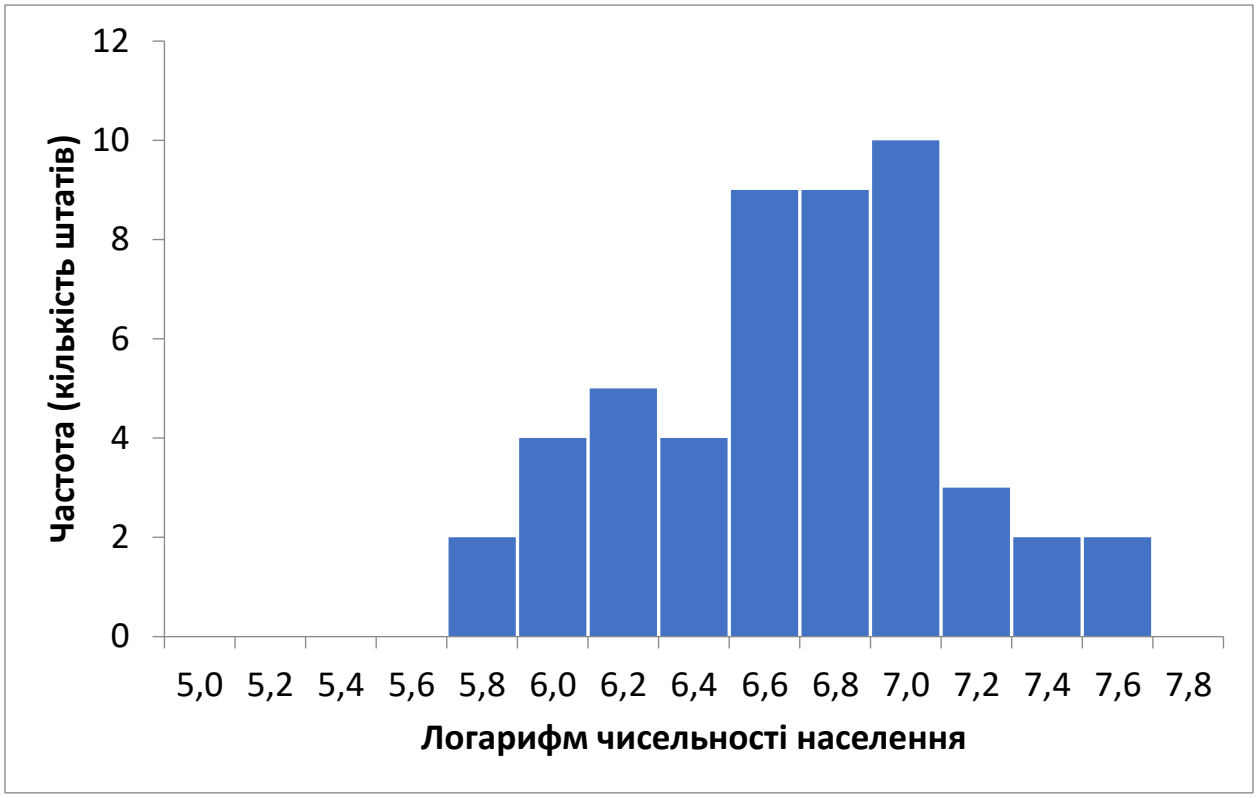
Один із способів впоратися з проблемою асиметрії полягає у використанні такого перетворення, яке переводить несиметричний розподіл в більш симетричний. Перетворення полягає в заміні кожного значення набору даних іншим числом (наприклад, логарифмом цього значення) з метою спростити статистичний аналіз. Найбільш поширеним типом перетворення даних в бізнесі та економіці є логарифмування, яке можна використовувати тільки для позитивних чисел (тобто для даних, які включають негативні значення або нуль, цей метод не підходить для нашого випадку). Логарифмування часто перетворює скошені (асиметричні) дані в симетричні, оскільки відбувається розтягування шкали біля нуля, що, в свою чергу, призводить до розподілу малих значень, згрупованих разом. У той же час логарифмування збирає разом великі значення, які розподілені на правому (позитивному) кінці шкали. Обидва типи найбільш часто використовуваних логарифмів («десятковий логарифм» по підставі 10 і «натуральний логарифм» по підставі «*e*») однаково ефективно можна використовувати для такого роду перетворень. У цьому розділі ми будемо використовувати десятковий логарифм.

Приклад. Перетворення даних про чисельність населення штатів

Порівнюючи гістограму чисельності населення на рис. 2.4.3 з гістограмою на рис. 2.4.4, побудованої для логарифмів тих же значень, можна побачити, що в результаті логарифмування асиметрія зникає. Хоча і в цьому випадку деякі результати випадають з загальної картини розподілу і крива не ідеально симетрична, більше немає різкого падіння на одній стороні і повільного зменшення значень на інший, як це має місце на рис. 2.4.3.

Логарифмічну шкалу можна інтерпретувати скоріше як мультиплікативну або процентну, ніж як адитивну. Як видно на рис. 2.4.4,

використання логарифмічною шкали призводить до того, що відстань по горизонталі 0,2 (ширина одного стовпчика) відповідає збільшенню (при русі зліва направо) населення на 58%. Відстань по горизонтальній осі, що дорівнює п’яти стовпчикам (наприклад, від точки 6,2 до точки 7,2), відповідає 10-кратному збільшенню чисельності населення штату. На початковій шкалі (відбиває фактичну чисельність населення в штатах) важко проводити порівняння відсотків. При русі зліва направо на рис. 2.4.3 перехід до кожного нового стовпчика означає збільшення населення на 5 мільйонів осіб, але в лівій частині малюнка ця різниця становить значно більший відсоток, ніж у правій.



Мал. 3.4.4. Перетворення може дозволити перейти від асиметрії до симетрії

Гістограма логарифмів значень чисельності населення штатів за 2020 рік в основному симетрична, за винятком випадкових відхилень. По суті, ніякої систематичної асиметрії не залишилося.

Інтерпретація і обчислення логарифма

Різниця на 1 для значень логарифма за основою 10 відповідає десятикратній різниці для вихідних значень. Наприклад, значення 392,1 і 3921

(співвідношення 1:10, різниця в 10 разів) після логарифмування перетворюються відповідно в значення 2,59 і 3,59 (різниця на 1). У табл. 2.4.2 містяться приклади декількох чисел і їх логарифмів.

З таблиці видно, як логарифм «стягує разом» дуже великі числа, зменшуючи різницю між ними і іншими значеннями в наборі даних (наприклад, замість різниці в 100 мільйонів отримуємо різницю в 8 одиниць). Крім того, зверніть увагу, що логарифм приблизно показує кількість розрядів в цілій частині числа. Наприклад, населення Каліфорнії становить 37691912 людини, а логарифм цього числа дорівнює 7,5762 (це відповідає стовпчику в правій частині рис. 2.4.3 і 2.4.4).

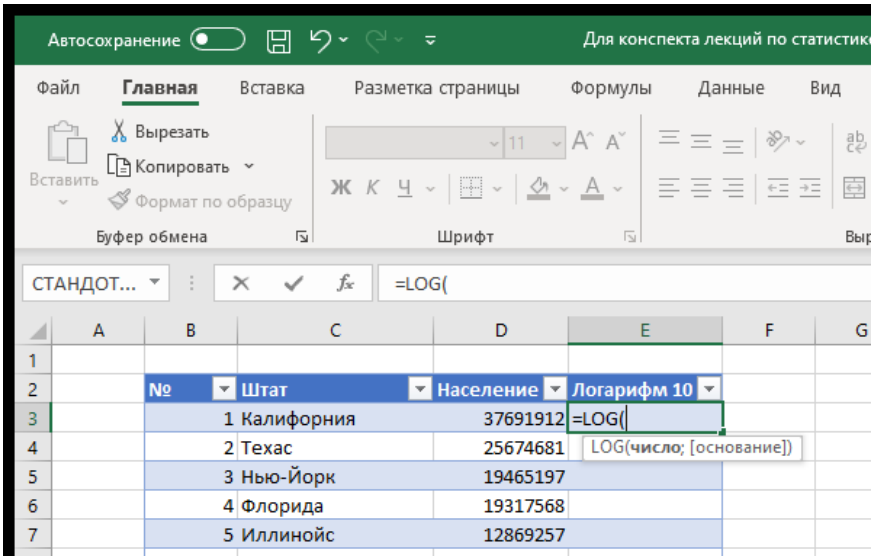
Найчастіше використовують логарифми двох видів. Ми розглянули логарифми по підставі 10. Логарифми другого виду називають натуральними, їх позначають 1π і обчислюють за основою числа *e=2,71828*... Натуральний логарифм часто використовують при обчисленні складних відсотків, темпів зростання, економічної еластичності та ін. В перетвореннях даних обидва види логарифмів призводять до однакового ефекту, тобто «стягують разом» на числовій осі великі числа і «розтягують» малі.

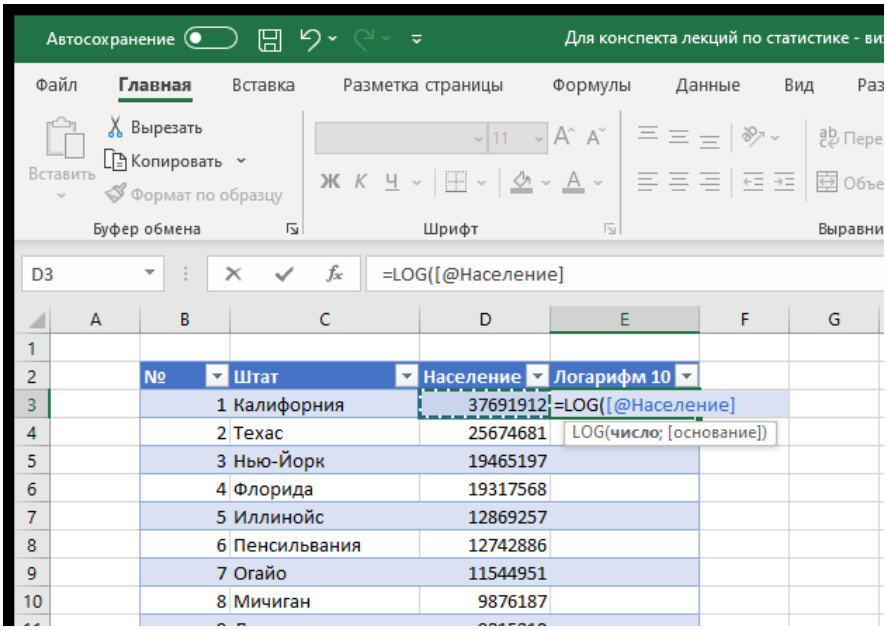
Таблиця 2.4.2. Результати логарифмування по підставі 10

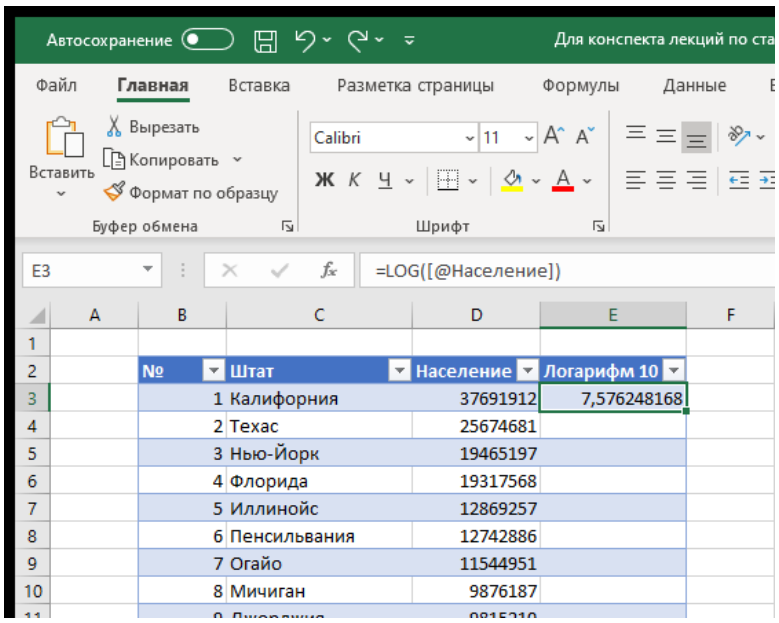
Число	**Логарифм**
0,001	-3
0,01	-2
0,1	-1
1	0
2	0,301
5	0,699
9	0,954
10	1
100	2
10 000	4
20 000	4,301
100 000 000	8

Багато електронних таблиць, наприклад Microsoft Excel, мають вбудовані функції логарифмування. Ви можете ввести = LOG (5) в комірку, щоб обчислити логарифм (по підставі 10) числа 5, рівний 0,69897. Якщо ж ви

введете = LN (5), то отримаєте логарифм числа 5 по підставі *e*, рівний 1,60944. Щоб знайти логарифм набору даних в колонці таблиці, можна за допомогою команд «Копіювати» і «Вставити» скопіювати формулу логарифму з першої комірки в усі комірки колонки, що істотно скорочує час обчислення логарифмів для ряду чисел. Ще більш швидкий спосіб створення колонки перетворених значень показаний нижче: слід після введення формули перетворення двічі клацнути на «швидкому заповненні» (маленький квадрат, розташований праворуч нижче виділеної комірки) або просто протягнути «швидке заповнення» мишею.







	A	B	C	D	E	F
1						
2		№	Штат	Население	Логарифм 10	
3		1	Калифорния	37691912	7,576248168	
4		2	Техас	25674681		
5		3	Нью-Йорк	19465197		
6		4	Флорида	19317568		
7		5	Иллинойс	12869257		
8		6	Пенсильвания	12742886		
9		7	Огайо	11544951		
10		8	Мичиган	9876187		

2.5. Бімодальний розподіл

Важливо вміти визначати, коли набір даних складається з двох або більш чітких груп, що розрізняються між собою, щоб можна було при необхідності аналізувати ці групи окремо. На гістограмі такій ситуації відповідає розрив між двома сусідніми групами стовпчиків. Якщо на гістограмі чітко видно дві окремі групи, то це говорить про бімодальний розподілі даних. Бімодальний розподіл – це розподіл, що має дві моди або два різних кластери (блоки) даних.

Наявність бімодального розподілу може свідчити про те, що ситуація більш складна, ніж ви припускали, і тому вимагає серйозної уваги. Щонайменше, слід виявити причини наявності двох груп. Можливо, інтерес представляє лише одна група, тому іншу групу можна виключити з аналізу. А може бути, вам необхідно вивчити обидві групи, але слід внести деякі уточнення, щоб врахувати факт наявної відмінності.

Приклад. Доходи валютного ринку

Розглянемо доходи валютного ринку як щорічні дивіденди, виражені в процентах. У табл. 2.5.1 представлена частина відповідного набору даних.

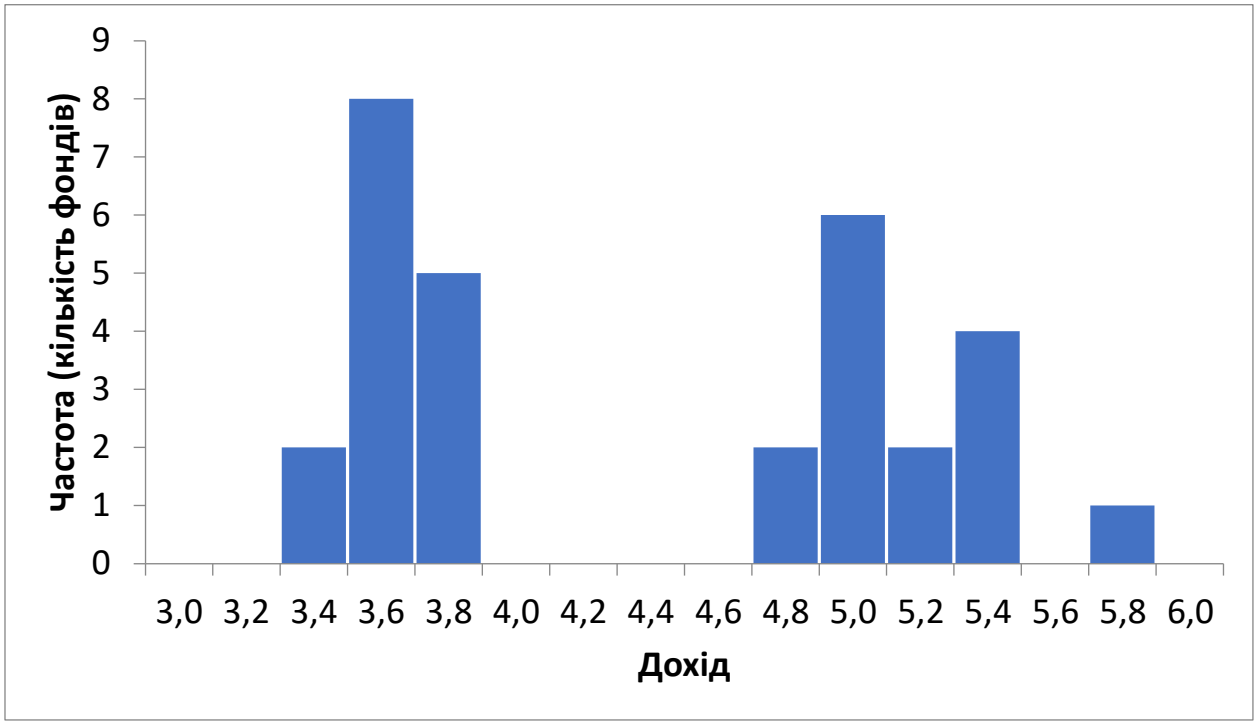
Гістограма повного набору даних, показана на рис. 2.5.1, виглядає як дві окремі гістограми. Одна група містить фонди з доходами від 3 до 4%, а інша –

від 6 до 8%. Малоймовірно, що такий поділ обумовлено простою випадковістю для одного однорідного набору даних. Мабуть, існує інша причина.

Таблиця 3.5.1. Доходи взаємних фондів на валютному ринку

(частина списку)

Фонд	Дохід за 7 днів,%	фонд	Дохід за 7 днів, %
Putnam MMA	5,3	DryMATR	3,71
QualivestA	4,86	DryCTMu	3,42
QualivestGvA	4,88	DreyCaUx	3,31
QualivestY	5,25	DrMI Mun	3,55
QftvCsh	4,81	DrNJMun	3.43
QuestCshGov	4,75	DrNYTE	3,44
QuestCshPr	4,85	DRPAMun	3,70
RNCUq	4,82	DreyTxExC	3,63
RegisDSI	5,61	DryMuResR	3,44
RemTreasTr	5,01	DryMAMun	3,39
RemGovTr	5,37	DryOHMu	3,76
RembMMTr	5,40	DryCATF	3,44
ResrevUSTrs	4,64	DryMATF	3,46
ResrveFdGvt	4,99	EatnVn	3,44
ReserveFd	5,01	EvrgnTEA	3,71



Мал. 3.5.1. Доходи валютного ринку

Це бімодальний розподіл з двома чіткими окремими групами, що, мабуть, не може бути пояснено тільки випадковістю

Приклад. Вартість одного дня перебування в лікарні

Розглянемо вартість одного дня перебування в місцевій лікарні в різних штатах (табл. 2.5.2).

Відзначимо значне варіювання вартості в різних штатах: вартість одного дня перебування в лікарні Каліфорнії майже вдвічі вище ($1134), ніж в лікарні Південної Дакоти ($457). Щоб уявити повну картину, подивіться на гістограму цього набору даних на рис. 2.5.2.

Це майже симетричний розподіл (тобто він не ідеально симетричний, але принаймні не є сильно асиметричним). Розподіл в основному нормальний, дані містять тільки одну групу значень.

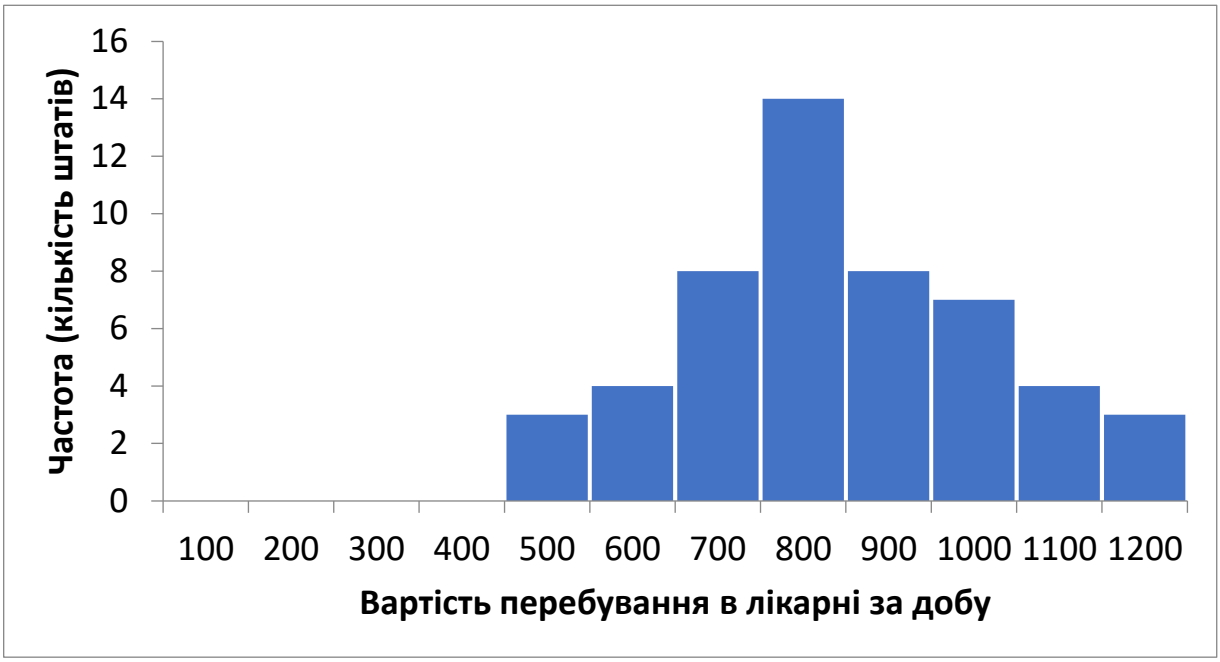
Однак якщо накреслити ту ж гістограму в зменшеному масштабі, з вужчими стовпчиками (рис. 2.5.3), то додатково в даних виявляться дві групи значень: 5 штатів з більш низькою вартістю перебування в лікарні (зліва) і інші штати (праворуч), а може бути, навіть і три групи.

Однак цей розподіл не є дійсно бімодальним з двох причин. По-перше, розрив малий у порівнянні з розкидом значень вартості в різних штатах. По-друге, і це більш важливо, стовпчики гістограми занадто малі, тому що багато хто з них представляють лише один штат. Пам’ятайте, що одна з основних цілей статистичних методів (таких як гістограма) полягає у виявленні загальної картини, а не окремих деталей.

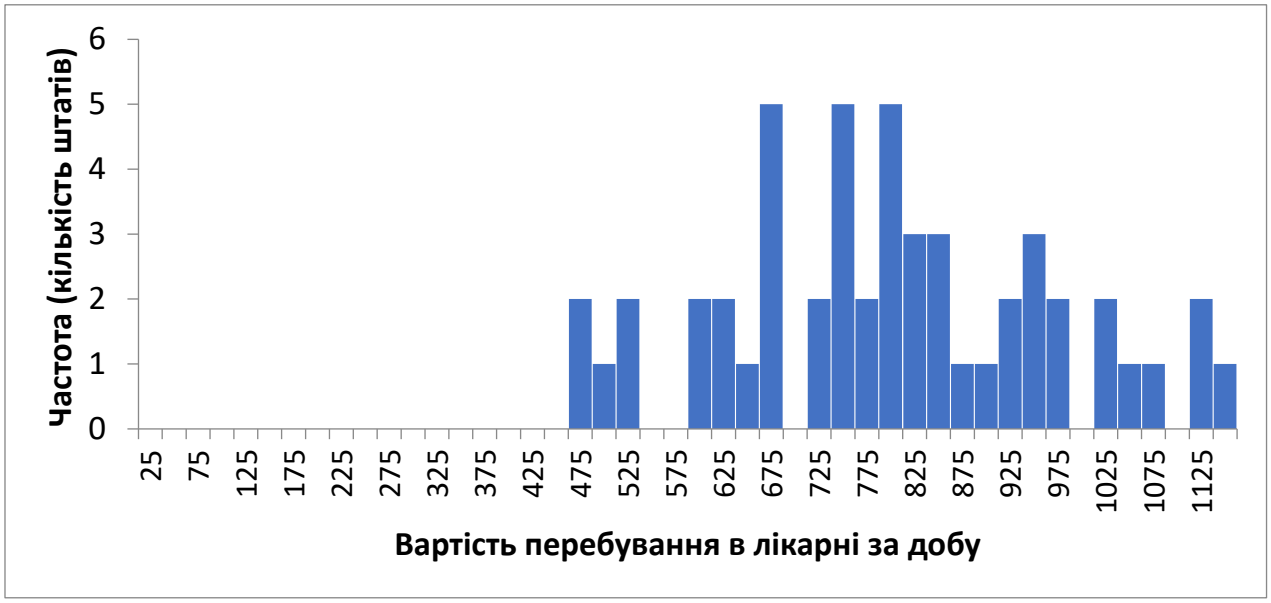
Таблиця 2.5.2. Середня вартість одного дня перебування в місцевій лікарні одного пацієнта (в доларах)

Alabama	729	Kentucky	674	North Dakota	484
Alaska	1 116	Louisiana	836	Ohio	875
Arizona	1 051	Maine	674	Oklahoma	740
Arkansas	633	Maryland	806	Oregon	1 011
California	1 134	Massachusetts	937	Pennsylvania	793
Colorado	904	Michigan	847	Rhode Island	801
Connecticut	1 012	Minnesota	618	South Carolina	782
Delaware	920	Mississippi	516	South Dakota	457
Dist. of Columbia	1 124	Missouri	792	Tennessee	796
Florida	886	Montana	474	Texas	933
Georgia	721	Nebraska	600	Utah	1 036
Hawaii	761	Nevada	952	Vermont	726

Idano	618	New Hampshire	776	Virginia	774
Illinois	849	New Jersey	737	Washington	974
Indiana	822	New Mexico	950	West 'Virginia	655
Iowa	588	New York	744	Wisconsin	674
Kansas	661	North California	711	Wyoming	515



Мал. 2.5.2. Вартість одного дня перебування в муніципальній лікарні в різних штатах. Це майже нормальний розподіл, який утворює одну цілісну групу



Мал. 2.5.3. Вартість одного дня перебування в муніципальній лікарні в різних штатах (ті ж дані, що і на попередній гістограмі, але тут використовуються більш вузькі стовпчики)

Оскільки на цій гістограмі представлено більше деталей, створюється враження (можливо, невірне), що в наборі даних присутні дві або навіть три групи. П’ять штатів зліва з найменшою вартістю трохи відокремлені від інших, як і три штати справа з найбільшою вартістю. Можливо, це просто випадковість і не є дійсною бімодальністю.

2.6. Викиди (значення, що сильно відхиляються)

Іноді в даних можна спостерігати викиди (значення, що сильно відхиляються), тобто такі значення, які, мабуть, не належать даному розподілу, оскільки вони або занадто великі, або занадто малі. Залежно від причин виникнення викидів, проблему вирішують по-різному. Існують два види викидів: помилки і коректні значення, які суттєво «відрізняються» від основних даних. Оскільки про викиди часто говорять при аналізі гістограм, ми їх обговоримо у цьому розділі. А в наступній лекції буде викладено формальний метод обчислень для визначення викидів (побудова блокової діаграми).

Робота з викидами (значеннями, що сильно відхиляються)

З помилками впоратися легко – потрібно просто відкоригувати значення. Наприклад, якщо значення, що відповідає об'єму продажів $1597,00, записано як 159700 через неправильно поставлену кому, то це значення буде сильно відрізнятися від інших значень на гістограмі. Побачивши таке дивне значення, потрібно перевірити ще раз дані і знайти помилку. Виправивши це значення на 1597, ви вирішите проблему.

На жаль, важче вирішувати проблеми викидів коректних даних. Якщо є переконливе підтвердження того, що викиди не відповідають тому, що вивчається, то їх можна просто видалити і аналізувати більш узгоджені між собою дані, що залишилися. Наприклад, в наборі даних щодо доходів грошового ринку може з'явитися кілька значень доходів фондів, не оподатковуваних податком. Якщо мета дослідження полягає в аналізі ринкової

ситуації для звичайних фондів, то ці викиди краще виключити із загальної картини. Як інший приклад припустимо, що ваша компанія оцінює новий лікарський препарат. В одному з дослідів лаборант чхнув в зразок перед його аналізом. Якщо ви не вивчаєте нещасні випадки з лабораторними матеріалами, то цей зразок годі й аналізувати.

Якщо ви вирішили не враховувати деякі викиди, ви повинні бути готові до того, що в правильності цього рішення потрібно переконати не тільки себе, але і того, кому призначений ваш звіт (хоча ця людина може мати й іншу думку). Таким чином, на питання, враховувати або не враховувати викиди, немає однозначної і єдиної вірної відповіді. Наприклад, для спрощення первинного внутрішньофірмового аналізу можна виключити деякі викиди. Однак, якщо дослідження призначене для громадськості або є державним дослідженням, то слід дуже обережно і з усією відповідальністю поставитися до виключення викидів.

При відсутності досить обґрунтованого аргументу для виключення викидів, як компроміс, можна виконати два різних аналізи: один з урахуванням викидів, а інший – з їх виключенням. Тоді ваш звіт буде містити всі результати. У кращому випадку, якщо результати обох аналізів будуть однаковими, тоді можна буде зробити висновок, що наявність викидів не має істотного значення. У більш складному випадку, коли ці два аналізи дадуть різні результати, ваші висновки і рекомендації будуть менш визначеними і однозначними. На жаль, немає вичерпного вирішення цієї досить тонкої проблеми.

При виключенні викидів з аналізу рекомендується керуватися одним важливим правилом, яке допоможе захистити себе від можливих звинувачень:

ПРИ ВИКЛЮЧЕННЯ ВИКИДІВ З АНАЛІЗУ
ЗАВЖДИ ПОЯСНЮЙТЕ, ЩО ВИ ЗРОБИЛИ І ЧОМУ!

Іншими словами, чітко поясніть в звіті (може бути достатньо виноски), що ваші дані містять викиди (значення, що сильно відхиляються). Опишіть ці значення. Поясніть і обґрунтуйте зроблені вами дії.

Чому проблеми з викидами потрібно обов’язково вирішувати? Є дві причини, за якими наявність викидів може призводити до проблем при аналізі даних. По-перше, важко інтерпретувати подробиці структури набору даних, якщо одне значення домінує в загальній картині і тому привертає до себе підвищену увагу. По-друге, як і в випадку асиметрії, багато з поширених сучасних статистичних методів не можна використовувати для аналізу тих даних, розподіл яких сильно відрізняється від нормального. Нормальний розподіл є симетричним і зазвичай не містить викиди. Отже, перш ніж зайнятися серйозними статистичними висновками, вам доведеться розібратися з викидами в даних.

Приклад. Зміни в витратах на телевізійну рекламу

У рекламному бізнесі, як і в більшості сфер економічної діяльності, час від часу відбуваються зміни. У табл. 2.6.1 наведені процентні зміни загальних витрат на телевізійну рекламу в 1994 році в порівнянні з 1993 роком для 25 найбільших рекламодавців 1994 року.

На гістограмі з рис. 2.6.1 є наявний викид (зміна витрат для Regal Communications складає 2353,7% – колосальне число, яке відповідає збільшенню витрат з 1,1 до 25,8 мільйона доларів) призвело до того, що всі інші компанії виявилися зведені в один стовпчик, який відповідає обсягу збільшення витрат рекламодавців десь між 0 і 500%. Винятком є одна компанія, яка збільшила витрати більш ніж на 500%, і одна компанія, яка зменшила витрати.

Зміна витрат компанії Regal Communications закриває картину розподілу процентних змін витрат інших підприємств (більшість значень лежать в межах від 0 до 100%). Навіть побудувавши гістограму з вузьких стовпчиків (рис. 2.6.2), можна побачити детальну картину. На жаль, гістограма в даному випадку не дуже корисна.

Таблиця 2.6.1. Зміна загальних витрат на телевізійну рекламу в 1994 році в порівнянні з 1993 роком

Рекламодавець	Зміна витрат на рекламу,%	Рекламодавець	Зміна витрат на рекламу,%
Procter & Gamble	43,2	Warner-Lambert	-22,7
Phillip Morris	27,5	AT & T	73,5
Kellogg	77,9	Grand Metropolitan	14,0
Time Warner	201,0	Johnson & Johnson	16,5
Unilever	16,7	National Education	217,3
Hasbro	54,5	Nestle	31,4
Mattel	47,7	Hershey	42,4
American Home Products	104,4	Regal Communications	2 353,7
General Motors	65,7	McDonald's	28,5
Wrigley	66,8	Sara Lee	16,4
Mars	33,3	Himmel Group	684,0
RJR Nabisco	65,9	Bayer Group	12,7
Sears, Roebuck	44,7		

Виняток компанії Regal Communications, яка, очевидно, представляє викид даних з найбільшим значенням (2353,7%), має на меті отримати більш структуровану картину щодо інших рекламодавців. Однак, як видно з рис. 2.6.3, більшість деталей все ще приховано, але на цей раз іншим викидом, рівним 684% (компанія Himmel Group).

Після виключення обох викидів можна нарешті побачити, що розподіл змін витрат на телерекламу для решти рекламодавців є приблизно нормальним, з центром близько 40% (рис. 2.6.4), але, можливо, з ще двома значеннями, що сильно відхиляються (викидами) – близько 200% (компанії Time Warner і National Education).

На основі цього аналізу змін витрат на телевізійну рекламу можна дати наступну оцінку ситуації.

Мал. 2.6.1. Гістограма процентного збільшення об'єднаних витрат на телевізійну рекламу.

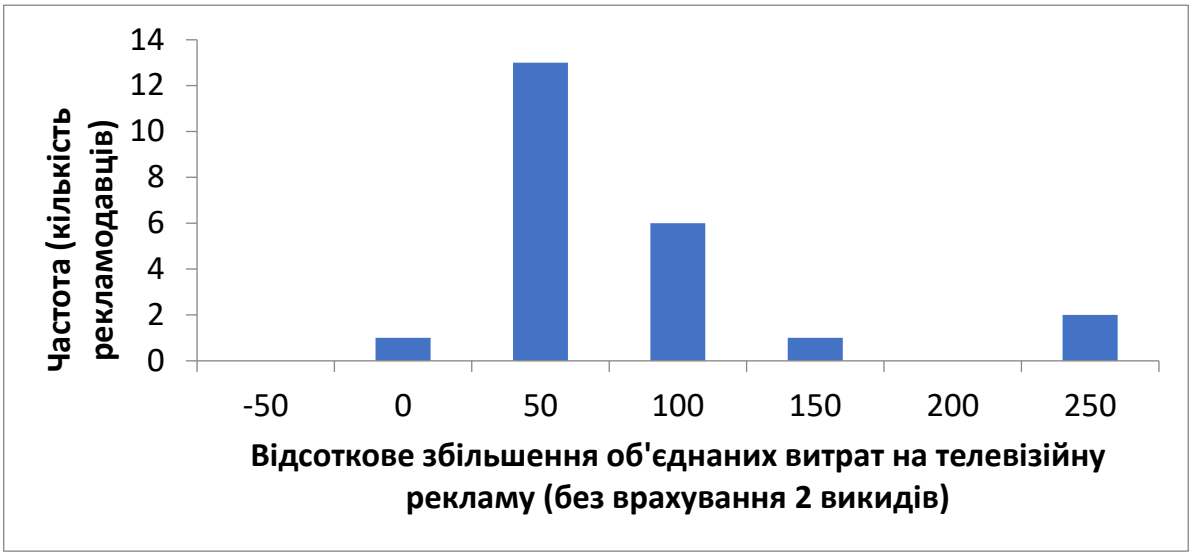
Зверніть увагу на наявність викиду в правій частині гістограми (компанія Regal Communications – 2353,7%), який затьмарює подробиці, пов’язані з більшістю інших рекламодавців, і, по суті, зводить майже всі компанії в один стовпчик – від 0 до 500%.

Мал. 2.6.2. Інша гістограма для даних про всіх 25 найбільших рекламодавців, але з більш вузькими стовпчиками.

Викид справа на гістограмі, як і раніше, приховує більшість даних, хоча тепер ми ясно бачимо, що значення лежать в основному в межах від 0 до 100%.

Мал. 2.6.3. Гістограма змін витрат 24 рекламодавців після виключення значення 2353,7 % – найбільшого викиду (компанія Regal Communications) і збільшення масштабу по горизонталі для отримання більш докладної картини.

Тепер видно другий викид – 684% (компанія Himmel Group), який продовжує приховувати деталі більшої частини набору даних



Мал. 2.6.4. Гістограма змін витрат 23 рекламодавців після виключення двох найбільших викидів значень (компанії Regal і Himmel).

Тепер видно окремі подробиці, які дають загальну ка ртину розподілу змін витрат (можливо, з двома новими викидами праворуч).

Дві компанії мають виключно велике збільшення витрат, а саме Regal Communications (2 353,7%) і Himmel Group (684%). Решта компаній дають типове збільшення витрат від 0 до 75% (можливо, трохи більше або менше), за винятком двох компаній з високим, близько 200%, зростанням витрат.

Дані цього аналізу свідчать про те, що витрати на рекламу сильно змінюються щороку. Великі рекламодавці не мають постійної стійкої стратегії, яка лише трохи коригується щороку. Більшість з 25 провідних рекламодавців для телебачення, мабуть, виявилися в цьому списку завдяки значному збільшенню своїх витрат на рекламу в порівнянні з попереднім роком.

2.7. Додатковий матеріал

Резюме

Набір даних найпростішого виду являє собою список чисел, що містять деяку інформацію (єдина статистична змінна), виміряну для кожного досліджуваного об'єкта (кожної елементарної одиниці). Такий список чисел може бути представлений у вигляді списку, або у вигляді таблиці, де записано, скільки разів кожне з значень повторюється в списку.

Першим кроком в аналізі списку чисел є вивчення гістограми, яка дає уявлення про основні властивості набору даних, таких як типові значення, особливі значення, концентрація, розподіл значень, характер даних і наявність в даних окремих груп значень. Гістограма частоти у вигляді стовпчикової діаграми, розташованої над числовою віссю і яка б показала, скільки раз різні значення зустрічаються в наборі даних. Числова вісь являє собою пряму лінію, зазвичай горизонтальну, з нанесеними під нею числами, що утворюють шкалу.

Нормальний розподіл являє собою теоретичну гладку гістограму в формі дзвона без випадкових відхилень. Їй відповідає ідеальний набір даних, в якому більшість значень сконцентровано в середній частині діапазону, а значення, що залишилися, симетрично з загасанням частоти розташовані по

обидва боки від вершини дзвона. Набір даних має нормальний розподіл, якщо форма його гістограми близька до ідеальної гладкої кривої в формі дзвона, можливо, з деякими випадковими відхиленнями. Нормальний розподіл відіграє важливу роль в теорії та практиці статистичного аналізу.

Асиметричний (скошений) розподіл не є ні симетричним, ні нормальним, оскільки значення даних з одного боку загасають більш різко, ніж з іншого. Асиметричні розподіли дуже часто зустрічаються в бізнесі. На жаль, більшість статистичних методів не застосовні до сильно скошених розподілів.

Перетворення полягає в заміні кожного значення іншим числом (наприклад, логарифмом цього значення) з метою спрощення статистичного аналізу. Логарифмування часто перетворює асиметрію в симетрію, оскільки дозволяє розтягнути шкалу в околиці нуля, розтягуючи за шкалою всі згруповані разом малі значення. Логарифмування також групує великі значення, розтягнуті на правому кінці вихідної шкали. Логарифмувати можна тільки позитивні числа. Для правильної інтерпретації результату логарифмування необхідно враховувати, що рівним відстаням на логарифмічній шкалі відповідають на вихідній шкалі рівні процентні збільшення, а не рівні збільшення значень (як, наприклад, обсяг фінансів в доларах).

Якщо на гістограмі чітко видно дві окремі групи, то це говорить про бімодальний розподіл даних. Важливо вміти визначати наявність бімодального розподілу, щоб робити відповідні дії при аналізі. Можливо, з’ясується, що вас цікавить тільки одна з цих груп даних, а другу можна не розглядати. Можливо, доведеться вносити в аналіз певні зміни, щоб впоратися з цією більш складною ситуацією.

Іноді дані можуть містити викиди (значення, що сильно відхиляються), тобто одне або кілька таких значень, які, мабуть, не належать даному розподілу, оскільки або занадто великі, або занадто малі. Викиди ускладнюють статистичний аналіз, тому їх слід ідентифікувати і обробити додатково. Якщо

викид являє собою просто помилку, то її слід виправити і продовжити аналіз. Якщо помилки немає, а значення сильно відрізняється від інших значень з набору даних, то цей викид можна або виключити, або не виключити з аналізу. Якщо ви переконаєте себе та інших, що викид не належить досліджуваній системі даних, його можна виключити. Якщо ви не можете обґрунтувати виняток викиду, може доведеться зробити два аналізи: з викидом і без нього. У будь-якому випадку в звіті вам необхідно чітко написати про наявність викиду і вжиті заходи.

Найкраще будувати гістограми за допомогою комп'ютерних пакетів програм статистичного аналізу.

Основні терміни

– Послідовність чисел (list of numbers)
– Числова вісь (number line)
– Гістограма (histogram)
– Нормальний розподіл (normal distribution)
– Несиметричний (скошений) розподіл (skewed distribution)
– Перетворення (transformation)
– Логарифм (logarithm)
– Бімодальний розподіл (bimodal distribution)
– Викид (outlier)

Контрольні питання

1. Що таке список чисел?
2. Назвіть шість властивостей набору даних, які можна побачити на гістограмі.
3. Що таке числова вісь?
4. Чим відрізняється гістограма від стовпчикової діаграми?
5. Що таке нормальний розподіл?
6. Чому нормальний розподіл відіграє важливу роль в статистиці?

7. Якщо реальний набір даних розподілено нормально, чи можна очікувати, що гістограма буде мати ідеально гладку форму у вигляді дзвону? Обґрунтуйте свою відповідь.

8. Всі набори даних підкоряються нормальному розподілу чи ні?

9. Що таке асиметричний розподіл?

10. У чому основна проблема асиметрії? Як цю проблему можна вирішити в багатьох випадках?

11. Як ви можете проінтерпретувати логарифм числа?

12. Що таке бімодальний розподіл? Що слід зробити у разі бімодального розподілу?

13. Що таке викид?

14. Чому важливо описати в звіті, які дії робилися щодо викидів?

15. Які проблеми виникають при наявності викидів значень?

16. В яких випадках викиди можна не враховувати і аналізувати тільки інші дані?

17. Припустимо, що в вашому наборі даних є викиди. Ви плануєте проаналізувати дані двічі: з викидами і без них. Який результат вас більше влаштує? Чому?