ЛЕКЦІЯ.
УЗАГАЛЬНЮЮЧІ ПОКАЗНИКИ

У складних ситуаціях один з найефективніших способів «побачити всю картину» полягає в узагальненні, тобто використанні одного або декількох відібраних або розрахованих значень для характеристики набору даних. Докладне вивчення кожного окремого випадку саме по собі не є статистичною діяльністю, але виявлення та ідентифікація особливостей, які в цілому характерні для розглянутих випадків, є такою, оскільки вся інформація при цьому розглядається в цілому.

Одна з цілей статистики полягає в тому, щоб звести набір даних до одного числа (або двох, або декількох), яке виражає найбільш фундаментальні властивості даних. Методи, які найбільше підходять для аналізу одного списку чисел (тобто одновимірного набору даних), включають визначення таких показників.

Середнє, медіана і мода – це різні способи вибору єдиного числа, яке краще всього описує всі числа в наборі даних. Такий показник, представлений одним числом, називається типовим значенням, або центром (також використовують термін міра центральної тенденції).

Перцентіль (також використовують термін процентіль). Узагальнює інформацію про ранги, характеризуючи значення, що досягається заданим відсотком загальної кількості даних, після того, як дані впорядковуються (ранжуються) по зростанню.

Стандартне відхилення – характеристика відмінностей між значеннями в наборі даних. Це поняття також називають розкидом, або мінливістю.

Як бути, якщо набір даних містить окремі значення, які неадекватно описуються цими показниками? Такі викиди (значення, що сильно відхиляються) можна просто описати окремо. Таким чином, можна надати характеристику великому набору даних, узагальнивши основні властивості більшості його елементів і потім створивши список винятків. Це дозволяє

досягти статистичної мети ефективного опису великого набору даних з урахуванням особливої природи окремих елементів.

3.1. Чому дорівнює найбільш типове значення?

Найпростіше узагальнення будь-якого набору даних є єдине число, яке найкращим чином представляє всі значення даних. Таке число можна було б назвати типовим значенням для даного набору даних. Якщо не всі значення в наборі даних однакові, то думки про «найбільш типове» можуть бути різними. Існують три види такої узагальнюючої міри.

1. Середнє, яке можна обчислювати тільки для чисел, що мають змістовний сенс (для кількісних даних).

2. Медіана, або серединна точка, яку можна обчислювати як для впорядкованих категорій (порядкові дані), так і для чисел.

3. Мода, або категорія, що найбільш часто зустрічається, яку можна визначити для неупорядкованих категорій (для номінальних даних), для впорядкованих категорій і для чисел.

Середнє: типове значення для кількісних даних

Середнє найчастіше використовують як типове значення списку чисел і визначається шляхом складання всіх чисел списку і ділення отриманої суми на кількість чисел в списку (кількість елементарних одиниць). Формула обчислення вибіркового середнього (тобто середнього вибірки даних) має наступний вигляд:

$$\text{Вибіркова середнє} = \frac{\text{Сума значень елементів даних}}{\text{Кількість елементів даних}},$$

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

де n – загальне число елементів в списку даних, X_1, X_2, ... X_n – безпосередньо самі значення даних. Грецька прописна буква *сигма*, $\sum$, вказує на необхідність скласти всі значення, які записані за нею, замінюючи

при цьому індекс i значеннями від 1 до n. Символ для запису середнього $\overline{X}$ вимовляється як «ікс з рискою».

Наприклад, середнє набору даних з трьох чисел, 4, 9, 8, дорівнює

$$\frac{4+9+8}{3} = \frac{21}{3} = 7.$$

Поняття середнього не залежить від того, представляє ваш список чисел всю генеральну сукупність або ж репрезентативну вибірку з більшої сукупності. У той же час позначення не однакові. Для всієї генеральної сукупності кількість елементів позначають буквою *N*, а середнє – буквою μ (грецька буква «мю»). Процес обчислення середнього однаковий як для генеральної сукупності, так і для вибірки.

Оскільки при обчисленні середнього значення дані підсумовують, ясно, що середнє не можна визначити для якісних даних (не можна складати кольори або рейтинги боргових зобов'язань).

Середнє можна інтерпретувати як рівномірний розподіл суми всіх значень між елементарними одиницями. Таким чином, якщо кожне значення даних замінити середнім, то загальна сума не зміниться. Наприклад, з бази даних працівників можна обчислити середню заробітну плату співробітників у Х'юстоні. Це середнє можна інтерпретувати таким чином: якби ми виплачували всім службовцям Х'юстона однакову заробітну плату, не змінюючи при цьому загальний фонд заробітної плати, то значення цієї заробітної плати дорівнювало б середньому. Зверніть увагу, що не слід розглядати структуру рівня заробітної плати, яка отримана виходячи з середнього, як індикатор типової заробітної плати (особливо, коли ви маєте справу з фондом заробітної плати як частини бюджету).

Оскільки середнє зберігає незмінною суму при рівномірному розподіленні значень, воно найбільш корисне в якості узагальнюючого показника при відсутності екстремальних значень (викидів), коли набір даних представляє собою більш-менш однорідну групу з елементами випадковості. Якщо один працівник заробляє набагато більше інших, то середнє не можна

використовувати в якості узагальнюючого показника. Хоча середнє і зберігає незмінною загальну суму заробітної плати, воно не буде належним показником величини заробітної плати окремих службовців, оскільки середнє буде занадто високим для більшості працівників і занадто низьким для цього високооплачуваного працівника.

Середнє є лише узагальнюючою характеристикою, яка зберігає загальну суму. Ця властивість середнього особливо корисна в тих ситуаціях, коли необхідно планувати загальну суму для великої групи. Спочатку обчислюють середнє для меншої вибірки даних, що представляє велику групу. Потім отримане середнє можна помножити на кількість окремих елементів в цій більшій групі. В результаті отримують оцінку або прогноз суми для більшої за розміром сукупності. В цілому, якщо необхідно визначити загальну суму, можна використовувати середнє.

Приклад. Скільки грошей витратять споживачі?

Фірма цікавиться, скільки в цілому витрачають на медичні товари мешканці Клівленду. Аналіз випадкової вибірки з трьохсот осіб, що живуть в даному регіоні, показав, що в минулому місяці кожен з них витратив в середньому $6,58.

Зрозуміло, хтось витратив більше, а хтось менше цього середнього кількості грошей. Замість того, щоб працювати з усіма 300 числами, ми використовуємо середнє, щоб визначити типове значення індивідуальних витрат кожного споживача. Що особливо важливо, помноживши середнє значення витрат на чисельність населення Клівленду, ми отримали прийнятну оцінку сумарних витрат на медичні товари мешканців всього міста:

Оцінка витрат на медичні товари мешканців Клівленду = (середнє значення витрат одної особи з вибірки)*(чисельність населення Клівленду) = ($6,58)*(503 000) = $3 309 740.

Цей прогноз сумарних продажів, що дорівнює $3 300 000, є прийнятним і, ймовірно, корисним. Однак це значення не є точним (в тому сенсі, що воно не відображає точну суму витрачених грошей). Далі, при вивченні довірчих

інтервалів, ми дізнаємося, як враховувати статистичну помилку, що виникає при поширенні результату, отриманого для вибірки з 300 осіб, на все населення, що складається з 503 000 осіб.

Приклад. Скільки є бракованих деталей?

Кожна партія виробів компанії Globular Ball Bearing Company містить 1000 виробів. Для проведення контролю якості виробів з вироблених за день 253 партій була взята випадковим чином вибірка, що включає 10 партій. Кількість бракованих виробів в кожній партії склало:

3, 8, 2, 5, 0, 7, 14, 7, 4, 1.

Середнє для цього набору даних:

$$\frac{3 + 8 + 2 + 5 + 0 + 7 + 14 + 7 + 4 + 1}{10} = \frac{51}{10} = 5{,}1$$

демонструє, що в середньому кожна партія містить 5,1 бракованих виробів. Іншими словами, рівень браку становить 5,1 виробів на 1000, або 0,51% (приблизно піввідсотка). Якщо поширити отримане середнє на всі випущені за день 253 партії, то можна очікувати

5,1 * 253 = 1290,3

бракованих виробів в денному випуску продукції, який складає 253 000 виробів.

Щоб показати, наскільки середнє дійсно є прийнятною узагальнюючою характеристикою списку чисел, на рис. 3.1.1 приведена гістограма для цього набору даних з 10 чисел з позначеним середнім. Зверніть увагу, наскільки добре в середині даних розташоване середнє, воно достатньо близько до всіх значень даних.

Мал. 3.1.1. Гістограма кількості бракованих деталей в кожній з 10 партій (одна партія міститься 1000 виробів) із зазначеним середнім значенням (5,1)

Зважене середнє: врахування важливості

Зважене середнє (використовують також термін середньозважене) схоже на середнє, але дозволяє привласнити різну важливість (значимість), або «вагу», кожному елементу даних. Зважене середнє дає можливість гнучко визначати систему важливості окремих елементів даних в тому випадку, коли їх не можна розглядати рівноцінними.

Якщо у підприємства є три заводи, при аналізі пенсійних витрат не потрібно брати просте середнє типових розмірів пенсійних витрат на кожному з трьох заводів як типове значення загальних пенсійних витрат, особливо, якщо заводи відрізняються за розмірами. Якщо чисельність працівників на одному заводі в два рази перевищує чисельність на іншому заводі, мабуть, буде доречним при обчисленні узагальнюючого показника врахувати пенсійний фонд першого заводу двічі. Середньозважене дозволить узагальнити дані, використовуючи ваги, що визначені у відповідності до розміру кожного заводу.

Ваги – це, зазвичай, позитивні числа, сума яких дорівнює 1. Не хвилюйтеся, якщо спочатку обчислена сума не буде дорівнювати 1. Ви завжди зможете відкоригувати значення ваг, розділивши кожну вагу на суму всіх інших ваг. Вихідні ваги можна було б визначити виходячи з чисельності працівників, ринкової вартості або будь-якого іншого об’єктивного показника, а також можна скористатися суб’єктивним методом (особистою думкою або думкою експерта). Іноді легше обрати ваги, не піклуючись, щоб їх сума дорівнювала 1, а потім перетворити їх, розділивши кожну на загальну суму.

Припустимо, ви вирішили обчислити середньозважене пенсійних витрат для трьох заводів, присвоївши ваги відповідно до чисельності працівників. Якщо чисельність службовців дорівнює 182, 386 та 697, то ваги відповідно дорівнюють:

$$182 / 1265 = 0{,}144;$$
$$386 / 1265 = 0{,}305;$$
$$697 / 1265 = 0{,}551.$$

Зверніть увагу, що значення ваги отримано шляхом ділення чисельності працівників на відповідному заводі на загальну кількість працівників всіх (трьох) заводів – 182 + 386 + 697 = 1265. Сума отриманих ваг, як і належить, дорівнює 1: 0,144 + 0,305 + 0,551 = 1.

Для обчислення зваженого середнього кожен елемент даних множать на привласнену йому вагу і підсумовують отримані значення. Відповідна формула має такий вигляд.

$$\text{Зважене середнє} = \text{Сума(вага помножена на значення елементу)} =$$

$$= \omega_1 X_1 + \omega_2 X_2 + \cdots + \omega_n X_n = \sum_{i=1}^{n} \omega_i X_i$$

Де $\omega_1, \omega_2, \omega_n$ – відповідні ваги, сума яких дорівнює 1. Ви можете вважати звичайне (не виважене) середнє також середньозваженими, в якому всі елементи даних мають однакову вагу, що дорівнює $1/n$.

Середньозважене значень 63, 47 і 98 з вагами, що дорівнюють 0,144; 0,305 і 0,551, відповідно, так само:

$$(0{,}144 * 63) + (0{,}305 * 47) + (0{,}551 * 98) =$$

$$= 9{,}072 + 14{,}335 + 53{,}998 = 77{,}405$$

Зверніть увагу, що, як і слід було очікувати, середньозважене відрізняється від звичайного (не зваженого) середнього цих трьох значень (53+47+98)/3=69,333. При обчисленні середньозваженого найбільше значення має вага 0,551 (що більше, ніж одна третина сумарної ваги). Ось чому в нашому випадку середньозважене більше, ніж звичайне не зважене середнє.

Середньозважене найкраще інтерпретувати як середнє, яке використовується в ситуаціях, коли одні елементи більш важливі, ніж інші. Більш важливі елементи вносять більший внесок в значення середньозваженого.

Приклад. Ваш середній бал

Середній бал (GPA – grade point average) ваших результатів навчання в університеті обчислюється як зважене середнє. Це пов’язано з тим, що деякі

курси оцінюються великою кількістю балів (кредитів) і, отже, є більш важливими в порівнянні з іншими. Цілком розумно, якщо курсу, який оцінюється в два рази більше, ніж інший, присвоюється в двоє більша вага, і середній бал це відображає.

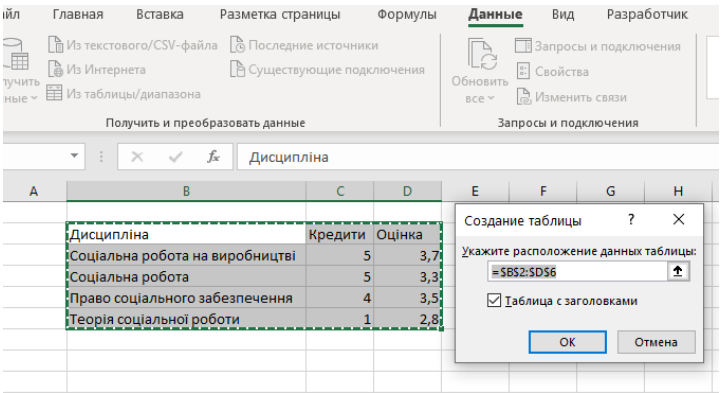
У різних університетах використовують різні системи оцінок. Припустимо, що система оцінок у вашому університеті включає оцінки від 0,0 (незалік) до 4,0 (відмінно) і в кінці семестру ваша картка з оцінками має такий вигляд.

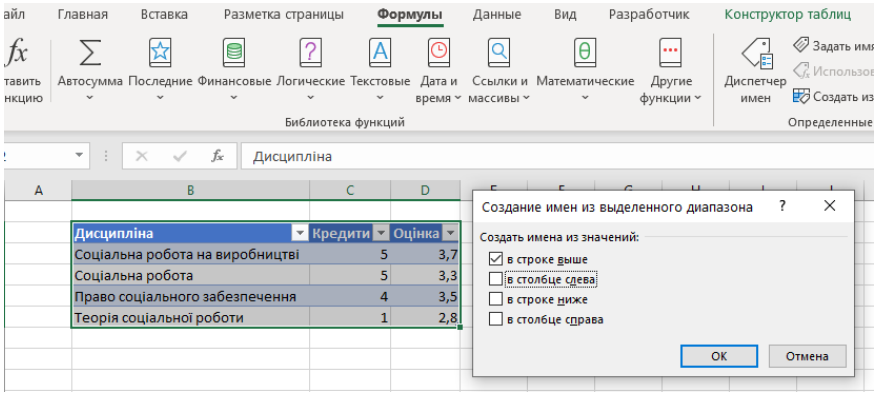
Дисципліна	Кредити	Оцінка
Соціальна робота на виробництві	5	3,7
Соціальна робота	5	3,3
Право соціального забезпечення	4	3,5
Теорія соціальної роботи	1	2,8
Разом	**15**	

Ваги можна обчислити, розділивши кількість кредитів за поточним курсом на 15 – загальну суму кредитів. Ваш середній бал розраховують як середньозважене ваших оцінок, зважене відповідно до кількості кредитів кожного з курсів:

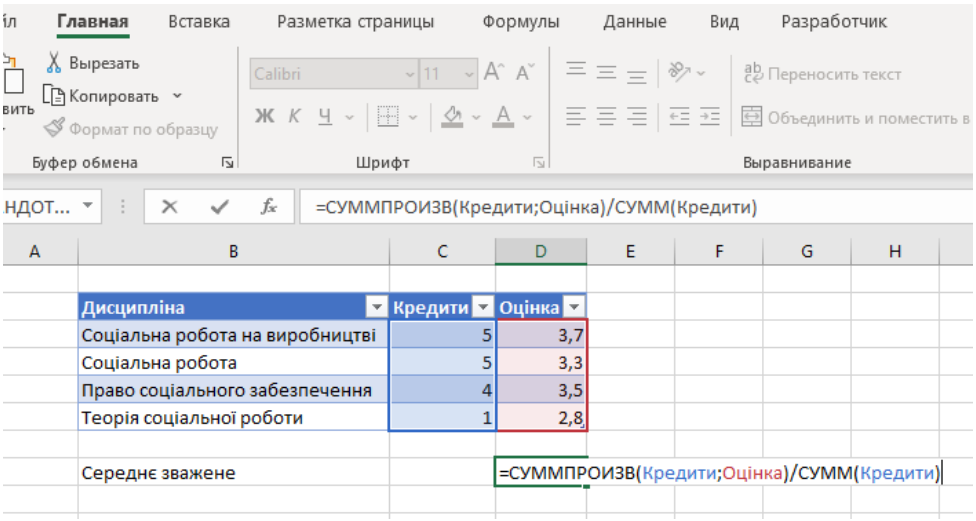
$$\left(\frac{5}{15} * 3{,}7\right) + \left(\frac{5}{15} * 3{,}3\right) + \left(\frac{4}{15} * 3{,}5\right) + \left(\frac{1}{15} * 2{,}8\right) = 3{,}45$$

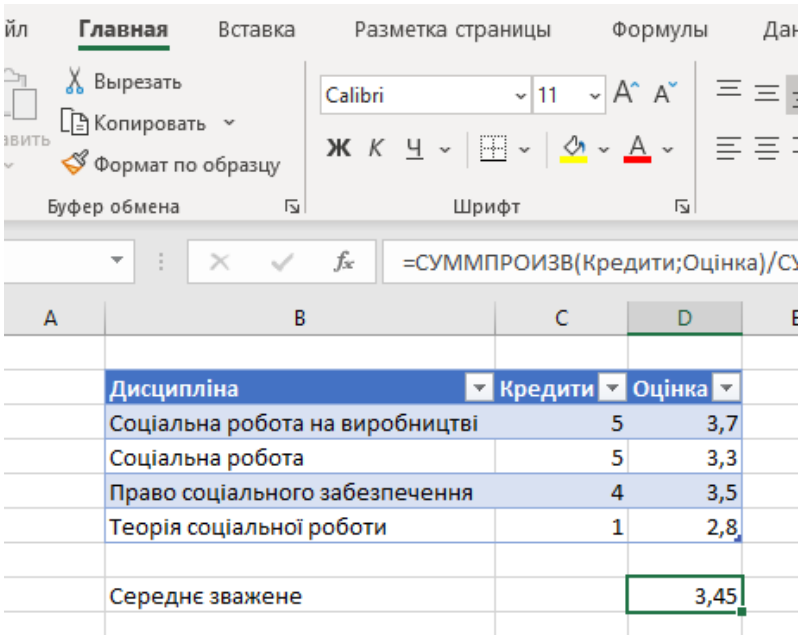
Щоб знайти середньозважене за допомогою Excel, спочатку дайте назву кожній з колонок чисел. Найпростіше це зробити, виділивши всі колонки, завершуючи верхнім заголовки, потім використати спочатку команду Ctrl+T (команда створення динамічної таблиці) та Ctrl+Shift+F3 (створення імен з означеного діапазону) і клацнути на кнопку ОК. Це виглядає так.





Тепер потрібно скористатися функцією Excel SUMPRODUCT (СУММПРОИЗВ), яка множить кількість кредитів курсу на відповідну оцінку і складає отримані результати, а потім результат ділить на загальну кількість кредитів (щоб сума ваг дорівнювала 1). Ви отримаєте середньозважене значення, рівне 3,45.





На щастя, низька оцінка з Теорії соціальної роботи не сильно вплинула на ваш середній бал (GPA), що дорівнює 3,45, оскільки вага цієї оцінки мала (усього 1 кредит). Якби ці чотири оцінки були просто усереднені, то результат був би нижче (3,33). Велика вдача, що у вирішальний момент в кінці семестру оцінки не дуже постраждали через захоплення практичними курсами!

Приклад. Вартість капіталу фірми

Вартість капіталу фірми, поняття з області корпоративного фінансування, обчислюють як зважене середнє. Суть в тому, що фірма збільшує свої грошові кошти за допомогою продажу різних цінних паперів: акцій, облігацій, векселів тощо. Оскільки кожен вид цінного паперу має свою власну прибутковість (вартість капіталу), корисно об'єднати і узагальнити різні рівні прибутковості в одне значення, що представляє собою сукупну вартість капіталу для цього набору цінних паперів.

Вартість капіталу фірми є простим середнім зваженим вартості капіталу по кожному цінному паперу (прибутковість або процентна ставка), причому ваги визначають відповідно до повної ринкової вартості цінних паперів. Наприклад, якщо вартість привілейованих акцій становить лише 3% від ринкової вартості представлених фірмою у обіг цінних паперів, то їй слід привласнити низьку вагу.

Розглянемо ситуацію для Leveraged Industries, Inc., гіпотетичної фірми з безліччю боргових зобов'язань, що утворилися в результаті недавньої діяльності по злиттю і придбанню.

Вид цінних паперів	**Ринкова вартість, тис. дол.**	**Прибутковість (норма прибутку), %**
Звичайна акція	100	18,5
Привілейована акція	15	14,9
Облігації (ставка 9%)	225	11,2
Облігації (ставка 8,5%)	115	11,2
Разом	**455**	

Для кожного виду цінних паперів розділіть відповідну ринкову вартість на загальну суму, щоб знайти ваги, які дають пропорцію ринкової вартості цього виду цінних паперів.

Вид цінних паперів	Вага
Звичайна акція	0,220
Привілейована акція	0,033
Облігації (9%)	0,495
Облігації (8,5%)	0,253

Для звичайних акцій вага дорівнює 0,220, що в термінах ринкової вартості означає, що на 22% фірма фінансується за рахунок звичайних акцій. Вартість капіталу можна обчислити, помноживши значення ринкової прибутковості на ваги і склавши отримані значення:

(0,220 * 18, 5) + (0,033 * 14, 9) + (0,495 * 11,2) + (0,253 * 11,2) = 12,94

Вартість (прибутковість) капіталу Leveraged Industries, Inc. становить 12,9%. Таке середньозважене об'єднує значення вартості (прибутковості) окремих видів цінних паперів (18,5%, 14,9% і 11,2%) в одне число.

Результат (12,9%) не змінився б, якби ви об'єднали два види облігацій, розглянувши їх як один елемент із загальною ринковою вартістю $340 000 і прибутковістю 11,2% (що це так, можна перевірити, провівши розрахунки). Це пов'язано з тим, що відмінність між облігаціями не має практичних результатів: два випуски облігацій мають різні купонні ставки, оскільки вони випущені в різний час, але з часом їх ринкова ціна змінилася таким чином, що прибутковість стала однаковою.

Середньозважена вартість акціонерного капіталу можна пояснити наступним чином. Якщо Leveraged Industries, Inc. вирішить збільшити додатковий капітал без зміни своєї основної бізнес-стратегії (тобто типу проектів, ризику проектів) і зберегти той же набір цінних паперів, то необхідно буде виплачувати в рік 12,9%, або $129 на $1000. Ці $129 будуть виплачені по різним типам цінних паперів відповідно до їх ваги.

Приклад. Коригування недостатньою репрезентативності

Крім того, зважене середнє використовують, щоб скорегувати недоліки репрезентативності вибірки по відношенню до генеральної сукупності, що цікавить вас. Оскільки середнє вибірки враховую всі елементи однаково, а вам відомо, що (в порівнянні з генеральною сукупністю) деякі групи елементів

представлені надлишково, а інші, навпаки, – недостатньо, то більш точний результат можна отримати, використовуючи зважене середнє. Зважене середнє буде точніше, оскільки в ньому відома інформація про кожну групу (взята з вибірки) буде об'єднана з додатковою інформацією про представництво кожної групи (у генеральній сукупності, а не у вибірці).

Знову розглянемо вибірку 300 мешканців Клівленду, яку ми аналізували раніше з точки зору витрат людей на медичні товари. Припустимо, що відсоток молодих людей (до 18 років) в цій виборці (21,7%) не відповідає відомому відсотку для всього населення міста (25,8%) і що середні грошові витрати, підраховані для кожної групи окремо, складають:

середні грошові витрати для людей молодше 18 років – $4,86;

середні грошові витрати для людей старше 18 років – $7,06.

При обчисленні середньозваженого цих витрат будемо використовувати ваги не вибірки, а відомі нам ваги генеральної сукупності, тобто вважатимемо, що маємо справу з 25,8% молодих людей і 74,2% людей старше 18 років (різниця 100% – 25,8%). Звичайно, якби були відомі оцінки витрат для міста в цілому, то ви б їх також використовували. Але такі дані вам недоступні. Вам відомі витрати лише для 300 осіб з вибірки. Після перетворення відсотків в ваги зважене середнє ви обчислюєте наступним чином:

Зважене середнє витрат = (0,258 x $ 4,86) + (0,742 x $ 7,08) = $ 6,49.

Зважене середнє $6,49 дає кращу оцінку середнього значення витрат на медичні товари в Клівленді, ніж звичайне (не виважене) середнє ($6,58). Зважене середнє краще, оскільки воно містить поправку на занадто великий відсоток людей у віці старше 18 років в нашій вибірці з 300 осіб. Оскільки люди такого віку витрачають більше, то без поправки середня оцінка витрат виходить завищеною ($6,58 у порівнянні з $6,49).

Звичайно ж, навіть ця нова зважена оцінка може бути помилковою. Але вона заснована на більшому обсязі інформації, тому очікувана помилка буде менше, що можна довести за допомогою математичних моделей. Нова оцінка не обов'язково кожен раз буде кращою (тобто і в даному прикладі звичайне,

не зважене, середнє може насправді бути ближче до істини), але ймовірність того, що зважена оцінка буде ближче до істини, набагато більше.

Медіана: типове значення для кількісних і порядкових даних

Медіана – це значення, яке розташоване посередині; половина елементів в наборі даних більше цього значення, а друга половина – менше. Таким чином, медіана розташовується в центрі даних і дає уявлення про список значень. Щоб знайти медіану, дані розташовують в порядку зростання, а потім визначають середнє значення. Зверніть увагу, що якщо в наборі даних немає одного центрального значення, то слід усереднити ті два значення, які розташовані посередині ряду.

Медіану можна визначити в термінах рангів. Ранги пов'язують числа 1, 2, 3, ... *n* зі значеннями даних таким чином, що найменше значення має ранг 1, наступне за величиною значення – ранг 2 і так далі до найбільшого значення, яке має ранг *n*. В основу визначення медіани покладено наступний принцип.

З урахуванням всіх можливих особливих випадків медіана для списку з *n* елементів обчислюється таким чином.

1. Розташуйте елементи даних в порядку зростання (або зменшення – це не має значення).

2. Визначте середнє значення отриманого ряду. Можливі два варіанти.

а) Якщо *n* – непарне число, то медіаною буде середнє значення даних, яке має номер $\frac{1+n}{2}$, якщо відраховувати від будь-якого з двох кінців впорядкованого списку. Наприклад, медіана списку 15, 27, 14, 18, 21 з *n*=5 значень дорівнює:

медіана (15, 27, 14, 18, 21) = медіана (14, 15, 18, 21, 27) = 18.

Слід зазначити, що медіана, 18, це третє за порядком значення в упорядкованому списку, що відповідає формулі, оскільки $\frac{1+n}{2} = \frac{1+5}{2} = 3$.

Як приклад порядкових даних розглянемо список рейтингів облігацій AAA, A, B, AA, A. Для цього списку медіана буде обчислюватися так:

медіана (AAA, A, B, AA, A) = медіана (B, A, A, AA, AAA) = A.

б) Якщо n – парне число, то ряд має не одне, а два середніх значення. Ці значення розташовані на відстані $\frac{1+n}{2}$ від кожного з двох кінців впорядкованого списку даних.

в) Якщо набір даних кількісний (тобто складається з чисел), то медіаною є середнє цих двох значень, розташованих в середині ряду. Наприклад, медіану списку 15, 27, 14, 18 з n=4 чисел обчислюється таким чином:

$$\text{медіана (15, 27, 14, 18)} = \text{медіана (14, 15, 18, 27)} = \frac{15+18}{2} = 16{,}5.$$

У цьому випадку за формулою $\frac{1+n}{2}$ маємо: $\frac{1+4}{2} = 2{,}5$; що говорить про необхідність пройти в упорядкованому списку половину шляху між другим і третім числом, усереднивши ці два числа.

г) Якщо набір даних є порядковим (тобто містить впорядковані категорії) і якщо два розташованих в середині ряду значення представляють одну і ту ж категорію, то ця категорія є медіаною. Якщо ці два значення представляють різні категорії, то обидві ці категорії будуть медіанами. Наприклад, для списку рейтингів облігацій A, B, AA, A медіана буде дорівнювати:

$$\text{медіана (A, B, AA, A)} = \text{медіана (B, A, A, AA)} = \text{A},$$

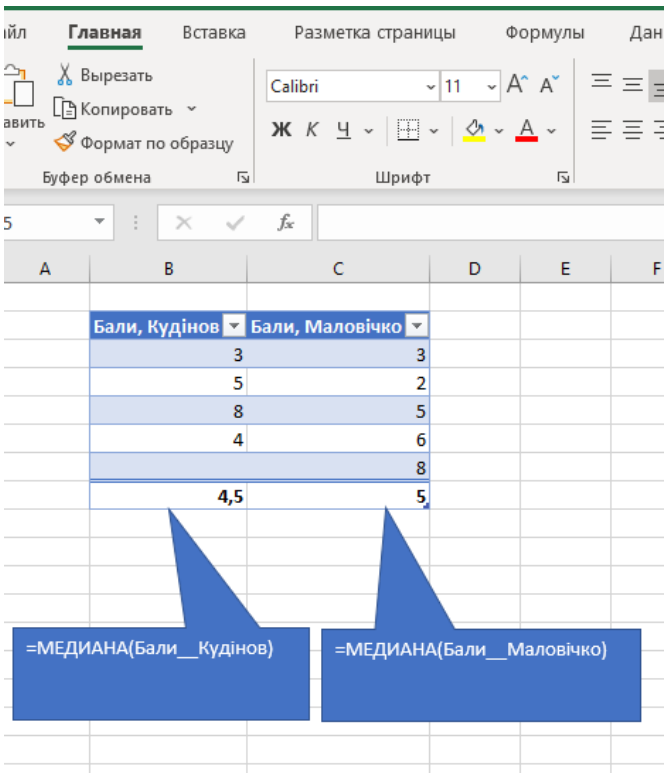
оскільки обидва розташованих посередині значення рівні A.

В іншому прикладі для списку рейтингів облігацій A, AAA, B, AA, AAA, B медіана буде обчислюватися так:

$$\text{медіана (A, AAA , B, AA, AAA, B)} = \text{медіана (B, B, A, AA, AAA, AAA)} =$$
$$= \text{A і AA}.$$

Це найкраще, що можна зробити в даній ситуації, так як для порядкових даних можна обчислити середнє двох значень.

Для обчислення медіани в Excel можна використовувати функцію MEDIAN (МЕДИАНА) наступним чином.



Чим відрізняється медіана від середнього? Якщо набір даних розподілений нормально, то значення медіани і середнього близькі між собою, оскільки нормальний розподіл симетричний і має чітко виражену середню точку. Однак навіть при нормальному розподілі (тут мова йде про «практично нормальний» розподіл, а не про теоретично нормальний розподіл) середнє і медіана дещо відрізняються один від одного, оскільки кожна з цих величин визначається по-своєму і, крім того, в реальних даних майже завжди присутня деяка випадковість. Якщо набір даних не підкоряється нормальному розподілу, то медіана і середнє можуть сильно відрізнятися, тому що у асиметричного розподілу немає чітко вираженої центральної точки. Зазвичай середнє по відношенню до медіані зрушено в напрямку більш довгого хвоста або в напрямку викиду, оскільки середнє реально враховує значення таких екстремальних спостережень, в той час як для медіани важливо лише, по який бік від неї лежить те чи інше значення.

Приклад. Особисті доходи

Розподіл таких кількісних даних, як особисті доходи окремих людей і сімей (як і розподіл продажів, витрат, цін тощо), часто скошено в сторону

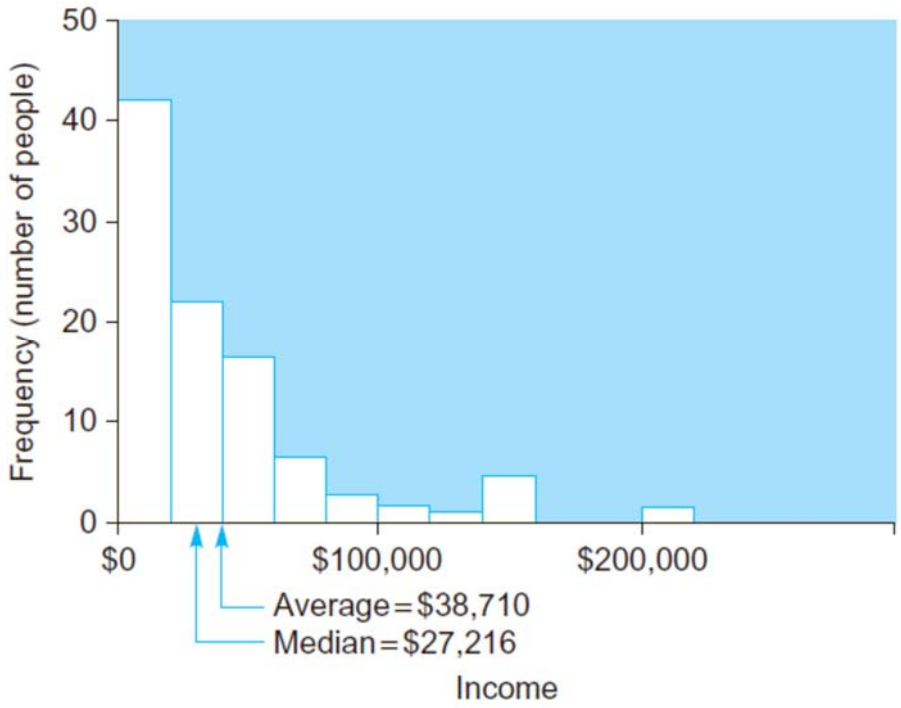
більш високих значень, оскільки такі набори даних містять багато невеликих значень, кілька середніх значень і трохи великих і дуже великих значень. Таким чином, зазвичай середнє більше, ніж медіана. Це пов’язано з тим, що на значення середнього, одержуваного складанням всіх елементів, сильно впливають великі значення. Розглянемо доходи домогосподарств в США в 2019 році:

середнє доходів домогосподарств – $39 020;

медіана доходів домогосподарств – $30 786.

Середнє доходу вище, ніж медіана, тому що на значення середнього сильно впливає відносно невелика кількість дуже високоприбуткових домогосподарств. Згадаймо, що при обчисленні середнього ці високі доходи входять в суму, а при обчисленні медіани вони є просто «високими доходами» (при цьому кожному домогосподарству з високими доходами відповідає домогосподарство з низькими доходами).

Гістограма на рис. 3.1.1 показує вид розподілу доходів для вибірки з 100 осіб. Розподіл сильно скошено в напрямку високих доходів, оскільки є багато людей з низькими доходами (на це вказують високі стовпчики зліва на гістограмі) і відносно небагатих людей, що мають середні і високі доходи (короткі стовпчики в середині і праворуч на гістограмі). Середнє значення доходу: $38 710 вище, ніж медіана $27 216. Медіана (точка, яка ділить кількість об’єктів навпіл) нижче середнього, тому що на даній гістограмі більшість людей мають низькі доходи, а наявність людей з високими доходами значно збільшує значення середнього.



Мал. 3.1.1. Гістограма розподілу даних про доходи 100 осіб.

Це асиметричний розподіл і середнє значно більше, ніж медіана.

Приклад. Стадії складання комп'ютерних системних блоків

Розглянемо процес виробництва комп'ютерів, що складається з наступних стадій:

– Виробництво материнської плати.

– Установка роз'ємів на материнську плату.

– Установка в роз'єми електронних мікросхем.

– Тестування зібраної материнської плати.

– Установка зібраної материнської плати в системний блок комп'ютера.

– Тестування зібраного системного блоку.

Якщо у вас є набір даних, в якому для кожного системного блоку зазначено, на якій виробничій стадій виготовлення він знаходиться, то такий одновимірний набір порядкових даних може мати наступний вигляд:

A, C, E, F, C, C, D, C, A, E, E, ...

Цей набір даних є порядковим, оскільки для категорій існує природний порядок – порядок проходження виробу через всі стадії виробничого процесу

від початку збирання до завершення. Такий набір даних можна представити у вигляді списку частот такого вигляду.

Стадія виробництва	Кількість комп'ютерних системних блоків
A	57
B	38
C	86
B	45
E	119
A	42
Разом	**387**

Оскільки це порядкові дані, для них можна обчислити медіану, але не середнє. Медіаною буде системний блок з рангом $\frac{1+387}{2} = 194$ в списку всіх системних блоків, упорядкованих у відповідності до стадії виробництва. Нижче показаний спосіб визначення медіани.

Блоки з рангами від 1 до 57 знаходяться на стадії A. Таким чином, медіана (яка має ранг 194) знаходиться за межами стадії A.

Блоки з рангами від 58 (57+1) до 95 (57+38) знаходяться на стадії B. Значить, медіана знаходиться за межами стадії B.

Блоки з рангами від 96 (95+1) до 181 (95+86) знаходяться на стадії C. Отже, медіана знаходиться за межами стадії C.

Блоки з рангами від 182 (181+1) до 226 (181+45) знаходяться на стадії D. Отже, медіана знаходиться на стадії D, оскільки ранг медіани (194) лежить між рангами 182 і 226.

Таким чином, близько половини системних блоків знаходяться на стадіях, що передують стадії D, і приблизно половина – на стадіях, що йдуть після стадії D. Тому стадія D є середньою точкою (з точки зору готовності збірки) для всіх системних блоків, які перебувають на даний момент у виробництві.

Мода: типове значення навіть для номінальних даних

Мода є найбільш поширеною категорією, тобто категорією, яка найчастіше зустрічається в наборі даних. Це єдина характеристика, яку можна визначити для номінальних якісних даних, оскільки невпорядковані категорії

не можливо складати (як це вимагається для середнього) і не можна ранжувати (як це вимагається для медіани). Моду можна легко знайти для порядкових даних, якщо просто проігнорувати впорядкованість категорій і виконувати всі дії так само, як для набору номінальних даних з невпорядкованими категоріями.

Мода також визначена для кількісних даних (чисел), хоча при цьому може мати місце певна невизначеність. Для кількісних даних моду можна визначити як значення, що відповідає найвищій точці на гістограмі, можливо, на середині самого високого стовпчика. Джерела не визначеності можуть бути різними. На гістограмі може бути два «найвищих» стовпчика. Або, що значно гірше, визначення моди може залежати від того, яким чином побудована діаграма: зміна ширини стовпчиків і їх розташування може привести до невеликих (або помірних) змін форми розподілу, в результаті чого може змінитися і мода. Для кількісних даних мода є дещо невизначеним поняттям.

Моду знайти легко. Незалежно від того, представляють наявні у вас числа кількість об’єктів в кожній категорії або відповідні відсотки, необхідно просто вибрати категорію з найбільшою кількістю або відсотком. Якщо на перше місце претендують дві або більше категорій, то необхідно вказати всі ці категорії під загальною назвою «мода» для цього набору даних.

Приклад. Голосування на виборах

Оскільки під час виборів підраховують кількість відданих голосів, то ці голоси можна вважати за набір номінальних якісних даних. У вас може бути своя думка щодо впорядкування кандидатів, але так як спільної думки в цьому питанні немає, то ви можете вважати цей набір даних неврегульованим. Список даних може виглядати так:

Ситнік, Зеленов, Усачов, Ситнік, Ситнік, Усачов, Ситнік...

Результати виборів можна записати в такий спосіб.

Прізвище	Кількість голосів	Відсоток
Усачов	7175	15,1
Зеленов	18956	39,9
Запорожченко	502	1,1

Ситнік	20817	43,9
Разом	**47450**	**100,0**

Ясно, що модою в цьому наборі даних буде Ситнік, оскільки він набрав найбільшу кількість голосів (20817) і найбільший відсоток голосів (43,9%). Зверніть увагу, що мода не обов'язково представляє більше половини (більшість) об'єктів, хоча іноді може бути і так. Мода просто представляє більше об'єктів, ніж будь-яка інша категорія.

Приклад. Контроль якості: відхилення в виробництві

Важливим видом діяльності при створенні якісних виробів є аналіз відхилень у виробничих процесах. Одні відхилення від виробничого процесу неминучі, але допустимі (через невелику величини), в той час як інші виводять процес з-під контролю і призводять до виробництва низькосортних виробів. Едвардс Демінг (W. Edwards Deming) вперше ввів контроль якості в Японії в 50-их роках. Деякі з його методів коротко можна узагальнити наступним чином.

Запропонований Демінгом метод в основі своїй є статистичними. Будь-яка виробнича діяльність, в цеху або в офісі, має відхилення від ідеалу. Демінг запропонував систематичний метод вимірювання відхилень виробничого процесу, виявлення причин цих відхилень та їх зменшення, вдосконалення за рахунок цього процесу, а значить, і підвищення якості продукції.

Збір і подальший аналіз даних – це ключовий компонент належного контролю якості. Припустимо, що підприємство реєструє причину браку кожен раз при появі виробу неприпустимої якості.

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

Ясно, що модою в цьому наборі даних є блок живлення, оскільки ця причина браку зустрічається частіше за інших. Мода допомагає зосередити увагу на найважливішій категорії (той, що найбільш часто зустрічається). Немає необхідності розробляти додаткові заходи з підтримки чистоти на

робочому місці або по недопущенню падіння коробок, оскільки ці причини мало впливають на загальну частоту браку. В першу чергу слід звернути увагу на модальну категорію.

У розглянутій ситуації фірма могла б спробувати розібратися з проблемою «блок живлення» і вжити відповідних заходів. Можливо, цей блок живлення має недостатню потужність для даного виробу і необхідне більш потужне джерело. Можливо, потрібно знайти більш надійного постачальника. У будь-якому випадку мода допомагає уточнити наявну проблему.

Приклад. Повторний розгляд стадій складання системних блоків

Розглянемо ще раз описаний раніше приклад даних про стан збирання системних комп'ютерних блоків. Нижче наведено набір даних.

Стадія виробництва	**Кількість системних блоків**
A	57
B	38
C	86
D	45
E	119
F	42
Разом	**387**

Раніше ми вже визначили, що медіана припадає на стадію виробництва D, оскільки ця стадія відмежовує половину системних блоків, які перебувають на початкових стадіях складання, від другої половини системних блоків на кінцевій стадії складання. Однак в даному випадку медіана не збігається з модою (хоча в деяких інших прикладах мода може збігатися з медіаною).

Тут мода являє собою стадію E, на якій знаходиться 119 системних блоків, тобто більше, ніж на будь-який інший стадії. У такій ситуації керівництво повинно бути проінформоване про моду, тому що найбільш «вузьке місце» у виробничому процесі, швидше за все, виявиться саме як мода.

У розглянутому прикладі стадія E – це установка материнської плати в системний блок. Наявність великої кількості системних блоків на цій стадії може бути пов'язано з більшою трудомісткістю даної операції. Але, з іншого боку, це може бути і свідченням наявності проблем у працівників, які працюють на цій стадії (можливо, причина в недостатній кількості людей або

великій кількості відсутніх працівників). В такому випадку керівництву необхідно звернути на це увагу.

Які показники потрібно використовувати

Який з трьох показників (середнє, медіану або моду) слід використовувати в конкретних обставинах? Є два варіанти відповіді. Перший залежить від того, що можна обчислити, а другий залежить від того, який з показників більш корисний.

Моду можна обчислити для будь-якого одновимірного набору даних (хоча в разі кількісних даних проблемою може бути деяка невизначеність). Середнє можна обчислити тільки для кількісних даних (чисел), а медіану – для всіх типів даних, крім номінальних (невпорядкованих категорій). Таким чином, ваш вибір обмежений, а в разі номінальних даних у вас взагалі немає іншого вибору, крім як використовувати моду. Рекомендації по вибору характеристики в залежності від типу даних можна уявити таким чином.

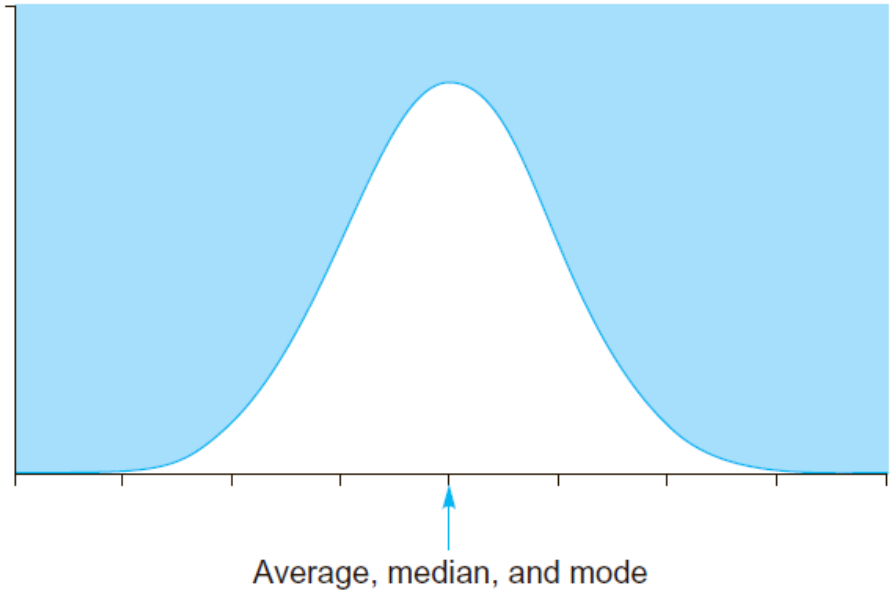
	Кількісні	**Порядкові**	**Номінальні**
Середнє	Так		
Медіана	Так	Так	
Меду	Так	Так	Так

У разі кількісних даних, для яких можна обчислити всі три характеристики, наскільки вони відрізняються між собою? Якщо розподіл близький до нормального, різниця невелика, оскільки кожна з характеристик прагне до чітко вираженої середини, що має криву розподілу у формі дзвона (рис. 2.1.2).

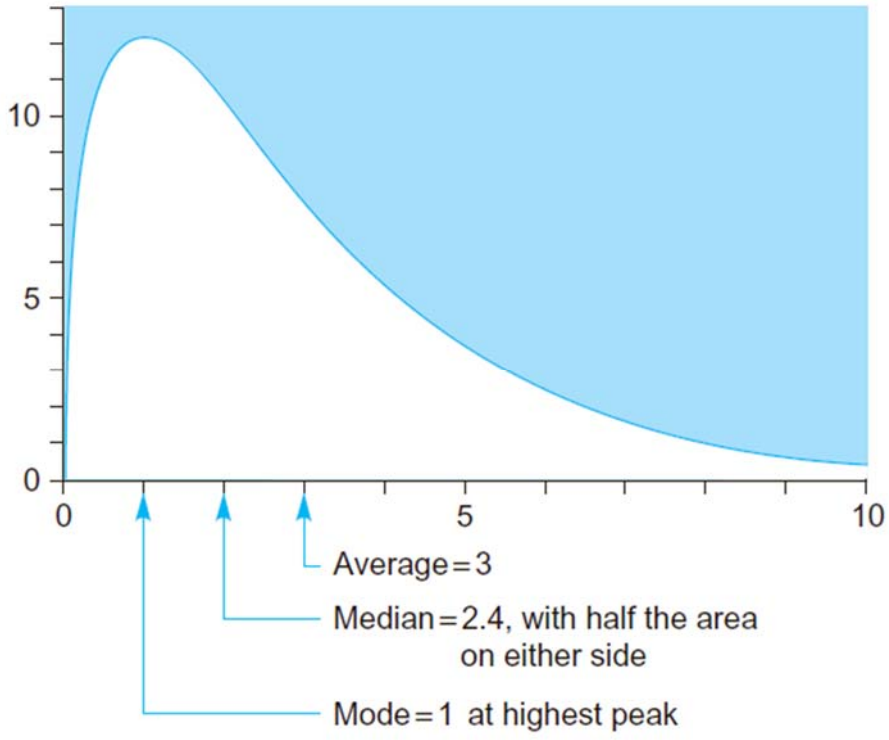
Однак у випадку асиметричного розподілу даних ці характеристики можуть помітно відрізнятися (як ми вже відзначали для середнього і медіани). На рис. 2.1.3 показані визначені характеристики для даних, що не підпорядковуються нормальному розподілу.

Середнє слід використовувати, коли набір даних розподілений нормально (принаймні приблизно), оскільки в цьому випадку середнє є найефективнішою характеристикою. Середнє також слід обчислювати і в тих

ситуаціях, де необхідно зберегти або передбачити загальну суму значень даних, так як інші характеристики не дозволяють це зробити.



Мал. 2.1.2. Для ідеального нормального розподілу середнє (average), медіана (median) і мода (mode) збігаються. Для реальних даних, де завжди присутня випадковість, ці характеристики будуть приблизно, але не точно, рівні між собою



Мал. 4.1.3. Для скошеного розподілу середнє, медіана і мода розрізняються.

Мода відповідає найвищій точці на кривій розподілу. По обидва боки від медіани знаходиться половина області під кривою розподілу. Середнє знаходиться в точці центру ваги розподілу, як точки опори дошки дитячих гойдалок

Медіана служить хорошою характеристикою асиметричного розподілу, оскільки на неї не впливає невелике число даних з високими значеннями.

У разі сильної асиметрії медіана значно краще середнього характеризує більшість даних. Медіана також корисна при наявності викидів значень, так як вона стійка до їх впливу. Медіана корисна для порядкових даних (впорядковані категорії), хоча в залежності питання, що вирішується, можна використовувати і моду.

Моду використовують при наявності номінальних даних, так як в цьому випадку не можна обчислювати середнє і медіану. Вона також корисна для порядкових даних, коли важливо визначити найбільш поширену категорію.

Крім розглянутих існує багато інших характеристик. Перспективним є використання так званих «робастних» (стійких) оцінок, які поєднують в собі кращі властивості середнього і медіани. Для нормально розподілених даних такі оцінки представляють досить ефективний вибір і в той же час вони, як і медіана, стійкі до впливу викидів.

3.2. Що таке перцентіль

Перцентілі – це характеристики набору даних, що виражають ранги елементів у вигляді відсотків від 0 до 100%, а не у вигляді чисел від 1 до *n*, таким чином, що найменшим значенням відповідає нульовий перцентіль, найбільшому – 100-й перцентіль, медіані – 50-й перцентіль і т.д. Перцентілі можна розглядати як показники, що розбивають набори кількісних і порядкових даних на певні частини.

Зверніть увагу, що перцентіль це елемент даних, що має певний ранг і виражений в тих же одиницях, що і одиниці набору даних. Наприклад, 60-й

перцентіль ефективності продажів може дорівнювати $385 062 (виміряно НЕ у відсотках, а в доларах, як і елементи набору даних). Якщо цей 60-й перцентіль, що дорівнює $385 062, характеризує діяльність певного агенту з продажу (наприклад, Аделіни), то це означає, що приблизно 60% інших агентів мають результати нижче, ніж у Аделіни, а 40% агентів мають більш високі результати.

Перцентілі використовують для двох цілей.

1. Щоб показати значення елементу в даних при заданому перцентільному ранзі (наприклад, «10-й перцентіль дорівнює $156 293»).

2. Щоб показати перцентільний ранг значення даного елемента в наборі даних (наприклад, «ефективність продажів агенту по збуту (Олександра) складають $296 994, що відповідає 55-му перцентілю»).

Екстремуми, квартилі і блокові діаграми

Перцентілі відграють важливу роль в якості опорних характеристик. Щоб узагальнити основні риси розподілу, достатньо кількох значень перцентілів. Так, 50-й перцентіль – це медіана, оскільки 50-й перцентіль знаходиться посередині між найбільшим і найменшим значеннями ряду. Інтерес представляють екстремуми – найбільше і найменше значення даних, тобто 0-й і 100-й перцентілі відповідно. Доповнюють набір базових характеристик квартилі, що визначаються як 25-й і 75-й перцентілі.

Дивно, але статистики досі сперечаються щодо точного визначення квартилів, оскільки їх можна обчислювати різними способами. Ідея квартилів зрозуміла. Квартилі – це значення рангового ряду, які знаходяться на відстані однієї четвертої на шляху від найменшого і найбільшого значень. Однак це формулювання не вказує точно, як обчислювати квартилі.

Джон Т’юкі, один з творців практичного аналізу даних, визначає квартилі таким чином.

1. Обчислюємо ранг медіани за формулою $\frac{1+n}{2}$ і відкидаємо дробову частину. Наприклад, при *n*=13 отримуємо $\frac{1+13}{2} = 7$. При *n*=24 відкидаємо дробову частину у $\frac{1+24}{2} = 12{,}5$ і отримуємо 12.

2. Додаємо до отриманого значення 1 і ділимо на 2. Отримане значення є рангом нижнього квартилю. Наприклад, при *n*=13 ранг нижнього квартилю дорівнює $\frac{1+7}{2} = 4$. При *n*=24 ранг нижнього квартилю дорівнює $\frac{1+12}{2} = 6{,}5$, що свідчить про необхідність усереднити значення з рангами 6 і 7.

3. Віднімаємо отримане значення від (*n*+1). Результатом буде ранг верхнього квартилю. Наприклад, при *n*=13 отримаємо (13+1) – 4 = 10. При *n*=24 отримуємо (1+24) – 6,5 = 18,5, що свідчить про необхідність усередники значення з рангами 18 і 19.

Значення квартилів знаходять виходячи з цих рангів. Нижче наведена загальна формула визначення рангів квартилів, яка представляє зазначені вище кроки обчислень.

$$\text{Ранг нижнього квартилю} = \frac{1 + int\left[\frac{1+n}{2}\right]}{2};$$

$$\text{Ранг нижнього квартилю} = n + 1 - \text{ранг нижнього квартилю},$$

де *int* – означає функцію отримання цілого, що відкидає дробову частину числа.

П’ять базових показників включають найменше значення, нижній квартиль, медіану, верхній квартиль, найбільше значення.

П’ять базових показників

Найменше значення даних (0-й перцентіль).

Нижній квартиль (25-й перцентіль на чверть відстані від найменшого значення).

Медіана (50-й перцентіль, середина).

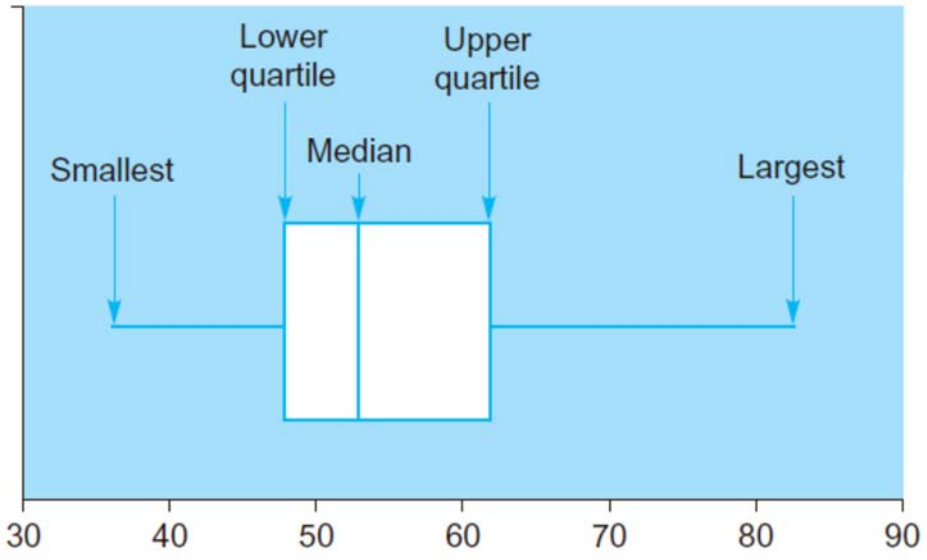
Верхній квартиль (75-й перцентіль, на три чверті відстані від найменшого значення або на чверть відстані від найбільшого значення).

Найбільше значення (100-й перцентіль).

Разом ці характеристики дають досить чітке уявлення про особливості ще необробленого набору даних. Два екстремуми характеризують розмах (діапазон) даних, медіана показує центр, два квартилі визначають межі «розташованої в центрі половини даних», а положення медіани щодо квартилів дає грубе уявлення про наявність чи відсутність асиметрії.

Блокова діаграма – це зображення всіх п’яти зазначених показників (рис. 3.2.1).

Блокова діаграма, як і гістограма, дає візуальне уявлення про розподіл, але використовує інший спосіб графічного відображення. Блокова діаграма не містить дрібних деталей, що дозволяє охопити всю картину в цілому і порівнювати кілька груп чисел, не вдаючись у деталі кожної з груп. При необхідності детально розглянути форму розподілу краще використовувати гістограму.



Мал. 3.2.1. Блокова діаграма містить п’ять базових показників одновимірного набору даних і дозволяє швидко визначити характер розподілу

Детальна блокова діаграма – це блокова діаграма, яка також має помічені мітками викиди (мітки також використовують для показу екстремальних спостережень, які не є викидами). Мітки виділяють ті

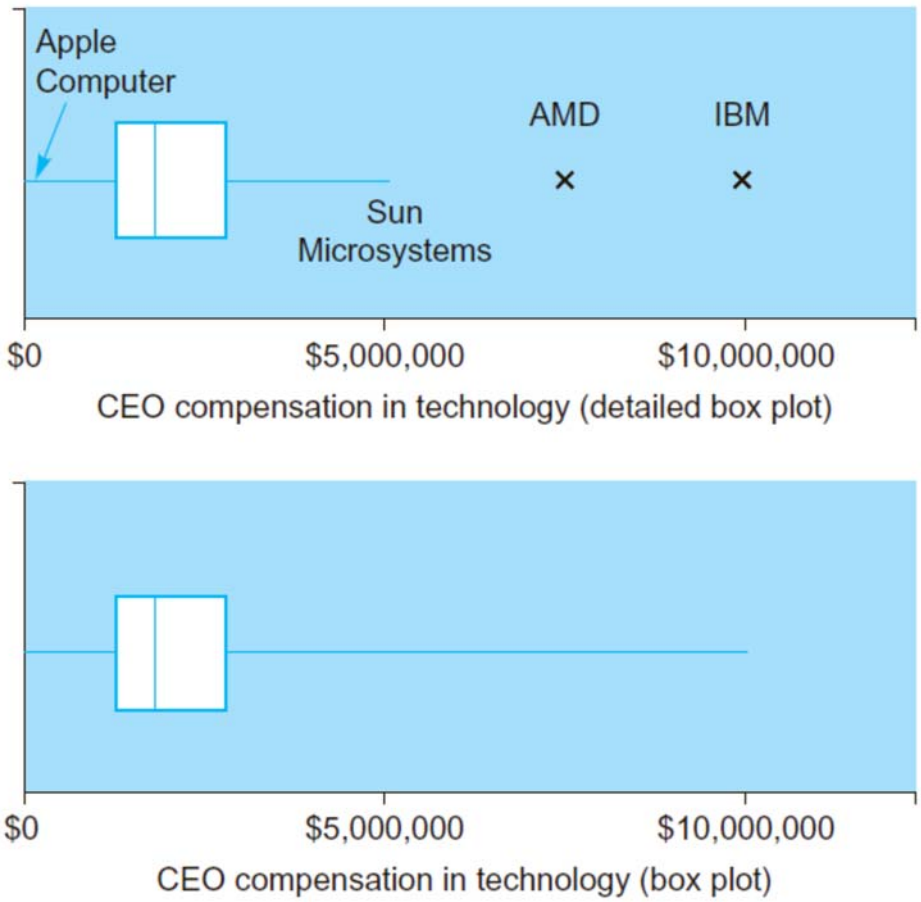
спостереження, що потребують особливої уваги. При створенні детальної блокової діаграми викиди визначають як ті значення даних (якщо вони є), які розташовані далеко від центру розподілу. Зокрема, велике значення в наборі даних розглядається як викид, якщо воно перевищує

верхній квартиль + 1,5 * (верхній квартиль – нижній квартиль).

Мале значення в наборі даних розглядається як викид, якщо воно менше, ніж

нижній квартиль – 1,5 * (верхній квартиль – нижній квартиль).

На додаток до нанесення на діаграму викидів з відповідними мітками можна також відзначити екстремальні значення, які викидами НЕ є (по одному з кожного боку), оскільки часто вони також заслуговують на особливу увагу. На рис. 2.2.2 для порівняння показані блокова і детальна блокова діаграми.



Мал. 3.2.2. Блокова діаграма (знизу) і докладна блокова діаграма (зверху) винагород керівників IT-компаній.

Обидві діаграми містять п’ять базових показників, але докладна блокова діаграма дає цінну інформацію про викиди (а також показує екстремальні значення, які викидами не є). У цьому прикладі викидами є фірми, які виплатили виключно високу винагороду своїм керівникам.

Приклад. Виплати керівникам

Розглянемо виплати (заробітну плату і премії) керівників фінансових компаній в 2019 році. Табл. 3.2.1 містить впорядкований список розмірів виплат, їх ранги і відповідні п’ять характеристик розподілу.

Таблиця 3.2.1. Виплати керівникам фінансових компаній

Фірма	**Керівник**	**Зарплата і премії в 2019 році, дол.**	**Ранг**
Equitable *	RH Jenrette	7 730 000	38 – Найбільше значення дорівнює $7 730 000
Bear Stearns *	JE Cayne	7 666 000	37
First Financial Mgmt. *	PH Thomas	6 910 000	36
Merrill Lynch *	DP Tully	4 840 000	35
Travelers *	SI Weill	3 903 000	34
American Inti. Group	MR Greenberg	3 750 000	33
Schwab (Charles)	CR Schwab	3 273 000	32
Dean Witter Discover	PJ Purcell	3 200 000	31
American Express	H. Golub	3 077 000	30
Marsh & McLennan	AJ Smith	2 101 000	29 – Верхній квартиль = $2 101 000
Progressive	PB Lewis	2 063 000	28
American General	HS Hook	1 960 000	27
Loews	PR Tisch	1 937 000	26
Torchmark	RK Richey	1 936 000	25
Household International	DC Clark	1 877 000	24
Aflac	D. P. Amos	1 726 000	23
Cigna	WH Taylor	1 723 000	22
Great Western Financial	JF Montgomery	1 674 000	21
Transamerica	FC Herringer	1 537 000	20
			Медіана = $1 497 500
General RE	RE Ferguson	1 458 000	19
Chubb	DR O’Hare	1 393 000	18
AON	PG Ryan	1 384 000	17
St. Paul	DW Leatherdale	1 294 000	16
CAN Financial	DH Chookaszian	1 242 000	15

Providian	IW Bailey II	1 190 000	14
Jefferson-Pilot	DA Stonecipher	1 119 000	13
Aetna Life & Casualty	RE Compton	1 075 000	12
First USA	JC Tolleson	1 040 000	11
Salomon	RE Denham	1 000 000	10 – Нижній квартиль = $1 000 000
Golden West Financial	HM Sandler	901 000	9
Cincinnati Financial	RB Morgan	896 000	8
Allstate	WE Hedien	767 000	7
Block (H & R)	TM Bloch	746 000	6
Franklin Resources	C. B. Johnson	743 000	5
Safeco	RH Egsti	601 000	4
Equifax	C. B. Rogers, Jr.	554 000	3
Unintrin	RC Vie	481 000	2
Berkshire Hathaway	WE Buffet	100 000	1 – Найменше значення = $100 000

* – викиди

Таблиця містить дані про n=38 фірм, отже, медіана ($1 497 500) має ранг (1+38) / 2 = 19,5 і являє собою середнє значення виплат керівникам фірм Transamerica (ранг 19) і General RE (ранг 20). Нижній квартиль ($1 000 000) має ранг (1+19) / 2 = 10 і являє собою виплати, отримані R.E. Denham з фірми Salomon. Верхній квартиль ($2 101 000) має ранг 38+1 – 10 = 29 і представляє собою виплати A.J.C. Smith з фірми March & McLennan. Нижче наведено п’ять базових показників для набору даних про розміри виплат керівникам цих 38 фірм.

Найменше значення $100 000.

Нижній квартиль $1 000 000.

Медіана $1 497 500.

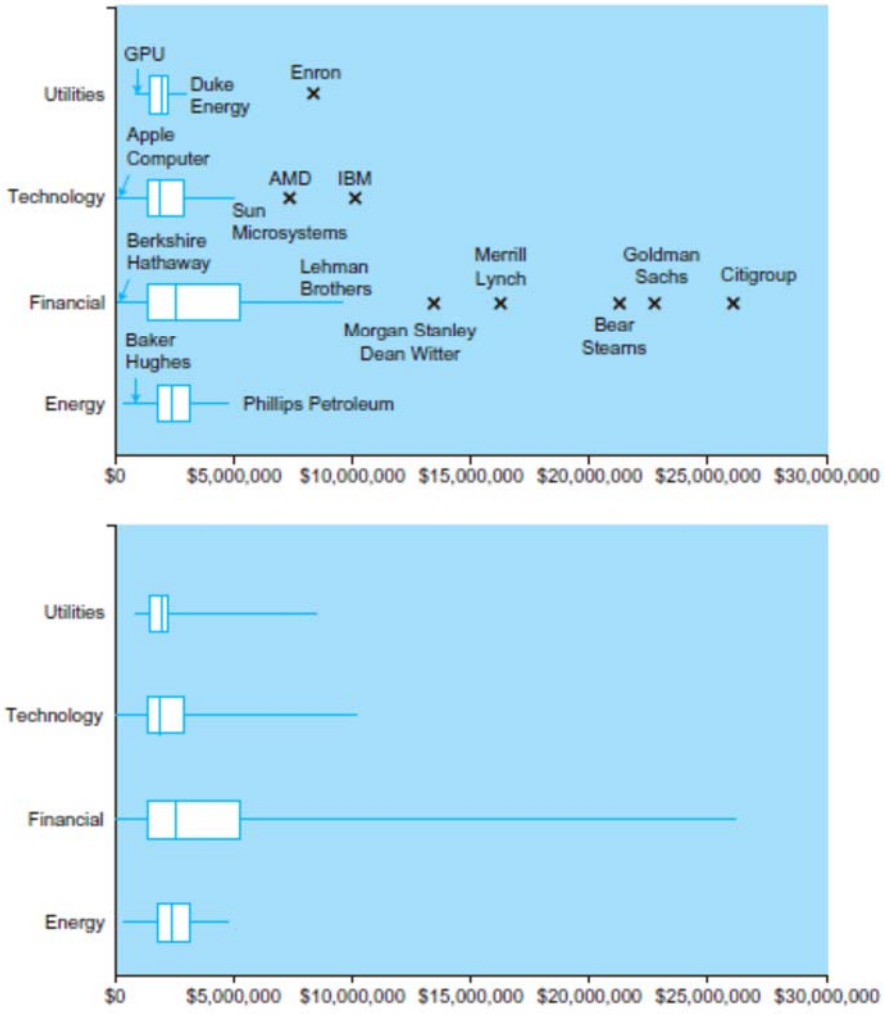
Верхній квартиль $2 101 000.

Найбільше значення $7 730 000.

Чи є серед значень викиди? Якщо розраховувати викиди з використанням квартилів, то виплати, розмір яких перевищує 2 101 000 + 1,5 * (2 101 000 – 1 000 000) = 3 752 000, будуть викидами. Таким

чином, п’ять найвищих виплат (виплачені фірмами Equitable, Bear Stearns, First Financial Mgmt., Merrill Lynch, and Travelers) є викидами в верхній частині. З іншого боку, будь-які виплати, розмір яких менше ніж 1 000 000 – 1,5 * (2 101 000 – 1 000 000) = – 651 500, також будуть викидами. Оскільки розмір найменшої виплати дорівнює 100 000, то в нижній частині розподілу викидів немає.

Блокові діаграми для аналогічних 23 фірм у 2000 році наведені на рис. 3.2.2. Хоча зазвичай використовують одну діаграму (ймовірно, з великою кількістю подробиць), ми для порівняння наводимо тут обидві діаграми.



Мал. 3.2.2. Побудовані в одному масштабі блокові діаграми розмірів виплат керівників великих фірм з деяких галузей дозволяють легко порівнювати галузі між собою. Верхній рисунок додатково містить також викиди і екстремуми, а на нижньому зображені тільки п’ять базових показників

Дані для діаграми представлено у табл. 3.2.2.

Таблиця 3.2.2. Виплати керівникам IT-компаній у 2000 році

Компанія	Керівник	Зарплата і премії, $	Ранг	П'ять базових показників
IBM *	Louis V. Gerstner Jr.	10,000,000	23	Найбільше значення $10,000,000
Advanced Micro Devices *	W.J. Sanders III	7,328,600	22	
Sun Microsystems	Scott G. McNealy	4,871,300	21	
Compaq Computer	Michael D. Capellas	3,891,000	20	
Applied Materials	James C. Morgan	3,835,800	19	
EMC	Michael C. Ruettgers	2,809,900	18	
				Верхній квартиль $2,792,350
Micron Technology	Steven R. Appleton	2,774,800	17	
Hewlett-Packard	Carleton S. Fiorina	2,766,300	16	
Motorola	Christopher B. Galvin	2,525,000	15	
National Semiconductor	Brian L. Halla	2,369,800	14	
Texas Instruments	Thomas J. Engibous	2,096,200	13	
Qualcomm	Irwin Mark Jacobs	1,723,600	12	Медіана $1,723,600
Unisys	Lawrence A. Weinbach	1,716,000	11	
Pitney Bowes	Michael J. Critelli	1,519,000	10	
NCR	Lars Nyberg	1,452,100	9	
Harris	Phillip W. Farmer	1,450,000	8	
Cisco Systems	John T. Chambers	1,323,300	7	
				Нижній квартиль $1,211,650
Lucent Technologies	Richard A. McGinn	1,100,000	6	
Silicon Graphics	Robert R. Bishop	692,300	5	
Microsoft	Steven A. Ballmer	628,400	4	
Western Digital	Matthew E. Massengill	580,500	3	
Oracle	Lawrence J. Ellison	208,000	2	
Apple Computer	Steven P. Jobs	0	1	Найменше значення $0

* – викиди

Одна з переваг блокових діаграм полягає в тому, що вони дозволяють сконцентрувати увагу на основних особливостях декількох наборів даних одночасно, не відволікаючись на деталі. Розглянемо виплати, отримані керівниками комунальних підприємств, енергетичних, IT та фінансових компаній. Тепер ми маємо чотири самостійних набори даних: по одному

одновимірному набору даних (набору значень) для кожної з чотирьох галузей. Це означає, що для кожної з галузей можна обчислити п’ять основних показників і побудувати блокову діаграму.

Розташувавши побудовані в одному масштабі блокові діаграми на одному малюнку (рис. 3.2.2), можна легко порівняти між собою типові розміри виплат керівникам в різних галузях. Прийнято розташовувати блокові діаграми вертикально, як зроблено на цьому малюнку, хоча не буде помилкою і горизонтальне розташування.

Зверніть увагу, наскільки більш інформативний верхній малюнок, що містить помічені виняткові значення виплат керівникам окремих фірм, в порівнянні з нижнім малюнком, на якому показано тільки п’ять базових показників. Хоча найвище оплачуються керівники деяких фінансових компаній (викиди), в цілому розміри виплат в цій галузі не дуже відрізняються від виплат керівникам в енергетичній сфері і у IT-галузі. З малюнка також видно, що керівникам комунальних служб, за деякими винятками, робота оплачуються нижче, ніж в інших галузях. Непогано працювати в галузі, де нижній квартиль виплат становить один мільйон доларів на рік!

Яка з діаграм краще? Є сенс витрачати час і енергію на побудову докладної блокової діаграми (з показом окремих викидів), тільки якщо це дає дійсно необхідну додаткову інформацію. Стратегічно розумно спочатку швидко нанести на діаграму п’ять базових показників, а потім вже вирішувати, чи варто витрачати час і зусилля на додаткові подробиці. Звичайно, якщо побудова діаграми виконується за допомогою комп’ютера, завжди (або майже завжди) слід віддавати перевагу докладній блоковій діаграмі.

Функція кумулятивного розподілу показує перцентилі

Функція кумулятивного розподілу даних представляється у вигляді графіка, який показує перцентилі шляхом встановлення відповідності між даними і відсотками. Оскільки на вертикальній осі відкладаються відсотки від 0% до 100%, а на горизонтальній – самі перцентилі (тобто значення даних), то, використовуючи цей графік, можна легко знаходити або значення перцентиля

при заданому значенні відсотка, або значення відсотка, що відповідає певному значенню даних.

Функція кумулятивного розподілу складається з вертикальних стрибків заввишки 1 / *n* для кожного з *n* значень даних і горизонтальних, що з'єднують точки значень даних. На рис. 3.2.3 показана функція кумулятивного розподілу для невеликого набору даних, що складається з *n* = 5 значень (1, 4, 3, 7, 3), одне з яких (3) зустрічається двічі.

Якщо задано значення і необхідно знайти його перцентильний ранг, необхідно діяти таким чином.

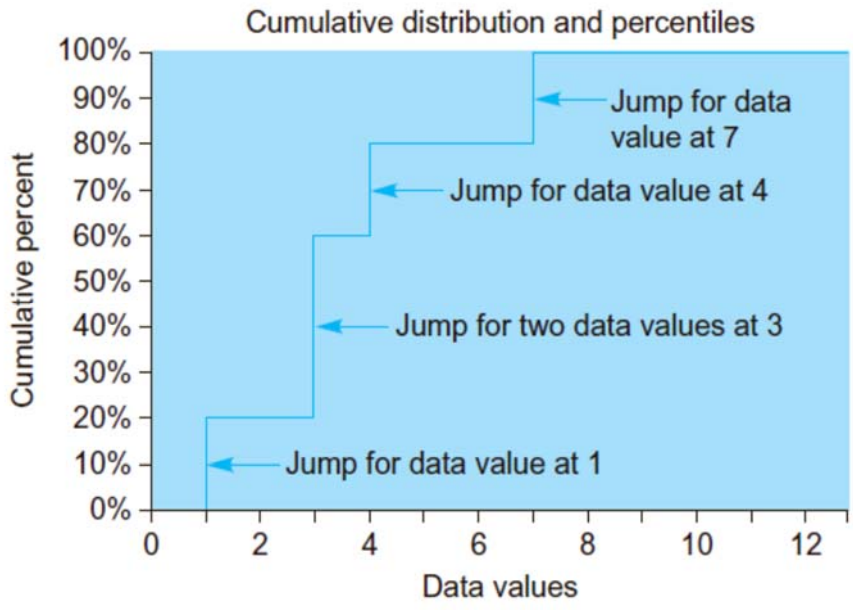
Визначення перцентильного рангу для заданого значення

1. Рухаючись по горизонтальній осі графіка функції кумулятивного розподілу, знайдіть задане значення.

2. Рухайтесь вертикально вверх до перетину з графіком функції кумулятивного розподілу. Якщо ви потрапили на вертикальну ділянку, то переміститеся вверх на його середину.

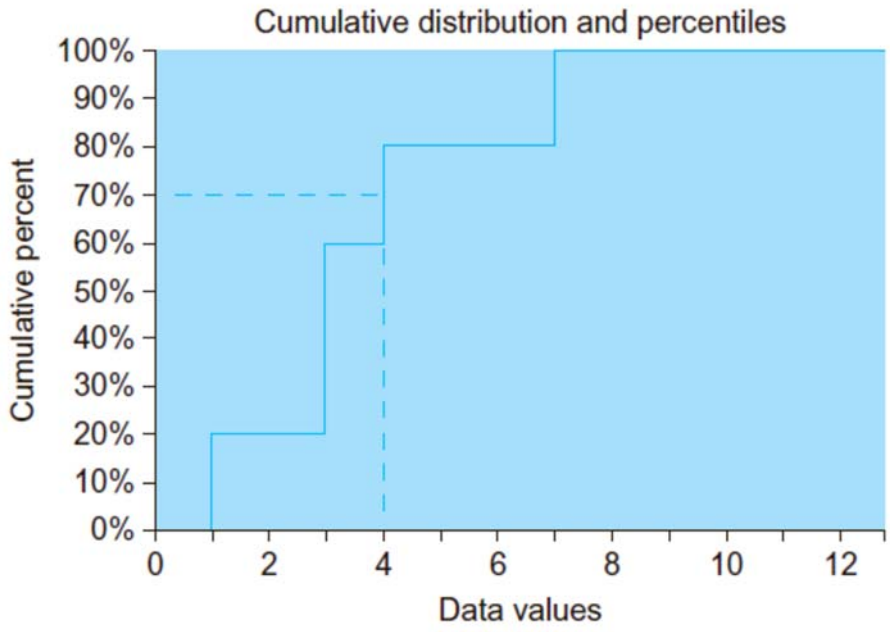
3. Рухайтесь по горизонталі вліво до перетину з вертикальною віссю, і ви отримаєте перцентильний ранг.

На цьому прикладі числу 4 відповідає 70-й процентиль, оскільки перцентильний ранг цього значення розташовано між 60 та 80% (рис. 3.2.4).



Мал. 4.2.3. Щоб знайти 44-й перцентиль, рухайтеся від 44% по горизонталі

вправо до перетину з графіком функції кумулятивного розподілу і потім вертикально вниз, де отримаєте необхідне значення 3



Мал. 3.2.4. Значення 4 представляє 70-й перцентиль. Рухайтеся вертикально вверх від значення 4, оскільки ви потрапили на вертикальну ділянку, пересіститеся вверх на середину цієї ділянки. Потім рухайтесь по горизонталі вліво до перетину з вертикальною віссю, і отримаєте результат 70%

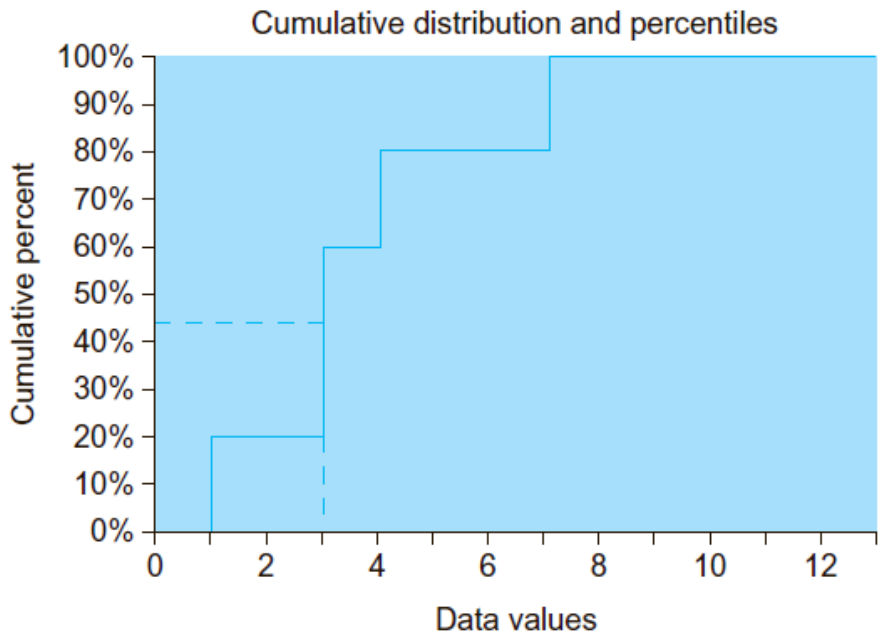
Знаходження перцентилю для заданого відсотка

1. Рухаючись по вертикальній осі графіка функції кумулятивного розподілу, знайдіть точку, що відповідає заданому відсотку.

2. Рухайтеся вправо по горизонталі до перетину з графіком функції кумулятивного розподілу. Якщо ви потрапили на горизонтальну ділянку, то перемістіться до його середини.

3. Від цієї точки рухайтеся вертикально вниз. Точка перетину з горизонтальною віссю дасть значення перцентилю.

У цьому прикладі 44-му перцентилю відповідає число 3 (рис. 3.2.5).



Мал. 3.2.5. Для знаходження 44-го перцентилю, рухайтесь від 44% по горизонталі вправо до перетину з графіком функції кумулятивного розподілу, а потім вертикально вниз, де отримаєте необхідне значення 3

Приклад. Банкрутства

Розглянемо значення показника кількості банкрутств на мільйон чоловік в окремих штатах. У табл. 3.2.3 містяться відповідні дані, впорядковані по зростанню.

На рис. 4.2.6 представлена функція кумулятивного розподілу для цього набору даних. З графіка видно, що в більшості штатів (від 10% до 90%) число банкрутств знаходиться в діапазоні від 150 до 400 банкрутств на мільйон населення.

На рис. 4.2.7 показано, як, використовуючи функцію кумулятивного розподілу, знайти перцентилі. Так, 50-й перцентиль дорівнює 260,2 (штат Гаваї, як видно з даних таблиці), що відповідає значенню медіани 260,2 банкрутства на мільйон чоловік. У той же час 90-й перцентиль дорівнює 432,4 (штат Колорадо), а 95-й дорівнює 524,4 (штат Арізона).

Для візуалізації даних ви можете вибрати будь-який з трьох графіків: гістограму, блочну діаграму або графік функції кумулятивного розподілу. Всі вони відображають одну і ту ж інформацію (значення даних), але в різному

вигляді. На рис. 4.2.8 наведені всі три типи графічного представлення даних про кількість банкрутств, що дозволяє порівняти їх між собою.

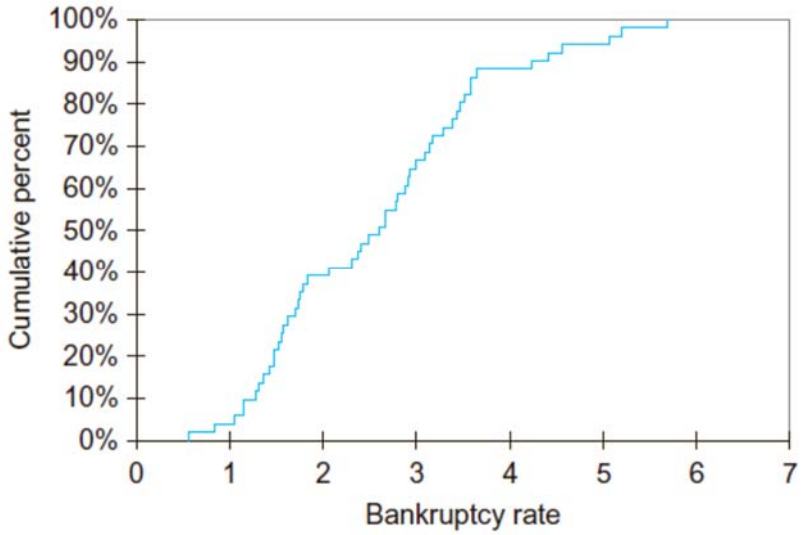
Областям високої концентрації даних (тобто тим, де знаходиться велика кількість значень) відповідають піки на гістограмі і крута функція кумулятивного розподілу. Зазвичай, як і в нашому випадку, область високої концентрації даних знаходиться в середині. Областям низької концентрації даних відповідають низькі стовпчики на гістограмі і пологі ділянки кумулятивної кривої.

Блокова діаграма містить п’ять базових показників, які можна побачити і на функції кумулятивного розподілу: найменше значення (для 0%), нижній квартиль (для 25%), медіана (для 50%), верхній квартиль (для 75%) і найбільше значення (для 100%).

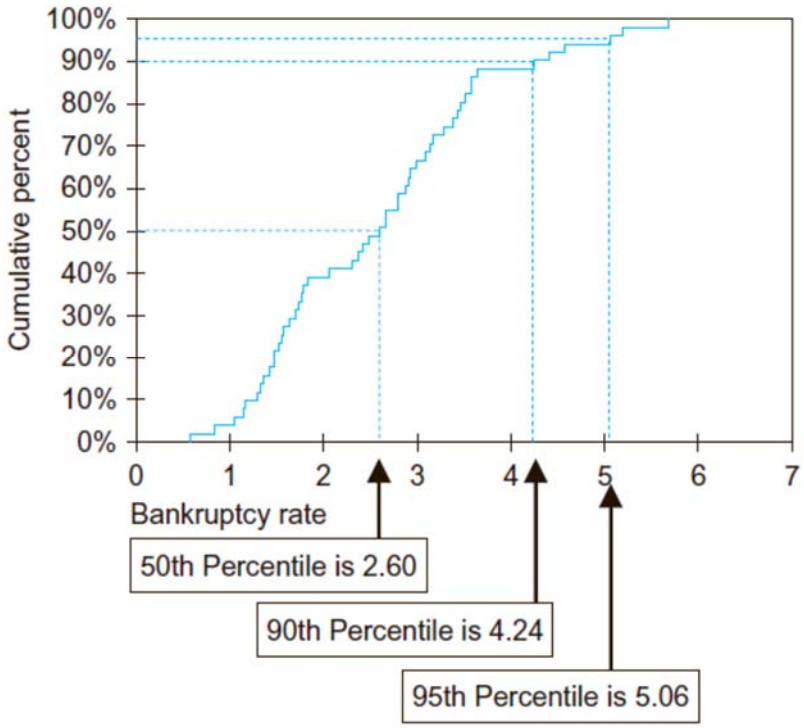
Таблиця 3.2.3. Кількість банкрутств на мільйон чоловік в окремих штатах (дані впорядковані по зростанню)

Штат	**Кількість банкрутств**	**Штат**	**Кількість банкрутств**
Арканзас	76,7	Віргінія	267,8
Південна Кароліна	107,6	Мічиган	268,6
Міссісіпі	121,8	Нью-Мексико	277,2
Луїзіана	154,6	Вермонт	300,3
Північна Кароліна	171,9	Мен	309,1
Західна Віргінія	173,1	Меріленд	310,2
Іллінойс	179,0	Айдахо	318,5
Айова	180,2	Орегон	319,6
Аляска	180,3	Коннектикут	333,5
Юта	188,7	Джорджія	339,7
Індіана	191,0	Род-Айленд	344,0
Вайомінг	191,5	Округ Колумбія	346,0
Огайо	191,8	Нью Джерсі	360,8
Делавер	195,7	Флорида	372,0
Алабама	200,9	Нью Йорк	380,1
Міннесота	203,9	Вашингтон	385,3
Монтана	206,2	Техас	393,5
Кентуккі	222,0	Невада	408,9
Північна Дакота	228,3	Канзас	422,4
Міссурі	235,0	Колорадо	432,4
Теннесі	237,1	Оклахома	445,7
Вісконсін	243,0	Массачусетс	452,4
Південна Дакота	244,8	Арізона	524,4
Небраска	248,3	Нью-Гемпшир	548,4
Пенсільванія	259,3	Каліфорнія	631,0
Гаваї	260,2		

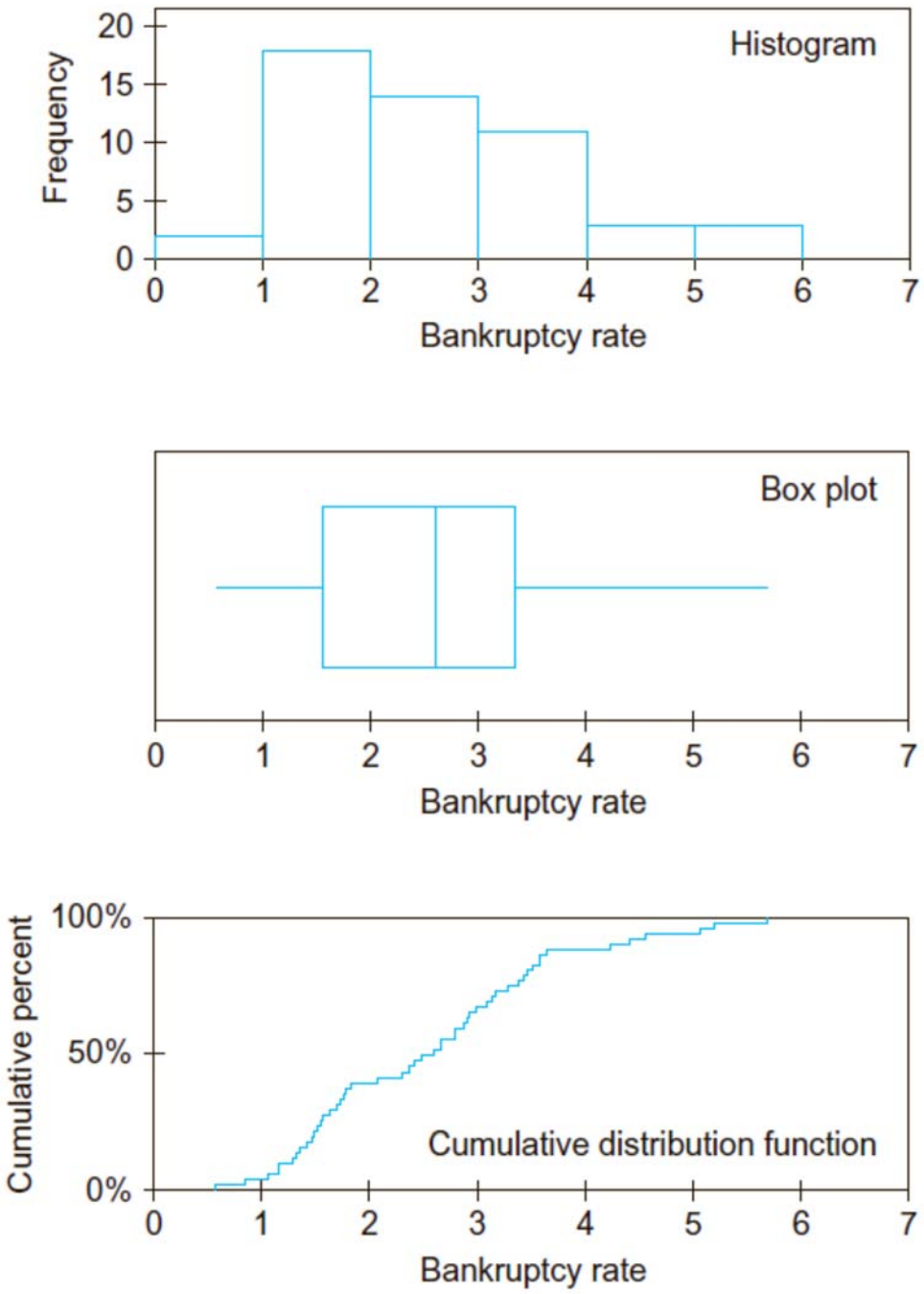
Зверніть увагу, що з представлених тут графічних зображень даних тільки функція кумулятивного розподілу містить всю інформацію про дані. При побудові гістограми частина інформації втрачається, так як гістограма відображає тільки кількість штатів в кожній з груп (наприклад, група з кількістю банкрутств від 100 до 200). При використанні блочної діаграми також втрачається частина інформації, оскільки діаграма містить тільки п’ять базових показників. І лише функції кумулятивного розподілу містять достатньо інформації для того, щоб можна було відновити кожне число вихідного набору даних.



Мал. 4.2.6. Функція кумулятивного розподілу для кількості банкрутств на мільйон осіб населення (по штатам) в 2019 році



Мал. 4.2.7. Функція кумулятивного розподілу для кількості банкрутств із зазначеними 50-м, 90-м і 95-м перцентилями



Мал. 4.2.8. Три типи графіків для даних про кількість банкрутств: гістограма, блокова діаграма і графік функції кумулятивного розподілу відповідно. Зверніть увагу, що в області високої концентрації даних функція кумулятивного розподілу круто йде вгору

4.3. Додатковий матеріал

Резюме

Узагальнення полягає в тому, щоб використовувати один або кілька відібраних або розрахованих значень для характеристики набору даних. При виконанні процедури узагальнення спочатку слід описати основну структуру більшості значень даних, а потім всі виключення або викиди значень.

Середнє є найбільш часто використовуваним показником типового значення в переліку значень даних. Обчислюють середнє шляхом складання всіх значень і діленням отриманої суми на кількість доданків. Формула розрахування середнього має наступний вигляд:

$$\text{Вибіркова середнє} = \frac{\text{Сума значень елементів даних}}{\text{Кількість елементів даних}},$$

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

Якщо мова йде про генеральну сукупність, то кількість елементів прийнято позначати N і значення середнього генеральної сукупності позначати μ (грецька буква «мю»). Середнє розподіляє загальну суму значень рівномірно між усіма спостереженнями, і використовувати його доцільно тоді, коли в даних відсутні екстремальні значення (викиди) і загальна сума значень важлива для аналізу. Середнє обчислюють тільки для кількісних даних.

Зважене середнє (середньозважене) схоже на середнє, проте цей показник дозволяє привласнити кожному елементу даних свою «вагу» (характеристику його важливості). Це дозволяє обчислювати середнє в ситуаціях, коли одні спостереження важливіші, ніж інші, а значить, повинні робити більший внесок в результат. Формула обчислення зваженого середнього (середньозваженого) має наступний вигляд:

$$\text{Зважене середнє} = \text{Сума(вага помножена на значення елементу)} =$$

$$= \omega_1 X_1 + \omega_2 X_2 + \cdots + \omega_n X_n = \sum_{i=1}^{n} \omega_i X_i$$

Зазвичай ваги вибирають так, щоб їх сума дорівнювала 1 (якщо це не так, то можна кожну вагу розділити на загальну суму ваг). Середньозважене можна обчислювати тільки для кількісних даних.

Медіана – це значення елемента, що припадає на середину сукупності; половина елементів в наборі даних більше медіани, а друга половина – менше. Ранги пов'язують числа 1, 2, 3, ... n зі значеннями даних таким чином, що найменше значення має ранг 1, наступне за величиною значення – ранг 2 і т.д. до максимального значення, яке має ранг n. Ранг медіани $(1+n) / 2$ показує, скільки спостережень слід відрахувати від найменшого (або від найбільшого) значення, щоб отримати медіану. Якщо ранг медіани виражається не цілим числом (наприклад, 13,5 для n=26), то усереднюють два значення, розташованих по обидві сторони від цього значення рангу (наприклад, в нашому випадку – значення з рангами 13 і 14). Медіану можна обчислити як для кількісних, так і для порядкових даних (впорядкованих категорій).

Медіана інакше, ніж середнє, узагальнює «типове» значення. У той же час ці два значення близькі між собою або збігаються, коли розподіл симетричний (як, наприклад, нормальний розподіл). Якщо розподіл асиметричний або містить викиди, то медіана і середнє можуть відрізнятися дуже сильно.

Мода є найбільш поширеною категорію, тобто такою, яка частіше за інших зустрічається в наборі даних. Моду можна обчислити для даних будь-якого типу: кількісних, порядкових і номінальних (невпорядкованих категорій). Для номінальних даних мода є єдиною узагальнюючою характеристикою. Для кількісних даних моду часто визначають як значення, що відповідає середині найвищого стовпчика на гістограмі. Однак таке визначення не зовсім однозначне, оскільки середина стовпчика може залежати від масштабу, в якому побудована гістограма.

Вибір узагальнюючої характеристики для конкретного набору даних необхідно здійснювати наступним чином. Для номінальних даних можна використовувати лише моду. Для порядкових даних можна використовувати і

моду, і медіану; мода представляє категорію, що найбільш часто зустрічається, а медіана вказує категорію, розташовану в центрі впорядкованого ряду значень. Для кількісних даних можна використовувати всі три показники. Якщо дані розподілені приблизно нормально, то значення всіх трьох показників близькі між собою і найкраще використати середнє. При асиметричному розподілі три показника можуть істотно відрізнятися. В цілому належний результат дає медіана, оскільки вона менш чутлива до наявності екстремальних значень в області довгого хвоста кривої розподілу. Однак, якщо важлива загальна сума значень, то краще використовувати середнє.

Перцентілі висловлюють ранги як відсотки від 0 до 100%, а не як числа від 1 до *n*; 0-й перцентіль відповідає найменшому значенню, 100-й перцентіль – максимальному значенню, 50 й – медіані і т.д. Відзначимо, що перцентіль визначений в тих же одиницях, що і значення вихідного набору даних (тобто в доларах, галонах і т.п.). Перцентілі можна використовувати для визначення значень даних при заданому перцентільному ранзі або, навпаки, для визначення перцентільного рангу по заданому значенню. Представляють також інтерес екстремуми – найбільше і найменше значення даних. Квартилі – це 25-й і 75-й перцентілі, ранги яких визначають за такими формулами:

$$\text{Ранг нижнього квартилю} = \frac{1 + int\left[\frac{1+n}{2}\right]}{2};$$

$$\text{Ранг нижнього квартилю} = n + 1 - \text{ранг нижнього квартилю},$$

де *int* – функція взяття цілого значення, яка відкидає дробову частину числа.

П’ять базових показників набору даних включають найменше і найбільше значення, нижній і верхній квартилі і медіану. На блоковій діаграмі ці п’ять показників зображені в графічній формі. Викиди визначаються як такі точки даних (якщо вони є), значення яких лежать далеко від тих значень, які знаходяться в середній частині набору даних. Детальна блокова діаграма містить значення викидів з відповідними позначками, а також найбільш екстремальні з тих спостережень, які не є викидами. Для порівняння декількох

наборів даних, виміряних в однакових одиницях, можна, використовуючи один масштаб для кожного з них, побудувати блокову діаграму і розташувати ці діаграми на одному малюнку.

Функція кумулятивного розподілу даних представляється у вигляді графіку, який показує перцентілі шляхом встановлення відповідності між даними і відсотками. Цей графік має вертикальний стрибок величиною 1 / *n* для кожного з *n* значень даних. Знаючи відсоток, можна знайти перцентіль, рухаючись по графіку вправо, а потім вниз. Знаючи значення, можна визначити перцентільний ранг (відсоток), рухаючись по графіку вгору і потім вліво. Таким чином, функція кумулятивного розподілу відображає перцентілі і дозволяє їх обчислити. Це єдина графічна форма представлення даних, яка «архівує» дані, зберігаючи достатньо інформації для відновлення всіх значень набору даних. Функція кумулятивного розподілу круто зростає в ділянках високої концентрації даних (там, де високі стовпчики на гістограмі).

Основні терміни

- Узагальнення (summarization)
- Усереднення (average)
- Середнє (mean)
- Виважене середнє (weighted average)
- Медіана (median)
- Ранг (rank)
- Мода (mode)
- Перцентіль (percentile)
- Екстремуми (extremes)
- Квартилі (quartiles)
- П’ять базових показників (five-number summary)
- Блокова діаграма (box plot)
- Детальна блокова діаграма (detailed box plot)
- Викид (outlier)
- Функція кумулятивного розподілу (cumulative distribution function)

Контрольні питання

1. Що являє собою процес узагальнення набору даних? Чому так важливо узагальнювати дані?
2. Перелічіть і коротко опишіть різні показники, що узагальнюють дані.
3. Що необхідно робити з винятками при узагальненні набору даних?
4. Що означає типове значення для переліку чисел? Назвіть три різні способи визначення типового значення.
5. Що таке середнє? Поясніть середнє з точки зору суми всіх значень набору даних.
6. Що таке зважене середнє? У яких випадках цей показник використовують замість звичайного середнього?
7. Що таке медіана? Як можна знайти медіану виходячи з її рангу?
8. Як знайти медіану для набору даних:

а) з парною кількістю значень?

б) з непарною кількістю значень?

9. Що таке мода?
10. Як зазвичай визначають моду для кількісного набору даних? Чому таке визначення містить неоднозначність?
11. Який узагальнюючий показник (або показники) можна використовувати для:

а) номінальних даних?

б) порядковий даних?

в) кількісних даних?

12. Які показники краще використовувати при:

а) нормальному розподілі даних?

б) при плануванні загальної кількості (суми)?

в) при асиметричному розподілі, коли загальна сума не важлива?

13. Що таке перцентіль? Зокрема, чи є він відсотком (наприклад, 23%) або виражений в тих же одиницях, що і дані (наприклад, $35,62)?
14. Назвіть два способи використання перцентілей.

15. Що таке квартилі?

16. Назвіть п’ять базових характеристик розподілу.

17. Що таке блокова діаграма? Які деталі часто додатково зображають на блокової діаграмі?

18. Що таке викид (значення, що сильно відхиляється)? Як можна визначити, чи є дана точка викидом?

19. Розгляньте функцію кумулятивного розподілу:

а) що вона собою являє?

б) як її накреслити?

в) для чого її використовують?

г) порівняйте її з гістограмою і блоковою діаграмою.