2. БАЗОВІ ПОНЯТТЯ СТАТИСТИКИ В ТЕСТУВАННІ

Тестування як вид вимірювання спирається на теорію математичної статистики. У цьому курсі передбачається, що студенти вже вивчали нормативний курс теорії ймовірностей та математичної статистики. Мета цього параграфу – нагадати читачу базові поняття математичної статистики та прив'язати ці поняття до потреб теорії тестування.

Далі розглядатимемо наступну типову ситуацію. Нехай група учнів чи студентів (далі всіх називатимемо учнями) пройшла тест з певної дисципліни, наприклад, біології. Нехай за правильну відповідь на кожне завдання тесту учень отримує певну визначену наперед кількість балів, за неправильну відповідь – 0 балів. Загальним попереднім результатом тестування учня є сума балів, отриманих ним за всі його відповіді. Очевидно, що у різних учнів сума балів може відрізнятися, і вона є наперед невідомою. З точки зору викладача ця сума є випадковою величиною.

З іншого боку, суть процедури тестування полягає в тому, щоб ця сума балів якимось чином відображала вираженість в учня риси або конструкту, що вимірюється (наприклад, рівень успішності засвоєння шкільного курсу біології).

Для цього викладач повинен знати, наскільки тест є валідним (тобто адекватним за рядом важливих аспектів), наскільки він є надійним, яка величина похибки вимірювання тощо. Для відповіді на подібні запитання якраз і використовується апарат математичної статистики.

Вибірковий метод. Найбільш повно поняття і факти математичної статистики використовуються при широкомасштабному стандартизованому тестуванні. У цьому випадку вважається, що тест розробляється для надійного і валідного тестування великої за чисельністю категорії учнів, наприклад, випускників загальноосвітніх середніх шкіл України. Така велика сукупність учнів називається *генеральною сукупністю*, або, простіше, *популяцією*. Під час розробки тесту неможливо визначити і перевірити його характери-

стики на всіх об'єктах популяції. Ця робота проводиться лише на відносно невеликій групі її представників, яка називається *вибіркою*. Отримані на вибірці характеристики тесту можуть, з певною долею достовірності, вважатися справедливими для всієї популяції, якщо вибірка є *репрезентативною*, тобто вона правильно, без спотворень, представляє популяцію щодо вимірюваного конструкту. Зауважимо, що завдання конкретного тесту є, в свою чергу, вибіркою з генеральної сукупності всіх можливих тестових завдань, придатних для вимірювання даної риси чи конструкту.

Загальноприйнятим методом отримання репрезентативної вибірки є *випадковий відбір* представників популяції. Але популяція може бути неоднорідною, тобто складатися з менших популяцій, які істотно відрізняються між собою щодо обставин, які впливають на вираженість риси чи конструкту. Ці менші популяції називають *стратами*. Наприклад, популяція всіх випускників середніх шкіл України може поділятися на страти за типом населеного пункту, у якому розташована школа (місто, село, селище), або за типом школи (звичайна, ліцей, гімназія), або за роком випуску. В подібних випадках для отримання репрезентативної вибірки слід подбати, щоб у вибірці були представлені всі страти у тих долях, у яких вони представлені в генеральній сукупності.

За рівних інших умов, вибірка дозволяє тим точніше визначити характеристики тесту, чим більший об'єм (кількість об'єктів) вона має. Це можна пояснити на такому прикладі: якщо в пологовому будинку одного дня народилося вдвічі більше дівчат, ніж хлопчиків, то це не можна вважати характерним для популяції всіх новонароджених, адже ми знаємо, що у популяції хлопчиків і дівчаток народжується приблизно порівну. Тим не менше, подібна «аномалія» досить часто трапляється у невеликих пологових будинках, де зазвичай народжується лише кілька дітей за день. У великих пологових будинках, де дітей народжується за день десятки або сотні, випадок співвідношення дівчаток до хлопчиків, рівне 2:1, є практично неможливим, натомість, воно зазвичай є ближчим до «правильного» співвідношення 1:1.

З іншого боку, з двох вибірок однакового об'єму, у яких вимірюються різні конструкти, більш інформативною є та, чий конструкт в популяції є менш мінливим. Наприклад, якщо тест перевіряє рівень успішності з окремої теми деякої дисципліни, то для

отримання висновків щодо якості тесту потрібна загалом менша вибірка, ніж для тесту з усієї дисципліни. Пізніше навчимося визначати необхідний для заданих характеристик якості вимірювання об'єм вибірки.

Таблиці частот та діаграми частот. Надалі користуватимемося наступним прикладом. Нехай 50 учнів склали тест з певної навчальної дисципліни. Тест складається з 10 завдань. Відповіді учнів на кожне завдання оцінювалася за *дихотомічною шкалою* (1 бал за правильну відповідь, 0 балів за неправильну). Всі результати зазвичай оформляються у вигляді матриці (таблиці) результатів, у якій кожен окремий рядок містить результати відповідей одного учня на всі завдання тесту, а кожен окремий стовпець – результати відповідей всіх учнів на одне завдання тесту (таблиця 1.1).

Таблиця також містить у крайньому правому стовпці та нижньому рядку суми балів. Ці суми слід розглядати як змінні – реалізації відповідних випадкових величин. Числа у крайньому правому стовпці – це так звані «сирі» бали учнів, отримані ними при проходженні всього тесту. Позначимо цей стовпець-змінну літерою X. Оскільки кожне значення цієї змінної є сумою набраних відповідним учнем балів, тобто сумою нулів і одиниць, то вона може набувати значень від 0 (всі відповіді учня неправильні) до 10 (всі відповіді правильні), всього 11 різних значень. Зокрема, учень за номером 1 відповів правильно лише на половину завдань і отримав відповідно 5 балів, а учень №33 відповів правильно на всі завдання тесту і отримав максимальні 10 балів.

Таблиця 1.1. Матриця результатів тестування

Номер учня	**Бали за завдання 1-10**										**Сума (X)**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
1	0	0	0	1	1	1	0	0	1	1	**5**
2	0	0	0	0	0	0	0	0	0	1	**1**
3	0	0	0	0	1	1	1	1	1	1	**6**
4	0	0	0	1	1	1	0	1	1	1	**6**

5	0	1	0	1	1	0	1	1	1	1	**7**
6	1	0	0	1	1	1	0	1	1	0	**6**
7	0	0	0	0	0	0	0	0	0	1	**1**
8	0	1	0	1	1	1	1	1	1	1	**8**
9	0	0	0	0	1	1	0	1	1	1	**5**
10	0	0	0	1	1	1	1	0	0	1	**5**
11	0	0	0	1	1	1	1	0	0	1	**5**
12	0	0	0	0	1	1	0	0	1	1	**4**
13	0	1	0	1	0	0	1	1	0	1	**5**
14	0	0	0	1	1	1	1	1	0	1	**6**
15	0	0	0	1	1	1	1	0	1	1	**6**
16	0	0	0	0	1	1	1	1	1	1	**6**
17	0	1	0	1	1	1	1	1	1	1	**8**
18	0	0	0	1	1	1	1	0	1	0	**5**
19	0	0	0	1	1	0	1	1	1	1	**6**
20	0	0	0	1	1	1	1	1	1	1	**7**
21	0	0	0	1	0	0	0	1	0	1	**3**
22	0	0	0	1	1	1	1	1	1	1	**7**
23	0	0	0	1	0	0	0	1	1	1	**4**
24	0	0	1	0	1	1	1	1	1	1	**7**
25	0	0	0	0	1	0	1	1	1	1	**5**
26	0	0	1	1	0	1	1	1	1	1	**7**
27	0	0	0	1	1	1	1	1	1	1	**7**
28	0	0	0	1	0	1	1	1	1	1	**6**
29	0	0	0	0	1	1	1	1	1	1	**6**
30	0	1	1	1	1	1	1	1	1	1	**9**
31	0	0	0	0	0	0	1	0	1	0	**2**
32	0	0	0	0	1	1	1	1	1	1	**6**
33	1	1	1	1	1	1	1	1	1	1	**10**
34	0	0	1	1	1	1	1	1	1	1	**8**
35	0	1	1	1	1	0	1	1	1	1	**8**

36	0	1	1	1	1	1	1	1	1	1	**9**
37	0	0	1	0	1	1	1	1	1	1	**7**
38	0	0	0	1	0	1	1	1	1	1	**6**
39	0	0	0	1	1	1	1	1	1	1	**7**
40	0	0	1	1	1	1	1	1	1	1	**8**
41	0	0	0	0	0	0	0	1	1	1	**3**
42	0	0	0	0	1	1	1	1	1	1	**6**
43	0	0	1	1	1	1	1	1	1	1	**8**
44	0	0	1	1	1	1	1	1	1	1	**8**
45	0	1	1	1	1	1	1	1	1	1	**9**
46	0	1	1	1	1	1	1	1	1	1	**9**
47	0	0	1	0	1	1	1	1	1	1	**7**
48	0	0	0	1	0	1	1	1	1	1	**6**
49	0	0	0	1	1	1	1	1	1	1	**7**
50	0	0	0	0	0	0	1	1	1	1	**4**
Сума (*Y*)	**2**	**10**	**14**	**34**	**38**	**38**	**40**	**41**	**43**	**47**	**307**

Наскільки сумарний результат кожного з учнів є типовим для даної вибірки учнів? Іншими словами, як часто зустрічається кожне з можливих значень змінної *X* у стовпці результатів? Відповідь на це питання міститься у другому стовпці *таблиці частот* (таблиця 1.2).

Таблиця 1.2. Таблиця частот

X	**Частоти *f(X)***	**Накопичені частоти *cf(X)***	**Відносні частоти *p(X)***	**Накопичені відносні частоти *cp(X)***
0	0	0	0	0
1	2	2	0,04	0,04
2	1	3	0,02	0,06
3	2	5	0,04	0,1
4	3	8	0,06	0,16
5	7	15	0,14	0,3

6	13	28	0,26	0,56
7	10	38	0,2	0,76
8	7	45	0,14	0,9
9	4	49	0,08	0,98
10	1	50	0,02	1

З таблиці видно, що, наприклад, результат 0 балів не зустрічається у вибірці жодного разу, а результат у 6 балів зустрічається найбільш часто (13 разів), тобто є якоюсь мірою типовим для даної вибірки. Ми не даремно тут кажемо «якоюсь мірою». Припустимо, що в деякій іншій вибірці учнів результат 6 балів зустрічається 14 разів, а всі інші 36 результатів розподілені так: 0 балів, 1 бал, 2 бали – по 12 разів, решта – по 0 разів. У такому випадку справедливо буде вважати тест загалом важким для учнів, а результат найбільш частий результат 6 балів – не дуже характерним. Зауважимо, що подібний розподіл результатів тестування був би, в загальному випадку, дуже підозрілим. Він міг би вказувати, наприклад, на сильну неоднорідність вибірки. Скажімо, так могло б бути, якби тест перевіряв знання математики за 5 клас, а вибірка складалася з 36 першокласників і 14 п'ятикласників. Для нашого ж прикладу результат тестування у 6 балів є більш типовим, оскільки близькі результати, зокрема 5 і 7 балів – теж зустрічаються у вибірці з подібними частотами.

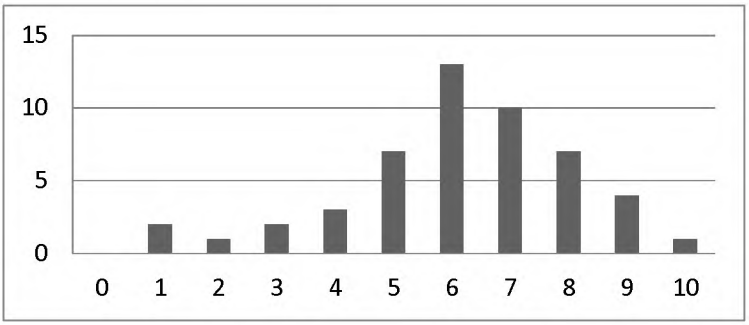


Рис. 1.1. Гістограма частот

Для кращого розуміння результатів тестування частоти $f(X)$ з таблиці 1.2 корисно зобразити у вигляді гістограми (стовпчикової діаграми). Для кожного значення суми балів X, відкладеної вздовж горизонтальної осі, будуємо стовпець, висота якого відповідає частоті цього результату (рис. 1.1).

З малюнка, зокрема, видно, що тест для даної вибірки учнів був загалом нескладним, оскільки найбільші частоти зосереджені в області значень змінної X від 5 до 9. Загалом 41 учнів з 50 отримали за тест від 5 до 9 балів.

Четвертий стовпець таблиці 1.2 містить *відносні частоти* $p(X)$. Для отримання відносної частоти потрібно кожну з частот поділити на об'єм вибірки (у нашому прикладі – на 50). Відносна частота пов'язана з теоретичним поняттям *ймовірності*: якщо вибірка учнів є репрезентативною, то для кожного нового учня з цієї ж популяції ймовірність отримати певну кількість балів за тест приблизно дорівнює відносній частоті цієї кількості балів у вибірці. Наприклад, ймовірність того, що новий учень отримає за тест 7 балів, дорівнює приблизно 0,2.

Числові вибіркові характеристики. Для описання результатів тестування використовуються числові характеристики, які ще називають описовими статистиками. Кожна окрема характеристика – це число, яке характеризує усю вибірку у відповідному аспекті. Всі числові характеристики можна поділити на три групи: міри положення, міри мінливості, міри форми. Далі розглянемо лише деякі, найважливіші, характеристики.

Міри положення, або, як їх ще називають, міри центральної тенденції, вказують на найбільш характерні значення випадкової величини. До цієї групи характеристик відносяться мода, медіана, середнє, а також більш загальне поняття процентиля.

Мода (англ. *mode*) – це значення випадкової величини, яке зустрічається у вибірці найбільш часто. Для нашого прикладу модою є значення $X = 6$, оскільки цей результат зустрічається найбільш часто – 13 разів. Вище ми вже зазначали, що модальне значення в одних випадках добре характеризує типові значення змінної, в інших випадках – гірше. З причин, які розглядатимемо далі, нормо-орієнтований тест слід вважати добре збалансованим по трудності завдань, якщо модальним є значення змінної, близьке до середнього. Так, для нашого прикладу це повинне було б бути

значення, близьке до $X = 5$. Іншими словами, слід було б очікувати, що середній результат у 5 балів набере найбільша кількість учнів.

Медіана (англ. *median*) – це таке значення випадкової величини X, яке ділить вибірку на дві частини приблизно порівну за кількістю об'єктів так, щоб у одній частині опинилися об'єкти із меншими за медіану або рівними їй значеннями, а в іншій частині – з більшими значеннями. Наприклад, якщо у класі 21 учень, то медіана їх зросту дорівнює зросту 11-го учня у вишикуваній за зростом шерензі учнів. Для нашого прикладу тестування 50 учнів медіанним слід вважати результат $X = 6$. Знайти медіану допомагає третій стовпець таблиці 1.2 – стовпець *накопичених частот* $cf(X)$, або п'ятий стовпець цієї таблиці – стовпець *накопичених відносних частот* $cp(X)$. Числа у цих стовпцях отримуються послідовним додаванням кожного нового значення з тих стовпців, які знаходяться зліва (відповідно, частот і відносних частот), до суми попередніх значень. Так, для значення $X = 3$ накопчена частота дорівнює $0 + 2 + 1 + 2 = 5$. Слід пам'ятати, що для відшукання накопичених частот значення змінної X повинні бути розташовані у таблиці за зростанням. Медіанне значення змінної знаходиться у тому рядку таблиці, для якої значення накопиченої частоти досягло 25 (половина вибірки), або значення відносної накопиченої частоти досягло 0,5.

Медіана є частинним випадком *процентиля*: p-й процентиль – це таке значення змінної, яке ділить вибірку так, що p відсотків об'єктів вибірки мають значення, менше або рівне даному. У нашому прикладі учень, який отримав за тест 9 балів, відповідає 98-му процентилю (значення 0,98 у стовпці накопичених відносних частот), що означає, що 98 відсотків учнів склали тест з результатом, не вищим від його результату. При нормоорієнтованому тестуванні результати інколи можуть повідомлятися учням саме у вигляді процентилів. Особливо інформативним є такий підхід у випадку, якщо до нього також додаються інші дані, які характеризують вибірку в цілому, наприклад, у якій школі (спеціалізованій чи звичайній, сільській чи міській) навчаються учні.

Медіана є 50-м процентилем. Вживаються також такі поняття, як *квартилі* – 25-й та 75-й процентилі.

Медіана використовується частіше для описання змінних, виміряних у порядковій шкалі. Для метричних змінних більш інформативним може виявитися поняття середнього.

Середнє вибіркове (англ. *mean*) – це середнє арифметичне усіх значень змінної:

$$\overline{\mathrm{X}} = \frac{\sum_{i=1}^{N} X_i}{N},$$

де N – об'єм вибірки. У нашому прикладі $N = 50$,

$$\bar{X} = \frac{0 \times 0 + 1 \times 2 + 2 \times 1 + \cdots + 9 \times 4 + 10 \times 1}{50} = \frac{307}{50} = 6{,}14.$$

Середнє вибіркове є *оцінкою* теоретичного значення математичного очікування випадкової величини для всієї популяції. Чим більш репрезентативною є вибірка, тим точнішою є ця оцінка.

Слід обережно трактувати значення середнього вибіркового. В літературі часто згадується такий жартівливий приклад: середня температура пацієнтів лікарні може дорівнювати 36,7, проте це зовсім не означає, що пацієнти здорові, і цю середню температуру не можна трактувати як показник успішності роботи лікарів. Середнє вибіркове зазвичай розглядається у парі з характеристикою мінливості змінної – статистичною дисперсією або середнім квадратичним відхиленням.

Статистична дисперсія (англ. *variance)*є характеристикою, яка входить до групи мір мінливості змінної. Ця характеристика є середнім арифметичним квадратів відхилень значень змінної від середнього вибіркового:

$$D(X) = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N}.$$

Тобто значення дисперсії показує, як часто і на скільки сильно відхиляються значення змінної від свого центра. Відхилення тут підносяться до квадрату для того, щоб усі доданки були невід'ємними, тобто щоб відхилення з різними знаками не компенсували одне одного. Більш природним було б використання модулів відхилень замість їх квадратів, однак використання модулів є

більш складним. Чим більшим є розсіювання значень змінної навколо середнього, тим більшою є дисперсія. Якщо змінна набуває лише одного значення, то воно ж і є центром, відхилення не спостерігаються, і дисперсія у цьому випадку дорівнює нулю. Якщо змінна набуває більше ніж одного значення, дисперсія завжди є додатною. Розглянемо для прикладу три учнівські класи, які отримали підсумкові оцінки з української мови за 12-бальною шкалою. Нехай у класі A всі учні отримали оцінку 8, у ласі B – половина учнів отримали 7, інша половина – 9, у класі C – половина отримали 6, інша половина – 10. Середня оцінка у кожному з цих класів однакова – 8 балів. Проте очевидно, що за рівнем успішності з предмету окремих учнів ці класи істотно відрізняються. В оцінках класу A ніякої мінливості не спостерігається, дисперсія цих оцінок дорівнює нулю. У класі B дисперсія вже не нульова, а у класі C вона є більшою, ніж у класі B.

Статистична дисперсія є оцінкою теоретичної дисперсії змінної для всієї популяції. Проте ця оцінка є *зміщеною*: її математичне очікування не дорівнює теоретичній дисперсії. Виправлену (незміщену) оцінку отримаємо, якщо у знаменнику формули для статистичної дисперсії замінимо N на $N-1$:

$$s^2 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N-1}.$$

У нашому прикладі тестування 50 учнів отримаємо таке значення незміщеної вибіркової дисперсії:

$$s^2 = \frac{0 \times (0-6{,}14)^2 + 2 \times (1-6{,}14)^2 + \cdots + 1 \times (10-6{,}14)^2}{50-1} \approx 3{,}9188.$$

Оскільки у формулі для обчислення дисперсії використовується операція піднесення до квадрату, отримуємо величину, обчислену в квадратних одиницях. Для того, щоб повернутися до тієї ж самої розмірності, яку має вимірювана змінна, візьмемо корінь квадратний з дисперсії:

$$s = \sqrt{s^2}.$$

Ця величина називається *середнім квадратичним*, або *стандартним, відхиленням* (англ. *standard deviation*) випадкової величини X. Для нашого прикладу маємо

$$s \approx \sqrt{3{,}9188} \approx 1{,}98.$$

Оскільки операція добування кореня квадратного з невід'ємного числа є монотонним перетворенням, стандартне відхилення має таку ж інтерпретацію, як і дисперсія: воно є мірою мінливості змінної у вибірці.

До групи *характеристик форми* розподілу змінної у вибірці відносяться, в першу чергу, асиметрія і ексцес. Обидві ці характеристики використовуються для описання розподілів, близьких до нормального, тому розглянемо їх пізніше.

Нормальний розподіл. Нормальний, або Гаусів, закон розподілу випадкової величини відіграє особливу роль в теорії і практиці вимірювань. Це пов'язано з тим, що ті випадкові величини, які є сумою багатьох випадкових величин (іншими словами, сформовані під впливом багатьох випадкових факторів), за умови, що вплив кожного з доданків на всю суму не є визначальним, має розподіл, близький до нормального. Оскільки ті випадкові величини, які зустрічаються в природі, на виробництві чи в інших сферах людської діяльності, якраз і формуються під впливом багатьох факторів, їх розподіли часто виявляються близькими до нормального. Зокрема, слушно вважати, що рівень інтелекту розподілений для популяції людей за нормальним законом. Так само з великою мірою впевненості можна стверджувати, що рівень навчальних досягнень учнів з деякого предмету теж добре описується нормальним законом розподілу. Разом з тим, слід пам'ятати, що нормальний розподіл – це лише математична модель, яка в одних випадках добре узгоджується з спостережуваними на практиці величинами, в інших випадках – погано.

Нормальним законом розподілу ймовірностей випадкової величини називається розподіл, щільність якого виражається формулою

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

де μ і σ – параметри розподілу (відповідно, математичне очікування і стандартне відхилення). На малюнку 1.2 схематично зображено графік щільності нормального розподілу.

Щільність розподілу – це функція, графік якої разом з віссю абсцис обмежує площу, рівну ймовірності того, що реалізація випадкової величини потрапить в результаті випробування у заданий інтервал.

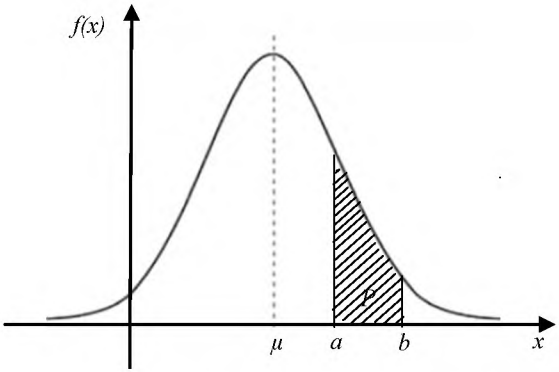


Рис. 1.2. Щільність нормального розподілу

Площа заштрихованої області на рисунку 1.2 дорівнює ймовірності того, що реалізація випадкової величини з таким розподілом потрапить у інтервал $[a, b]$. Очевидно, що ймовірність потрапляння нормально розподіленої випадкової величини в інтервал заданої довжини є найбільшою, якщо цей інтервал є симетричним відносно центра розподілу μ. Тобто реалізації нормально розподіленої випадкової величини зосереджені найбільше поблизу центра. Чим більше відрізняється значення випадкової величини від μ, тим рідше воно трапляється. Іншими словами, люди з середнім зростом, чи з середнім рівнем інтелекту зустрічаються частіше, а з відхиленнями від середнього в ту чи іншу сторону – тим рідше, чим більшим є це відхилення. Зокрема, слід очікувати, що учнів з середнім рівнем навчальних досягнень з певного предмету має бути

більше, ніж учнів з слабким або, навпаки, високим рівнем досягнень.

Варто пам'ятати такі значення ймовірностей потрапляння нормально розподіленої випадкової величини X у заданий симетричний відносно центра розподілу інтервал:

$$P\{X \in [\mu - \sigma, \mu + \sigma]\} \approx 0{,}68.$$

$$P\{X \in [\mu - 2\sigma, \mu + 2\sigma]\} \approx 0{,}95.$$

$$P\{X \in [\mu - 3\sigma, \mu + 3\sigma]\} \approx 0{,}997.$$

Тобто приблизно 68% реалізацій X відрізняються від середнього μ не більше, ніж на σ, 95% - не більше ніж на 2σ. І майже напевне (у 99,7% випадках) відхилення буде не більшим, ніж на 3σ. Останнє твердження називають «правилом трьох сигм».

Важливою властивістю нормального розподілу є той факт, що його математичне очікування, мода та медіана збігаються.

Як ми вже зазначали, нормальний розподіл є у багатьох практичних застосуваннях моделлю, яка більшою чи меншою мірою описує реальні дані. Відхилення реальних даних від нормального закону описується, зокрема, такими мірами форми розподілу, як асиметрія і ексцес.

Асиметрія (точніше, коефіцієнт асиметрії, англ. *skewness*) обчислюється за формулою:

$$As = \frac{E(X - EX)^3}{\sigma^3},$$

де символ E означає математичне очікування. Величина у чисельнику – це центральний момент третього порядку (математичне очікування куба відхилення. Дисперсія є центральним моментом другого порядку). Виправленою оцінкою асиметрії за даними вибірки є величина

$$As = \frac{N}{(N-1)(N-2)} \frac{(\sum_{i=1}^{N}(X_i - \bar{X})^3)}{s^3}.$$

Значення асиметрії дорівнює нулю, якщо щільність розподілу симетрична відносно центра (математичного очікування). Зокрема, якщо випадкова величина розподілена нормально, то асиметрія дорівнює нулю. Якщо асиметрія додатна, то це означає, що правий хвіст розподілу є довшим за лівий (рис. 1.3), а якщо асиметрія від'ємна – то, навпаки, лівий хвіст довший.

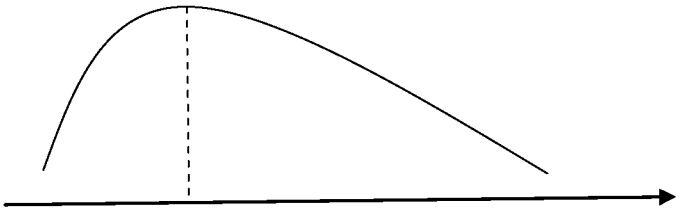


Рис. 1.3. Випадок додатної асиметрії розподілу

Якщо асиметрія результатів тестування популяції учнів з нормальним розподілом рівня успішності додатна, це може означати, що тест є надто складним для учнів, якщо асиметрія від'ємна – то, навпаки, надто легким.

У нашому прикладі тестування 50 учнів гістограма розподілу частот (рис. 1.1) показує, що крива, яка огинає гістограму, має вершину, зсунуту вправо від середнього значення частоти, і лівий хвіст розподілу є довшим. Отже, слід очікувати, що асиметрія розподілу результатів тестування є від'ємною. Дійсно, підставивши значення у формулу для вибіркової оцінки асиметрії, знайдемо, що вона дорівнює приблизно $-0{,}663$.

Ексцес (англ. *kurtosis*) є мірою гостровершинності розподілу у порівнянні з нормальним. Якщо розподіл більш крутий, ніж нормальний, ексцес є додатним. Якщо розподіл більш пологий, то ексцес від'ємний. Величина теоретичного коефіцієнта ексцесу обчислюється за формулою:

$$Ex = \frac{E(X - EX)^4}{\sigma^4} - 3.$$

Перший доданок є відношенням центрального моменту четвертого порядку до четвертого степеня стандартного відхилення. Для нормального розподілу цей доданок дорівнює 3. Тому у формулу вводиться другий доданок, щоб визначений таким чином коефіцієнт дорівнював для нормального розподілу нулю. Незміщена оцінка ексцесу обчислюється за формулою:

$$Ex = \frac{N(N+1)}{(N-1)(N-2)(N-3)} \sum_{i=1}^{N} \left(\frac{X_i - \bar{X}}{s} \right)^4 - \frac{3(N-1)^2}{(N-2)(N-3)}.$$

Для нашого прикладу тестування 50 осіб (таблиця 1.1) величина вибіркової незміщеної оцінки ексцесу дорівнює приблизно 0,619, тобто розподіл тестових балів у нашому випадку має більш гостру вершину у порівнянні з нормальним.

Як ми вже зазначали, багато випадкових величин, з яким має справу людина у своїй практичній діяльності, добре моделюються нормальним розподілом. Разом з тим, ми добре розуміємо, що розподіл, наприклад, зросту чоловіків відрізняється від розподілу зросту жінок, який в свою чергу, відрізняється від розподілу рівня інтелекту у жінок. Усі ці розподіли відрізняються значеннями параметрів середнього μ і стандартного відхилення σ.

Як впливають значення параметрів розподілу на положення і форму кривої щільності розподілу? Якщо змінювати лише параметр μ, то форма кривої не змінюватиметься, але сама вона буде зсуватися вздовж горизонтальної осі так, щоб її вершина була у точці з новим значенням μ. Якщо, навпаки, змінювати параметр σ, то крива буде стискатися або розтягуватися вздовж вертикальної осі: при збільшенні значення σ крива буде ставати більш пологою, а при зменшенні – більш крутою. Це стає зрозумілим, якщо пригадати, який зміст має стандартне відхилення: чим більше його значення, тим частіше зустрічаються відхилення від центра, тобто щільність у самому центрі зменшується, зростаючи натомість у більш віддалених точках.

Будь-який нормальний розподіл можна легко перетворити на *стандартний нормальний розподіл*, тобто розподіл з середнім 0 і

стандартним відхиленням 1, за допомогою заміни змінної, яку називають z-перетворенням:

$$z = \frac{x - \mu}{\sigma}.$$

Це перетворення є надзвичайно важливим з двох причин. По-перше, таблиці значень щільності і функції нормального розподілу існують тільки для випадку $\mu = 0$ і $\sigma = 1$. По-друге, для порівняння схожих за природою, але виміряних у різних шкалах, довільно розподілених величин, їх потрібно спочатку привести до єдиної шкали. Як порівняти, наприклад, оцінки з двох предметів, якщо один з них оцінювався за шкалою 100-200 балів, а інший – за шкалою 0-10 балів? Якщо взяти за базову першу шкалу, то оцінки, отримані за другою шкалою, слід перетворити, помноживши їх на 10 (перехід до шкали 0-100) і потім додавши до них 100. У випадку зведення нормальних розподілів вигідно обидва звести до стандартного, оскільки для нього існують таблиці значень щільності та функції розподілу.

Якщо потрібно знайти ймовірність, з якою нормально розподілена випадкова величина потрапить у заданий інтервал $[a,\ b]$, використовується формула:

$$P\{X \in [a, b]\} = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right),$$

де $\Phi(x)$ - функція Лапласа, значення якої можна знайти за спеціальною таблицею. Ця функція відрізняється від функції *стандартного нормального розподілу* (тобто нормального розподілу з середнім 0 і стандартним відхиленням 1) у кожній точці на величину 0,5, але нею користуватися більш зручно, тому що вона є непарною (тобто $\Phi(-x) = -\Phi(x)$) і тому немає потреби заносити у таблицю її значень значення для від'ємних x. Вирази в дужках є z-перетворенням відповідних кінців відрізка. На малюнку 1.4. схематично зображені графіки функції Лапласа та функції стандартного нормального розподілу.

З малюнка видно, що вже при $X = 3$ значення функції Лапласа наближається до 0,5.

Коли потрібно аналізувати вибірку, зокрема, дані тестування групи осіб, то для того, щоб використати нормальний розподіл у якості моделі, необхідно вирішувати, наскільки добре модель підходить, і з якими значеннями параметрів. Гіпотезу про відповідність розподілу вибіркових даних нормальному для заданого рівня значущості слід перевіряти відповідними статистичними методами. Для цього краще скористатися спеціальним програмним забезпеченням.

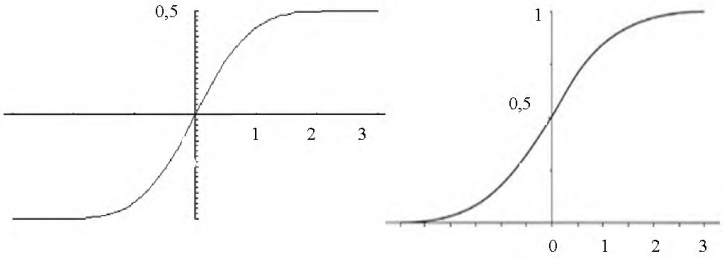


Рис. 1.4. Зліва – функція Лапласа, справа – функція стандартного нормального розподілу

Статистичний зв'язок між змінними. Розглянемо тепер ситуацію, коли в кожному випробуванні спостерігається не одна, а дві випадкові змінні. Наприклад, група учнів складає два тести з різних дисциплін. Тепер поняття вибірки ми не будемо ототожнювати з об'єктами (у нашому випадку – учнями), а з самими змінними – характеристиками об'єктів, які вимірюються. Таким чином, під вибіркою розумітимемо просто сукупність чисел, яка має певний статистичний розподілу, тобто для однієї і тієї ж групи N досліджуваних будемо розглядати дві, за кількістю змінних, вибірки об'єму N.

При одночасному спостереженні двох змінних з'являється принципово нове і надзвичайно важливе питання: чи існує зв'язок між змінними? Припустимо, що перед нами учень, про якого відомо, що він складав два тести – з алгебри та з геометрії. Припустимо також, що нас цікавить результат тестування цього учня з геометрії. Якщо розподіл тестових балів з геометрії для групи, до якої

належить наш учень, відомий, ми можемо на основі цього розподілу робити припущення щодо оцінки цього учня. Наприклад, якщо відомо, що дві третини учнів отримали менше 70 балів, ми з упевненістю в майже 67 відсотків очікуємо, що й наш учень набрав менше 70 балів. Але нехай нам стало відомо, що учень отримав найвищий бал з алгебри. Швидше за все, це змусить нас переглянути свої припущення щодо його оцінки з геометрії: тепер ми з меншою ймовірністю, ніж 67 відсотків, очікуємо, що у нього менше 70 балів, натомість зростає наша віра у те, що його оцінка є вищою. Це відбувається тому, що ми припускаємо наявність зв'язку між оцінками з алгебри та геометрії: чим вищою є оцінка з алгебри, тим вищою вона має бути й з геометрії.

Але чи можемо ми, погоджуючись з існуванням подібного зв'язку, і знаючи оцінку учня з алгебри, зі 100-відсотковою упевненістю назвати його оцінку з геометрії? Очевидно, ні. На обидві оцінки впливає багато випадкових факторів, які ми не можемо врахувати. Хоча ми й очікуємо, що висока оцінка з алгебри означає також високу оцінку з геометрії (і навпаки), цілком ймовірною є, наприклад, ситуація, коли одна з оцінок учня є високою, а інша – низькою, і цьому може бути багато причин. Іншими словами, припускаючи існування зв'язку між змінними, ми, разом з тим, розуміємо, що цей зв'язок не є однозначним, іншими словами, він є ймовірнісним.

Взагалі кажучи, між двома змінними може існувати кілька різних видів зв'язку. Для нас надалі важливо буде чітко розрізняти три види зв'язку: функціональний, статистичний, кореляційний.

Функціональним називається такий зв'язок між змінними X та Y, коли кожному можливому значенню X відповідає одне і тільки одне значення Y. Часто функціональний зв'язок може задаватися математичним рівнянням, наприклад: $Y = 2X + 3$. У цьому прикладі значенню $X = 5$ відповідає єдине значення $Y = 13$. Якби між результатами тестування з двох дисциплін існував функціональний зв'язок, ми, знаючи одну з оцінку, знали б і іншу. Зауважимо, що у такому разі одне з тестувань виявилося б просто зайвим.

В реальних ситуаціях, спостерігаючи дві змінні, кожна з яких є випадковою величиною (а саме такими і є результати двох різних психометричних вимірювань), ми ніколи не побачимо між ними функціонального зв'язку. Втім, це не обов'язково означає,

що зв'язку немає взагалі. Між змінними може існувати *ймовірнісний* (*стохастичний*, *статистичний*) зв'язок. Тут важливо розуміти, як співвідноситься теорія з практикою, математична модель – з даними спостережень. Розглядаючи рівні навчальних досягнень групи учнів з алгебри і геометрії, цілком природно припустити, що між ними *в теорії* існує функціональний зв'язок у формі, наприклад, лінійної залежності виду $Y = aX + b$, чи квадратичної залежності виду $Y = aX^2 + b$ тощо, але реальні результати вимірювання шляхом тестування показуватимуть відхилення від цього зв'язку в той чи інший бік, унаслідок впливу різноманітних сторонніх випадкових факторів. Також ми в теорії можемо допускати існування зв'язку між рівнями навчальних досягнень з алгебри і української мови, але природно очікувати, що цей зв'язок є *менш тісним*, ніж у випадку алгебри і геометрії, тобто відхилення від функціонального зв'язку між результатами тестування з алгебри і української мови очікуються в середньому більшими. Таким чином, важливою характеристикою статистичного зв'язку є його *сила* (*тіснота*), як міра відхилення від теоретичного функціонального зв'язку.

Розглянемо реальний приклад – результати тестів ЗНО 2010 року з математики (X) і української мови та літератури (Y) п'ятдесяти вступників до фізико-математичного факультету Ніжинського державного університету імені Миколи Гоголя. Для більшої виразності бали з української мови та літератури переведені з шкали 100-200 у шкалу 0-10 балів лінійним перетворенням: від кожної оцінки ЗНО відняли 100 балів і результат поділили на 10. Подамо результати у вигляді таблиці 1.3.

Таблиця 1.3. Результати ЗНО 50 вступників

Укр. мова	**Математика**
10	181
9	174 194 190 187
8	184 180 177 194 174 178 188 173 158 165
7	180 188 174 168 177 171 160 163 168 173 168
6	179 175 169 190 156 170 177 147 158 154
5	162 174 129 166 147 157
4	155 162 158 147 147 125
3	142 140

У таблиці дані згруповані за оцінкою з української мови: один вступник мав 10 з мови і 181 з математики, чотири вступники мали 9 з мови і 174, 194, 190 та 187 з математики відповідно і так далі. Як бачимо для кожного із значень оцінки з мови та літератури існує свій набір оцінок з математики. Цей набір називається умовним розподілом. Наприклад, у передостанньому рядку бачимо умовний розподіл оцінок вступників з математики *за умови*, що оцінка з мови дорівнює 4. Іншими словами, *умовний розподіл* змінної Y за умови $X = a$ – це набір значень, які набула змінна Y при фіксованому значенні змінної $X = a$. Позначимо тепер оцінки кожного вступника точкою на декартовій площині з координатами (українська мова та література, математика) (рис. 1.5).

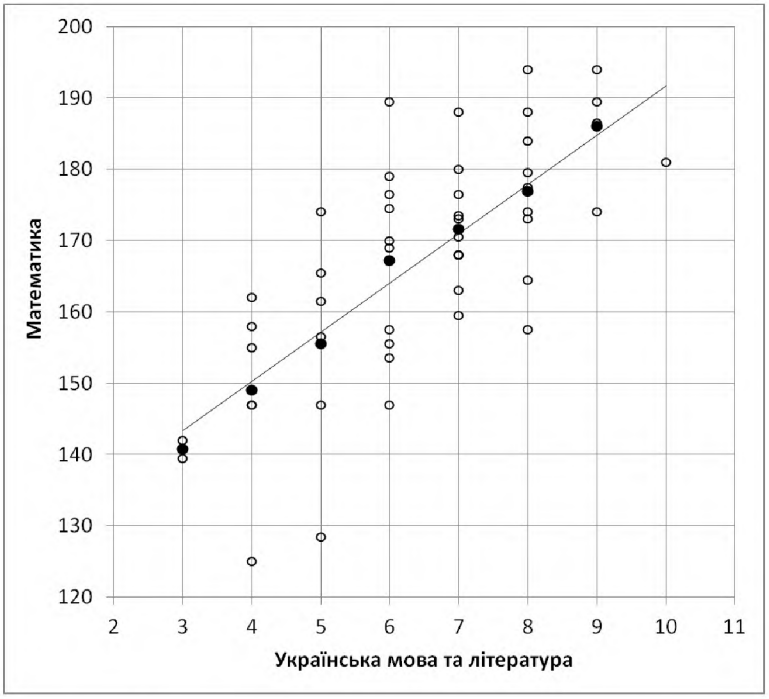


Рис. 1.5. Розподіл оцінок 50 вступників за двома предметами

Для кожної оцінки з української мови та літератури знайдемо середнє значення відповідних оцінок з математики і позначимо його на рисунку чорним кружечком. Як бачимо, ці середні значення розташовані досить близько до деякої прямої. Це дає змогу припустити, що умовні математичні очікування оцінок з математики усієї популяції вступників, до якої належить вибірка, знаходяться на цій прямій, тобто існує теоретична лінійна залежність між оцінками. Сама пряма називається *прямою регресії* змінної Y по змінній X або *лінією тренду*. Теоретичне рівняння прямої регресії Y по X має вигляд:

$$y - EY = r_{xy}\frac{\sigma_Y}{\sigma_X}(x - EX).$$

Тут EX, EY – математичні очікування випадкових величин X та Y; σ_X, σ_Y – їх середні квадратичні (стандартні) відхилення. Нова для нас величина r_{XY} – це коефіцієнт кореляції Пірсона. Ця величина відіграє надзвичайно важливу роль, тому далі розглянемо її детально. Оскільки, володіючи лише вибірковими даними, ми не зможемо дізнатися істинних значень математичних очікувань змінних, їх стандартних відхилень та коефіцієнта кореляції, то замість теоретичного рівняння прямої регресії ми можемо використовувати *вибіркове рівняння*, яке отримаємо, замінивши у наведеній вище формулі теоретичні величини їх вибірковими оцінками:

$$y - \bar{Y} = \rho_{XY}\frac{S_Y}{S_X}(x - \bar{X}).$$

Тут ми замінили математичні очікування їх вибірковими оцінками – середніми арифметичними відповідних вибірок, а стандартні відхилення – вибірковими стандартними відхиленнями. Формула для вибіркової оцінки коефіцієнта кореляції Пірсона:

$$\rho_{XY} = \frac{\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^{N}(x_i - \bar{X})^2 \sum_{i=1}^{N}(y_i - \bar{Y})^2}.$$

Усі вибіркові оцінки отримують зазвичай за допомогою спеціальних комп'ютерних програм. Наприклад, у середовищі

Microsoft Excel для вибіркового рівняння прямої регресії, записаного у стандартному вигляді $y = ax + b$, для обчислення параметра b, який називається інтерцептом і вказує на точку перетину прямої з віссю Oy, використовується стандартна функція INTERCEPT, а для обчислення параметра a, який є коефіцієнтом нахилу прямої до осі Ox – функція SLOPE. У нашому прикладі отримаємо рівняння прямої:

$$y = 6{,}906x + 122{,}637.$$

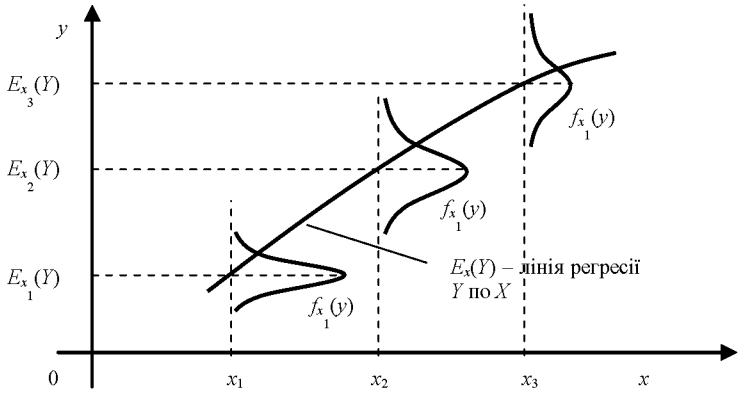
Навіщо потрібна пряма регресії? Знаючи її рівняння, можна прогнозувати оцінку деякого вступника з математики на основі його оцінки з мови та літератури. Припустимо, що вступник має оцінку 8 з мови та літератури. Підставивши це значення у рівняння, отримаємо:

$$y = 6{,}906 \times 8 + 122{,}637 \approx 178.$$

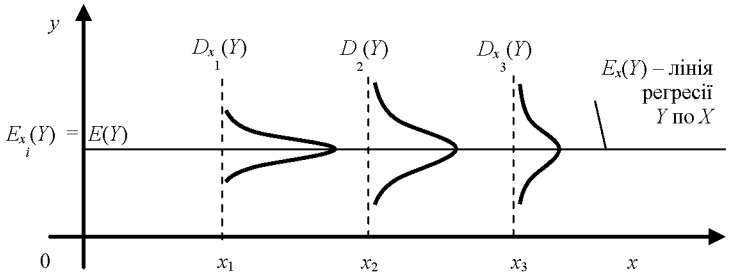
У програмі Excel є спеціальна функція TREND для обчислення прогнозованого значення Y за даним X.

Таким чином, очікувана оцінка цього вступника з мови та літератури – 178 балів. Зауважимо, що зазвичай реальна оцінка відрізняється від прогнозованої. Точність прогнозу залежить від тісноти статистичного зв'язку між змінними. Розглянемо тепер більш детально величину, яка характеризує силу лінійного зв'язку між змінними – коефіцієнт кореляції Пірсона.

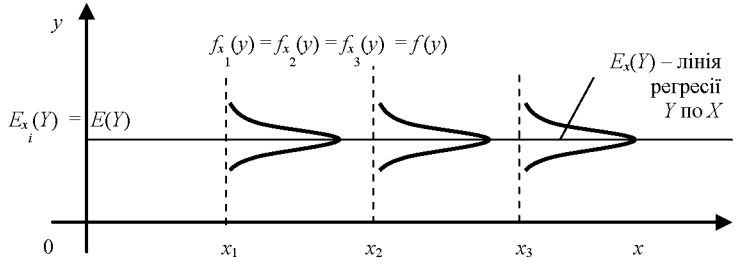
Кореляційний зв'язок між змінними. Коефіцієнти кореляції Пірсона та Спірмена. Перш за все, зауважимо, що статистичний та кореляційний зв'язки – це різні поняття. Відсутність кореляційного зв'язку не означає, що між змінними немає статистичного зв'язку. Розглянемо три різні випадки, зображені на рис. 1.6. У випадку а) при зростанні змінної x в цілому спостерігається також і зростання y, причому із зміною x також змінюються і умовні розподіли y – збільшується дисперсія розподілу. Це свідчить про наявність кореляційного зв'язку між змінними, який, проте, не є лінійним. У випадку б) лінія регресії є горизонтальною прямою, що вказує на відсутність лінійного кореляційного зв'язку (коефіцієнт a у рівнянні прямої дорівнює нулю).



а)



б)



в)

Рис. 1.6. Співвідношення статистичного та кореляційного зв'язку

Проте між змінними все ж спостерігається статистичний зв'язок: із зростанням x змінюються умовні розподіли y. Нарешті, випадок в) вказує на відсутність і кореляційного, і статистичного зв'язку між змінними – умовні розподіли змінної y при зміні x залишаються однаковими.

На рис. 1.7 зображено випадок, коли між змінними існує функціональний зв'язок у вигляді залежності $x = y^2 + 10$, а проте лінія тренду є горизонтальною, тобто лінійна кореляція не спостерігається.

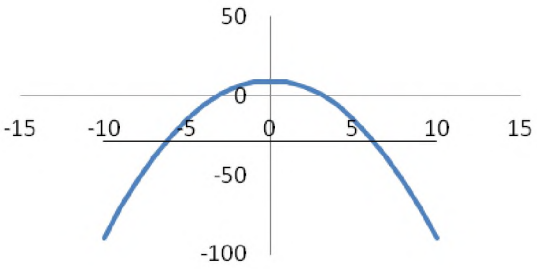


Рис. 1.7. Між змінними існує строга функціональна залежність, але відсутній лінійний кореляційний зв'язок

Кореляційний зв'язок (або, простіше, кореляція) між змінними, найчастіше має лінійну форму. Тому часто, взагалі кажучи, не коректно, говорять, що змінні корелюють між собою, маючи на увазі саме лінійний зв'язок.

Коефіцієнтом кореляції Пірсона називається величина, яка виражається формулою

$$r_{XY} = \frac{E((X - EX)(Y - EY))}{\sigma_X \sigma_Y}.$$

Математичне очікування добутку відхилень X та Y, яке знаходиться у чисельнику – це так звана *коваріація* X та Y. Відхилення $X - \bar{X}$ – це випадкова величина, які отримуються зсувом усіх значень випадкової величини X на число EX вздовж осі. Математичне очікування величини X переміститься при цьому у точку нуль, і

тому відхилення є *центрованою* випадковою величиною, тобто його математичне очікування вже дорівнює нулю.

Величина коефіцієнта кореляції Пірсона вказує як на напрям лінійного зв'язку між змінними, так і на силу (тісноту) цього зв'язку. Це число завжди лежить в інтервалі [-1, 1]. Якщо коефіцієнт додатній, то зв'язок між змінними прямо пропорційний (із зростанням однієї змінної спостерігається в цілому зростання іншої, як у розглянутому вище прикладі). Якщо коефіцієнт від'ємний, то це означає, що між змінними спостерігається обернено пропорційний лінійний зв'язок (із зростанням однієї змінної спостерігається в цілому спадання іншої), у цьому випадку графіком пряма регресії є спадною. Якщо коефіцієнт кореляції за модулем близький до нуля, то зв'язок є слабким. Якщо коефіцієнт кореляції близький до одиниці або мінус одиниці, то зв'язок є сильним. Нарешті, якщо коефіцієнт кореляції дорівнює плюс або мінус одиниці, це означає, що між змінними існує функціональний лінійний зв'язок (усі точки лежать на прямій). Останній випадок на практиці не спостерігається через наявність випадкових впливів на значення змінних. Слід добре розуміти, що вибіркова оцінка коефіцієнта кореляції практично завжди відмінна від нуля. Для малих вибірок значення, наприклад 0,1 може означати, що насправді для усієї популяції об'єктів вимірювання кореляційний зв'язок відсутній, а відмінність отриманого числа від нуля є наслідком обмеженості вибірки. У цьому випадку кажуть, що отримане значення вибіркової оцінки коефіцієнта кореляції не є *значущим*. Значущість отриманої оцінки перевіряється за допомогою спеціальних методів перевірки статистичних гіпотез.

Якщо змінні виміряні у порядковій шкалі, то у якості міри лінійного кореляційного зв'язку між ними використовується *коефіцієнт кореляції Спірмена*. Цей коефіцієнт отримаємо, замінивши у формулі вибіркового коефіцієнта Пірсона значення змінних їх порядковими номерами – рангами. Альтернативною мірою для порядкових змінних є коефіцієнт «тау» Кендалла. Зауважимо, що порядок у обох змінних має бути однаковим – за зростанням вираженості вимірюваної ознаки або за спаданням.

Стандартна похибка вимірювання та довірчий інтервал. Якщо рівняння регресії використовувати для передбачення значення однієї оцінки особи за значенням її іншої оцінки, постає

питання про точність такого прогнозу. Відповісти на нього допомагає поняття стандартної похибки вимірювання. Якщо ми будемо багаторазово робити вибірку одного і того ж об'єму з популяції і для кожної вибірки застосовувати вимірювання, то, очевидно, ми щоразу отримуватимемо різні значення середнього вибіркового, тобто середнє вибіркове є випадковою величиною. Стандартне відхилення цього розподілу називається *стандартною похибкою вимірювання*. Нехай по даному значенню оцінки x потрібно спрогнозувати значення оцінки y'. Тоді стандартна похибка прогнозу обчислюється як

$$s_{YX} = s_Y\sqrt{1-\rho_{XY}^2}.$$

Якщо припустити, що похибки прогнозу нормально розподілені навколо кожного значення y' з однаковою умовною дисперсією, то це дає змогу визначити, з якою упевненістю можна стверджувати, що істинне значення оцінки y потрапляє у той чи інший окіл оцінки y', який називається *довірчим інтервалом*. Так, з упевненістю (надійністю) приблизно 0,68 (або 68%), істинна оцінка потрапляє в інтервал $y' \pm 1s_{XY}$, а з надійністю приблизно 95% - у інтервал $y' \pm 2s_{XY}$.

Кореляційний та причинний зв'язки. Наявність статистичного чи кореляційного зв'язку нічого не говорить про причинно-наслідкові зв'язки між змінними. У одних випадках зовнішня інформація дозволяє легко вказати, що є причиною, а що наслідком. Наприклад, очевидно, що між часом, який учень щодня витрачає на вивчення математики, та оцінками з математики існує пряма кореляція, причому збільшення часу є причиною зростання оцінки. Змінну-причину називають *незалежною змінною*, а змінну-наслідок – *залежною*.

У інших випадках обидві змінні корелюють внаслідок впливу якоїсь третьої змінної. Сюди можна віднести розглянутий вище приклад кореляції між оцінками вступників з математики та мови і літератури. Нехтування причинно-наслідковими зв'язками може привести до помилок в інтерпретації зв'язку між змінними. У психологічній літературі часто цитується наступний приклад. Деякий

психолог помітив, що між довжиною стопи учнів та їх математичними здібностями існує пряма лінійна залежність. З цього наш психолог зробив сенсаційний висновок: довжина стопи *впливає* на математичні здібності дітей! Насправді ж у дослідженні брали участь діти різного віку. Зрозуміло, що саме збільшення віку є причиною як зростання довжини стопи, так і математичних здібностей дитини. Можуть навіть зустрічатися ситуації, коли між двома змінними існує додатна кореляція, але внаслідок впливу третьої змінної спостережена кореляція є від'ємною!

Усунути ефект впливу третьої змінної Z можна за допомогою так званого *частинного коефіцієнта кореляції*:

$$r_{XY|Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}.$$

Наприклад, якщо $r_{XY} = 0{,}63, r_{XZ} = 0{,}9,\ r_{YZ} = 0{,}7,$ то за цією формулою отримаємо $r_{XY|Z} = 0$, тобто змінні X та Y не корельовані!

Коефіцієнт детермінації. Піднісши коефіцієнт кореляції до квадрату, отримаємо величину, яку називають *коефіцієнтом детермінації*. Ця величина вказує на частку дисперсії залежної змінної, яка обумовлена впливом незалежної змінної. Розглянемо приклад. Різні дослідження вказують на те, що між коефіцієнтом IQ та показниками успішності школярів існує кореляція в межах від 0,5 до 0,7. Отже, коефіцієнт детермінації лежить у межах від 0,25 до 0,49. Тобто можна стверджувати, що дисперсія середнього бала успішності може бути передбаченою за результатами тестування IQ не більше ніж на 25%-49%. Іншими словами, якби ми розглядали учнів лише з певним однаковим показником інтелектуального розвитку, мінливість показника їх успішності була би на 25%-49% меншою у порівнянні з усією популяцією учнів. Зауважимо, що отримані межі коефіцієнта детермінації є досить низькими і це вказує на те, що для прогнозування рівня успішності учнями замало користуватися лише значеннями їх IQ. Для прогнозування краще використовувати множинну регресію, яка розглядає залежність змінної від більше ніж однієї незалежних змінних. Тому, наприклад, для прийому абітурієнтів до ВНЗ в Україні, тобто для про-

гнозування майбутньої успішності навчання студентів, враховуються кілька оцінок зовнішнього незалежного тестування, а також середній бал шкільного атестату. Щоправда, на момент 2012 року кожна з цих оцінок враховується з однаковим ваговим коефіцієнтом, тоді як зрозуміло, що не всі вони однаково добре підходять для прогнозу. Оптимальні вагові коефіцієнти для кількох незалежних змінних можна отримати з рівняння множинної лінійної регресії, для якої розглянуте нами вище рівняння парної регресії є частинним випадком.

Тестова оцінка як сума. Зазвичай на початкових етапах тестування оцінка екзаменованого за тест є простою сумою балів, отриманих ним за кожне з тестових завдань. Також часом тест є *батареєю тестів*, тобто складається з окремих субтестів, кожен з яких призначено для вимірювання деякої окремої якості. В усіх випадках оцінка за тест складається з окремих оцінок і розробнику тесту важливо знати, як залежать статистичні властивості тестової оцінки від властивостей оцінок за окремі завдання чи субтести.

З усіх схем оцінювання окремого тестового завдання важливо виділити *дихотомічну*. Це оцінювання, яке допускає лише дві оцінки – 0 за відповідь, яка підтверджує вимірювану якість, 1 – за відповідь, яка цю якість не підтверджує. Такими як правило є завдання закритого типу з однією правильною відповіддю або завдання з пропущеним словом. Не дихотомічні схеми оцінювання допускають кілька градацій оцінки, наприклад, 0 – за повністю неправильну відповідь, 1 – за частково правильну, 2 – за повністю правильну відповідь. Такі схеми використовуються в завданнях закритої форми з кількома правильними відповідями, зокрема, завданнями на відповідність, а також есе. Для групи екзаменованих можна визначати, як характеризує вимірювану якість кожне з завдань, розглядаючи розподіл оцінок за дане завдання у групі.

Для дихотомічних завдань відношення кількості правильних відповідей до кількості всіх відповідей у групі екзаменованих є одночасно середнім арифметичним і показником *трудності* завдання (позначають як p_j для j-го завдання). Зауважимо, що термін «трудність» є технічним, його зміст прямо протилежний розмовному значенню цього слова – чим більше отримано правильних

відповідей на завдання, тим воно є легшим для екзаменованих, але тим більшою є його трудність як психометрична величина.

Для статистичних характеристик дихотомічних завдань існують формули, які полегшують обчислення, у порівнянні з традиційними.

Дисперсію розподілу відповідей на дихотомічне завдання легко обчислити, помноживши трудність завдання на частку неправильних відповідей (тобто «легкість» завдання $q_j = 1 - p_j$):

$$s_j^2 = p_j q_j.$$

Зокрема, якщо трудність завдання дорівнює 0,5, дисперсія його розподілу дорівнюватиме 0,25, а стандартне відхилення – 0,5.

Для того, щоб обчислити коефіцієнт кореляції між двома дихотомічними завданнями з номерами j і k, зручно скористатися формулою так званого φ-коефіцієнта:

$$\rho_\varphi = \frac{p_{jk} - p_j p_k}{\sqrt{p_j q_j p_k q_k}},$$

де p_{jk} – частка тих екзаменованих, які дали правильні відповіді одночасно на обидва завдання. Хоча цей коефіцієнт і має своє позначення, він є все тим же коефіцієнтом кореляції Пірсона.

Розглянемо тепер особливості оцінки за тест як суми оцінок за завдання.

Нехай тест X складається з двох завдань X_1 та X_2 і його склала група з 10 осіб. Для відшукання середнього значення по кожному завданню потрібно суму балів за це завдання поділити на 10. Для відшукання ж середнього оцінки за весь тест потрібно спочатку додати оцінки за два завдання для кожного екзаменованого, щоб знайти його оцінку за тест, а потім поділити суму усіх оцінок за тест на 10. Таким чином, *середнє значення оцінок за тест дорівнює сумі середніх за кожне завдання*:

$$\bar{X} = \overline{X_1} + \overline{X_2}.$$

За індукцією отримаємо аналогічний результат і для тесту з будь-якою кількістю завдань.

Дисперсія оцінок за тест обчислюється за формулою:

$$s_X^2 = \sum_{i=1}^{N} s_i^2 + 2\sum_{i<j} \rho_{ij} s_i s_j.$$

Згідно з даним раніше означенням, під знаком другої суми стоять коваріації різних пар завдань. Таким чином, кожна додатна кореляція між парою завдань збільшує дисперсію оцінки за тест, а кожна від'ємна – зменшує. Цей факт можна частково перевірити, розглянувши дві вибірки гіпотетичних оцінок десяти учнів за два завдання, з яких одна – це числа 1, 2, …, 10, а друга – ті ж самі числа, але у зворотному порядку: 10, 9, …, 1. Тут є лінійний обернено пропорційний функціональний зв'язок між змінними: $X_2 = 11 - X_1$, тобто коефіцієнт кореляції дорівнює -1. Дисперсії (і стандартні відхилення) вибірок очевидно однакові: $s_1^2 = s_2^2$. Тоді

$$s_X^2 = s_1^2 + s_2^2 - 2s_1 s_2 = 0.$$

Тобто мінливість результатів тестування відсутня. Це й зрозуміло, адже кожен з учнів отримає одну і ту ж оцінку за тест 11 балів. Такий тест був би абсолютно непотрібним, оскільки він не дозволяє диференціювати екзаменованих. З іншого боку, якби оцінки за друге завдання були такими ж, як і за перше, то оцінка за тест коливалася б від 2 до 20, і такий розмах є найбільш можливим для даних множин чисел.

Отримані результати дозволяють зробити ряд важливих висновків. По-перше, збільшення кількості завдань приводить до збільшення мінливості результатів тестування лише у випадку, коли між кожною парою завдань існує додатна кореляція.

По-друге, для забезпечення максимальної мінливості результатів тестування (а отже, й роздільної здатності тесту) між завданнями повинна бути не тільки достатньо висока кореляція, але й трудність завдань має бути близькою до середньої. Наприклад, для завдань з дихотомічним оцінюванням дисперсія $s^2 = pq$ є найбільшою (0,25) при $p = q = 0{,}5$. З іншого боку, слід пам'ятати, що

для диференціації дуже слабких або дуже сильних екзаменованих тест повинен містити, відповідно, деяку кількість дуже простих та складних завдань. Також слід розуміти, що якщо на два завдання кожен з учнів дає однаково правильну або неправильну відповідь, то, взагалі кажучи, одне з цих завдань є лишнім, адже воно нічого не додає до уже наявної інформації про диференціацію екзаменованих, хоча й ідеально збільшує дисперсію оцінок за тест.

По-третє, намагання розробника тесту включити до нього завдання з різних частин цільової області, щоб охопити тестом якомога більше матеріалу, може призвести до слабкої кореляції між відповідями на ці завдання, що може призвести до недостатньо високої мінливості тестових оцінок.

Але збільшення дисперсії результатів тестування хоча й бажане, все ж не є саме по собі показником якості тесту. Перш за все, тест має бути валідним та надійним, і штучне збільшення дисперсії тестових оцінок не повинне погіршувати цих характеристик. Фундаментальні поняття валідності та надійності вимірювання розглянемо пізніше.