ЛЕКЦІЯ.
ОПИСОВА СТАТИСТИКА

Розподіли змінних є основним матеріалом при проведенні досліджень. Оскільки соціальні дослідження зазвичай містять велику кількість спостережень, безліч методів пов'язано з представленням даних для їх їх найбільш інформативної та осмисленої візуалізації. До цього були розглянуті кілька способів представлення даних, включаючи частотні розподіли, таблиці і графічні форми візуалізації. Тепер обговоримо способи чисельного опису змінних, способи отримання простих числових значень, що описують розподіл даних.

3.1. Вимірювання центральної тенденції

Вимірювання центральної тенденції включає три характеристики: моду, медіану і середнє значення. Ці характеристики можуть бути легко обчислені і використані для подальшого аналізу. Інші характеристики рідше зустрічаються в соціологічних дослідженнях, і тому тут не будуть розглянуті.

Кожна характеристика – це унікальна інформацію про розподіл. Є обмеження при їх використанні, оскільки не всі характеристики можуть бути обчислені для різних типів шкал. Наприклад, мода може бути обчислена для номінальної, порядкової або інтервального шкали, середнє може бути отримано для даних, вимірюваних за інтервальною шкалою, і з деякими застереженнями, для порядкової шкали.

Вимірювання центральної тенденції (measure of central tendency) полягає у виборі одного числа, яке найкращим чином описує всі значення ознаки з набору даних. Таке число називають центром, типовим значенням для набору даних, мірою центральної тенденції.

Чому це потрібно? Отримавши таке число – одне-єдине, дослідник отримує інформацію про розподіл ознаки «в стислій формі». При цьому дослідник може порівнювати за допомогою цього числа два і більше різних

розподілів. Головний недолік полягає в тому, що тут втрачається , в порівнянні з розподілом частот, багато інформації.

Мода

Мода – значення, що найбільш часто зустрічається у вибірці або наборі даних. У разі, якщо дані згруповані і побудовано розподіл частот, модою є значення, що має найбільшу частоту.

Моду будемо позначати *Мо*. Мода цілком придатна для вимірювання центральної тенденції хоча б тому, що це єдиний спосіб описувати номінальний розподіл не гірше порядкового або інтервального. Обмеження в застосуванні пов'язані з тим, що мода розглядає лише одну особливість розподілу, а саме, розташування найбільш частого значення. Інші важливі особливості, такі як число спостережень вище або нижче моди, відстань між модальними значеннями і інші характеристики, залишаються поза увагою.

У таблиці 3–1 представлено категоріальний розподіл даних про вибір дисциплін спеціалізації 641 студентом. Мода для цих даних – соціологія, оскільки її вибрали в якості дисципліни спеціалізації найбільша кількість студентів – 149. Наступні за популярністю спеціалізації – політика і соціальна робота.

Таблиця 3–1. Дисципліни спеціалізації

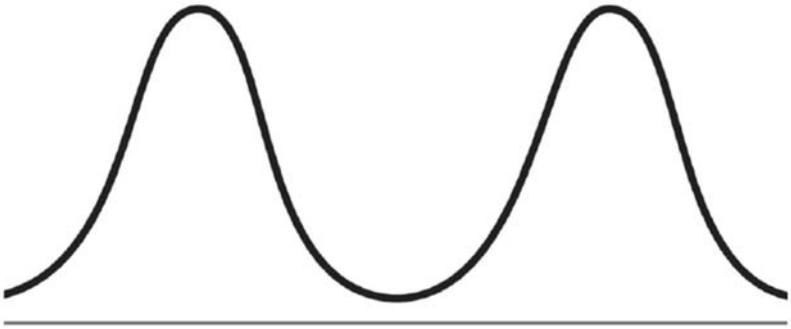
Дисципліни	***f***
Антропологія	97
Економіка	104
Політика	110
Психологія	72
Соціальна робота	109
Соціологія	149
Разом	**641**

Мода для згрупованого розподілу частот є середина інтервалу, в який потрапляє найбільша кількість спостережень. У таблиці 3–2 мода для групи 1 є 144,5, оскільки це є середина інтервалу, що містить найбільшу кількість спостережень – 23 з 80 випадків.

Таблиця 3–2. Вага тіла двох груп людей

Вага (фунти)	Група 1 – Б1	Група 2 – Б2
190-199	3	3
180-189	2	8
170-179	4	21
160-169	10	7
150-159	13	9
140-149	23	6
130-139	12	7
120-129	7	20
110-119	3	3
100-109	3	2
Разом	**80**	**86**

Розподіл може мати більш ніж одну моду, як видно по групі 2 таблиці 3-2. Інтервал 170-179 має найбільшу кількість спостережень, 21, але майже таке ж значення 20 потрапляє в інтервал 120-129. Оскільки різниця є незначною – всього одне спостереження, цей розподіл може бути описано як той, що має дві моди – 124,5 і 174,5. Його можна назвати бімодальним. Обидві категорії є рівною мірою популярними. Розподіл може мати більше двох популярних значень, але якщо він має більше трьох мод, опис такого розподілу в термінах найбільш частих значень втрачає будь-який сенс.



Малюнок 3–1. Вид бімодального розподілу

Медіана

Медіана визначається як серединне значення вибірки, або значення, вище і нижче якого розташовується однакове кількість спостережень. Для знаходження медіани обов'язково впорядкувати дані.

Медіана є точною серединою вибірки. Позначається *Me* і визначається по-різному для вибірок з парним і непарним числом елементів. Для непарної кількості спостережень медіана є спостереженням з номером $(n+1)/2$. Для парної кількості спостережень медіана обчислюється як середнє значення спостережень з номерами $n/2$ і $(n+2)/2$.

У разі непарної кількості спостережень медіана є просто серединою вибірки, вище і нижче якої розташовується однакова кількість спостережень.

Таблиця 3–3. Вибірки парного і непарного розміру

Вибірка 1 (N=5)	**Вибірка 2 (N=6)**
198	197
179	193
172	189
167	187
154	183
	179

У Таблиці 3–3 перша вибірка містить п'ять спостережень. Застосовуючи формулу, отримуємо, що медіана дорівнює $(5+1)/2 = 3$, то є третє спостереження, перелічене знизу або зверху, а саме, 172. Друга вибірка містить парну кількість спостережень, 6, це означає, що медіана є середнє значення спостережень з номерами $6/2 = 3$ і $(6+2)/2 = 4$, тобто значень 187 і 189. Тим самим, медіана дорівнює $(187+189)/2 = 188$. В обох випадках, для парної і непарної кількості спостережень, медіана є серединою вибірки.

«Вище» і «нижче» може мати сенс лише по відношенню до даних, які впорядковані за зростанням або за зменшенням. Медіана може бути визначена також в термінах ранжування варіаційного ряду і знаходження рангу серединного елементу, про що буде описано трохи пізніше.

Медіана може бути визначена для числових даних і даних, вимірюваних порядковою шкалою. Для номінальної шкали медіану неможливо відшукати через неможливість упорядкувати категорії номінальної шкали.

Відзначимо кілька очевидних властивостей, які є медіани. Крайні значення, що сильно відрізняються від інших даних, не впливають на

величину медіани. Значення медіани є єдиним для кожного набору даних, на відміну від моди. Медіана може бути визначена не з повного набору даних. Досить знати їх порядкове розташування, загальне кількість і кілька значень, розташованих в середині.

Середнє

Середнє визначається як середнє арифметичне вибірки, тобто як сума всіх значень вибірки, поділена на її обсяг.

Дотримуючись визначенню, будемо знаходити середнє значення за формулою:

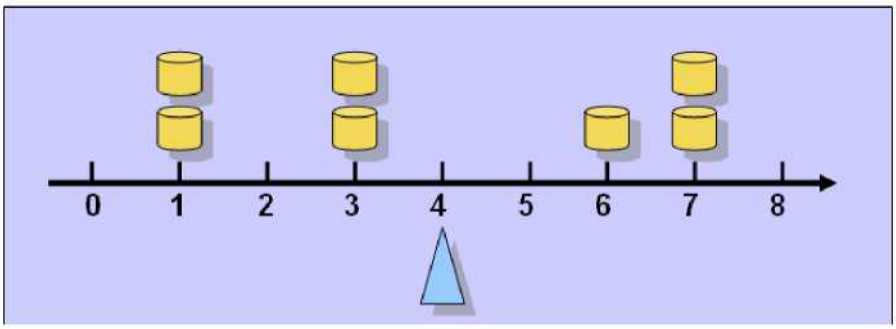
$$\overline{x} = \frac{\sum x}{n}$$

де $\sum x$= сума всіх значень вибірки, n = обсяг вибірки.

Наприклад, для вибірки з семи значень: 1, 1, 3, 3, 6, 7, 7, середнє значення буде обчислюватися так:

$$\overline{x} = \frac{1+1+3+3+6+7+7}{7} = \frac{28}{7} = 4.$$

Середнє значення може розумітися як «точка балансу». Якщо уявити, що числова пряма це ваги, то спробуємо «зважити» наші спостереження. Викладемо всі наявні спостереження на числову вісь. Середнє значення буде тією самою точкою, для якої права частина і ліва не переважують один одного. Це показано для нашого прикладу на малюнку 3.1.



Малюнок 3–1. Середнє значення є «точкою балансу»

Середнє значення, що розуміється як точка балансу, дуже активно використовується для опису вибіркового розподілу і застосовується в статистичних моделях і обчисленнях високого рівня.

Середнє для згрупованих даних

Середнє значення для згрупованих даних можна знаходити за більш зручною формулою:

$$\overline{x} = \frac{\sum(f * x)}{n}$$

У цій формулі $\sum(f * x)$ позначає суму добутків кожного з значень ознаки на його частоту, n є загальною кількістю спостережень, що дорівнює також сумі всіх частот.

Таблиця 3–6. Середнє для згрупованих даних

Оцінка	**Частота**	**Добуток**
x	*f*	*f*x*
5	17	85
4	41	164
3	20	60
2	7	14
Разом	**85**	**323**

Для прикладу, обчислимо середній бал для даних, представлених в таблиці розподілу частот 3.6. У першому стовпчику перераховані всі можливі значення ознаки: 5, 4, 3, 2. У другому стовпці вказані частоти, з якими ці значення зустрілися в наборі даних або вибірці. Всього по другому стовбцю 85 спостережень. У третьому стовпці ми обчислюємо для початку в кожному рядку добуток оцінки на частоту, а потім знаходимо суму по стовпцю. Сума дорівнює 323. Тепер скористаємося формулою:

$$\overline{x} = \frac{\sum(f * x)}{n} = \frac{323}{85} = 3{,}8$$

В результаті отримали, що середній бал становить 3,8.

Такий спосіб розрахунку за допомогою таблиці, істотно спрощує обчислення, що проводяться вручну і навіть на комп'ютері. Є ще кілька

формул для обчислення середнього. Розглянемо формулу обчислення середнього для розподілу, складеного за інтервалами, а також зваженого середнього.

Середнє для інтервального розподілу

У разі, якщо ми маємо інтервальний розподіл, самі значення спостережень, що потрапляють всередину кожного з інтервалів, невідомі. Знаходження середнього в цьому випадку відбувається наступним чином. У таблицю інтервального розподілу слід додати стовпець, в який проставляються середини інтервалів. Середини є «представниками» всього інтервалу і множаться на частоту для знаходження середнього.

Таблиця 3–7. Знаходження середнього по інтервалах

Інтервал	**Частота**	**Середина**	**Добуток**
	f	*m*	*f*m*
0-99	11	49,5	544,5
100-199	12	149,5	1794,0
200-299	14	249,5	3493,0
300-399	1	349,5	349,5
400-499	2	449,5	899,0
Разом	$\sum \boldsymbol{f}$**=40**		$\sum(\boldsymbol{f} * \boldsymbol{m})$**=7080,0**

Для прикладу розглянемо таблицю 3–7. У перший інтервал потрапляє 11 спостережень, хоча конкретні їх значення не відомі. Виберемо в якості представника інтервалу 0-99 його середину 49,5.

Формула для знаходження середнього значення в цьому випадку буде виглядати таким чином:

$$\overline{x} = \frac{\sum(f * m)}{n}$$

де *f* = частота потрапляння в інтервал, *m* = середина інтервалу, *n* = обсяг вибірки.

Формула означає, що ми перемножуємо частоти на середини інтервалів, складаємо добутки і ділимо на загальну кількість спостережень.

Для нашого прикладу:

$$\overline{x} = \frac{\sum(f * m)}{n} = \frac{7080}{40} = 177{,}0$$

Зважене середнє

Часто ми маємо кілька груп спостережень, середні значення всередині кожної з яких нам відомі. Виникає питання – як обчислити групове середнє, тобто середнє значення за всіма спостереженнями, складеним з усіх наявних груп спостережень.

Зважене середнє – середнє значення, що отримується при об'єднанні кількох груп спостережень.

Таблиця 3–8. Знаходження зваженого середнього

Група	**Середнє по групі**	**Обсяг групи**
	$\overline{x}$	n
A	87	65
B	92	110
C	89	85
D	96	200
E	84	60
Разом		**520**

Якщо групи мають однаковий обсяг, то групове середнє можна обчислити як середнє арифметичне наявних середніх значень по кожній групі. Якщо ж групи мають різний обсяг, то групове середнє можна знайти за такою формулою:

$$\overline{X} = \frac{\sum(\overline{x} * n)}{N}$$

де $\sum(\overline{x} * n)$ = сума добутків середніх в групі на кількість елементів в цій групі, N = загальна кількість спостережень у всіх групах.

Наприклад, в таблиці 3.8 клас A з 65 спостереженнями вносить менший внесок в групове середнє в порівнянні з групою B, що має 110 спостережень. Використовуємо формулу для нашого прикладу:

$$\overline{X} = \frac{65 * 87 + 110 * 92 + 85 * 89 + 200 * 96 + 60 * 84}{65 + 110 + 85 + 200 + 60} = \frac{47580}{520} = 91{,}5$$

Назва «зважене середнє» використовується тому, що для його знаходження враховуються ваги, які мають середні значення по групах.

Середнє для дихотомічної шкали

Як уже зазначалося, дихотомічна шкала має унікальну властивість. Для неї може обчислюватися середнє значення, незважаючи на те, що для номінальних шкал середнє не обчислюється, оскільки в них заборонені арифметичні операції. Слід пам'ятати, що в дихотомічній шкалі є лише два значення: так – ні, знаю – не знаю і т.п. Якщо два значення ознаки кодуються 0 і 1, то обчислене середнє покаже частку одиниць у вибірці.

Наприклад, для вибірки {1, 0, 0, 0, 1, 1, 1, 1, 1, 0} середнє дорівнюватиме: число одиниць / число елементів вибірки = 6 / 10 = 0,6. Це означає, що 60% значень вибірки приймають значення, що дорівнює одиниці.

Середнє не означає краще

Відзначимо деякі очевидні властивості середнього.

– Середнє обчислюється тільки в числових шкалах. На жаль, середнього не існує для номінальних і порядкових шкал. Дихотомічна шкала – це приємний виняток.

– При обчисленні середнього необхідно використовувати всі дані. Пропущені значення в даних не допускаються.

– Для кожного набору даних може бути обчислено лише одне значення середнього. Цією властивістю володіє медіана і не володіє мода.

– Середнє є єдиною мірою центральної тенденції, для якої сума відхилень кожного значення від нього дорівнює нулю: $\sum(x - \overline{x}) = 0$.

Середнє, незважаючи на наявність переваг перед іншими мірами центральної тенденції, має серйозні недоліки. Вони стають зрозумілі після деякої практики використання середнього в якості міри.

Наведемо лише один приклад. В якомусь селі Запорізької області проживає 50 мешканців. Серед них 49 осіб – селяни з місячним доходом в 4 тис. гривень, а один мешканець – заможний власник будівельної фірми, з

місячним доходом 750 тис. гривень. Обчислимо середнє. Воно дорівнює майже 19 тис. гривень. Однак, навряд чи в цьому випадку можна стверджувати, що це число адекватно представляє дохід жителів села. В цьому випадку, більш розумно взяти в якості міри центральної тенденції моду або медіану (обидві дорівнюють 4 тис. гривень).

Тоді виникає логічне питання, а яка міра центральної тенденції є все-таки найкращою? Відповідь на це питання існує і залежить від критеріїв. Кожна міра є найкращою в своєму, цілком певному сенсі.

Якщо вважати, що дані найкращим чином представляє елемент, що найбільш часто зустрічається, тоді це Мода.

Якщо вважати, що найкращим представником даних є значення, для якого сума абсолютних відхилень від нього всієї решти значень буде найменшою, тоді це Медіана.

Якщо вважати, що найкращим представником для даних є таке значення, для якого сума квадратів відхилень від нього всіх значень буде найменшою, тоді це Середнє.

Залишилося узагальнити відповідність різних мір центральної тенденції типам шкал, для яких вони можуть бути застосовані. Така відповідність описана в таблиці 3.9.

Таблиця 3–9. Типи шкал і міри центральної тенденції

Типове значення	Номінальні дані	Порядкові дані	Інтервальні дані
Мода	O	O	O
Медіана		O	O
Середнє			O

Резюме

Міри центральної тенденції це єдине число, яке може розглядатися в якості представника набору даних. Таке єдине значення має переваги і недоліки. Крім того, три різні міри – мода, медіана і середнє – можуть бути застосовні не завжди. У деяких випадках вони просто не можуть бути обчислені, а іноді не відображають головну мету – представляти набір даних для аналізу і порівнянь.

3.2. Вимірювання варіації

Вимірювання центральної тенденції не дає уявлення про відмінності даних всередині вибірки. Для цього існує вимір варіації вибірки або набору даних.

Вимірювання варіації (measure of variation) полягає в знаходженні чисел, які характеризують ступінь розкиду даних щодо центру розподілу.

Варіація може бути проілюстрована наступним прикладом. Припустимо, у нас є вимірювання рівня достатку в декількох містах області та відповідних приміських зонах. Кількість сімей нижче рівня бідності на 100 сімей відображено показником і дані представлені в таблиці 3.10.

Таблиця 3–10. Сім'ї нижче рівня бідності

Околиця міста	Центр міста
24,5	27,4
23,8	24,6
23,1	23,0
22,4	22,5
21,7	21,8
21,0	21,6
21,0	20,9
20,3	19,7
19,6	18,1
19,6	17,4
217,0	**217,0**
$\overline{X}$ **= 21,7**	$\overline{X}$ **= 21,7**

В обох випадках середнє значення бідних сімей на кожні 100 склало 21,7. Це означає, що якщо для опису вибірок використовувати виключно середнє, ми прийдемо до висновку, що вибірки ідентичні. Однак, навіть поверхового погляду достатньо, щоб зрозуміти відмінності між ними. Для першої вибірки характерна не надто сильна різниця результатів вимірювань від середнього значення 21,7. Найменшим значенням є 19,6, найбільшим – 24,5. Вибірка для центральної частини міст має зовсім іншу картину. Діапазон зміни значень йде від 17,4 до 27,4. В цьому випадку середнє значення 21,7 не настільки добре описує кількість бідних сімей в випадково обраному місці в

центрі міста. Вийшло, що сім'ї на околицях міст більш однорідні в сенсі кількості бідних сімей, ніж у центральній частині міст.

Розмах (відстань)

Розмах або **відстань (range)** – різниця між найбільшим і найменшим значеннями.

Першою характеристикою варіації є розмах. Для знаходження розмаху перш рекомендується упорядкувати дані в порядку зростання. Можна записати розмах за допомогою формули:

$$R = x_{max} - x_{min}$$

У нашому прикладі в таблиці 3.10 різниця становить для першої вибірки 24,5 –19,6 = 4,9, а для другої 27,4 – 17,4 = 10. Як говорилося раніше, відмінність між двома вибірками існує, але описати цю відмінність кількісно було важко.

Різниця дає нам цілком придатну інформацію про вибірку шляхом опису відстані між найбільшим і найменшим значеннями. Очевидно, ця інформація має пряме відношення до характеристики розкиду досліджуваного нами розподілу.

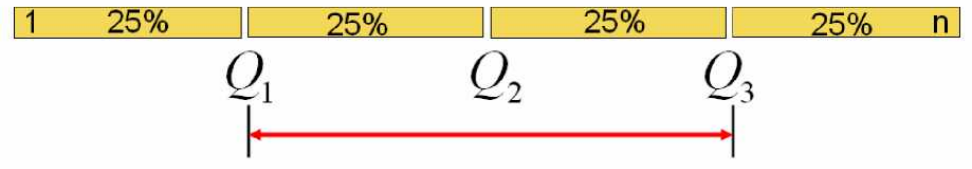
Квартильний розмах

Квартилі (quartile) – значення, які ділять варіаційний ряд на чотири рівні за обсягом частини.

Квартильний розмах або **міжквартильний розмах** (Inter Quartile Range – IQR) – різниця між третім і першим квартилями.

Таких значень повинно бути три: перша, друга і третя квартиль відповідно. Для початку дані слід упорядкувати. Після цього знаходиться медіана, яка є другим квартилем за визначенням. Після цього знаходяться перший і третій квартилі.

Існує кілька варіантів формального визначення квартилей. Для навчальних цілей ми спростимо задачу їх пошуку аналогічно знаходженню медіани. Якщо в якості серединного елемента виступить два претендента – ми знайдемо їх середнє арифметичне і його назвемо відповідним квартилєм.



Малюнок 3–2. Квартилі і квартильний розмах

Квартальний розмах знаходиться за формулою:

$$IQR = Q_3 - Q_1$$

Якщо при обчисленні розмаху використовуються тільки найбільше і найменше значення ознаки, а розподіл даних між ними повністю ігнорується, то при обчисленні квартильного розмаху ігноруються «крайні» дані, розташовані за межами першого і третього квартилів. Між Q_1 і Q_3 розташовано 50% всіх даних.

Дисперсія

Дисперсія для набору даних або вибірки – середнє арифметичне квадратів відхилень значень від їх середнього.

Дисперсія позначається як s^2 . Основна формула (за визначенням) для знаходження дисперсії:

$$s^2 = \frac{\sum(x - \overline{x})^2}{n - 1}$$

Формула означає, що нам слід віднімати середнє з кожного значення вибірки, підсумувати квадрати різниці, а потім розділити отриману суму на кількість спостережень мінус 1. Чому в знаменнику при знаходженні середнього арифметичного квадратів відхилень використовується $(n - 1)$ замість n, буде роз'яснено пізніше, при обговоренні логіки побудови оцінок параметрів генеральної сукупності.

Обчислимо дисперсію в наведеному прикладі. Розрахунок проведемо шляхом добудовування таблиці 3.10 в таблицю 3.11. Отримаємо:

Околиця міста: $s^2 = \frac{\sum(x-\overline{x})^2}{n-1} = \frac{26{,}46}{10-1} = 2{,}94$

Центр міста: $s^2 = \frac{\sum(x-\overline{x})^2}{n-1} = \frac{79{,}34}{10-1} = 8{,}82$

Результати підтверджують наше інтуїтивне уявлення про те, що в центральній частині міст дисперсія кількості бідних сімей більше, ніж на околицях міст. Це видно за результатами підрахунку: 8,82 для центру міста більше 2,94 для околиць. Стандартне відхилення, яке ми розглянемо нижче, дозволяє привести дисперсії, що аналізуються, до стандартного вигляду, який більш зрозумілий для розуміння і порівнянь.

Таблиця 3–11. Розрахунок дисперсії для рівня бідності

Околиця міста (*N* = 10)			**Центр міста (*N* = 10)**		
x	$x - \overline{x}$	$(x - \overline{x})^2$	x	$x - \overline{x}$	$(x - \overline{x})^2$
24,5	2,8	7,84	27,4	5,7	32,49
23,8	2,1	4,41	24,6	2,9	8,41
23,1	1,4	1,96	23,0	1,3	1,69
22,4	0,7	0,49	22,5	0,8	0,64
21,7	0	0	21,8	0,1	0,01
21,0	-0,7	0,49	21,6	-0,1	0,01
21,0	-0,7	0,49	20,9	-0,8	0,64
20,3	-1,4	1,96	19,7	-2,0	4,00
19,6	-2,1	4,41	18,1	-3,6	12,96
19,6	-2,1	4,41	17,4	-4,3	18,49
Σ=217,0		**Σ=26,46**	**Σ=217,0**		**Σ=79,34**

Є друга формула для знаходження дисперсії вибірки:

$$s^2 = \frac{n * \sum x^2 - (\sum x)^2}{n * (n - 1)}$$

Вважається, що ця формула більш придатна для ручного розрахунку, оскільки вимагає меншої кількості арифметичних операцій. Проілюструємо цю формулу на наступному прикладі. Розглянемо вибірку з 4 значень: 2, 3, 6, 9. Обчислимо дисперсію. Допоміжна таблиця 3-12 матиме всього два стовпці: стовпець значень вибірки і квадрати значень. Сума значень дорівнює 20, а їх квадратів 130. Це все, що необхідно нам для обчислень по другій формулі для дисперсії:

Таблиця 3–12. Розрахунок дисперсії

x	x2
2	4
3	9
6	36
9	91
Σ=20	**Σ=130**

Дисперсія для згрупованих даних

Для згрупованих даних дисперсія обчислюється за такою формулою:

$$s^2 = \frac{n * \sum(f * x^2) - [\sum(f * x)]^2}{n * (n - 1)}$$

Для обчислення за вказаною формулою нам також буде потрібно допоміжна таблиця 3–13.

Таблиця 3–13. Допоміжна таблиця для обчислення дисперсії

Стаж работы	***f***	***x***	***f*x***	***f*x²***
2-4	2	3	6	18
5-7	5	6	30	180
8-10	10	9	90	810
11-13	4	12	48	576
14-16	2	15	30	450
	Σ=23		**Σ=204**	**Σ=2034**

Розглянемо як приклад дані, в яких 23 співробітника вказали свій стаж роботи в компанії і ці дані розміщені в двох перших шпальтах таблиці. У третьому стовпці ми поставимо середній стаж для кожного інтервалу (середину). Четвертий і п’ятий стовпець є допоміжними для обчислень. Отримані значення підставляємо в формулу:

$$s^2 = \frac{n * \sum(f * x^2) - [\sum(f * x)]^2}{n * (n - 1)} = \frac{23 * 2034 - 204^2}{23 * (23 - 1)} = 10{,}2$$

Стандартне відхилення

Стандартне відхилення – квадратний корінь з дисперсії вибірки.

Позначається *s* і обчислюється за формулою:

$$s^2 = \sqrt{\frac{\sum(x - \overline{x})^2}{n - 1}}$$

Коефіцієнт варіації

Коефіцієнт варіації обчислюється як відношення стандартного відхилення до середнього значення вибірки.

Формула для коефіцієнту варіації:

$$CV = s/\overline{x}$$

Коефіцієнт варіації корисний, якщо порівнюються кілька сукупностей, вимірюваних в різних величинах, або порівнюються сукупності, вимірювані в однакових величинах, але мають середні, які сильно відрізняються.

Як приклад з'ясуємо, які дані мають велику варіацію: ті, що мають стандартне відхилення 20 при середньому 200 або мають стандартне відхилення 3 при середньому 30? Скористаємося формулою для коефіцієнту варіації:

$$CV = s/\overline{x} = 20/200 = 0{,}1$$

$$CV = s/\overline{x} = 3/30 = 0{,}1$$

В результаті отримуємо, що в обох випадках коефіцієнти варіації рівні. Це означає, що варіація однакова.

Резюме

Слідом за вивченням числових характеристик центральної тенденції були розглянуті основні числові характеристики варіації даних. Найбільш простими з точки зору знаходження є розмах і квартальний розмах. Більш складні для обчислень, але надзвичайно корисні і важливі для подальших стадій статистичного аналізу такі характеристики як дисперсія і стандартне відхилення. У розрахунках використовувалися кілька різних формул для обчислення дисперсії, які тим не менше дають однакові результати для однакових даних.

Наступний параграф допоможе використовувати всі розглянуті числові характеристики для так званого дослідницького аналізу даних. Цей аналіз завершить вивчення основ описової статистики.

3.3. Дослідницький аналіз даних

Вивчення описової статистики завершує дослідницький аналіз даних, який дозволяє провести комплексний аналіз найважливіших характеристик розподілу.

Дослідницький аналіз даних (Exploratory Data Analysis) – це застосування статистичних методів для подання, упорядкування даних і розуміння їх найважливіших характеристик.

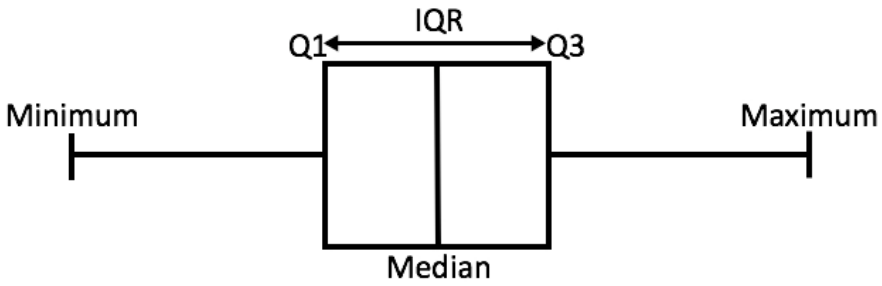
Основними розділами аналізу є:

1. Вимірювання центральної тенденції. Обчислення і аналіз середнього, моди, медіани.
2. Вимірювання варіації. Знаходження мінімуму і максимуму, розмаху і квартильного розмаху, обчислення дисперсії і стандартного відхилення.
3. Знаходження і аналіз викидів. Виділення меж для викидів, аналіз екстремальних і помірних викидів.
4. Аналіз форми розподілу. Обчислення і аналіз коефіцієнтів асиметрії і куртозісу.

Дослідницький аналіз даних дозволяє аналізувати числові значення в якості ключових характеристик і робити висновки на основі цього аналізу, що мають відношення до наявних даних. На жаль, такий аналіз, при всій його комплексності та повноті не дозволяє робити обґрунтованих висновків щодо генеральної сукупності, з якої отримані дані. Це можливо буде зробити в рамках іншої теорії.

Коробкова (вусикова) діаграма

При проведенні аналізу дуже корисна так звана коробкова діаграма, яка має такий вигляд (рис. 3–3):



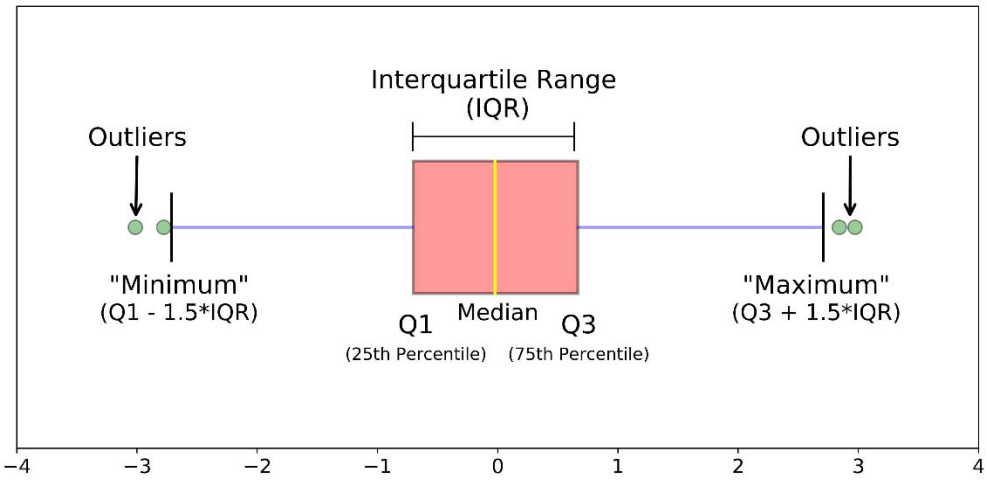
Малюнок 3–3. Коробкова діаграма (Box Plot)

По виду коробкової діаграми можна зробити висновок, де розташована медіана по відношенню до мінімуму і максимуму, по виду і розміру «коробки» можна судити, де розташовані 50% даних.

Викиди

Викидами (outliers) називаються дані, які сильно віддалені від основної кільності даних.

З коробковою діаграмою тісно пов’язаний аналіз викидів. Для знаходження та аналізу викидів крім звичайної коробкової діаграми будується також розширена коробкова діаграма, яка містить позначки помірних і екстремальних викидів.



Малюнок 3–4. Розширена коробкова діаграма

Щоб відшукати викиди нам повинно бути відомо значення IQR – квартильний розмах, який знаходиться як різниця між третім і першим квартилями. Зауважимо, що IQR – це довжина «коробки».

Помірні викиди (mild outliers) – значення ряду даних, що віддалені нижче першого квартилю або вище третього на відстань від 1,5 IQR до 3 IQR.

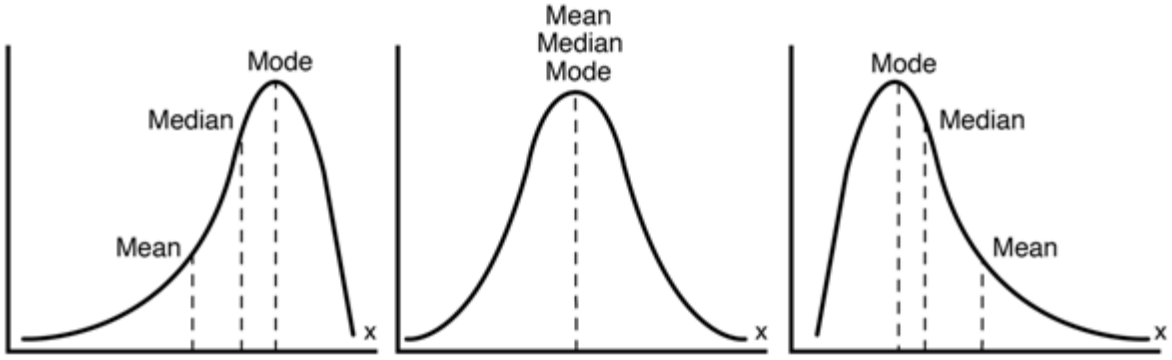
Екстремальні викиди (extreme outliers) – значення ряду даних, віддалені нижче першого квартилю або вище третього на відстань більше 3 IQR.

При аналізі викидів потрібно приймати рішення – або відмовитися від викидів і вести подальші дослідження без них, або залишити викиди для подальшого аналізу. Якщо викиди виключаються, це повинно бути детально аргументовано і описано в звіті про дослідження. Якщо викиди залишаються, слід провести два паралельних дослідження: з ними і без них, а потім зіставити результати і зробити додаткові висновки.

Асиметрія має тісний зв'язок з розташуванням моди, середнього і медіани. Якщо розподіл симетричний, асиметрія дорівнює нулю. В цьому випадку збігаються значення моди, медіани і середнього значення (середній графік).

Якщо одне або кілька значень істотно перевищують інші, є позитивна асиметрія. Середнє більше моди і медіани (лівий графік).

Якщо одне або кілька значень істотно менше за інших, є негативна асиметрія. Середнє менше моди і медіани (правий графік).



Малюнок 3–5. Види асиметричних графіків

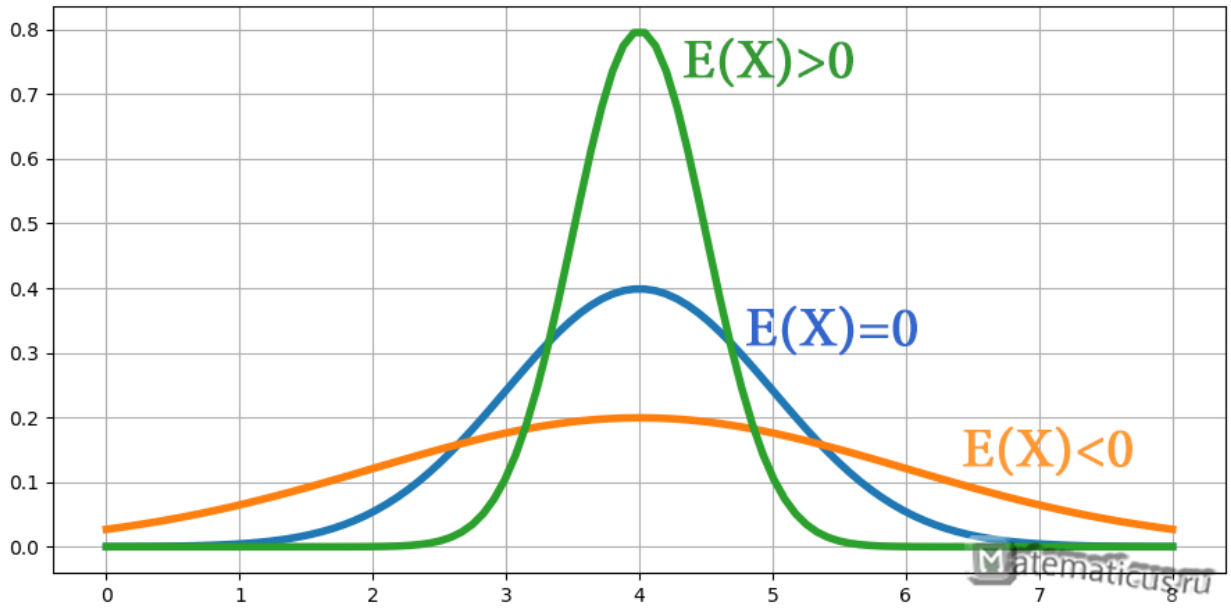
Асиметрія вимірюється за допомогою коефіцієнта, який обчислюється за формулою:

$$Sk = \frac{\overline{x} - Me}{s}$$

Іноді коефіцієнт асиметрії використовується з коефіцієнтом 3 в чисельнику і тоді він змінюється в межах від -3 до +3. У наведеному випадку коефіцієнт змінюється від -1 до +1. Позитивні значення характеризують позитивну асиметрію. У разі, коли медіана і середнє збігаються, коефіцієнт дорівнює нулю.

Ексцес

Під ексцесом розуміється крутість кривої розподілу, яка визначається зіставленням кривої з кривою стандартного нормального розподілу (малюнок 3–6). Обмежимося цим і не будемо вивчати більш детально знаходження відповідних числових характеристик.



Малюнок 3–6. Види розподілу, що відрізняються крутизною

Резюме

Дослідницький аналіз даних включає сукупність методів чисельного аналізу даних. Висновки аналізу не можуть відноситися до генеральної сукупності, а виключно до самих даних, їх розподілу. Він називається дослідницьким, оскільки використовується для отримання первинних висновків, формування гіпотез щодо генеральної сукупності.

Використовуємо комп'ютер

Матеріал цієї глави потребує значної роботи з комп’ютером, оскільки вивчені поняття і характеристики важливо вміти обчислювати в одному зі статистичних пакетів (в нашому випадку в статистичній системі R). Кілька завдань по дослідницькому аналізу даних слід виконати вручну, а потім отримати результати на комп’ютері і порівняти. Далеко не завжди результати виявляться однаковими. Буде потрібна деяка наполегливість, щоб зрозуміти, в чому полягає причина розбіжностей.