ЛЕКЦІЯ.
ДОСЛІДЖЕННЯ РІЗНОМАНІТНОСТІ ДАНИХ

Одна з причин, у зв'язку з якою виникає необхідність проведення статистичного аналізу, полягає в тому, що дані мінливі. Якби дані не змінювалися, то відповіді на багато питань були б просто очевидними і нам не потрібно було б звертатися до методів статистичного аналізу. Ситуація, в якій присутня мінливість, часто містить ризик, оскільки навіть використання всієї доступної інформації не дозволяє точно передбачити, що ж відбудеться в майбутньому. Для адекватної роботи з ризиком необхідно розуміти його природу і вміти вимірювати мінливість (часто також використовують термін «варіація»), яка є його наслідком. Наведемо кілька ситуацій, в яких мінливість має важливе значення.

Випадок перший. Розглянемо мінливість продуктивності праці працівників. Цілком очевидно, що ефективність роботи відділу визначається загальною продуктивністю праці всіх його співробітників. Однак будь-які зусилля, спрямовані на підвищення продуктивності праці, повинні враховувати індивідуальні особливості працівників. Наприклад, деякі програми підвищення продуктивності праці можуть бути орієнтовані на всіх працівників, в той час як інші – приділяти особливу увагу самим «швидким» або самим «повільним» з них. Визначення мінливості продуктивності праці дає можливість виявити розкид таких індивідуальних відмінностей і отримати корисну інформацію для планування заходів щодо підвищення загальної продуктивності праці.

Деякі фахівці в області статистики в приватних бесідах відзначають, що саме наявність мінливості даних забезпечує їм роботу!

Випадок другий. Фондова біржа в середньому забезпечує більш високу прибутковість вкладених коштів, ніж, наприклад, фонди грошового ринку. Однак робота на фондовій біржі пов'язана з великим ризиком, і інвестування в акції може привести до реальних втрат. Таким чином, середня, або

«очікувана», прибутковість не відображає повністю всю картину. Міра мінливості дохідності окремих інвестицій буде відображати рівень ризику, пов'язаного з кожним конкретним вкладенням коштів.

Випадок третій. Припустимо, що ви порівнюєте маркетингові витрати своєї фірми з аналогічними витратами фірм, що працюють у вашій галузі промисловості, і виявляєте, що витрати вашої фірми менше витрат, типових для даної галузі. Для того щоб оцінити витрати на майбутнє, дуже корисним може виявитися облік розкиду відповідних даних по галузі. Знайшовши різницю між значенням витрат своєї фірми і середнім значенням по галузі і порівнявши отриману величину з мірою мінливості витрат в галузі, можна зробити висновок про те, чи знаходиться маркетингова діяльність вашої фірми в порівнянні з іншими аналогічними фірмами лише на дещо нижчому рівні або ж ваша фірма є деяким винятком із загальної картини. Така інформація може допомогти в стратегічному плануванні витрат на маркетинг в наступному році.

Мінливість можна визначити як ступінь відмінностей між окремими значеннями. Подібний сенс мають також такі поняття, як різноманітність, невизначеність, розсіювання і розкид. Далі ми побачимо, що існують три різні способи опису ступеню мінливості набору даних, причому кожен з них вимагає відповідних числових значень.

1. *Стандартне відхилення* (в україномовній літературі за статистикою часто також використовують терміни «середньоквадратичне відхилення» і «середнє квадратичне відхилення») використовують найчастіше. Цей показник описує, наскільки сильно результат спостережень зазвичай відрізняється від середнього значення. При зведенні стандартного відхилення в квадрат отримуємо *дисперсію*.

2. *Розмах* легко обчислюється, проте дає кілька поверхневе уявлення про мінливість даних і має обмежене застосування. Ця величина описує межі зміни даних в наборі і являє собою відстань між мінімальним і максимальним значеннями.

3. *Коефіцієнт варіації* зазвичай обирається в якості відносної (на противагу абсолютній) міри мінливості. Цей показник використовується досить часто. Він показує, наскільки сильно зазвичай відрізняється результат конкретного спостереження від середнього значення, в процентному відношенні до середнього; при цьому використовується відношення стандартного відхилення до середнього значення.

Ми також розглянемо, як впливає на мінливість даних зміна шкали вимірювання (наприклад, перехід від японської єни до доларів США або перехід від кількості одиниць випущеної продукції до грошової вартості цієї продукції).

4.1. Стандартне відхилення: традиційний вибір

Стандартне відхилення – це число, яке описує, наскільки значення даних зазвичай відрізняються від середнього. Поняття стандартного відхилення є дуже важливим в статистиці, оскільки воно являє собою основний інструмент визначення ступеня випадковості в досліджуваній ситуації. Зокрема, цей показник є мірою випадковості відхилень окремих значень від їх середнього.

Якщо всі величини однакові, як, наприклад, у наведеному нижче простому наборі даних

5,5; 5,5; 5,5; 5,5

то середнє матиме значення $\overline{X} = 5{,}5$, а стандартне відхилення складе $S = 0$. Останнє відображає той факт, що в цьому тривіальному наборі дані не схильні до мінливості.

У реальному житті більшість даних характеризується більшою або меншою мірою мінливості. Окремі значення набору даних розташовуються на деякій відстані від середнього, а стандартне відхилення характеризує ступінь мінливості. Розглянемо тепер інший набір даних, яким властива деяка мінливість:

43,0; 17,7; 8,7; –47,4

Ці числа є значеннями ставки прибутковості (наприклад, 43%) акцій чотирьох компаній (Maytag, Boston Scientific, Catalytica і Mitcham Industries), обраних випадковим чином (випадковість забезпечувалася шляхом метання стріли гри дарт в газетну сторінку з котируваннями акцій). Середнє значення в цьому випадку таке ж, $\overline{X} = 5{,}5$, тобто ці акції мають середню ставку прибутковості 5,5% (це означає, що портфель рівних в грошовому вираженні інвестицій в названі вище акції матиме цю середню прибутковість 5,5%). Незважаючи на те що середнє значення тут таке ж, як і в попередньому випадку, окремі значення даних істотно розрізняються між собою. Перша величина, 43,0, розташовується на відстані $X_1 - \overline{X} = 43{,}0 - 5{,}5 = 37{,}5$ від середнього значення. З цього випливає, що ставка прибутковості акцій Maytag перевищує середню ставку прибутковості на 37,5%. Останнє значення, –47,4, розташоване від середнього на відстані $X_4 - \overline{X} = -47{,}4 - 5{,}5 = -52{,}9$; таким чином, ставка прибутковості акцій Mitcham Industries виявляється на 52,9% нижче середнього рівня (нижче – оскільки величина негативна). У табл. 4.1.1 показано, наскільки кожне з значень відрізняється від середнього.

Таблиця 4.1.1. Розрахунок відхилень від середнього значення

Компанія	Ставка прибутковості	Відхилення від середнього
Maytag	43,0	$43{,}0 - 5{,}5 = 37{,}5$
Boston Scientific	17,7	$17{,}7 - 5{,}5 = 12{,}2$
Catalytica	8,7	$8{,}7 - 5{,}5 = 3{,}2$
Mitcham Industries	–47,4	$-47{,}4 - 5{,}5 = -52{,}9$
Сума	**22,0**	
Середнє	**5,5**	

Описані вище відстані від середнього значення називаються відхиленнями, або різницею. Вони показують, наскільки вище середнього значення (в разі позитивної різниці) або нижче середнього (якщо різниця негативна) лежить кожне значення даних. Відхилення в свою чергу утворюють набір даних, розташованих навколо нуля, що схоже на вихідний набір даних, значення в якому розташовані навколо середнього.

В якості узагальнюючої характеристики відхилень використовують стандартне відхилення. Просто усереднити відхилення не можна, оскільки частина з них виявиться негативними, а частина – позитивними, в результаті чого результат такого усереднення завжди буде дорівнює нулю і не буде містити ніякої додаткової інформації. Замість цього використовують стандартний прийом, що полягає в тому, що кожне значення спочатку зводять в квадрат (тобто його множать на себе), щоб позбутися від знака «мінус», потім складають, ділять на n–1 і витягають квадратний корінь (це зворотна операція по відношенню до виконаного раніше зведення в квадрат).

Визначення і формула для стандартного відхилення і дисперсії

Стандартне відхилення визначається як величина, яка обчислюється таким чином. Зверніть увагу на те, що при цьому обчислюється також дисперсія (квадрат стандартного відхилення). Дисперсію іноді використовують в якості міри мінливості в статистиці, особливо коли працюють безпосередньо з формулами. Однак часто в якості міри мінливості краще брати стандартне відхилення. Дисперсія не несе ніякої додаткової (у порівнянні зі стандартним відхиленням) інформації, і в той же час, в практичних застосуваннях її складніше інтерпретувати, ніж стандартне відхилення. Так, наприклад, під час набору даних, що містить витрачені суми грошей (виміряні в гривнях), дисперсія буде виражатися в «гривня в квадраті», – це одиниця виміру, яку важко собі уявити; в той же час стандартне відхилення для цього набору даних буде виражено в звичних для всіх гривнях.

Обчислення стандартного відхилення для вибірки

Для того щоб знайти стандартне відхилення для вибірки, необхідно виконати наступні дії.

1. Знайти відхилення, віднімаючи з кожного значення середнє.
2. Звести отримані значення в квадрат, скласти і розділити отриману суму на n–1. Результатом буде дисперсія.
3. Витягти з отриманого значення квадратний корінь. Це і буде стандартне відхилення.

У табл. 4.1.2 описана вище процедура проілюстрована на прикладі акцій компаній, обраних вище випадковим чином. В результаті розподілу суми зведених в квадрат відхилень, 4363,74, на *n*–1 отримуємо дисперсію 4363,74 / 3 = 1454,58. Витягуючи квадратний корінь, одержуємо стандартне відхилення 38,14. Це значення дійсно є розумним описом власне відхилень (якщо не враховувати знаки і розглядати в першу чергу величину відхилень).

Таблиця 4.1.2. Обчислення суми квадратів відхилень, дисперсії і стандартного відхилення

Компанія	**Ставка прибутковості**	**Відхилення від середнього**	**Відхилення в квадраті**
Maytag	43,0	$43{,}0 - 5{,}5 = 37{,}5$	1406,25
Boston Scientific	17,7	$17{,}7 - 5{,}5 = 12{,}2$	148,84
Catalytica	8,7	$8{,}7 - 5{,}5 = 3{,}2$	10,24
Mitcham Industries	–47,4	$-47{,}4 - 5{,}5 = -52{,}9$	2798,41
Сума	**22,0**	**0**	**4363,74**
Середнє	**5,5**		
Дисперсія	$\mathbf{\frac{4363,74}{4-1} = 1454,58}$		
Стандартне відхилення	$\mathbf{\sqrt{1454,58} = 38,14}$		

Формула для обчислення стандартного відхилення є коротким математичним записом описаної вище процедури. Стандартне відхилення для вибірки даних позначається буквою S, і формули обчислення стандартного відхилення і дисперсії мають наступний вигляд.

Стандартне відхилення для вибірки

$$S = \sqrt{\frac{\text{Сума квадратів відхилень}}{\text{Кількість елементів вибірки} - 1}} =$$

$$\sqrt{\frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

Дисперсія для вибірки

$$\text{Дисперсія} = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Обчислення стандартного відхилення для нашого простого прикладу з використанням відповідної формули дає той же результат

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = \sqrt{\frac{4363,74}{4-1}} = \sqrt{1454,58} = 38,14$$

Використання калькулятора або комп'ютера

Звичайно, існує інший, значно простіший спосіб обчислення стандартного відхилення: використовувати калькулятор або комп'ютер. Саме так дійсно обчислюють стандартне відхилення, покладаючи проведення всіх обчислень на спеціально призначений для цього електронний пристрій. Однак дії, з якими ми ознайомилися вище, і наведені вище формули для нас все одно важливі, оскільки вони забезпечують розуміння тих чисел, які обчислюються автоматично. Так, при інтерпретації стандартного відхилення важливо розуміти, що це типова (чи стандартна) величина відхилення.

Вибір одного з багатьох різних способів розрахунку стандартного відхилення за допомогою комп'ютера залежить від того, яке програмне забезпечення використовується: електронні таблиці, програми для роботи з базами даних, компілятори мов програмування або спеціальні програми для статистичного аналізу даних.

Інтерпретація стандартного відхилення

Стандартне відхилення має просту і зрозумілу інтерпретацію: ця величина описує типову відстань від середнього значення для окремих значень набору даних. Таким чином, стандартне відхилення виступає в якості міри мінливості для цих окремих значень. Оскільки стандартне відхилення відображає типову величину відхилення, то можна припустити, що для одних значень відхилення буде менше, ніж стандартна, а для інших – більше. Таким чином, ми можемо очікувати, що для деяких значень їх відстані від середнього будуть менше стандартного відхилення, в той час як для інших значень ця відстань буде перевищувати величину стандартного відхилення.

На рис. 4.1.1 показано, як можна зобразити стандартне відхилення у вигляді відстані від середнього значення. Оскільки середнє показує центр всього набору даних, окремі значення будуть, мабуть, розташовуватися по обидві сторони від середнього.

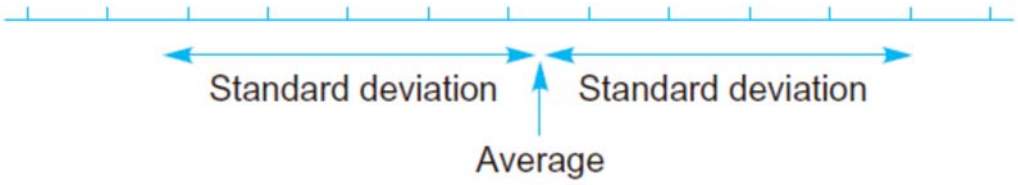

Рис. 4.1.1. Числова вісь з показаними на ній середнім значенням (average) і стандартним відхиленням (standard deviation). Зверніть увагу на те, що середнє – це деяка фіксована точка на числовій осі (вона показує абсолютну величину), а стандартне відхилення задає певну відстань, а саме типову відстань від середнього значення

Приклад. Витрати на рекламу

Припустимо, що ваша фірма витрачає 19 мільйонів доларів в рік на рекламу і керівництво фірми бажає знати, чи відповідає ця сума реальним потребам. Незважаючи на те що існує досить багато способів оцінки цієї стратегічно важливої величини, завжди корисно порівняти себе з конкурентами. Припустимо інші фірми, що працюють у вашій галузі та мають приблизно такий же розмір, в середньому витрачають на рік на рекламні цілі 22,3 мільйона доларів. Ви можете скористатися стандартним відхиленням для того, щоб виходячи з різниці (22,3 – 19 = 3,3 мільйона доларів) оцінити, наскільки витрати на рекламу вашої фірми менше, ніж в інших аналогічних фірмах.

Розглянемо витрати на рекламу (в мільйонах доларів) групи з $n = 17$ фірм, схожих на вашу:

8; 19; 22; 20; 27; 37; 38; 23; 12; 11; 23; 20; 18; 23; 35; 11.

Легко переконатися, що середнє складає 22,3 мільйона доларів (результат округлення величини 22,29411 мільйона доларів) і стандартне

відхилення дорівнює 9,18 мільйона доларів (результат округлення значення 9,177177).

Оскільки різниця між витратами на рекламу у вашій фірмі і середніми витратами на рекламу в групі фірм (3,3 мільйона доларів) навіть менше одного стандартного відхилення (9,18 мільйона доларів), то можна зробити висновок, що бюджет рекламної діяльності вашої фірми досить типовий. Незважаючи на те, що він менше середнього значення, він ближче до цього середнього, ніж бюджет типової фірми з даної групи.

Уявімо положення вашої фірми в групі більш наочно.

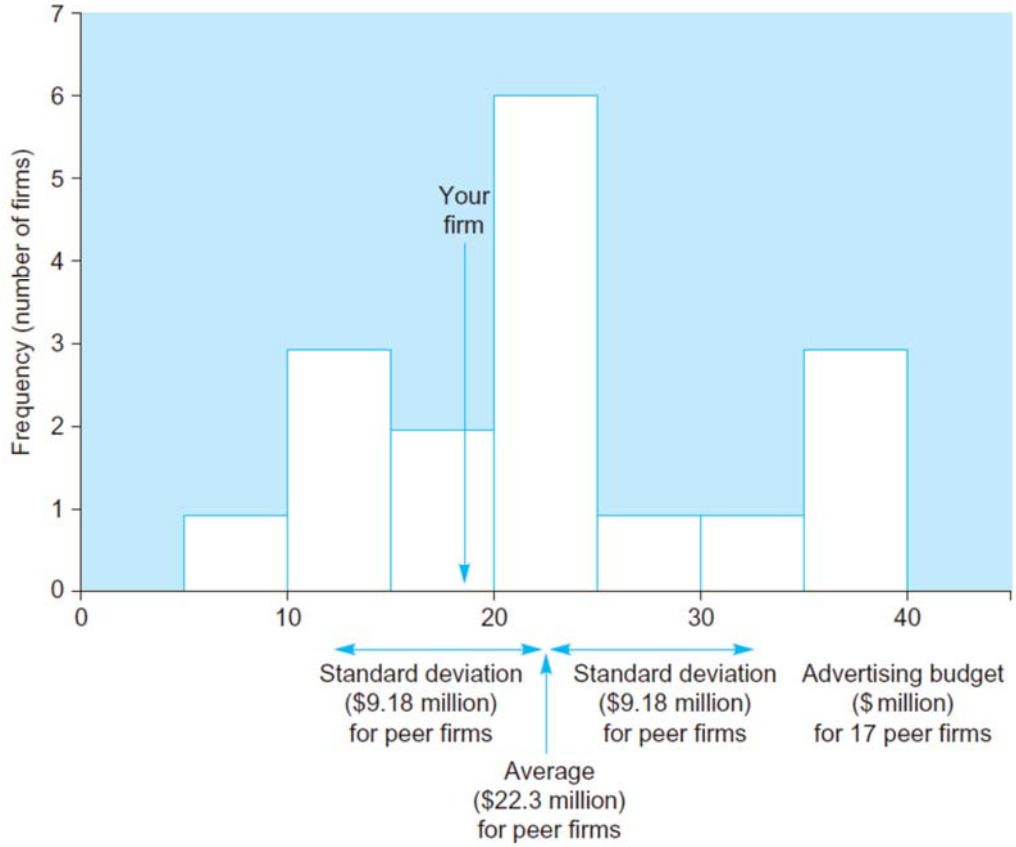

Рис. 4.1.2. Гістограма розмірів витрат на рекламу для групи 17 подібних фірм. Показані середнє значення і стандартне відхилення. Бюджет вашої фірми, що становить 19 мільйонів доларів, досить типовий характер у порівнянні з аналогічними фірмами. Він відрізняється від середнього навіть менше, ніж на одну величину стандартного відхилення

На рис. 4.1.2 ви бачите гістограму розмірів витрат на рекламу фірм даної групи. На цій гістограмі наведено середнє значення і стандартне відхилення (зверніть увагу на те, наскільки дійсно ефективно стандартне відхилення відображає величину розкиду даних по обидва боки від середнього значення). Ваша фірма з бюджетом реклами в 19 мільйонів доларів дійсно виявляється досить типовою. Незважаючи на те що різниця в 3,3 мільйона доларів між бюджетом вашої фірми і середнім значенням здається в грошовому вираженні досить значною, вона невелика в порівнянні з індивідуальними розходженнями, що існують між бюджетами фірм, що входять в дану групу. З точки зору обсягу бюджету реклами положення вашої фірми не набагато нижче середнього.

Приклад. Облік відмінностей між клієнтами

Ваші клієнти відрізняються один від одного; між ними існують відмінності в розмірах замовлень, виборі різних товарів, циклічності роботи протягом року, потреби в інформації, прихильності роботі з вами і т.п. Однак у вас, мабуть, є певне судження про «типового клієнта», а також деяке уявлення про ступінь відмінностей між клієнтами.

Ви можете узагальнити інформацію про заявки клієнтів за допомогою середнього і стандартного відхилення.

Загальний річний обсяг замовлення на одного клієнта	
Середнє значення	$85 600
Стандартне відхилення	$28 300

Таким чином, за останній рік кожен клієнт в середньому замовив товарів на суму $85 600. Як характеристика відмінностей між клієнтами виступає стандартне відхилення. Його величина в $28 300 показує, що зазвичай ваші клієнти робили замовлення на суми, менші або більші приблизно на $ 28300, ніж середнє значення в $85 600. Слово «приблизно» має тут велике значення: деякі клієнти можуть робити замовлення, розмір яких дуже близький до середнього рівня, в той час як для інших спостерігатимуться відмінності, що значно перевищують $28 300. Середнє значення показує типовий обсяг

замовлень, що надійшли протягом року від одного клієнта, а стандартне відхилення ілюструє типове відхилення від середнього.

Зверніть увагу також на те, що стандартне відхилення вимірюється в тих же одиницях, що і середнє значення; в даному випадку обидві ці величини визначені в доларах. Якщо говорити точніше, то одиниця вимірювання тут – це «долар на рік на одного клієнта». Це пов'язує одиниці виміру з вихідним набором даних, який представляє собою послідовність обсягів «доларів в рік», причому кожному клієнту відповідає одне число.

Інтерпретація стандартного відхилення для нормального розподілу

У тому випадку, коли набір даних має приблизно нормальний розподіл, стандартне відхилення набуває особливого змісту. Приблизно дві третини значень з такого набору даних знаходяться в межах одного стандартного відхилення по обидва боки від середнього значення, як показано на рис. 4.1.3.

Так, наприклад, якщо здатності ваших працівників розподілені приблизно нормально, то ви можете очікувати, що оцінки здібностей приблизно двох третин з них потрапляють на відстані не більше однієї величини стандартного відхилення від середнього значення – або вище, або нижче середнього. Це означає, що приблизно третина працівників має здатності, що лежать в межах однієї величини стандартного відхилення вище середнього, а приблизно третина – у відповідній області нижче середнього. Решта працівників, яких також приблизно третина, розподіляться таким чином: близько половини цієї однієї третини (шоста частина всіх працівників) має здібності, що перевищують середнє більш ніж на величину одного стандартного відхилення, і приблизно шоста частина всіх працівників (на жаль!) виявиться нижче середнього далі, ніж на відстані одного стандартного відхилення.

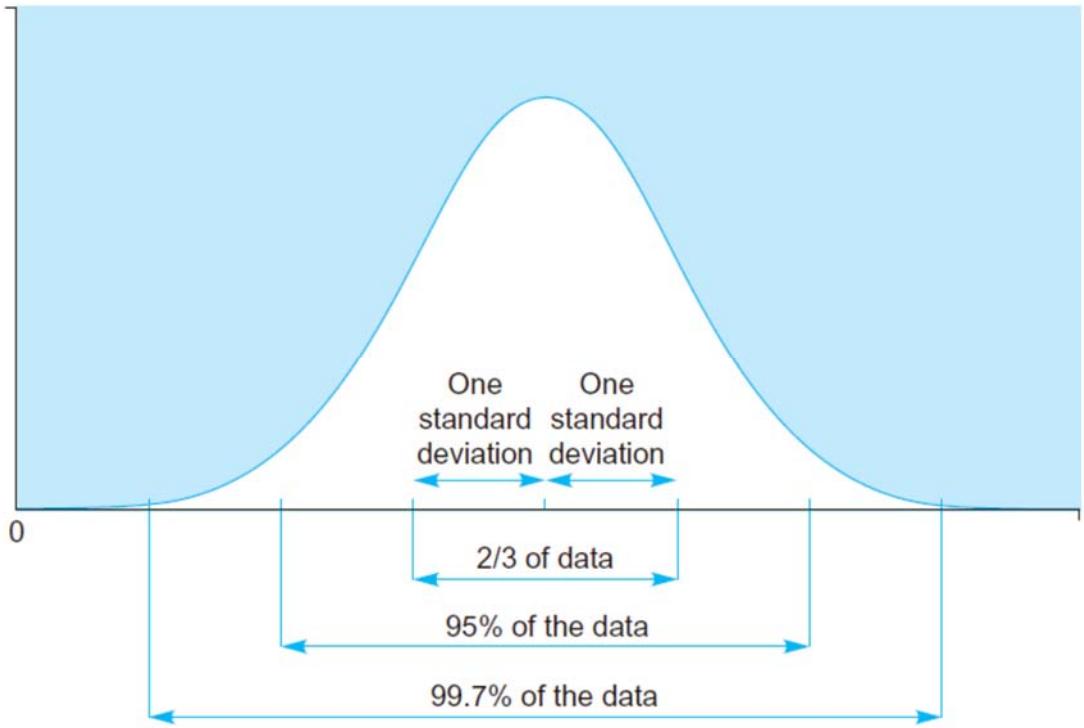

Рис. 4.1.3. У разі нормального розподілу значення в наборі даних можна легко розділити в залежності від вираженого в значеннях стандартного відхилення їх відстані від середнього значення. Приблизно дві третини всіх значень знаходяться не далі одного стандартного відхилення від середнього. Близько 95% всіх значень знаходяться в межах двох стандартних відхилень від середнього. І, нарешті, ми можемо очікувати, що виявимо майже всі дані (99,7%) на відстані не більше трьох стандартних відхилень від середнього

З рис. 4.1.3 також видно, що в разі нормального розподілу слід очікувати, що приблизно 95% всіх даних виявляться в межах двох величин стандартного відхилення від середнього значення[1]. Цей факт буде мати велике значення при розгляді статистичних висновків, оскільки допустимі похибки оцінок часто обмежуються величиною 5%.

І, нарешті, ми можемо припустити, що майже всі дані (99,7%) будуть знаходитися в межах трьох величин стандартного відхилення від середнього значення. При цьому тільки 0,3% всіх значень набору даних виявляються від

[1] У разі ідеального нормального розподілу в точності 95% всіх даних потрапляють в область поблизу середнього значення в межах 1,96 стандартного відхилення. Оскільки величина 1,96 досить близька до значення 2, ми використовуємо опис «дві величини стандартного відхилення» в якості зручного і гарного наближення.

середнього на більшій відстані. На рис. 4.1.3 можна бачити, що графік нормального розподілу на відстанях близько трьох стандартних відхилень від середнього опускається майже до нуля. У картах контролю, які широко використовують для контролю якості продукції, межі часто встановлюються таким чином, щоб в якості проблеми, що заслуговує на увагу, виступали саме ті результати спостережень, які знаходяться від середнього на відстані, більшій, ніж три стандартних відхилення.

Що ж відбувається в тому випадку, якщо набір даних не підпорядковується нормальному розподілу? В такому випадку описані вище відсотки застосовувати неможливо. На жаль, оскільки існує безліч скошених (або інших, що відрізняються від нормального) розподілів, можна вказати єдине правило визначення таких відсотків для довільного розподілу.

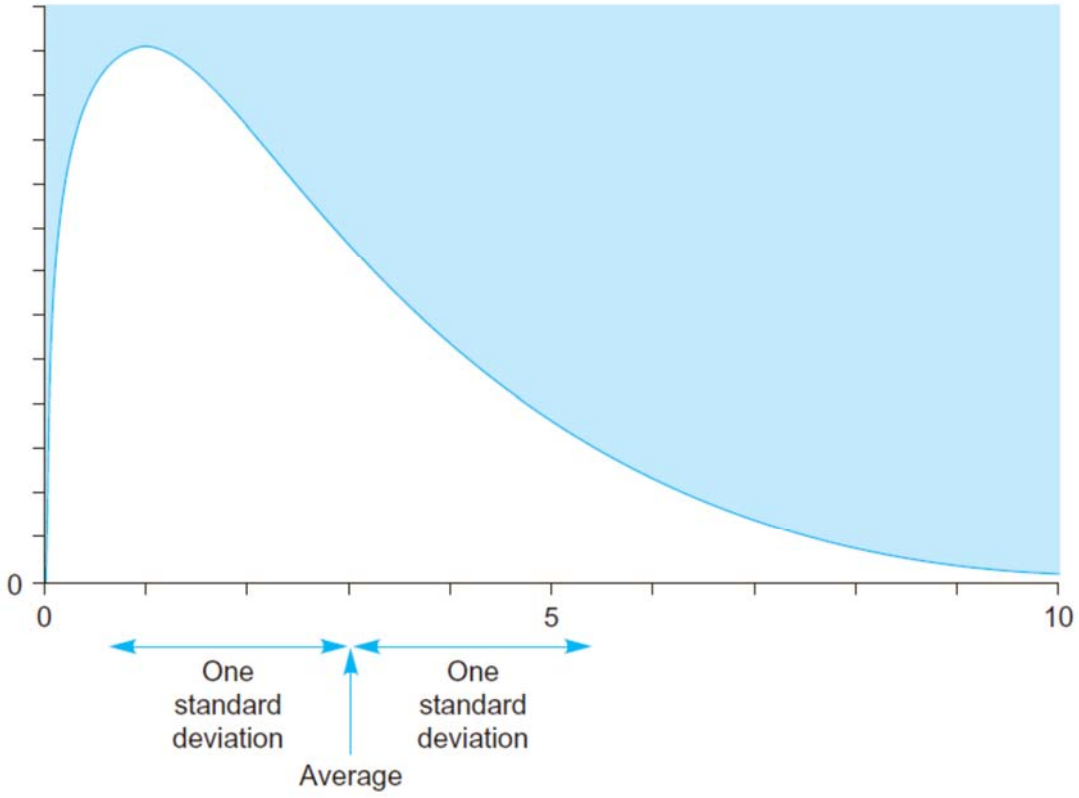

Рис. 4.1.4. У разі скошеного розподілу не існує простих правил для визначення частини даних, що потрапляють в межі одного (двох або трьох) стандартних відхилень від середнього значення

На рис. 4.1.4 наведено приклад скошеного (асиметричного) розподілу. У цьому випадку на відстані, що не перевищує одне стандартне відхилення, знаходиться не дві третини, а три чверті всіх даних. Крім того, більшість цих даних розташована зліва від середнього (оскільки тут графік розподілу проходить вище).

Приклад. Контрольні карти для контролю якості зображень

Підприємство випускає екрани моніторів і використовує для контролю і поліпшення якості продукції карти контролю якості (або, як ще кажуть, контрольні карти). Зокрема, розмір однієї точки екрану повинен бути настільки малий, щоб користувач міг бачити найдрібніші деталі зображення. Карта контролю містить результати окремих вимірювань розміру точки (які відрізняються для різних моніторів), середні значення (які, як можна бачити, проходять через центр даних) і контрольні кордони (які встановлюються вище і нижче середнього значення на відстані трьох стандартних відхилень).

На рис. 4.1.5 показаний приклад карти контролю якості для набору пристроїв, у яких результати всіх вимірювань знаходяться в межах контрольних меж. На рис. 4.1.6 наведено інший приклад, який демонструє, що у монітора №22 спостерігається вихід за контрольні межі. Карти контролю якості допомагають виявити проблему. Подальше дослідження і виправлення ситуації залежить від менеджера.

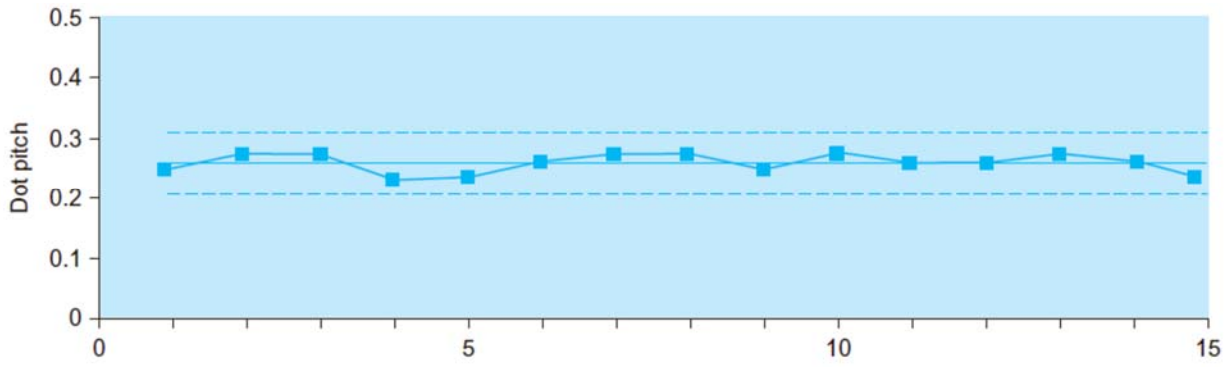

Рис. 4.1.5. Карта контролю якості з результатами вимірювань для екранів моніторів. Лінії нижньої і верхньої контрольних меж проведені на відстані трьох стандартних відхилень. Показано також середнє значення, яке

проходить через центр даних. Система знаходиться під контролем, і є тільки випадкові відхилення від середнього, оскільки в відхиленнях немає чітких тенденцій і результатів вимірювань, які виходять за межі контрольних меж

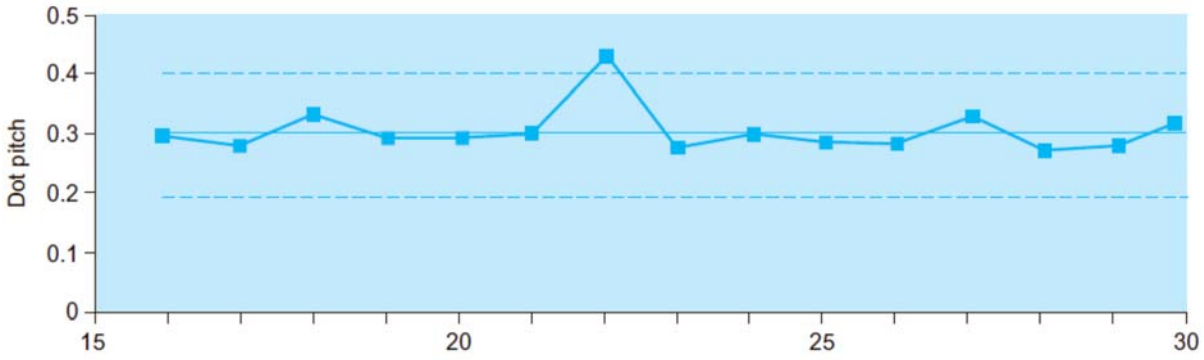

Рис. 4.1.6. Ця система вийшла з-під контролю. Зверніть увагу на те, що результат вимірювання для екрану №22 відхиляється від середнього значення більш ніж на три стандартних відхилення. Це виходить за межі звичайної варіації для системи, виникає необхідність у вивченні причин і усунення подібних проблем в подальшому

Стандартне відхилення вибірки і генеральної сукупності

Існують два різних (проте пов'язаних між собою) виду стандартного відхилення: стандартне відхилення вибірки (для вибірки, зробленої з більшою генеральної сукупності, позначається буквою *S*) і стандартне відхилення генеральної сукупності (для всієї генеральної сукупності, позначається буквою σ – мала грецька буква «сигма»).

Назви цих величин відображають правила їх використання. У разі роботи з вибіркою даних, взятих випадковим чином з більшої генеральної сукупності, використовується стандартне відхилення вибірки. Якщо ж вивчається вся генеральна сукупність, слід використовувати стандартне відхилення генеральної сукупності (часто також використовують терміни «вибіркове стандартне відхилення» та «генеральне стандартне відхилення» відповідно). Стандартне відхилення вибірки трохи більше, що забезпечує поправку на випадковість самої вибірки.

У деяких випадках ситуація може бути не зовсім однозначною. Так, наприклад, набір даних про заробітну плату всіх працівників певної компанії можна розглядати і як генеральну сукупність (оскільки розглядаються всі працівники цієї компанії) і як вибірку (якщо розглядати співробітників компанії як представників більшої генеральної сукупності подібного роду фахівців). Така неоднозначність є наслідком оцінки ситуації, що розглядається, а не наслідком характеру самих даних. Якщо вважати, що дані охоплюють повністю коло вирішуваних завдань, то ці дані, безумовно, є генеральною сукупністю. Якщо ж ми ставимо за мету провести деяке узагальнення (наприклад, перейти від розгляду співробітників даної компанії до розгляду співробітників, що працюють в аналогічних компаніях), то ті ж дані можна вважати вибіркою з деякої (можливо, гіпотетичної) генеральної сукупності.

Щоб покінчити з рештою неясності, приймемо наступне правило: при наявності сумнівів використовувати стандартне відхилення для вибірки. Ця величина дещо більше, і вибрати її – значить, діяти більш обережно і консервативно, і в кінцевому підсумку не допустити систематичної недооцінки невизначеності.

Що стосується обчислень, то єдина відмінність між цими двома показниками полягає в тому, що при обчисленні стандартного відхилення вибірки віднімають 1 (тобто ділять на n–1), а при обчисленні стандартного відхилення генеральної сукупності не вираховують 1 (тобто ділять на N). У зв'язку з цим, використання формули стандартного відхилення для вибірки дає трохи більше значення для невеликих розмірів вибірки, що відображає збільшення невизначеності, обумовленої використанням вибірки замість всієї множини даних. Існують також деякі загальноприйняті відмінності в позначеннях. Середнє для вибірки з n елементів позначається $\overline{X}$, а середнє генеральної сукупності з N елементів позначається грецькою буквою μ («мю»). Формули для розрахунку стандартного відхилення мають такий вигляд.

Стандартне відхилення для вибірки

$$S = \sqrt{\frac{\text{Сума квадратів відхилень}}{\text{Кількість елементів вибірки} - 1}} =$$

$$\sqrt{\frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

Стандартне відхилення для генеральної сукупності

$$\sigma = \sqrt{\frac{\text{Сума квадратів відхилень}}{\text{Кількість елементів генеральної сукупності}}} =$$

$$\sqrt{\frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{N}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2}.$$

Чим менше кількість елементів (*n* або *N*), тим сильніше виявляються відмінності між цими двома формулами. У випадку 10 елементів стандартне відхилення вибірки перевищує стандартне відхилення генеральної сукупності на 5,4%. При 25 елементах відмінність становить 2,1%. Зі збільшенням числа елементів ця розбіжність зменшується, доходячи до 1,0% для 50 елементів і 0,5% для 100 елементів. Таким чином, у випадку досить великої кількості даних відмінність між двома розглянутими методами обчислень виявляється несуттєвою.

4.2. Розмах: швидка і поверхнева оцінка

Розмах, або інтервал, який обіймають значення даних, дорівнює різниці між найбільшим і найменшим значеннями. Він визначає, до якої міри окремі значення відрізняються між собою. Нижче показано обчислення розмаху для невеликого набору даних, що представляють кількість отриманих за останній час замовлень на п’ять різних видів товару:

Розмах набору даних (185, 246, 92, 508, 153) =

= Максимальне – мінімальне = 508 – 92 = 416.

Зверніть увагу, що розмах дуже легко вирахувати. Для цього потрібно лише переглянути список значень, вибрати зі списку саме велике і найменше значення, а потім відняти від більшого менше. Раніше, до появи електронних калькуляторів і комп'ютерів, простота обчислення розмаху була причиною того, що цей показник часто використовувався в якості міри мінливості. Зараз, коли обчислювати стандартне відхилення стало набагато простіше, розмах використовують не так часто.

Коли важливі екстремальні значення (тобто найбільше та найменше), розмах може бути гарною мірою розкиду. Прикладом може бути необхідність описати межі зміни значень даних. Така характеристика може виявитися корисною для двох цілей: по-перше, для опису меж зміни даних і, по-друге, для пошуку помилок в значеннях. При наявності в наборі даних дуже великих (або дуже малих) помилково записаних значень розмах має тенденцію зростати і відразу ж стає, з позицій здорового глузду, дуже великим. Така особливість робить розмах корисним для пошуку помилок і редагування значень даних.

З іншого боку, в силу чутливості до граничних значень розмах виявляється не дуже корисним в якості такої статистичної заходи розкиду, яка характеризує набір даних в цілому. Розмах не відображає типову мінливість даних, а скоріше фокусується лише на двох значеннях. Стандартне відхилення більш чутливе до всіх даних, завдяки чому ця величина дає більш чітке уявлення про загальну картину. Розмах завжди більше, ніж стандартне відхилення.

Приклад. Заробітна плата персоналу

Розглянемо заробітну плату найманих працівників технічного відділу фірми, що виробляє товари споживчої електроніки. Необхідні нам значення наведені в табл. 4.2.1. Не будемо звертати увагу на ідентифікаційні коди і зосередимося на стовпці заробітної плати. Легко побачити, що найбільш високооплачуваний працівник отримує в рік $44 500 (посада керівника технічного відділу), а найменш оплачувана посада у молодого перспективного

співробітника, який ще не має освіти – $16 500. Розмах становить $28 000 =44500 – 16500. Ця величина показує відмінність в доларовому вираженні між найбільш і найменш оплачуваними співробітниками.

Таблиця 4.2.1. Зарплата персоналу

Ідентифікаційний код співробітника	Зарплата, дол.	Ідентифікаційний код співробітника	Зарплата, дол.
918886653	34000	743594601	33000
771631111	26000	731866668	16500
148609912	27000	490731488	44500
742149808	22500	733401899	35000
968454888	21000	589246387	20000

Зверніть увагу, що розмах розраховується виходячи з двох граничних значень: максимальної і мінімальної зарплати. Величина розмаху не відображає типову варіацію зарплати в відділі, для цього необхідно використовувати стандартне відхилення.

Для більш повного аналізу (щоб задовольнити інтерес читача або дати йому можливість перевірити обчислені самостійно значення) можна вказати, що середня зарплата в відділі становить $27 950 і стандартне відхилення (яке краще характеризує типову мінливість зарплати) дорівнює $8 581. Це стандартне відхилення вибірки. При цьому інженери фірми розглядаються як вибірка з величезної кількості всіх інженерів, що виконують аналогічну роботу.

Коротко описану вище ситуацію можна охарактеризувати наступним чином. Величина $28 000 (розмах) розділяє мінімальну і максимальну зарплати. Стандартне відхилення, $8 581, показує, наскільки (приблизно) зарплати окремих співробітників відрізняються від $27 950 – середньої зарплати для даної групи.

Приклад. Тривалість перебування пацієнтів у лікарні

Робота лікарень зараз більше схожа на комерційну діяльність, ніж це було раніше. Почасти це пов’язано з проникненням конкурентної боротьби в сферу охорони здоров’я. Багато організацій, які надають послуги в галузі

охорони здоров’я, просто наймають лікарів як службовців, в той час як в традиційних лікарнях лікарі мають велику незалежність. Ще одна причина комерціалізації охорони здоров’я полягає в тому, що згідно з програмою охорони здоров’я Medicare в даний час є тенденція виробляти фіксовані виплати на основі діагнозу, а не гнучкі виплати в залежності від тривалості лікування. Це сприяє виникненню сильної тенденції до скорочення тривалості лікування в разі конкретної хвороби пацієнта.

В якості одного з показників інтенсивності лікування виступає кількість днів перебування пацієнта в лікарні. У табл. 4.2.2 відображено кількість днів перебування в лікарні для вибірки пацієнтів в минулому році. Розмах цього ряду даних становить 386 – 1 = 385 днів, що представляє собою дуже велике значення, оскільки в році лише 365 (або 366) днів, а цей набір даних, як передбачається, відноситься тільки до одного року. Даний приклад ілюструє користь застосування поняття розмаху для редагування набору даних з метою виявлення помилок перед початком аналізу даних. Для цього також корисно уважно дослідити найменші і найбільші значення.

При ретельній перевірці даних була виявлена помилка. Реальне значення 286 було помилково записано як 386. У виправленому наборі даних розмах став дорівнює 285 (тобто 286 – 1).

Таблиця 4.2.2. Тривалість перебування в лікарні для вибірки пацієнтів

Ліжко-днів за минулий рік		
17	33	5
16	5	6
1	1	12
1	7	16
7	4	386
74	13	2
2	6	7
163	33	28
51		

4.3. Коефіцієнт варіації: міра відносної мінливості

Коефіцієнт варіації, який визначається як результат ділення стандартного відхилення на середнє значення, являє собою відносну міру коливання і виражається у відсотках або частках середнього значення. Такий підхід особливо корисний в тому випадку, коли набір даних не містить від'ємних значень. Формула для обчислення коефіцієнта варіації має наступний вигляд:

Коефіцієнт варіації

$$\text{Коефіцієнт варіації} = \frac{\text{Стандартне відхилення}}{\text{Середнє}}$$

Для вибірки

$$\text{Вибірковий коефіцієнт варіації} = \frac{S}{\overline{X}}$$

Для генеральної сукупності

$$\text{Коефіцієнт варіації генеральної сукупності} = \frac{\sigma}{\mu}$$

Зверніть увагу на те, що стандартне відхилення знаходиться в чисельнику. Таким чином, результат ділення характеризує мінливість.

Так, наприклад, якщо в середньому покупець витрачає в супермаркеті \$35,26, а стандартне відхилення становить \$14,08, то коефіцієнт варіації дорівнює 14,08 / 35,26 = 0,399, або 39,9%. Це означає, що зазвичай суми, які покупець витрачає при відвідуванні супермаркету, відрізняються від середнього значення приблизно на 39,9%. В абсолютному вираженні ця типова відмінність від середнього розміру витрат дорівнює \$14,08 (стандартне відхилення), що становить 39,9% (коефіцієнт варіації) від середнього.

Коефіцієнт варіації – безрозмірна величина. Це просто число, частка або відсоток. При обчисленні коефіцієнта варіації розмірність зникає в результаті поділу стандартного відхилення на середнє значення. Коефіцієнт варіації корисний в тих випадках, коли важлива не абсолютна величина відмінностей значень даних, але їх відносна мінливість.

Використовуючи коефіцієнт варіації, можна порівняти варіацію обсягів продажів для великої і малої фірми «з поправкою на розмір фірми». Зазвичай у фірми, оборот якої складає сотні мільйонів доларів, відмінності в обсягах продажів також досить великі – наприклад, вони можуть досягати десятків мільйонів доларів. Для іншої фірми, обсяг продажів якої обчислюється мільйонами доларів, відмінності можуть становити сотні тисяч. Однак в кожному з цих двох випадків варіація становить близько 10% середнього значення загального обсягу продажів. Для більшої фірми абсолютне значення варіації виявиться більше (більше стандартне відхилення), проте відносна, або та, що враховує обсяг, величина варіації (коефіцієнт варіації) виявляється однаковою для обох фірм.

Слід також зазначити, що коефіцієнт варіації може перевищити 100% навіть в тому випадку, якщо всі значення позитивні. Це, зокрема, може бути в разі сильно скошеного розподілу або при наявності значень, що сильно відрізняються від середнього. Такий результат означає, що в досліджуваній ситуації спостерігається дуже сильна варіація по відношенню до величини середнього значення.

Приклад. Невизначеність прибутковості портфеля інвестицій

Уявіть собі, що ви вклали $10 000 в 200 акцій деякої корпорації, акції якої продаються по $50 за штуку. Ваш знайомий придбав 100 акцій цієї ж корпорації за $5 000. Ви обоє очікуєте, що вартість акцій зросте в наступному році до $60 за акцію, що відповідає ставці прибутку 20%, (60–50) / 50. Обидва ви також вважаєте маркетингову стратегію цієї корпорації досить ризикованою, оскільки вона характеризується стандартним відхиленням курсу акцій $9. Це означає, що, хоч ви і очікуєте, що вартість однієї акції становитиме в майбутньому році $60, для вас не виявиться несподіваним, якщо вона буде приблизно на $9 більше або менше цього значення.

Ви припускаєте, що обсяг ваших інвестицій зросте наступного року до $12 000 ($60 * 200), зі стандартним відхиленням $1 800 ($9 * 200). Інвестиції

вашого знайомого, як очікується, в наступному році виростуть до $6 000, зі стандартним відхиленням $900.

Складається враження, що ваш ризик (стандартне відхилення в $1 800) в два рази більше, ніж ризик вашого знайомого ($900). І це дійсно так, оскільки ваші інвестиції в абсолютному вираженні в два рази більше. Однак обидва ви робите вкладення в одні і ті ж цінні папери, а саме в акції однієї і тієї ж корпорації. Таким чином, у всіх відносинах, за винятком обсягу інвестицій, ваша схильність до ризику буде однаковою. У відносному вираженні (щодо обсягу початкових вкладень) ризики виявляються однаковими. У цьому можна переконатися, обчисливши коефіцієнт варіації (стандартне відхилення для вартості акцій в майбутньому році, поділене на середнє або очікуване значення). Коефіцієнт варіації в вашому випадку буде дорівнює $1 800 / $12 000 = 0,15, і він рівний коефіцієнту варіації для інвестицій вашого знайомого $900 / $6 000 = 0,15. При цьому і ви, і ваш знайомий будете вважати, що невизначеність (або ризик) становить близько 15% від очікуваної в наступному році вартості портфеля інвестицій.

Приклад. Продуктивність праці у відділі торгівлі по телефону

Розглянемо відділ торгівлі по телефону, в якому працюють 19 співробітників, що займаються продажем квитків на концерт симфонічної музики. В середньому кожен співробітник продає 23 квитка на годину. Стандартне відхилення становить 6 квитків на годину. Це означає, що будь-який із співробітників може продавати в годину в середньому на 6 квитків більше або менше середнього значення (23 квитка).

Якщо скористатися коефіцієнтом варіації і висловити відмінності в роботі співробітників в відносних величинах, можна легко вирахувати, що ця величина складає 6 / 23 = 0,261, або 26,1%. Це означає, що варіація продуктивності праці співробітників складає приблизно 26,1% від середнього рівня продажів.

Для цілей проведення аналізу на високому рівні і для формування використовуваної менеджерами вищої ланки стратегії число 26,1%

(коефіцієнт варіації) може виявитися значно корисніше, ніж інформація про відмінності в 6 квитків на годину (стандартне відхилення). Менеджери можуть розглядати окремо рівень продуктивності праці (23 квитка на одного співробітника в годину) і варіацію продуктивності (зазвичай продуктивність праці співробітників відрізняється від середньої не більше ніж на 26,1% в сторону більших або менших значень).

Використання коефіцієнта варіації виявляється особливо корисним при проведенні порівнянь в умовах різних обсягів. Розглянемо ще один відділ торгівлі по телефону, що займається продажем квитків в театри, в якому середній рівень продажів складає 35 квитків на годину, а стандартне відхилення дорівнює 7. Оскільки продуктивність праці при продажу театральних квитків виявляється в цілому вище продуктивності при продажу квитків на концерти симфонічної музики (середні значення відповідно 35 і 23), природно, що і варіації тут виявляються вищими (7, а не 6). Однак коефіцієнт варіації для відділу, що працює з театральними квитками, становить 7 / 35 = 0,200, або 20,0%. Порівнюючи цю величину з коефіцієнтом 26,1%, що характеризує варіацію продажів квитків на симфонічні концерти, менеджери можуть зробити висновок про те, що група, яка працює з театральними квитками, фактично більш однорідна (з точки зору продуктивності окремих співробітників), ніж група, зайнята продажем квитків на концерти симфонічної музики.

4.4. Результати додавання константи або зміни шкали

Якщо ситуація змінюється певним систематичним чином, то необхідність в перерахунку таких характеристик, як типові значення (середнє значення, медіана, мода), перцентілі або міри мінливості (стандартне відхилення, розмах, коефіцієнт варіації) не виникає. Існує кілька основних правил швидкого обчислення відповідних показників для ситуації, що змінилася.

Якщо до кожного значення даних додається фіксована величина, для отримання відповідних характеристик отриманого таким чином нового набору даних цю ж величину необхідно додати до середнього, медіані, моді і перцентілей вихідного набору даних. Так, наприклад, додавання нового збору в $5 до рахунків, рівним $38, $93, $25 і $89, означає, що ці рахунки виявляться рівними $43, $98, $30 і $94. Середня величина рахунку виросла рівно на $5, з $61,25 до $66,25. Замість того щоб розраховувати середнє значення для нових рахунків, можна просто додати $5 до знайденого раніше середнього значення. Це правило може бути застосовано і для інших типових значень. Так, наприклад, медіана в даному випадку зростає на $5, з $63,50 до $68,50. Однак стандартне відхилення і розмах залишають свої попередні значення, оскільки зсув значень зберігає між ними попередні відстані. Коефіцієнт варіації змінюється, але його можна легко розрахувати, виходячи зі стандартного відхилення і середнього значення нового набору.

Якщо кожне значення даних множиться на фіксовану кількість, для отримання середнього, медіани, моди, перцентілів, стандартного відхилення і розмаху нового набору даних відповідні показники вихідного набору необхідно помножити на це ж число. Коефіцієнт варіації залишається без змін.

Якщо всі величини, що входять в набір даних, множаться на множник c і до них додається величина d, наведені вище правила діють спільно: величина X перетворюється в $cX + d$. Нове середнє значення при цьому виявляється рівним c * (старе середнє) + d. Аналогічні зміни зазнають медіана, мода і перцентілі. Нове стандартне відхилення дорівнює $|c|$ * (старе стандартне відхилення). Аналогічним чином коригується і розмах (зверніть увагу на те, що величина значення d, що додається, в цьому випадку зовсім не впливає). Новий коефіцієнт варіації легко можна обчислити на основі нових значень середнього і стандартного відхилення.

У табл. 4.4.1 представлені описані вище правила. Новий коефіцієнт варіації легко обчислити, скориставшись новими значеннями стандартного відхилення і середнього.

Таблиця 5.4.1. Результати додавання константи або зміни шкали

	Вихідні дані	**Додаток *d***	**Множення на *c***	**Множення і додаток**
Величини даних	X	$X + d$	cX	$cX + d$
Середнє значення (те ж для медіани, моди і перцентілів)	$\overline{X}$	$\overline{X} + d$	$c\overline{X}$	$c\overline{X} + d$
Стандартне відхилення (те ж для розмаху)	S	S	$\|c\|S$	$\|c\|S$

Приклад. Невизначеність розміру витрат в японських єнах і доларах США

Уявіть собі, що розташований за кордоном виробничий підрозділ вашої фірми сформувало бюджет витрат на майбутній рік в наступному вигляді.

Очікувані витрати	325 700 000 Японських єн
Стандартне відхилення	50 000 000 Японських єн

Для складання загального фінансового кошторису фірми необхідно перевести ці значення в долари США. Розглянемо для простоти тільки комерційний ризик (представлений стандартним відхиленням). Докладніший аналіз може зажадати обліку ризику, пов’язаного з можливими змінами курсів обміну валют.

Японська єна легко переводиться в долари США з використанням поточного значення курсу обміну валют. Для перекладу єни в долар необхідно помножити відому величину передбачуваних витрат на число 0,007224, що позначає, скільки доларів обмінюється за одну єну. Помноживши обсяг очікуваних витрат і стандартне відхилення на цей перекладний коефіцієнт, знаходимо очікувані витрати і ризик в доларовому вираженні (величини округлені до тисяч).

Очікувані витрати	$2 353 000
Стандартне відхилення	$361 000

В даному прикладі використання наведених нами основних правил дало можливість перевести дані, представлені в єнах, в доларовому виразі без повторного повного складання кошторису в доларах.

Приклад. Загальна вартість виробленого товару

При обчисленні собівартості виробничі витрати часто поділяють на фіксовані витрати і змінні витрати на одиницю продукції. Фіксовані витрати не залежать від кількості вироблених одиниць продукції, в той час як змінні витрати закладаються в кошторис для кожної одиниці виробленої продукції. Фіксовані витрати можуть включати орендну плату і інвестиції в виробниче обладнання, в той час як змінні витрати можуть становити вартість реально використаних для виробництва матеріалів.

Розглянемо виробництво шампуню, для якого фіксовані витрати складають $1 000 000 в місяць, а змінні витрати дорівнюють $0,50 на один флакон. На основі ретельного аналізу ринкового попиту менеджери передбачили в наступному місяці випуск 1 200 000 флаконів шампуню. Виходячи з попереднього досвіду невизначеність для прогнозованого обсягу виробництва можна оцінити на рівні близько 250 000 флаконів. Таким чином, очікується випуск в середньому 1 200 000 флаконів шампуню, зі стандартним відхиленням 250 000 флаконів.

Якщо для обсягу виробництва існує такий прогноз, то яким буде прогноз для витрат? Зверніть увагу на те, що обсяг виробництва перекладається в витрати шляхом множення кількості одиниць товару на $0,50 (змінні витрати) з додаванням $1 000 000 (фіксовані витрати). Таким чином, в нашому випадку

загальна вартість = $0,50 * 1 200 000 + $1 000 000 = $1 600 000,

стандартне відхилення вартості = $0,50 * 250 000 = $125 000

Отже, кошторис витрат складений. Очікуються витрати $1 600 000 зі стандартним відхиленням (невизначеністю) $125 000.

Коефіцієнт варіації для кількості одиниць виробленої продукції дорівнює 250 000 / 1 200 000 = 20,8%. Коефіцієнт варіації для витрат також легко обчислюється, він дорівнює $125 000 / $1 600 000 = 7,8%. Зверніть увагу, що відносна варіація в вартісному вираженні виявляється значно менше, оскільки великі фіксовані витрати призводять до збільшення бази порівняння і відповідно до помітного зниження значення варіації.

4.5. Додатковий матеріал

Резюме

Мінливість (яку також називають різноманітністю, невизначеністю, розсіюванням, розкидом і варіацією) являє собою міру відмінності окремих значень набору даних між собою. У той час як величини, що характеризують центр (такі як середнє значення, медіана, мода) вказують типову для набору даних величину значень, мінливість показує, наскільки близько до цього центру зазвичай розташовуються окремі значення набору даних. Якщо всі величини даних однакові, мінливість дорівнює нулю. Чим більше розкид величин, тим більше мінливість.

Стандартне відхилення, яке зазвичай використовують в якості характеристики мінливості, відображає типову відстань між середнім значенням і окремими значеннями набору даних. Стандартне відхилення показує ступінь випадковості в розташуванні окремих значень щодо їх загального середнього. Відхилення – це відстані між кожним із значень і середнім значенням набору даних. Позитивні відхилення відповідають значенням, що перевищують середнє, а негативні відхилення – значенням, які менші за середнє. Усереднення цих відхилень завжди дає результат, що дорівнює нулю. Стандартне відхилення показує типову величину таких відхилень (знак «мінус» при цьому не враховується) і являє собою число, що вимірюється в тих же одиницях, що і вихідні дані (наприклад, в доларах, в милях на один галон або в кілограмах).

Щоб обчислити стандартне відхилення, необхідно виконати наступне.

1. Знайти відхилення, віднімаючи з кожного значення набору даних середнє.

2. Звести отримані величини відхилень в квадрат, скласти їх і розділити отриману суму на n–1. Отриманий результат називається дисперсією.

3. Витягти квадратний корінь. Отримане значення і є стандартним відхиленням.

При роботі з даними всієї генеральної сукупності необхідно використовувати стандартне відхилення генеральної сукупності (позначається буквою σ. У тому випадку, якщо необхідно зробити узагальнення і перейти від наявного набору даних до деякої більшої безлічі (реальній або гіпотетичній), використовується стандартне відхилення вибірки (позначається літерою S). При виникненні сумнівів у тому, яку з цих величин застосувати, потрібно використовувати стандартне відхилення вибірки. Формули для знаходження названих величин мають такий вигляд:

$$S = \sqrt{\frac{\text{Сума квадратів відхилень}}{\text{Кількість елементів вибірки} - 1}} =$$

$$\sqrt{\frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

$$\sigma = \sqrt{\frac{\text{Сума квадратів відхилень}}{\text{Кількість елементів генеральної сукупності}}} =$$

$$\sqrt{\frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{N}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2}.$$

В обох цих формулах знаходять квадрати відхилень, ділять їх на відповідну величину, а потім витягають квадратний корінь, щоб від квадратів знову повернутися до вихідних відхилень. При обчисленні стандартного відхилення вибірки ділять на *n*–1, оскільки відхилення обчислюють на основі невизначеного середнього значення вибірки, а не на основі точного середнього значення генеральної сукупності.

Дисперсія – це квадрат стандартного відхилення. Ця величина несе ту ж інформацію, що і стандартне відхилення. Однак інтерпретація дисперсії ускладнена тим, що одиниці вимірювання дисперсії є квадрат одиниць виміру вихідних даних (наприклад, долар в квадраті, квадратні милі на один галон в квадраті або кілограми в квадраті, незалежно від змістовного сенсу таких

одиниць). У зв’язку з цим в якості характеристики мінливості частіше використовують стандартне відхилення.

Якщо дані мають нормальний розподіл, стандартне відхилення дорівнює приблизно половині довжини відрізка числової прямої, який містить дві третини всіх значень набору даних. Це означає, що приблизно дві третини всіх значень знаходяться на відстані не більше однієї величини стандартного відхилення від середнього (вище або нижче середнього). Приблизно 95% всіх значень знаходяться на відстані не більше двох величин стандартного відхилення від середнього, а близько 99,7% значень лежать в межах трьох стандартних відхилень від середнього. Однак не слід очікувати справедливості цих тверджень для інших (відмінних від нормального) розподілів.

Розмах дорівнює різниці між максимальним і мінімальним значеннями набору даних. Ця величина характеризує протяжність, або ширину набору даних. Розмах використовують як для опису даних, так і для пошуку проблем в даних (зокрема, для пошуку помилок при записі значень). Як статистична характеристика розмах має той недолік, що він акцентує увагу лише на екстремальних значеннях і не враховує типові значення. Для більшості цілей статистичного аналізу в якості міри мінливості більш корисно використовувати стандартне відхилення.

Коефіцієнт варіації дорівнює частці від ділення стандартного відхилення на середнє значення і характеризує відносну мінливість даних, виражену в частках або відсотках від середнього. Коефіцієнт варіації – безрозмірна величина. Він може бути корисний при порівнянні мінливості наборів даних, представлених в різних одиницях виміру.

Додавання фіксованого числа до всіх значень набору даних призводить до збільшення середнього, медіани, перцентілів і моди на таке ж число; стандартне відхилення і розмах при цьому не змінюються. При множенні кожного із значень набору даних на фіксоване число всі характеристики – середнє, медіана, перцентілі, мода, стандартне відхилення і розмах – множаться на це ж число, а коефіцієнт варіації не змінюється. При множенні

кожного із значень даних на деяке число і додаванні іншого фіксованого числа два описаних вище правила діють спільно. Коефіцієнт варіації можна легко визначити після того, як із застосуванням цих правил обчислюється середнє і стандартне відхилення.

Основні терміни

– Мінливість (variability), різноманітність (diversity), невизначеність (uncertainly), розсіювання (dispersion), розкид (spread)

– Стандартне відхилення (standard deviation)

– Відхилення (deviation)

– Дисперсія (variance)

– Стандартне відхилення вибірки (sample standard deviation)

– Стандартне відхилення генеральної сукупності (population standard deviation)

– Розмах (range)

– Коефіцієнт варіації (coefficient of variation)

Контрольні питання

1. Що таке мінливість?
2. а) Які показники зазвичай використовують в якості міри мінливості?

   б) Які ще показники використовують для цієї мети?
3. а) Що таке відхилення від середнього значення?

   б) Чому дорівнює середнє значення всіх відхилень?
4. а) Що таке стандартне відхилення?

   б) Яку інформацію про взаємозв’язок між окремими значеннями і середнім несе стандартне відхилення?

   в) В яких одиницях вимірюється стандартне відхилення?

   г) В чому полягає відмінність між стандартним відхиленням вибірки і стандартним відхиленням генеральної сукупності?
5. а) Що таке дисперсія?

   б) В яких одиницях вимірюється дисперсія?

в) Яку із мір мінливості легше інтерпретувати – стандартне відхилення чи дисперсію? Чому?

г) Якщо відомо стандартне відхилення, чи дає дисперсія істотну додаткову інформацію про мінливість?

6. Припустимо, що деякий набір даних має нормальний розподіл. Яку частину значень можна в такому випадку очікувати знайти?

а) На відстані не більше одного стандартного відхилення від середнього значення?

б) На відстані не більше двох стандартних відхилень від середнього значення?

в) У межах трьох величин стандартного відхилення від середнього значення?

г) На відстані більше одного стандартного відхилення від середнього?

д) На відстані більше одного стандартного відхилення вище середнього значення? (Будьте уважні!)

7. Припустимо, що деякий набір даних НЕ розподілено нормально. Яку частину значень можна в такому випадку очікувати знайти?

а) На відстані не більше одного стандартного відхилення від середнього значення?

б) На відстані не більше двох стандартних відхилень від середнього значення?

в) У межах трьох величин стандартного відхилення від середнього значення?

г) На відстані більше одного стандартного відхилення від середнього?

д) На відстані більше одного стандартного відхилення вище середнього значення? (Будьте уважні!)

8. а) Що таке розмах?

б) В яких одиницях вимірюється розмах?

в) В яких випадках використовують такий показник як розмах?

г) Чи корисний розмах як статистична міра мінливості? Обґрунтуйте свою відповідь.

9. а) Що таке коефіцієнт варіації?

б) В яких одиницях вимірюється коефіцієнт варіації?

10. Яку міру коливання краще використовувати при порівнянні мінливості в двох різних ситуаціях за умови, що середні в цих двох ситуаціях сильно відрізняються? Обґрунтуйте свій вибір.

11. Вкажіть, як зміняться в результаті додавання фіксованого числа до кожного значення такі характеристики набору даних.

а) Середнє, медіана, мода.

б) Стандартне відхилення і розмах.

в) Коефіцієнт варіації.

12. Вкажіть, як зміняться при множенні на фіксоване число кожного значення такі характеристики набору даних.

а) Середнє, медіана, мода.

б) Стандартне відхилення і розмах.

в) Коефіцієнт варіації.