МІНІСТЕРСТВО ВНУТРІШНІХ СПРАВ УКРАЇНИ
ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ВНУТРІШНІХ СПРАВ

Факультет № 6
Кафедра соціології та психології

ТЕКСТ ЛЕКЦІЇ

з навчальної дисципліни
«**Комп'ютерні методи практичної психології**»
обов'язкових компонент освітньої програми
першого (бакалаврського) рівня вищої освіти

053 Психологія (практична психологія)

Тема №4. Аналіз психологічних даних в SPSS, Excel, STATISTICA: описові статистики

Харків 2023

ЗАТВЕРДЖЕНО
Науково-методичною радою
Харківського національного
університету внутрішніх справ
Протокол від 30.08.2023 р. №7

СХВАЛЕНО
Вченою радою факультету №6
Протокол від 25.08.2023 р. №7

ПОГОДЖЕНО
Секцією Науково-методичної
ради ХНУВС з гуманітарних та
соціально-економічних дисциплін
Протокол від 29.08.2023 р. №7

Розглянуто на засіданні кафедри соціології та психології
Протокол від 15.08.2023 р. №8

Розробник:
Професор кафедри соціології та психології факультету №6
д-р соціол. н., професор Нечітайло Ірина Сергіївна

Рецензенти:
1. Керівник психологічної служби Харківського гуманітарного університету «Народна українська академія», доцент кафедри соціології та гуманітарних дисциплін, к. психол. н., Гога Н. П.;
2. Доцент кафедри соціології та психології факультету №6, к. психол. н., доцент Філоненко В. М.

ТЕМА №4: АНАЛІЗ ПСИХОЛОГІЧНИХ ДАНИХ В SPSS, EXCEL, STATISTICA: ОПИСОВІ СТАТИСТИКИ

План лекції

4.1. Міри середньої тенденції: теоретичні аспекти і алгоритми обчислення.
4.2. Міри варіації: теоретичні аспекти і алгоритми обчислення.

Рекомендована література

Основна

1. Нечітайло І. С., Бірюкова М. В. Математичні методи в соціології : підручник для студентів ВНЗ / Нар. укр. акад., [каф. соціології]. Харків : Вид-во НУА, 2012. 243 с.

2. Татьянчиков А. О. Математичні методи в психології: навчально-методичні рекомендації (в допомогу до самостійної роботи для здобувачів вищої освіти ступеня бакалавра факультету психології, політології та соціології) ; кафедра психології НУ «Одеська юридична академія». Одеса : Фенікс, 2021. 48 с.

Допоміжна

3. Застосування математично-статистичних методів аналізу у психологічних вимірюваннях [Електронний ресурс]. Режим доступу: http://surl.li/aghly . Дата звернення: 31.07.2023.

4. Катаєв Є.С. Використання статистичних методів обробки даних у дослідженнях “я-концепції” особистості. Вісник Національного університету оборони України. 2012. №2 (27) /2012. С. 171-176.

5. Салюк М. А Статистична обробка даних експериментального дослідження. Методичний посібник з курсу «Експериментальна психологія» / за ред. Е.Л. Носенко. Дніпропетровськ: Інновація, 2010. 26 с.

6. Татьянчиков А.О. Методичні рекомендації до виконання лабораторних робіт з курсу «Методи психологічного дослідження: математичні методи в психології». Одеса : Вид-во Університету Ушинського, 2019. 38 с.

7. Foster, G., Lane D.; Scott D., Hebl M. and other. An Introduction to Psychological Statistics. University of Missouri, St. Louis. 2018. 271 p.

ТЕКСТ ЛЕКЦІЇ

4.1. Міри середньої тенденції: теоретичні аспекти і алгоритми обчислення

*Основні числові характеристики одновимірного розподілу: максимум; мінімум; середні величини***.** Характеристики положення – це величини, які визначають положення центру емпіричного розподілу та встановлюють положення всіх одиниць розподілу уздовж шкали. Визначення цих характеристик може бути досить корисним і результативним у практичній психології. По завершенню етапу збору даних, як правило, дослідник стикається з величезними масивами даних. У всій своїй повноті ці дані можуть бути надмірно докладними (дробовими), занадто «незручними» для подальших операцій із ними, громіздкими для аналізу та порівняння з іншими аналогічними даними. Як ми вже говорили на початку цього розділу, для того щоб подолати громіздкість і надмірну дрібність отриманих даних, дослідник звертається до їх агрегування (укрупнення). Задля цього він апелює до математичної статистики, яка, відповідно до свого призначення, повинна: а) давати інформацію про взаємне розташування фактів і подій, 2) виключати ті величини, які в даний момент не відносяться до справи, 3) наочно представляти загальну картину розподілення.

Крайньою межею ефективної агрегації, очевидно, було б зведення безлічі подій до однієї єдиної величини, яка певним чином віддзеркалювала повну сукупність даних. Ясно, що ніяка окрема величина не може бути достатньо багатосторонньою, щоб відобразити кожну характеристику розподілу; вона може виділити тільки одну яку-небудь спільну властивість множини. Ця характерна величина не є точною копією всієї множини значень, вона лише приблизно описує їх розподіл.

Будь-яка величина в розподілі може описувати всю сукупність, якщо відомо її відносне положення в розподілі. Отже, необхідно проаналізувати всі події в сукупності для того, щоб дати оцінку представленості будь-якого з них. Але, на практиці, не всі величини однаково корисні в цьому сенсі. Найбільш зручними і корисними для виділення їх з таблиць є: 1) максимум; 2) мінімум; 3) центральні або типові величини, відомі як середні.

Максимум. В деяких випадках максимальна величина змінної в розподілі є єдиною представницькою величиною. При флуктуаціях розміру і ваги вантажного транспорту такою величиною є найбільший вантаж, який можна безпечно перевозити по шосе або по мосту, наприклад. Знання «середнього» показника у даному випадку зовсім недостатньо, оскільки міст, побудований з урахуванням лише середнього навантаження, обов’язково зруйнувався би при максимальному навантаженні транспорту. Подібно до цього, місткість шкіл, госпіталів та інших закладів планується з урахуванням очікуваного максимуму, а не передбачуваного середнього. Знання середньої величини в даному випадку не принесе ніякої користі. Тому максимум, як міра положення,

являє собою граничну величину, нижче якої розташовані величини, які не стосуються справи, не мають значення для спостерігача.

Мінімум. Багато практичних завдань соціальних та поведінкових наук зводяться до вибору мінімальної величини розподілу в якості діючої норми. Очевидно, що мінімум − це така величина, вище якої розташовуються всі інші величини в даному розподілі. Так для законодавчого проекту, що стосується норм добробуту населення, виходячи з розподілу доходів «нормальних» сімей виводиться мінімальний дохід (прожитковий мінімум). Виходячи з гіпотетичного розподілу зрілого населення за віком, встановлюється мінімальний вік для одруження, служби в армії, права голосування та права на обрання, та багато інших узагальнюючих показників, що певним чином слугують для здійснення соціального контролю та регуляції соціальної поведінки. Майбутній студент вузу може, наприклад, цікавитися тим, які мінімальні витрати необхідні для навчання в цьому навчальному закладі.

Ні максимум, ні мінімум не вимагають якихось серйозних обчислень а ні при визначенні їх місця положення, в ні при їх інтерпретації. Зміст цих показників є дуже простим і, отже, не потребує більш детального обговорення.

Середнє. Особливості вибору середнього. Найбільш загальною і поширеною мірою положення (і, взагалі-то, найбільш корисною) є середнє. Зазвичай − це центральна величина, навколо якої групується розподіл. Явна тенденція багатьох статистичних сукупностей концентруватися навколо центру часто називається «центральною тенденцією», а значення величини в цьому центрі − «мірою центральної тенденції» чи просто – «середнє». Однак, цей функціональний центр не обов'язково є ідентичним середині області даних спостереження. Область найбільшої концентрації значень може знаходитися як поблизу середньої точки, так і на значній відстані від неї. Розподіл оцінок у тестах на випробування розумових здібностей має дугоподібну форму з максимумом у середній точці області. З іншого боку, наприклад, розподіл доходів часто представляється U-подібними кривими, коли більшість одиниць розподілу сконцентровано в лівій та правій осі шкали, а не посередині, або J-подібними кривими, центр яких істотно зміщений в одну з осей шкали.

Як і більшість інших статистичних мір, уявлення про середнє уходить своїм корінням в звичайний здоровий глузд. Кожен політичний діяч, спираючись на результати дослідження «свого» електорату, зовсім звично говорить про «середнього виборця», «середню сім'ю», «середнього студента» і таке інше. Широке різноманіття рис, які зводяться до середнього, ілюструється відомим описом «середнього» чоловіка України, зріст якого складає сімдесят дев'ять сантиметрів, вага − 82 кілограми, який віддає перевагу брюнеткам, футболу, салу та борщу і вважає, що здатність вести домашнє господарство є найбільш важливою чеснотою дружини. Зрозуміло, що не існує жодного чоловіка, який є середнім по всіх параметрах; людина середнього зросту не обов'язково буде людиною середнього розуму або середньої краси. Тому популярне твердження, що «середньої людини» не існує, взагалі-то, є повністю виправданим. Такі фіктивні поняття, як «середня школа» і «середній українець», не відповідають ніякому строгому статистичному поняттю і

застосовується лише по відношенню до рядів вимірювання однієї змінної. Проте, який би сенс не закладався в це поняття, кожний політик, у випадках змальованих вище, на якомусь підсвідомому рівні, відчуває те, що визнається статистиками: середнє – є різновидом норми, навколо якої коливається змінна. Різниця між політичним діячем і дослідником полягає в тому, що останній вимагає більшої точності, ніж це дозволяє неофіційне народне слововживання. Тому соціолог-професіонал, звертається і до специфічної термінології, і до спеціальних математичних процедур для вимірювання середнього і тим самим обмежується змінними, у яких центральна тенденція допускає деякі види квантифікації. Різним типам центральних тенденцій відповідають різні середні, кожне з яких використовується у залежності від конкретної дослідницької проблеми. З цих численних типів середніх в рамках даної лекції ми будемо докладно обговорювати тільки три: середнє арифметичне, моду та медіану.

Середнє арифметичне (M [X]). Як і в звичайному слововживанні, на статистичній мові «середнє» означає «типове», «звичайне», «очікуване». Середнє арифметичне − таке значення ознаки, сума відхилень від якого всіх значень ознаки дорівнює нулю (з урахуванням знака відхилення). Якщо мова йде про вибіркове дослідження, в основі якого лежить випадкова вибірка, середнє арифметичне часто називається математичним очікуванням, оскільки в даному випадку мається на увазі середнє значення випадкової величини. Слід підкреслити, що умовне позначення середнього в даному випадку може відрізнятися. Як правило, якщо мається на увазі середнє за вибіркою, то замість *M[X]* використовуються $\bar{x}$, що необхідно запам'ятати, задля уникнення плутанини в формулах. Вважається, що будь-яке фізичне тіло, перебуваючи у невизначеному стані, буде прагнути до стану рівноваги (опори на свій центр ваги), так само як і середнє значення будь-якої випадкової величини, при досить великій кількості випробувань, буде прагнути до свого математичного очікування. Цей факт підтверджується теорією ймовірності. *Математичним очікуванням* випадкової величини називається сума добутків усіх можливих значень випадкової величини на ймовірності цих значень.

В цілому обчислення середнього арифметичного є необхідним для здійснення більш складних математичних операцій, пов'язаних з аналізом психологічних даних. Саме середнє арифметичне, як математичне очікування, є однією з основних характеристик, ознак вибірки, тому що за його допомогою можна прогнозувати значення деякої випадкової ознаки при досить довгому періоді випробувань. Наприклад, людина, пригнічена тривалою «смугою невдач», зазвичай, сподівається, що обставини повинні прийти в норму. Вона вважає, що існує щось на кшталт закону природи, згідно з яким смуга невдач, в решті решт, збалансується вдачами.

Процедури обчислення середнього арифметичного для незгрупованих і згрупованих даних дещо відрізняються.

Обчислення середнього арифметичного для незгрупованих даних. Для незгрупованих даних середнє обчислюється за простою формулою, шляхом підсумовування окремих величин і подальшого поділу на загальне число подій, що має наступний вигляд:

$$M[X] = \frac{\sum x_i}{N}$$ (формула обчислення простого середнього),

де

- M [X] − середнє арифметичне;
- $\sum x_i$ − сума змінних;
- x_i − значення змінної (її величина),
- N − число подій.

Обчислення середнього арифметичного для згрупованих даних. У першому підрозділі цього розділу ми говорили про те, що згруповані дані відрізняються від незгрупованих тим, що кожній групі подібних величин приписується частота або «вага». Тому для обчислення середнього для згрупованих подій кожне значення змінної (*xi*) множиться на свою частоту (*ni*), Отримані у результаті числа підсумовуються, і ця сума ділиться на суму всіх частот, що має такий вигляд:

$$M[X] = \frac{\sum x_i \times n_i}{N}$$ (формула обчислення середнього зваженого),

де

- x_i − значення змінної (її величина),
- *n i* − частота значення змінної,
- N − сума всіх частот, об’єм досліджуваної сукупності.

Якщо застосувати дві різні формули (простого та зваженого середнього) до одного й того ж розподілу, значення середніх, отримані у результаті, відрізнятися не будуть.

Обчислення середнього арифметичного для інтервальних рядів. Процедура обчислення середнього арифметичного для інтервальних рядів даних дещо відрізняється від процедур, описаних вище. Необхідно зрозуміти, що для неперервних інтервальних рядів частота інтервалу співпадає з його середньою точкою. Отже, перед тим, як обчислювати середнє арифметичне для інтервального ряду, необхідно знайти середні точки кожного інтервалу. З цією метою спочатку знаходяться межі інтервалів, а потім обчислюються їхні середні точки за формулою:

$$x_{ci} = \frac{(x_i + x_{i+1})}{2}$$

де

- *xi* − нижня межа інтервалу,
- *xi+1* − верхня межа інтервалу.

Після цього кожна середня точка зважується (тобто перемножується на частоту інтервалу) і обчислюється середнє арифметичне для всього інтервального ряду за формулою:

$$M[X] = \frac{\sum x_{ci} \times n_i}{N},$$

де

- x_{ci} – середня точка інтервалу,
- $n\,i$ – частота інтервалу,
- N – сума всіх частот інтервального ряду.

Обчислення середнього ряду середніх (комбіноване середнє). Дві або більше середні величини самі часто усереднюються, тобто можна отримувати середню ряду середніх. Середні підгруп, до того, як вони будуть скомбіновані, мають бути зважені у відповідності зі своїми N. Недооцінка необхідності зважування середніх може привести до абсурдних результатів. Можна навести приклад з гри в баскетбол. Гравець за 75 ударів набрав 25 очок, що дало в середньому 0,33. В цей же день він за п'ять ударів набрав п'ять очок, що дало в середньому – 1,00. Яке буде середнє (комбіноване)?

Дещо наївним було б припущення про те, що якщо ми складемо дві середні величини і розділимо потім на 2 (за принципом знаходження середнього арифметичного простого), то отримаємо очікуваний правильний результат. В такому разі, стосовно розглянутого прикладу, комбіноване середнє буде рівним 0,667 − явно неправильний результат. У таблиці 4.1 проілюстрована процедура правильного обчислення комбінованого середнього для наведеного вище прикладу.

Таблиця 4.1.

Розрахункова таблиця обчислення комбінованого середнього

Кількість ударів (n_i)	*Загальна кількість балів* ($\sum$xi)	*Середня кількість балів за кожний удар* ($\bar{x}i$)	$\overline{xi} \times n_i$	*Кінцевий результат− M [X]*
75	25	25/75=0,33	75×0,33=24,75	
5	5	5/5=1,00	5×1,00=5,00	
Всього: N =80			24,75+5,00=29,75	*M [X]* = =29,75:80= =0,37

Виходить, що в середньому за кожен удар в цей день баскетболіст отримав приблизно по 0, 37 балів (число, округлене до сотих).

Наведемо приклад знаходження середнього арифметичного для ряду середніх. Припустимо, нам відомі середні показники чисельності дітей, які виховуються в дошкільних установах кожного з районів міста. Але перед нами стоїть завдання вийти один середній показник по всьому місту. Порядок обчислень, в напрямку вирішення цієї задачі проілюстрований в таблиці 4.2.

Таблиця 4.2.

Мережа дошкільних закладів усіх відомств м. Харкова в 1999 р.[1]

№ п/п	*Назва району*	*Кількість д/у (n_i)*	*В них дітей в середньому ($\bar{x}i$)*	$\overline{Xi} \times n_i$
1	Дзержинський	33	161,06	33×161,06=5315
2	Жовтневий	13	140,62	1828
3	Київський	33	141,88	4682
4	Комінтернівський	17	182,82	3108
5	Ленінський	23	93,09	2141
6	Московський	35	208,46	7296
7	Орджонікідзевський	24	151,88	3645
8	Червонозаводський	18	126,00	2268
9	Фрунзенський	16	188,56	3017
10	Всього (*N*)	212		33300

Для визначення середньої кількості дітей (M [X]) в дошкільних закладах м. Харкова в 1999 р. необхідно насамперед дізнатися якою є загальна кількість дітей ($N = \sum \bar{x} \times n = 33300$), а потім отриманий результат поділити на кількість дошкільних установ – 212. Продовжуючи цей процес, знайдемо, що *M [X]* = 157,07, тобто в середньому на один дошкільний заклад м. Харкова в 1999 р. припадає 157,07 дитини. Дробовий характер дискретної величини (в реальному житті не буває 1,5 людини), звичайно, нікого не бентежить, якщо мова йде про середній показник.

Підводячи підсумки під всім сказаним щодо середнього арифметичного, виділимо його *основні властивості:*

1) сума відхилень різних значень ознаки від середнього арифметичного дорівнює нулю;

2) якщо від кожної варіанти відняти або до кожної варіанти додати будь-яке постійне число, то середнє збільшиться або зменшиться на те ж саме число;

3) якщо кожну варіанту помножити (розділити) на будь-яке постійне число, то середнє збільшиться (зменшиться) в стільки ж разів;

4) якщо ваги, або частоти, розділити чи помножити на якесь постійне число, то величина середньої не зміниться.

Необхідно усвідомити і запам'ятати, що *середнє арифметичне обчислюється і має сенс тільки для метричних та порядкових шкал.* Воно представляє величину кожної події в розподілі. У зв'язку з цим воно піддається впливу як дуже великих, так і дуже малих величин, що особливо помітно в несиметричних розподілах. Для таких розподілів більш інформативними можуть бути інші міри усереднення, такі як мода і медіана.

Мода чи ймовірнісне середнє. Мода являє собою найбільш часто

[1] Статистичний збірник. Показники роботи закладів освіти та наукових установ області за 2000 рік ; [За заг. Ред. О. Л. Сидоренко, Л. О. Бєлової, А. С. Доценка]. – Х., 2001 – 87 с.

повторювану величину в упорядкованому розподілі; вона характеризує те місце розподілу, де концентрація подій максимальна. Етимологічно вона пов'язана з уявленнями про найбільш підтримувану манеру одягатися чи з етикетом, до якого пристосовується більшість тієї чи іншої соціальної групи. Отже, моду (M_o) можна також визначити як найбільш ймовірну величину, і тому її називають імовірнісним середнім.

Дослідження розмовних виразів наводить на думку, що під модою в дійсності часто мають на увазі поняття середнього показнику. Частково це пояснюється тим, що ознаки, так само як і змінні, можуть мати переважні частоти в серії спостережень. Коли політичний діяч говорить про «середнього виборця», інтереси якого він відстоюватиме, то виходить з того, що більшість виборців керується своїми власними інтересами; або коли офіціантка зауважує, що «середній» клієнт не п'є чорну каву, вона, ймовірно, має в увазі, що більшість відвідувачів даного ресторану не п'ють чорної кави. Відвідувачі ресторану і виборці, в свою чергу, кажуть про «середню офіціантку» чи «середнього політичного діяча».

Мовою статистики *мода − це величина, з якою найбільш імовірно можна зустрітися в серії зареєстрованих спостережень.* Таким чином, мода є найбільш імовірною величиною, хоча знання однієї лише тільки моди не дозволяє визначити ступеня цієї ймовірності. Зрозуміло, що якби частоти всіх величин були однаковими, то не було б ніякого сенсу вводити це поняття.

Обчислення моди. У переважній більшості випадків, для визначення моди достатньо просто підрахувати частоту появи кожної величини, тому що мода є такою величиною, яка найбільш часто спостерігається. У випадку неперервних даних мається на увазі, що емпіричні вимірювання настільки докладні і короткотривалі, що ніякі два вимірювання не можуть дати тотожних результатів. Очевидно, що мода не може виявитися без угрупування. В разі наявності інтервалів, вони повинні бути однаковими по ширині; якщо цього правила не дотримуватися, можна для досить великих інтервалів отримати моду практично будь-якої бажаної величини – результат явно безглуздий.

Отже, необхідно пройти дві різні стадії у визначенні моди: 1) визначення модального інтервалу і місця в ньому переважаючої або «модальної» частоти; 2) знаходження величини, відповідної цій частоті. У якості прикладу розглянемо наступну таблицю:

Таблиця 4.3

Кількість вищих навчальних закладів України III – IV рівня акредитації, які з'явилися в певний навчальний рік

(відомо, що в 1992 році їх було 1580)

Роки $(x_i; x_{i+1})$	Кількість навч. закл. III–IV рівня акредитації (n_i)	Накопичені частоти **(!!!тільки для обчислення медіани!!!)** (n_i^{H})
1993-1994	1	1
1994-1995 *(Модальний інтервал)*	**73**	74
1995-1996 *(Медіанний інтервал)*	**23**	**97**
1996-1997	19	116
1997-1998	6	122
1998-1999	18	140
1999-2000	15	155
Всього	155	

У таблиці 4.3 найбільша частота дорівнює 73, отже, інтервал 1994/1995 рр. є модальним, а модальна величина або наближена мода дорівнює 1994,5, що є середньою точкою модального інтервалу.

Однак, дещо «свавільне» визначення ширини інтервалу вносить деяку невизначеність і в саму процедуру обчислення моди. Існує два можливі шляхи уникнення цієї невизначеності: 1) відмовитися від даного типу середньої величини і застосувати іншу міру середнього, 2) застосувати замість наближеного більш точний метод обчислення моди, який зменшив би невизначеність. Слід зазначити, що навіть за умови невизначеності в обчисленні моди, досить часто трапляються такі випадки, коли не можна відмовитися від моди на користь інших мір усереднення. Отже, краще звернути увагу на удосконалення техніки обчислення моди.

У наближеному методі, розглянутому вище, знаходиться середня точка інтервалу з найбільшою частотою, при цьому ігноруються суміжні інтервали і їх частоти. Однак ці суміжні інтервали вплинули б на величину моди, якби їх межі були інакше розташовані. Отже, необхідно розробити більш «чутливий» метод, що дозволить врахувати вплив суміжних інтервалів.

Метод різниць в обчисленні моди. Цей метод є більш досконалим, ніж вищеописаний метод обчислення моди, і зводиться він до наступного: 1) обчислюються різниці між модальною частотою та частотами суміжних інтервалів, 2) обчислюється відношення однієї з цих різниць (зазвичай з частотою попереднього інтервалу) до суми двох інших різниць, 3) це відношення множиться на ширину модального інтервалу, а потім отриманий

результат додається до істинної нижньої межі модального інтервалу. У підсумку отримуємо уточнене значення моди. Формула обчислення в даному випадку буде мати наступний вигляд:

$$Mo = x_0 + h\frac{n^{mo} - n^{-}}{2n_{mo} - n^{+} - n^{-}},$$

де

- x_o – нижня межа модального інтервалу,
- h – ширина модального інтервалу,
- n_{mo} – частота модального інтервалу,
- n^{-} – частота інтервалу, що передує модальному,
- n^{+} – частота наступного інтервалу.

Застосувавши формулу до таблиці 2.3.3, отримаємо наступні результати:

x_o =1994 h =1 n_{mo} =73 n^{-} =1 n^{+} =23

$$Mo = 1994 + 1 \times \frac{73 - 1}{2 \cdot 73 - 1 - 23} = 1994{,}59.$$

Бімодальність. Деякі розподіли виявляють два максимуми і тому називаються бімодальними, на відміну від унімодальних розподілів, які мають лише одне модальне значення. Бімодальний розподіл може бути наслідком накладення двох або більше популяцій з різними частотними максимумами. Так, наприклад, в полігоні розподілу зросту дорослого населення, завдяки об'єднанню груп чоловіків та жінок, які характеризуються двома різними розподілами зросту, може виникнути бімодальність. Зіткнувшись з бімодальністю, дослідник повинен спробувати або розділити розподіли, які її викликають, або (у випадку невдачі) йому нічого іншого не залишиться, окрім того, як прийняти бімодальність як характеристику, притаманну даному розподілу.

Медіана. У будь-якій впорядкованій сукупності кожна подія займає певне місце – перше, друге, десяте чи сімдесят п'яте (або ранг). Очевидно, що конкретна чисельна величина рангу набуває сенсу і значення в залежності від того, якою є загальна кількість рангів. Ранг, рівний 10, в ряду зі 100 рангів, є більш високим, ніж ранг 10 в групі з 20.

Точка, яка розсікає впорядковану (ранжирувану) сукупність на дві рівні частини так, що одна половина подій точно знаходиться нижче, а інша половина вище цієї точки, називається медіаною. Наприклад, в 1950 р. медіанний вік усього населення Сполучених Штатів дорівнював 30,4 року. Це означає, що одна половина населення була старше, а інша половина населення молодше цього віку. Оскільки медіана точно позначає положення величини в певній послідовності всіх величин сукупності, вона часто називається «середнім положенням».

Обчислення медіани. За аналогією із середнім арифметичним, існують різні формули обчислення медіани для не згрупованих та згрупованих даних. Якщо дані *не згруповані,* як правило, дотримуються однієї з наступних формул:

$\frac{N}{2}$ чи $\frac{N+1}{2}$, *де* N − загальне число рангів. При цьому, перша формула частіше використовується у випадку непарної кількості рангів, а друга – у випадку парної кількості.

Якщо ж *дані згруповані* − обчислення здійснюються за наступною формулою:

$$Me = x_0 + h\frac{\frac{1}{2}N - n_H}{n_{me}},$$

де

- x_o − нижня межа медіанного інтервалу,
- h – ширина медіанного інтервалу,
- N – обсяг вибірки (відповідно, $\frac{1}{2}N$ - «половинна» подія),
- n_H − частота, накопичена до медіанного інтервалу.
- n_{me} − частота медіанного інтервалу.

Якщо коротко описати алгоритм обчислення медіани для даних, згрупованих в інтервали, то слід почати з (1) ранжирування всіх подій або знаходження накопичених частот для кожного інтервалу, потім (2) знайти, так звану, «половинну подію» (адже медіана − точка, яка ділить упорядковану сукупність навпіл), (3) по ряду накопичених частот знайти інтервал, в який потрапляє ця подія і (4) здійснивши елементарні математичні обчислення, знайти саму медіану.

Застосуємо даний алгоритм обчислення медіани до таблиці 2.4.3, представленої вище: (1) зробимо ранжирування всіх подій шляхом находження накопичених частот для кожного інтервалу, (2) поділимо суму всіх частот навпіл: $N/2$=155/2=77,5 («половинна» подія), (3) знайдемо медіанний інтервал, тобто місце положення 77,5-ої події в ряду накопичених частот *(очевидно, що «половинна» подія не потрапляє в перші два інтервали накопичені частоти яких менше за 77,5; частота третього інтервалу, рівна 97, що істотно перевершує зазначену межу),* отримуємо, що 77,5-а подія знаходиться десь всередині інтервалу 1995/1996 рр. з частотою в 23, яка, за припущенням, рівномірно розподілена по всьому інтервалу. Тепер можна визначити всі невідомі компоненти формули обчислення медіани: x_o=1995, h =1, $\frac{1}{2}N$ = 77,5, n_H=74, n_{me}=23

Отримані значення підставимо у відповідну формулу:

$$Me = 1995 + 1 \times \frac{77{,}5 - 24}{23} = 1995{,}15$$

Інтерпретуючи цю цифру, слід проговорити, що рівно половина всіх навчальних закладів III – IV рівня акредитації з'явилася в Україні до1995,15-го року, а рівно половина – пізніше за цю дату.

Медіана дискретних даних. Деякі автори обмежують застосування медіани тільки неперервними даними з тієї причини, що дискретні дані, за визначенням, не можуть бути дробовими, як це потрібно для медіани. Але таке обмеження виявляється не таким вже і необхідним. У розподілі розмірів сімей відсутнє таке число, що відображає кількість членів сім'ї , щоб точно 50% сімей мали кількість членів більшу за це число, а 50% − меншу. Чи можна в такій ситуації взагалі відмовитися від медіани чи слід прагматично трактувати дані як неперервні і прийняти дробову величину за медіану? Як і раніше, приймемо останню альтернативу. Адже сутність медіани вкрай проста: 50% подій мають більшу величину, а 50% – меншу. Вона має також ще один показовий критерій: сумарна відстань між медіаною і кожною з величин розподілу завжди менше, ніж подібна величина, обчислена для будь-якої іншої точки. Саме тому медіана «ближче» до всіх подій даного розподілу, ніж будь-яка інша міра середнього. В цьому і полягає сенс того, що медіана займає центральне положення в розподілі.

Інші міри усереднення. Замість того щоб обчислювати медіану, можна було б для більшої точності локалізувати дані в меншому інтервалі – верхній чверті, десятій чи навіть сотій. Для таких додаткових уточнень потрібні менші підрозділи. Обчислення квартилів, децилів і центилів, які поділяють безліч подій на четверті, десяті й соті частини, виконуються абсолютно аналогічно обчисленню медіани, за винятком того, що в формулу для медіани, надану вище, підставляють замість n_H іншу величину, яка відповідає частоті або відсотками подій, що лежать нижче розглянутої точки. Наприклад, для знаходження точки, нижче якої припадає найменша чверть випадків, замінюють $N/2$ на $N/4$.

Таким чином, перший квартиль в розподілі дорівнюватиме:

$$Q_1 = x_o + \delta \frac{N/4 - n_{H1}}{n_{Q1}}.$$

Якщо потрібно буде виділити точку, нижче якої знаходяться 75% подій (чи Q_3), то обчислення проводяться за наступною формулою:

$$Q_3 = x_o + \delta \frac{3N/4 - n_{H3}}{n_{Q3}}.$$

90-й центіль (чи C_{90}) може бути знайдений за формулою:

$$C_{90} = x_o + \delta \frac{90N/100 - n_{H90}}{n_{C90}}.$$

Квантилі як нормуючий критерій. Медіана, квартилі, децілі та центилі, які, згідно зі своїм визначенням, вказують на частку подій, розташованих нижче або вище цієї величини, носять узагальнену назву квантилі. Ці міри усереднення можна застосовувати для фіксації відносного положення будь-якої величини в ряду інших величин. Можна локалізувати за допомогою 90-го центилю вагу в 80 кілограмів (C_{90}=180). Це буде означати, що 10% населення володіють вагою вище, а 90% – менше цієї ваги. Точно так же на абстрактній шкалі центилів можна локалізувати і такі події, як вік в 62 роки, зріст в 180 см, рівень інтелекту в 120 балів, тощо.

Є очевидним, що квантилі можна розглядати, як стандартизовані міри розташування незалежно від метричної системи або типу даних. Таким чином, людина, яка перебуває на 90-му центилі в розумовому розвитку, може знаходитися приблизно на 90-му центилі і за рівнем доходу, що вказує «підозрілу подібність» між цими двома соціальними явищами. Такий підхід дозволяє з успіхом зіставляти та порівнювати «незрівнянні» величини. Дані характеристики положення не залежать також від виду розподілу, тобто від того, чи є воно нормальним, скошеним та будь-яким іншим. Ця обставина ще більше підвищує цінність квантилів, дозволяючи за допомогою вищеописаних процедур вистроїти за рангом випадкові величини у впорядкованій сукупності.

Підводячи підсумок всьому сказаному у цьому параграфі, представимо стислий виклад характеристик середніх:

Середнє арифметичне

1. Це така величина в даній сукупності, яка спостерігалася б у тому випадку, якщо б всі величини були рівні.

2. Суми відхилень від середнього в будь-яку сторону є рівними; отже, алгебраїчна сума цих відхилень дорівнює нулю.

3. Середнє арифметичне репрезентує значення кожної величини розподілу.

4. Сукупність величин має одне і тільки одне середнє.

5. Над середнім можна робити алгебраїчні дії, середні підгруп можна комбінувати при належному зважуванні.

6. Середнє можна обчислити навіть в тому випадку, коли значення окремих величин невідомі, за умови, що відома сума всіх величин (N).

7. Задля обчислення середнього немає ніякої необхідності групувати і впорядковувати величини.

8. Середнє можна обчислити тільки для закритих інтервалів.

9. Середнє є фіксованим в тому сенсі, що процедури групування не справляють скільки-небудь серйозного впливу на його значення.

10. Середнє застосовується лише до кількісних денних.

Мода

1. Це величина, яка найбільш часто спостерігається в розподілі; точка найбільшої щільності.

2. Значення моди визначається переважаючою частотою, а не значеннями змінної в розподілі.

3. Це найбільш імовірна величина і, отже, найбільш типова.

4. Розподіл може не тільки одну, але й дві або більше моди. З іншого боку, в прямокутному розподілі не існує ніякої моди.

5. Мода не відображає ступень модальності.

6. Над модою не можна виробляти алгебраїчні маніпуляції; моди підгруп не можна комбінувати.

7. Вона є невизначеною в тому сенсі, що залежить від процедури угруповання.

8. Мода обчислюється як для відкритих, так і для закритих розподілів.

9. Мода – єдиний тип середнього, який може представляти якісні змінні.

Медіана

1. Це величина, розташована точно в середній точці упорядкованого розподілу (а не в області зміни змінної); половина подій має значення більші за медіальне значення, а половина – менші.

2. Значення медіани знаходиться, виходячи із її розташування в сукупності даних і не залежить від значення окремих величин.

3. Сума відстаней від медіани до інших величин сукупності менше, ніж подібна сума, обчислена для будь-якої іншої точки розподілу.

4. Кожна сукупність може мати одну і тільки одну медіану.

5. З медіаною можна робити алгебраїчні дії; медіани підгруп не можуть зважуватися і комбінуватися.

6. Медіана є фіксованою в тому сенсі, що процедура групування не справляє на неї помітного впливу.

7. Для обчислення медіани всі величини повинні бути впорядковані і згруповані.

8. Медіана обчислюється як для відкритих, так і для закритих інтервалів.

9. Якісні дані не дозволяють розрахувати медіану.

Підкреслимо, що визначення всіх характеристик положення є доцільним далеко не у всіх випадках. Існує ряд критеріїв, які допомагають вирішити питання про застосування того чи іншого типу середнього. Про ці критерії і буде йти мова в наступному підрозділі.

Порівнянність середніх. У зв'язку з тим, що нам часто доводиться вибирати між різними видами усереднення, необхідно викласти основні принципи вибору центрального значення. Вже було з'ясовано, що не існує універсальних типів середнього, які можна застосовувати абсолютно в усіх випадках. Необхідно завжди пам'ятати, що середнє − це єдиний «представник» розподілу, величина, яка дуже є дуже зручною унаслідок її компактності, і, в той же час, вона є незручною через короткостроковість. В кращому випадку, середня величина надає саме стільки інформації, скільки витягує з розподілу.

Відомі три критерії, які допомагають вирішити питання про застосовність типу середнього: (1) мета усереднення, (2) вид розподілу даних, (3) обмеження, пов'язані з «технічними» причинами і типом шкал.

Цілі усереднення. Будь-яке дослідження у галузі соціальних та поведінкових наук по суті є спробою дати відповідь на питання про природу

явища, і статистична процедура виступає лише в ролі інструмента. Дослідника при обчисленні того чи іншого типу середнього цікавлять наступні питання: «Якими є розміри сім'ї?», «Якою є тривалість життя?», «Яким є вік населення?». Якщо знання розміру сімей є необхідним з метою планування житлового будівництва, то наближена мода була б більш корисною величиною, ніж середнє арифметичне або медіана, навіть з урахуванням тієї обставини, що невідома точна ступінь модальності. Будинки будуються не для абстрактних арифметичних середніх сімей, а для реально існуючих. Якщо ж необхідно вивчити «плодючість» сімей, то середнє арифметичне було б більш корисною мірою, тому що воно представляє як великі, так і малі сім'ї.

Вид розподілу. Розподіли можуть мати самий різний вид, від ідеально симетричного до вкрай асиметричного. *Симетрія* означає, що величини розподілені ідентично по обидві сторони від середнього. Ступінь *асиметричності* впливає на типовість і репрезентативність середніх величин, і, отже, її необхідно брати до уваги при виборі типу середнього. Іноді стверджують, що, якщо крива симетрична, то взагалі не виникає ніяких проблем у виборі способу усереднення, так як середнє, медіана і мода більш-менш збігаються. Однак це справедливо лише в арифметичному розумінні, але не в теоретичному. Навіть якщо чисельні величини середніх тотожні кожному типу − середньому арифметичному, моді, медіані − при цьому, відповідають абсолютно різні «образи». Наприклад, «середній» студент вважає свою оцінку не сумою всіх оцінок, поділену на N, а скоріше, просто типовою оцінкою. Ці приклади приводять до висновку, що обирається такий тип середнього, який, перш за все, задовольняє цілям дослідження.

У міру того, як розподіл стає все більш і більш асиметричним, величини різних типів середнього починають помітно відрізнятися, і проблема вибору способу усереднення набуває серйозного значення. Перш за все, середнє арифметичне для несиметричного розподілу перестає бути типовим. В U-подібному розподілі середнього взагалі практично може і не бути, а тому його величина може бути абсолютно фіктивною. Багато які дані: заробітна платня, розміри міст, розміри сімей, тощо − досить часто мають асиметричну форму, що вимагає особливої уваги при виборі типу середнього.

Обмеження, пов'язані з «технічними» причинами і типом вимірювальних шкал. Існують певні чисто технічні особливості обчислень, які можуть змусити використовувати той або інший тип середнього. Так, наприклад, середнє арифметичне не можна обчислити для розподілів з відкритими інтервалами. В цьому випадку користуються медіаною, якщо розподіл є не дуже асиметричним, тобто коли значення середнього арифметичного та медіани не набагато відрізняються один від одного. З іншого боку, коли відомо тільки суму частот і величин, можна обчислити лише середнє арифметичне, хоча інші типи середнього були б більш цікавими.

Необхідно завжди брати до уваги технічні можливості обчислювальних методів. Оскільки не існує ніякого методу комбінування або зважування медіан та мод, то вони, зазвичай, є завершальним етапом обчислень. Тому, коли передбачаються додаткові обчислення, необхідно вибирати середнє

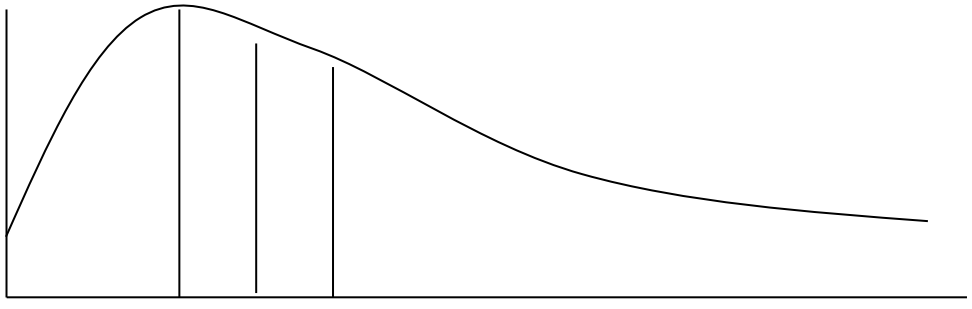
арифметичне чи його похідні.

Крім того, вибір тієї чи іншої міри усереднення може обмежуватися типом шкали, за допомогою якої вимірювався ознака. Згадаймо, що найбільші обмеження в цьому плані накладає номінальна шкала. Розподіл, отриманий за номінальною шкалою ми можемо охарактеризувати лише за допомогою моди.

Мінімум, максимум і проміжні міри. У багатьох випадках тільки середні величини розглядаються в якості мір положення. Однак необхідно підкреслити, що дана точка зору занадто ортодоксальна. Міра положення не обов'язково повинна співпадати з мірою типовості. Середні величини, як правило, свідчать про типовість, але так саме, як і максимум, і мінімум чи будь-яка проміжна величина можуть служити мірою положення.

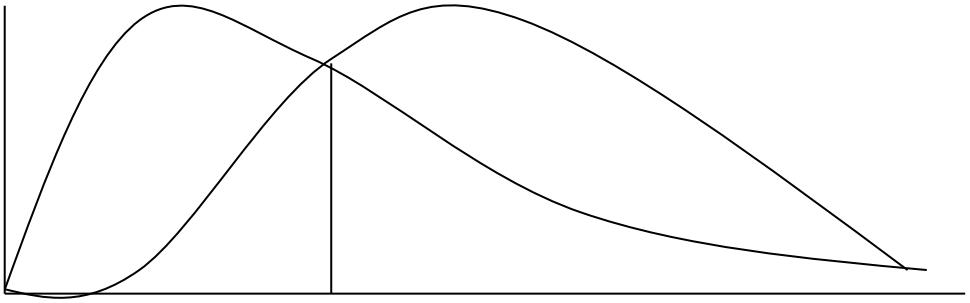
Характеристики середніх. Передумовою для належного застосування кожного типу середнього є знання відповідних описових характеристик. Останні часто аналізуються з точки зору їх «переваг і недоліків». Але такий підхід є скоріше оціночним, ніж описовим, і тому не надає строгого викладу суті кожної процедури. Переваги при вирішенні однієї проблеми можуть виявитися недоліками при вирішенні іншої. Тому можна формально викласти описові характеристики, властиві кожному типу середнього, незалежно від ситуації, в якій вони можуть застосовуватися.

Порівняння типів середнього. Подібно будь-якому статистичному показнику середні, насамперед, застосовуються з метою порівняння. Найчастіше середні величини використовуються для порівняння результатів вимірювання відповідних ознак окремих груп і розподілів, яке було здійснено за допомогою одних і тих самих одній шкал. Але ж, звичайно, бувають такі випадки, коли різні типи середніх величин принципово не можна порівнювати. Наприклад, середнє. Будучи схильним до впливу кожної величини, середнє арифметичне асиметричного розподілу буде зміщуватися в напрямку екстремальних величин. На моду краї розподілу не справляють аніякого впливу, в той час, як медіана зміщується переважно у напрямку ближче до «хвоста» асиметричного розподілу. Однак це зміщення не є дуже великим, оскільки концентрація подій у «хвості», як правило, невелика. Якщо одномодальний розподіл має правий ухил, порядок розташування різних типів середніх на базової лінії буде таким: мода, медіана, середнє арифметичне, а інтервали між ними будуть змінюватися в залежності від ступеня асиметричності (див. *Мал. 4.1*). Якщо розподіл має лівий ухил, порядок середніх буде зворотним.



Малюнок 4.1.Середнє, медіана і мода: правий ухил

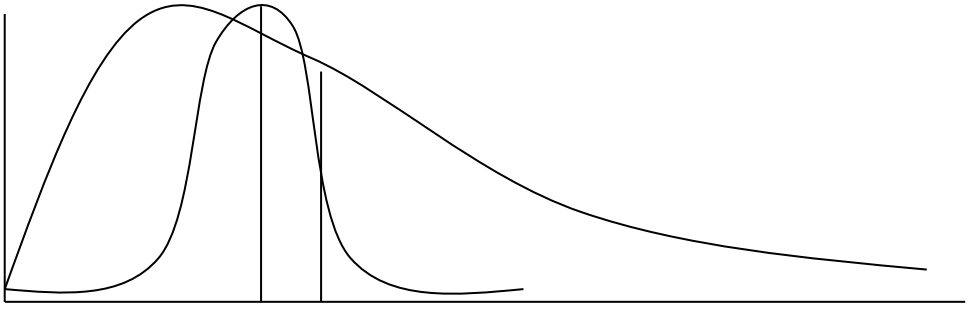
Але навіть однотипні середні можна порівнювати тільки за умови подібності розподілів, в тому випадку, коли «інші умови рівні». Наприклад, порівняння середнього зросту чоловіків і середнього зросту жінок є цілком обґрунтованим. Ступінь відмінності середнього зросту чоловіків і жінок відповідним чином вимірюється різницею їх середніх арифметичних. Однак коли розподіли помітно відрізняються один від одного, такі порівняння фактично можуть ввести в оману, особливо в разі порівняння середніх арифметичних. Наприклад, два ряди величин можуть мати рівне середнє арифметичні і абсолютно різні типи розподілу, що зображене на малюнку 4.2.



Малюнок 4.2. Протилежний ухил кривих розподілу з однаковими середніми

Ці розподіли займають однакову область. Проте їх максимуми не збігаються. Було б доцільнішим порівнювати дані цих розподілів за їх модами.

На малюнку 4.3. максимуми є розташованими приблизно в одній і тій самій області, проте середні арифметичні не рівні через несиметричність (значний ухил) одного з розподілів.



Малюнок 4.3. Вплив несиметричності розподілу на середнє

Роблячи загальний висновок, слід підкреслити, що основні характеристики положення, про які йшла мова в цьому пункті, не надають повного і всебічного опису даних. Їх слід розглядати не як автономні величини, а лише як певні способи репрезентації конкретних даних. Використання середніх ставить важливу проблему реконструкції характеру розподілу, виходячи з кількох «витягнутих» з нього величин. При вирішенні цієї проблеми важливо знати інші характеристики − характеристики розсіювання.

4.2. Міри варіації: теоретичні аспекти і алгоритми обчислення

*Міри варіації (протяжності) ознаки***.** Як уже говорилося вище, характеристики положення, хоча і є надзвичайно важливими при вивченні тих чи інших ознак, що варіюються, але все ж таки, вони не надають вичерпної інформації про цю ознаку. Неважко уявити собі два емпіричні розподіли, у яких середні однакові, але при цьому в одного з них значення ознаки розсіяні у вузькому діапазоні навколо середнього, а в іншого − в широкому. Тому поряд з характеристиками положення, нерідко визначаються і характеристики розсіювання досліджуваної сукупності, що показують, наскільки близько / далеко всі значення ознаки віддалені від середніх показників. Характеристики розсіювання виражаються в мірах варіації або мірах протяжності, найбільш уживаними з яких є дисперсія, відхилення (середнє лінійне і середнє квадратичне), коефіцієнти варіації (для лінійного і квадратичного відхилення).

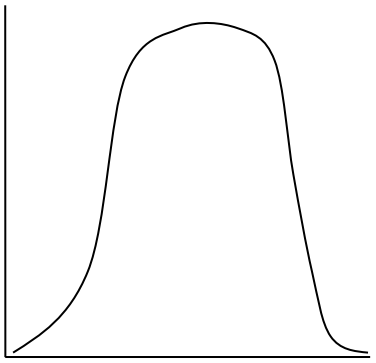
Взагалі, поняття варіації лежить в основі всіх статистичних розрахунків. Багато підручників не встановлюють відмінності між термінами «варіювання» і «варіація». Однак, загострімо увагу на цих відмінностях. Термін *«варіювання»* відображає здатність змінюватися. *Варіація* – прояв такої здатності, яку можна описати і виміряти.

Якби всі величини досліджуваної сукупності були ідентичними, то обчислення середнього значення або будь-яких інший статистичних величини стало б зайвим. Адже основна мета усереднення – отримання однієї величини, яка б репрезентувала (представляла) цілу групу неоднакових величин. Середні величини і були винайдені для виключення, так би мовити, «відволікаючих» відмінностей між величинами. Однак при певних обставинах саме характер відхилень, а не результат усереднення представляє більший інтерес для дослідника. Наприклад, у двох професійних групах, що мають приблизно однаковий середньорічний заробіток, наприклад, професорів і бізнесменів, можуть бути представлені найрізноманітніші заробітки. Серед професорів університетів оклади порівняно мало відхиляються від середнього, тоді як в бізнесі дохід є менш стійким. Оцінюючи загальні здібності і успішність студента, викладач повинен брати до уваги не тільки його середній бал, але і тенденцію цих балів. Студент з балами «5», «4» і «3», загалом, буде оцінений інакше, ніж студент з балами «4», «4», «4», хоча середній бал у них однаковий і дорівнює «4». Тренеру по баскетболу, який відбирає для університетської першості одного з двох гравців, що мають рівний середній бал, більше підходить стабільний гравець, який рідко «відхиляється від середнього», ніж нестійкий спортсмен, який в середньому показує низький рівень гри і лише іноді ефектно проявляє себе.

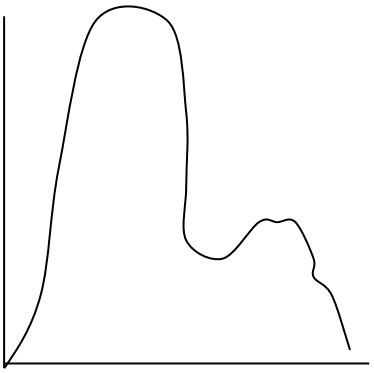
У загальному випадку знання *картини варіації – того, що статистики називають розкид, розсіюванням або дисперсією,* виявляється не менш важливим для дослідника, ніж знання середніх величин. Для вивчення цієї картини в арсеналі сучасної статистики є досить багато випробуваних прийомів і методів. Хоча вони різняться в деталях, їх можна розділити на три широкі категорії: 1) вимірювання області, яка містить всю або основну частину

розподілу; 2) вимірювання відхилень змінної від центрального значення; 3) вимірювання ступеня однорідності якісних змінних.

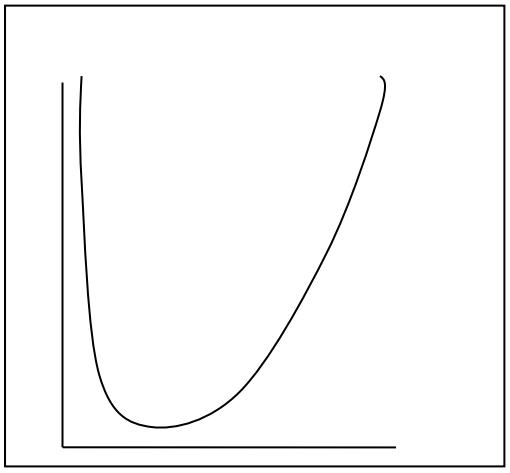
Деяке загальне уявлення про дисперсії кількісних даних можна отримати з графіків, представлених на малюнку 4.4. Графік частот відразу ж дозволяє прийти до висновку, чи є дисперсія симетричною; чи збільшуються частоти при наближенні до середнього арифметичного (унімодальний розподіл) або ж вони зростають при зміщенні до кінців інтервалу (U –подібний розподіл); чи розподіляються величини рівномірно по всій області варіації змінної або ж вони концентруються в двох точках, як от, наприклад, в бімодальному розподілі.



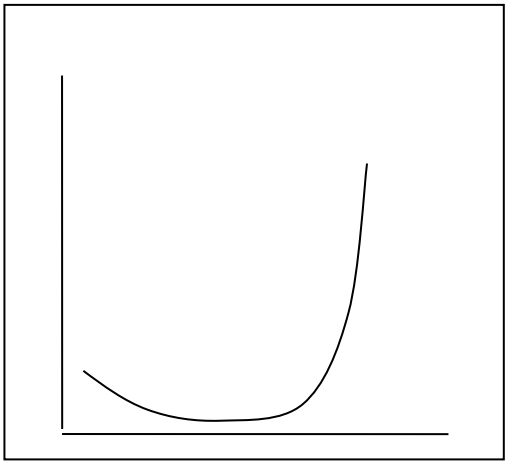
(Крива унімодального розподілу)



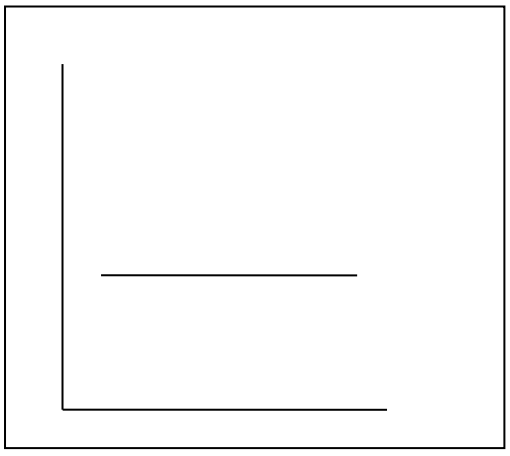
(Крива бімодального розподілу)



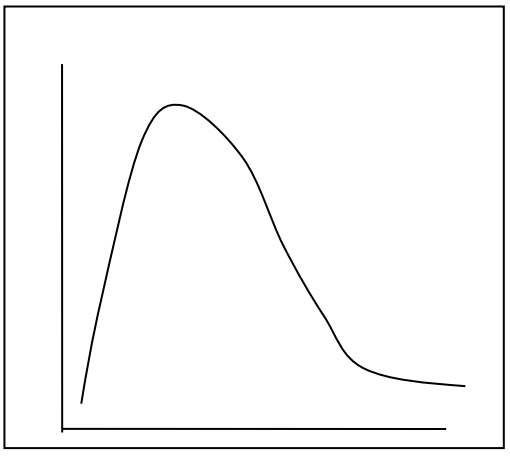
(U –подібна крива розподілу)



(J-подібна крива розподілу)



(Зображення лінійного розподілу)



(Лівий ухил кривої розподілу)

Малюнок 4.4. Графіки частот, що дозволяють судити про дисперсії

Однак подібні візуальні зображення є індивідуальними і суб'єктивними, а висновки про дисперсії, зроблені на їх основі, навряд чи можуть претендувати на науковість. В такому разі наукова обґрунтованість інформації забезпечується за рахунок об'єктивних показників, що відображають ту чи іншу міру варіації (протяжності, розкиду, розсіювання), які знаходяться за допомогою стандартних обчислювальних процедур

Вимірювання розмаху варіації. Найпростіша і найгрубіша міра варіації – розмах варіації (або «діапазон») *який. в загальних рисах, є розміром області варіації змінної. За визначенням, розмах варіації – це інтервал, що містить в собі всі значення змінної, що варіюється.* Отже, він знаходиться так само, як і звичайний інтервал: обчислюється різниця між істинними крайніми значеннями всієї сукупності змінних, які встановлюють межі розмаху варіації:

$$R = (X_{max} - X_{min}).$$

Наприклад, для визначення діапазону кількості студентів вищих навчальних закладів III – IV рівня акредитації знаходимо екстремальні істинні значення, а потім віднімаємо одне від іншого. Найменшим значенням є 159, найбільшим – 259, отже, діапазон дорівнює різниці між цими числами, тобто − 100. Це число означає значення відстань, необхідну для розташування всіх спостережуваних частот. Для іншої сукупності частот, безсумнівно, можна отримати іншу величину діапазону, інший розмах варіації. Збільшуючи кількість спостережень, можна або розширити цей розмах, або залишити його незмінним, проте скоротити його не можливо. З цієї причини два або більше діапазону можна порівняти тільки в тому випадку, якщо вони складаються з приблизно однакового числа спостережень. Наприклад, не слід порівнювати діапазони оцінок двох студентів, якщо кожен діапазон не містить приблизно однакову кількість цих оцінок.

Для дискретних даних процедура вимірювання діапазону є такою ж самою, як і для неперервних, за винятком того, що справжні межі стають при цьому, в силу необхідності, фіктивними. Так, розподіл розмірів сімей від 2 до 12 осіб має розмах варіації рівний 11, що є різницею між ***12, 5*** і ***1, 5***. Це означає, що змінна може приймати 11 і лише 11 послідовних значень: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 або 12. Ця величина знаходиться за формулою:

$$R = (X_{max} - X_{min}) + 1 = 11.$$

Застосувавши цю формулу в разі розглянутого вище прикладу, отримаємо: (12 – 2) + 1 = 11.

Значення діапазону визначається досить просто. Але, подібно будь-який інший статистичної величині, воно дасть лише обмежену інформацію. Оскільки ця величина визначається положенням на шкалі розподілу, необхідно вказувати не тільки цю величину , виражаючи її відповідним абсолютним числом, а й граничні точки. Повсякденна практика підтверджує справедливість цього положення в таких твердженнях, як: «Ціна на нові

автомобілі встановляться в діапазоні від 4000 до 5000 доларів» або «Температура завтра буде в діапазоні від 12 ^{0} С до +18 ^{0} С». Слід звернути увагу, що, наприклад, перелік окладів від 500 до 1000 гривень має той же абсолютний розмах варіації, що і перелік від 2500 до 3000 гривень, однак має зовсім інший зміст. А при виборі місця для відпочинку недостатньо знати, що діапазон температур там дорівнює 30 ^{0} С, важливо знати абсолютні значення крайніх температур.

Характерною особливістю діапазону є те, що він не враховує структури варіації в своїх межах, проте іноді саме ця структура представляє найбільший інтерес. Діапазон середньорічного доходу сім'ї в Україні становить величину понад 3500 гр., що не дає ніякої відповіді на питання, про те, згруповані доходи українців в середині, чи концентруються вони ближче до кінця інтервалу, який обмежує діапазон, або ж вони є рівномірно розподіленими по всьому діапазону.

Більш того, в більшості спостережуваних розподілів межі варіації відповідають досить малим частотам, отже, повний діапазон, який визначається за допомогою цих значень, може створити враження більшої величини варіації, ніж насправді. Встановлюючи віковий діапазон студентів вузу від 14-річного вундеркінда до 64-річної бабусі, яка бажає навчатися разом зі своїми внуками, отримуємо діапазон в 51 рік. Але цей результат затьмарює той факт, що більшість студентів відрізняються один від одного лише кількома роками, і, отже, повний діапазон, як показник варіації в даному випадку вводить в оману.

Проміжні діапазони. Така залежність від меж варіації може бути зменшена за допомогою проміжних діапазонів, які не враховують крайні значення змінної. Обмежуючи область найбільшої концентрації спостережень, такий діапазон забезпечує більшу стабільність і надійність. У звичайній практиці береться різниця між 90-м і 10-м центилями, тобто встановлюється діапазон, що включає 80% випадків. Ще більш обмеженим діапазоном є проміжний інтервал між першим і третім квартилями, який містить в середні 50% випадків. Зазвичай, він називається інтерквартильним діапазоном.

Інтерквартильний діапазон може бути зображений графічно шляхом нанесення квартилів на базову лінію графіка. Подібна процедура виявляє ступінь «згруповані» випадків, які відносяться до середнього інтервалу, навколо медіани. У розглянутому прикладі чотири інтервали, утворені квартилями, дуже виразно відрізняються по ширині. Хоча кожен з чотирьох інтервалів містить рівно до 25% загальної частоти. Така класифікація спостережень є протилежною по відношенню до ортодоксальної таблиці частот, в якій інтервали обираються рівними, а частоти при цьому різняться. Тут, навпаки, встановлюємо рівні частоти інтервалів та допускаємо змінну ширину інтервалів.

Неважко переконатися, що побудова проміжних діапазонів, таких, як інтервал 10–90%% або *інтерквартильний* діапазон, є необмеженою ніякими рамками. Проміжні діапазони в ряді випадків забезпечують більшу ясність у питанні про відносну концентрацію або дисперсію випадків, у порівнянні з

повним діапазоном. У багатьох випадках використання стратегічно розташованих квартилів дуже непогано описує дисперсію, що робить непотрібними звернення до більш складних методів.

Відхилення від середнього як міра варіації. Діапазон повний або проміжний дозволяє судити про область розподілу змінних. Хоча ця міра нерідко застосується, особливо, коли необхідно вказати абсолютні межі об’єкту вимірювання, вона дає інформацію лише про межі варіації, а не про варіацію в цих межах. Отже, необхідні деякі показники, які відображали б ступінь варіації величин повного розподілу.

Спробуємо визначити варіацію як відхилення від деякого значення. Можна, наприклад, отримати безліч парних різниць для всіх величин розподілу, а потім «конденсувати» ці різниці в деякий показник варіації. Однак кожен, кого цікавить варіація ряду значень, інтуїтивно розглядає їх у зв'язку з фіксованим стандартом, який побудований на сукупному соціальному досвіді. «Високий оклад», «надзвичайно високий рівень народжуваності» або «слабкі здібності» – всі ці судження є оцінкою відносної величини варіації, заміряної від деякої встановленої основи. Високий темп народжуваності можна розглядати як відхилення в позитивному напрямку, низький темп народжуваності – як відхилення в негативному напрямку від норми. Отже, ця норма стає центральною величиною, щодо якої вимірюється варіація. Так само можна судити і про рівень смертності, і про успіхи у навчанні, і про викладацькі окладах, особливо, коли вони є досить близькими до спостережуваних екстремумів. Тим не менш, більш обґрунтованим виявляється обрання в якості точки відліку не суб'єктивних «стандартів», а середньої величини. Іншими словами, як правило, дослідник прагне організувати спостереження навколо середнього значення, взятого в якості норми, вважаючи, що це середнє є характерним значенням. Залишається вирішити наступні проблеми: 1) від якого середнього обчислювати відхилення, 2) як представити ці відхилення в одному компактному показнику.

Вибір норми (від якої «відхиляється» відхилення). Взагалі-то, нормою в даному випадку вважається середній показник. Однак, нам відомі, принаймні, три різні середні міри *(M[X], Mo, Me),* а вибрати в якості норми потрібно тільки одну з них. Очевидно, що для симетричних унімодальних розподілів цей вибір не становить особливих труднощів, оскільки середнє арифметичне, мода і медіана більш-менш співпадають. Проте абсолютно симетричні розподіли зустрічаються вкрай рідко. Отже, дилема виникає при виборі норми для тих розподілів, для яких мода, медіана і середнє арифметичне відрізняються один від одного. Багато дослідників схильні частіше обирати моду, тобто максимальну частоту, в якості основи для порівняння. Однак більш широко в цих цілях використовуються середнє арифметичне або медіана, які вважаються більш репрезентативними.

Як уже говорилося, середнє арифметичне – це величина, варіації якої в обидві сторони є рівними. Отже, середня зручно використовувати в якості норми, однак і медіана настільки ж успішно може служити початком відліку варіації. Оскільки медіана ділить сукупність розподілу на рівні частини – сума

відхилень від медіани менше, ніж від будь-якої іншої точки. Говорячи іншими словами, медіана − це точка, для якої арифметичні «помилки» є мінімальними. Наведене твердження можна назвати «принципом мінімального відхилення».

Абсолютні показники варіації. Середнє лінійне відхилення $(\overline{d})$. Проста сума арифметичних відхилень є марною в якості показника варіації, оскільки вона безпосередньо залежить від кількості об'єктів в розподілі. Наприклад, вона буде великою для 1000 об'єктів та малою для 10. Щоб усунути впливу кількісного фактору, розділимо суму відхилень на N і виміряємо відхилення, що припадає на один об'єкт. Цей результат називається середнім лінійним відхиленням $(\overline{d})$ і обчислюється з використанням наступних формул:

1. *Для первинного ряду:*

$$\overline{d} = \frac{\sum(x_i - \mathrm{M[X]})}{n}.$$

2. *Для варіаційного ряду:*

$$\overline{d} = \frac{\sum |x_i - \mathrm{M[X]}| * n_i}{N}.$$

3. *Для інтервального ряду:*

$$\overline{d} = \frac{\sum |x_{ci} - \mathrm{M[X]}| * n_i}{N}.$$

При цьому

- x_i − значення змінної *(її величина)*;
- x_{ci} − середня точка інтервалу (*у разі інтервального ряду)*;
- M[X] − середнє арифметичне (причому замість *M [X]* ми можемо підставити в формулу *Me*, в тому випадку, коли розподіл є нерівномірним, асиметричним);
- n_i − частота значення змінної *(або частота інтервалу в інтервальнмуо ряді);*
- *n* - число варіант ознаки *(у разі первинного ряду)*
- *N* - сума всіх частот.

*Квадратичні відхилення***.** Вимірювання варіації за допомогою простих арифметичних (лінійних) відхилень є найбільш простою процедурою. Якщо дослідника цікавить лише наявність або відсутність дисперсії, то легко обчислюваний $d_{ср}$ цілком підходить для цієї мети. Тим не менш, як правило, варіація вимірюється через квадратичні відхилення від середніх величин. Логіка цієї операції ґрунтується на принципі мінімальних відхилень. Як сума квадратичних відхилень від середнього арифметичного, так і сума квадратичних відхилень від медіани є мінімальною величиною. Це положення отримало назву «принцип найменших квадратів» і є одним з найбільш важливих принципів статистичних розрахунків. Метод квадратичних відхилень може здатися, на перший погляд, штучним і надмірно ускладненим. Якщо варіація може бути задовільно виміряна посередництвом м обчислення лінійного відхилення, то які додаткові переваги дає зведення в квадрат?

Задовільну відповідь на це запитання можна отримати, лише поглибившись у вивчення математичної статистики. В рамках цього курсу доцільність такого поглиблення уявляється сумнівною, тому покладемося на досвід практиків, які прибігають до обчислення квадратичного відхилення набагато частіше, ніж простого середнього відхилення.

Інформація про квадратичні відхилення може бути отримані декількома способами, кожний з яких придатний для певної мети: *1) сума квадратів відхилень (дисперсія), 2) варіація, 3) середнє квадратичне відхилення.*

Сума квадратів відхилень (дисперсія). Дисперсія (δ^2 или *D*) – величина, яка дорівнює середньому значенню відхилень окремих значень ознаки від середнього значення. Для того, щоб знайти цю величину, необхідно зробити обчислення за наступними формулами:

1.Для первинного ряду:

$$\delta^2 = \frac{\sum(x_i - M[X])^2}{n}.$$

2. *Для варіаційного ряду:*

$$\delta^2 = \frac{\sum(x_i - M[X])^2 * n_i}{N}.$$

3. *Для інтервального ряду:*

$$\delta^2 = \frac{\sum(x_{ci} - M[X])^2 * n_i}{N}.$$

При цьому

- x_i – значення змінної (її величина);
- x_{ci} – середня точка інтервалу *(у разі інтервального ряду);*
- $M[X]$ – середнє арифметичне;
- *n* – загальне число варіант ознаки;
- n_i – частота значення змінної *(частота інтервалу в разі інтервального ряду);*
- *N* - сума всіх частот.

Для демонстрації процедури обчислення дисперсії, пропонуємо звернутися до прикладу, який ми розглядали в попередньому параграфі: «Мережа дошкільних установ усіх відомств у м. Харкові в 1999 р.». Ми обчислювали середнє арифметичне для згрупованих даних, намагаючись таким шляхом знайти середню чисельність дітей у харківських дошкільних установах. За результатами цих обчислень ми отримали *M [X] = 157,07*. Цей результат і процедура його знаходження представлені в третьому стовпці розрахункової (для дисперсії) таблиці, представленої нижче (див. *Табл. 4.5).* Коли середнє є відомим, здійснення подальших дій для знаходження дисперсії не складе великих труднощів, що також видно з даної таблиці.

Таблиця 4.5.

Мережа дошкільних закладів усіх відомств у м. Харкові в 1999 р.[2]

Кількість д/у в 1999 г. (n_i)	В них дітей в середньому ($\bar{x}_i$)	$n_i \times \bar{x}_i$	$(x_i - \text{M[X]})^2$	$(x_i - \text{M[X]})^2 \times n_i$
33	161,06	33×161,06= =5315	161,06 - *157,07* = =15,92	525,52
13	140,62	1828	270,75	3519,81
33	141,88	4682	230,77	7615,51
17	182,82	3108	663,24	11275,15
23	93,09	2141	4093,83	94158,09
35	208,46	7296	2640,64	92422,35
24	151,88	3645	26,99	647,71
18	126,00	2268	965,34	17376,21
16	188,56	3017	991,78	15868,44
N=212		33300/212 = *157,07* (M[X])		**243408,78**

Виходячи з формули обчислення дисперсії, нам залишилося виконати останню дію: підсумкове табличне значення 243408,78 розділити на суму всіх частот, рівну 212. Отримуємо: σ^2 = 243408,78 / 212 = 1148,15.

Середнє квадратичне відхилення (сигма − σ). Оскільки варіація заснована на квадратичних відхилення, вона не є лінійною мірою. Якщо потрібно лінійна міра, то необхідно витягти квадратний корінь з обох частин співвідношення. У такій формі ця величина відома як середнє квадратичне відхилення або сигма (σ). Середнє квадратичне відхилення показує, наскільки в середньому кожне значення ознаки відхиляється від середнього. Геометрично, при нанесенні на графік, сигма показує, наскільки крива розподілу «розмита» щодо середнього.

Сигма використовується як міра варіації, абсолютно аналогічно $\bar{d}$, від якого відрізняється головним чином тим, що відхилення є квадратичними, а їх середнє є лінійним. Однак витяг кореня не може повністю знищити вплив попереднього зведення в квадрат; ефект зважування частково зберігається.

Обчислення середнього квадратичного відхилення (σ). В принципі середньоквадратичне відхилення є трохи більш складним показником, ніж середнє лінійне відхилення, вимагаючи додаткового зведення відхилень в квадрат і витягання квадратного кореня з їх середнього. Взагалі, якщо нам відома дисперсія, то сигма знаходиться дуже просто, шляхом вилучення квадратного кореня, а саме: $\sigma = \sqrt{\delta^2}$.

[2] Статистичний збірник. Показники роботи закладів освіти та наукових установ області за 1999 рік; [За заг. редакцією О. Л. Сидоренка, А. С. Доценка, П. С. Дементьєва]. – Х., 2000. – 82 с.

Якщо ж дисперсія розподілу не відома, то середнє квадратичне відхилення для цього розподілу знаходиться за наступними формулами:

1. Для первинного ряду:

$$\sigma = \sqrt{\frac{\sum(x_i - M[X])^2}{n}} .$$

2. Для варіаційного ряду:

$$\sigma = \sqrt{\frac{\sum(x_i - M[X])^2 * n_i}{N}}.$$

3. Для інтервального ряду:

$$\sigma = \sqrt{\frac{\sum(x_{ci} - M[X])^2 * n_i}{N}}.$$

При цьому

- x_i – значення змінної *(її величина)*;
- x_{ci} – середня точка інтервалу *(у разі інтервального ряду)*
- $M[X]$ – середнє арифметичне;
- n – число варіант ознаки;
- n_i – частота значення змінної *(або частота інтервалу в інтервальному ряді);*
- N - сума всіх частот.

Середнє лінійне VS середньоквадратичне. Для описових цілей прийом відкидання знаків при обчисленні $\overline{d}$ є абсолютно законним. Оскільки $\overline{d}$ вимірює відхилення без зведення в квадрат, то воно за абсолютною величиною менше, ніж σ, яке непропорційно збільшує великі відхилення в результаті зведення в квадрат. Проте, нехтування знаками робить $\overline{d}$ непридатним для використання в наступних алгебраїчних обчисленнях, незалежно від його початку відліку. Саме тому одним з найбільш поширених засобів статистичного аналізу протягом майже століття була σ, причому, не тільки як міра дисперсії, але і як складова частина більш складних обчислень. Широке використання сигми в якійсь мірі пояснюється двома властивими їй перевагами. *По-перше,* σ двічі відображає величину кожної змінної розподілу: а) точка відліку, від якої заміряються відхилення *(M[X])* сама є репрезентацією всіх змінних; б) кожна величина, як така, представлена квадратичним відхиленням. *По-друге,* зведення відхилень в квадрат автоматично знижує проблему знака відхилення.

Відносні показники варіації

Лінійний коефіцієнт варіації. Як вже говорилося, будь-яке відхилення має смисл тільки тоді, коли відомо, від чого воно буде «відхилятися», тобто

лише після того, як буде задано початок відліку, норма. Цей принцип втілено в коефіцієнті варіації, який виражає міру варіації через процентне відхилення від початку відліку, незалежно від того, є воно медіаною чи середнім арифметичним. У тому випадку, коли $\overline{d}$ відраховується від середнього арифметичного, формула коефіцієнта варіації в знаменнику буде мати *M[X]*, якщо ж він заснований на медіані, формула коефіцієнта варіації в знаменнику буде мати *Me.*

$$V_d = \frac{\bar{d}}{\text{M[X]}} \times 100$$ *(якщо ми хочемо відобразити коефіцієнт у%%).*

При цьому

- $\overline{d}$ − середнє лінійне відхилення;
- M[X] − середнє арифметичне (в дільнику може бути і *Me,* в разі нерівномірного, асиметричного розподілу).

Подібно цьому прикладу, середньоквадратичне відхилення також може бути перетворено в міру відносної варіації за допомогою нормування його по відношенню до власного початку відліку, тобто середнього арифметичного:

$$V_\sigma = \frac{\sigma}{M[X]} \times 100$$ *(якщо ми хочемо відобразити коефіцієнт у%%)*

При цьому

- δ− середнє квадратичне відхилення;
- $M[X]$ − середнє арифметичне для даного розподілу.

Коефіцієнт варіації є показником мінливості ознаки щодо її середньої величини. Якщо, наприклад, в результаті відповідних обчислень, ми отримали $V_\sigma = 0{,}8$, а виразивши його у відсотках − 80%, це означає, що тільки 20% всього розподілу за цією ознакою є однорідним і наближеним до середнього значення, а інша частина розподілу неоднорідна, 80% всіх значень дуже сильно відрізняються від середнього і «далеко» розсіяні по відношенню до цього середнього.

Коефіцієнти варіації виявляються особливо корисним у процедурах порівняння, оскільки вони не залежать від абсолютних значень і від одиниць вимірювання, що використовуються. Коефіцієнт варіації дозволяє порівнювати безлічі малих та великих однорідних величин, а також якісно відмінних (до певної міри) об'єктів. Однак *V застосовується лише в тих випадках, коли: (1) спостерігаються значення мають нуль; (2) всі інтервали рівні.* Крім того, *V* більш доцільно використовувати при порівняннях між послідовностями пов'язаних даних. Наприклад, відносна варіація заробітної плати на сході може бути менше, ніж на заході України. При вимірюванні симпатії публіки до деякого композитору високий *V* буде виходити при великому розходженні в думках; низький коефіцієнт, навпаки, відбиватиме тенденцію згоди, переважного збігу думок.

Як зіставляються міри середньої тенденції та варіації і інтерпретуються результати такого співставлення. Слід підкреслити, що

мале значення σ при великому середньому вказує на більшу однорідність даних і, в силу цього, − на типовість середнього, що в деяких обставинах є вкрай суттєвим. Середнє, яке дорівнює 125, при σ=5 та V=4%, буде більш репрезентативним, ніж середнє в 125, при σ=25 і V рівним 20%. Якщо V дорівнює нулю − це вказує на повну відсутність варіації. Слід зазначити, що в тій мірі, в якій σ збільшує варіацію щодо середнього арифметичного, V, відповідно, збільшує відносну варіацію.

Варіація якісних змінних. Очевидно, що для номінальних ознак є некоректним використання всіх наведених вище мір розкиду. У якісних змінних не існує точки відліку, яка дорівнює нулю , і, отже, вони не мають величини. Разом із тим, не існує і середнього значення діапазону, а також і проміжних інтервалів. Отже, не існує і арифметичних відхилень. Це, однак, не означає, що будь-яка група якісних змінних складається з абсолютно ідентичних подій. Спробуємо зрозуміти, як можна інтерпретувати такий розкид. Дві події можна вважати різними, якщо вони володіють різними якостями. Замість обчислення величин, підраховуються відмінності в якостях. Чим більше число пар подій, що відрізняються, тим більш неоднорідною є сукупність, і, отже, тим більше варіація всередині цієї сукупності. Аналогічно, чим менше це число, тим більше однорідність всередині сукупності і менше варіація. Тому розумно встановити показник якісної варіації по повному числу різних пар подій даної множини. Питання тепер лише в тому, *по-перше,* як підрахувати повне число відмінностей і, *по-друге,* як перетворити це число в компактний показник.

Щоб знайти повне число відмінностей, підсумовуються всілякі розбіжності у групі подій. Наприклад, в множині з шести хлопчиків і шести дівчат, кожен з шести хлопчиків буде відрізнятися за своїми ознаками від кожної з шести дівчат, даючи в результаті 36 відмінностей по статі. Якби було дев'ять хлопчиків і три дівчинки, то кожен з дев'яти хлопчиків відрізнявся б від кожної з трьох дівчат, що давало б у підсумку 27 відмінностей. У групі з 12 хлопчиків очевидний результат відсутності відмінностей був би отриманий при множенні 12 на нуль.

Виходячи зі сказаного, можна підсумувати, що процедура визначення повного числа відмінностей зводиться до наступного правила: множимо частоту кожної ознаки на частоту кожної відмінної від неї ознаки і складаємо отримані результати. Наприклад, в сукупності з чотирьох католиків, п'яти християн і шести іудеїв будемо мати: (4 × 5) + (4 × 6) + (5 × 6) = 74 (відмінності).

Коефіцієнт якісної варіації (V_q). Число відмінностей, як показник варіації, можна порівняти тільки з максимально можливим числом відмінностей. Це максимальне число відмінностей буде спостерігатися в тому випадку, коли всі частоти різних ознак будуть рівними. Таким чином, максимум обчислюється шляхом прирівнювання частот (тобто обчислення середньої частоти), перемноження частот і додаванні результатів цих множень один до одного. Іншими словами, здійснюються наступні операції: (1) знаходиться середня частота, (2) цей результат зводиться в квадрат, 3) квадрат

множиться на число можливих пар ознак. У вищезгаданому прикладі, у випадку з дев'ятьма хлопчиками та трьома дівчатками, максимально можливе число відмінностей за статтю в групі з 12 людей було б наступним: 6 (хлопчиків) × 6 (дівчаток) = 36, або в цьому конкретному випадку середня частота множиться на саму себе.

Відносна величина варіації тепер може бути виміряна за допомогою співвідношення емпіричного числа відмінностей та його гіпотетичного (теоретично можливого) максимуму:

$$Коэффициент\_качественной\_вариации(V_q) = \frac{Полное\_число\_набл.\_различий}{Макс.\_возможн.\_число\_различий}$$

Результат обчислень за відповідною формулою представляє собою частку всіх значень ознаки, які сильно неоднорідні. Очевидно, що цей показник приймає значення від 0 до 1. Чим більше коефіцієнт варіації наближається до нуля, тим меншою є варіація значень ознаки. Як і у випадку з варіацією кількісних даних, отримане число, яке не може перевищувати одиниці і опускатися нижче нуля, можна представити у відсотках, помноживши отриманий результат на 100.

Проілюструємо застосування даної формули на попередньому прикладі з дев'ятьма хлопчиками та трьома дівчатами: $V_q = \frac{27}{36} = 0{,}75(\times 100) = 75\%$. Цей показник говорить про те, що варіація значень ознаки є досить високою, і лише 25% (100% мінус 75%) цих значень можна назвати відносно однорідними.

Середнє число членів кожної з трьох згаданих вище релігійних груп дорівнює п'яти. Помноживши «5» на «5» і підсумувавши результати множення, знайдемо, що максимальне число відмінностей має дорівнювати 75. Спостережувані відмінності, як уже було обчислено, дорівнюють 75. Отже: $V_q = \frac{74}{75} = 0{,}99(\times 100) = 99\%$. Отримана цифра говорить про вкрай високу неоднорідність всіх значень ознаки.

Приклади використання коефіцієнту варіації. Як вже говорилося, коефіцієнт, що розглядається нами, може бути використаний для порівняння тих чи інших відносних величин. Наприклад, спробуємо порівняти, наскільки виросла/знизилась кількість професорів та доцентів в системі вищій освіти України з 1996 по 2000 рр. (див. *Табл. 4.6*).

Таблиця 4.6

Динаміка чисельності основного професорсько-викладацького складу вищих навчальних закладів України III – IV рівня акредитації

Роки		Професори	Доценти
1995	1996	29597	5728
1999	2005	28540	6546

В Україні в 1995/1996 навчальному році у вищих навчальних закладах III – IV рівня акредитації працювало 29597 професорів і 5728 доцентів, отже, максимально можливе число відмінностей дорівнюватиме

$((29597+5728)/2)^2 = 17663^2$, тогда $V_q = \frac{29597 \cdot 5728}{17663^2} = 0{,}54(\times 100) = 54\%$.

У 1999/2000 навчальному році ситуація була наступною:

$$V_q = \frac{28540 \cdot 6546}{17543^2} = 0{,}61(\times 100) = 61\% .$$

Елементарне нормування. Необхідність нормування. Будь-яку подію дослідник розглядає не ізольовано, а в порівнянні з конкретною нормою, що випливає з соціальної основи даної події. Наприклад, річний дохід в 13500 гривень сприймається дослідником не як абстрактне число, а як соціальне явище, віднесене до певного стандарту; факт народження 100 чоловік у певній спільності має сенс лише у зв'язку з такими даними, як загальна кількість населення, період часу, число народжень у попередній рік або число народжень в інших спільнотах. Якщо норму, з якою проводиться порівняння, не встановлено, дослідник мимоволі встановлює її самостійно.

Ще більші труднощі виникають при порівнянні двох чи більше величин, взятих з різних сукупностей. Наприклад, порівняння розумових здібностей чоловіків і жінок (за результатами тестування) може бути помилковим, оскільки відомо, що результати таких тестів залежать від рівня освіти, який може бути розподілений нерівномірно між представниками різної статі.

Існує багато способів здійснення процедури подібного порівняння; деякі з них будуть розглядатися в цьому розділі. Сукупність таких процедур можна назвати операціями нормування, оскільки вони встановлюють певні стандарти для спостережуваних величин. Процес нормування вже відомий читачеві з досвіду обчислення варіацій, і в цьому підрозділі ми зупинимося на деяких інших аспектах цієї важливої процедури, зокрема в застосуванні до психологічних і соціологічних даних.

Можна здійснювати нормування приблизно в такому порядку складності: *1) процентні відношення, 2) пропорції, 3) ступені; 4) індекси; 5) підкласифікація; 6) стандартизація.*

Процентні відношення. Найпростіша форма нормування полягає у приведенні рядів абсолютних чисел до стандартної чисельної основі. Громіздкі абсолютні числа замінюються зазначенням їхнього ставлення до деякої основі, вираженої в відсотках. Замість того щоб вказувати, що зареєстроване число юнаків і дівчат − студентів вузу дорівнює, відповідно 9244 та 4622, ці значення перетворюють в 66,7 і 33,3%. Ця операція є настільки звичною, що основний принцип, на якому вона заснована, навіть не завжди повністю усвідомлюється.

Пропорції. Можна порівнювати дві величини у формі відношення або висловлювати одну з них як кратну інший. Відношення бувають різними за складом: можна виділити відношення «частина–частина» частот в межах однієї і того ж числа й відношення «ціле до цілого» між частотами двох взятих змінної. Таким чином, співвідношення статей може розглядатися як відношення «частина – частина», воно порівнює кількість чоловіків в даній сукупності з кількістю жінок.

У 1999 році в Україну було 34 млн. осіб міського населення і 16,1 млн.

осіб – сільського. Співвідношення між цими громіздкими числами легше запам'ятати та усвідомити існуючу різницю, якщо виразити це відношення в наступному вигляді: 2,1 міського жителя припадає на 1 сільського.

В антропометрії є таке поняття, як «цефалічний індекс», який являє собою відношення «цілого−до−цілого» двох показників розміру черепу – ширини до довжини, з метою виявлення сукупного кількісного показнику відмінності між круглою та овальною формою голови. Відношення перемножується на 100, задля того,щоб зробити запис більш зрозумілим:

$$Цефалический\_индекс = \frac{ширина}{длина} \times 100\% .$$

Аналогічно цьому ступінь розумового розвитку дорівнює відношенню між розумовим та хронологічним віком респондента, що дозволяє порівнювати людей різних вікових груп за розумовим розвитком. Таким чином, $Iq = \frac{Умственный\_возраст}{Хронологический\_возраст} \times 100\% .$

Це відношення дорівнює 100 в тому випадку, коли хронологічний і розумовий вік виявляються рівними. Інші загальноприйняті пропорції, що використовуються, наприклад, в соціально-економічному аналізі − це пропорції: люди – житло; населення – земля; діти – дорослі.

Ступені (коефіцієнти). Ступінь є, по суті, арифметичним середнім. Вона являє собою середнє число значень однієї змінної, виражене в одиницях іншої. Так, ступінь, рівна 20 кілометрам на 1 літр, тобто середнє споживання палива, при якому відстань в кілометрах буде приймати певне значення при заміні одного літра іншим. Хоча всі ступені засновані на спостереженнях, вони можуть забезпечувати передбачення майбутнього. Тому іноді обчислені ступені (та, до речі, середнє) можуть розглядатися як очікувані величини. Будучи в основному результатом великого числа спостережень, ступінь часто виявляється ефективним інструментом аналізу даних. Наразі багато ступенів вже набули статусу загальноприйнятих понять: ступінь злочинності, ступень народжуваності і смертності та багато інших.

Математико-статистичний смисл ступеня залежить в основному від двох змінних: від проблемної змінної і від нормованої змінної. При обчисленні ступенів, найбільш важливим є вибір нормованої змінної, з якою буде порівнюватися проблемна змінна. Наприклад, при обчисленні ступеня народжуваності необхідно вибрати нормуючу сукупність, з якою в подальшому буде порівнюватися абсолютне число актів народження. Для цієї мети можна було б використовувати або загальну кількість людей, або число жінок, що досягли зрілого віку, або число заміжніх жінок, які досягли зрілого віку. Найбільш часто для цієї мети використовується (хоча і не цілком виправдано) повна сукупність людей: чоловіків, жінок, дітей.

Наступним етапом побудови ступеня є вибір стандартної чисельної основи: 10, 100, 1000 або кратне їм значення. Призначення числової основи полягає просто у вказівці десяткового масштабу для зручності табулювання, для полегшення цитування і більш швидкого розуміння. Чисельний масштаб часто встановлюється за угодою, особливо коли не існує іншого виходу, окрім

простого підпорядкування усталеній традиції. Так, рівень народжуваності, якому присвоєно число 24, має інтернаціональне трактування і означає 24 випадки народжень на 1000 осіб всього населення в даному році і на даній території. Таке уявлення сприймається більш відчутно, ніж , наприклад, запис: 768 з 32462. Обчислення здійснюється наступним чином: число народжень = 768, повне населення = 32462, числовий масштаб = 1000:

$$Ступінь\_народжуваності = \frac{768}{32462} \times 1000 = 24$$

Узагальнена формула читалася б таким чином:

$$Степень = \frac{Частота\_проблемной\_переменной}{Частота\_нормировочной\_переменной} \times Числовой\_масштаб.$$

В умовних позначеннях:

$$Степень = \frac{r_{пп}}{r_{тп}} \times r_{т},$$

де

- $r_{пп}$ – частота (емпірична) проблемної змінної;
- $r_{тп}$ – частота нормованої змінної;
- $r_{пп}$ – числовий масштаб.

Ані нормуюча змінна, ані числовий масштаб не визначаються твердою угодою, як от, наприклад, для ступеня народжуваності або смертності. У таких випадках допускається деяка свобода дій, проте описання процедури обчислення слід обов'язково забезпечити пояснюючими примітками. Ступінь розлучень може бути обчислена як для всього населення, так і для числа подружніх пар в один і той же рік і на одній площі, або навіть для числа шлюбів протягом попередніх десяти років, як це іноді робиться для більшості розлучень. Ступені злочинності може обчислюватися відносно певного віку і конкретних статевих груп; ступінь шлюбів – лише для вікової групи від 14 років і більше, тощо.

З попередніх прикладів можна зробити висновок про те, що нормована змінна повинна мати ті ж характеристики, що й проблемна – вони об'єднуються під загальною назвою «*відкрита група*». Якщо смерть може трапитися з кожною людиною, то народження дитини, одруження або заміжжя і розлучення – не з кожним. Отже, ступінь, заснована на розумно обраній відкритій сукупності, менш схильна до спотворення через будь-які зовнішні чинники. Ступені народжуваності, шлюбів і розлучень, обчислені по відношенню до повної кількості населення, зазвичай, називаються *наближеними ступенями*, а ступені, обчислені для особливих груп, позначаються як «уточнені».

Індекси. Індекс – це термін, який використовується в розмовній мові і техніці задля досягнення самих різних цілей; відноситься до більш складних ступенів або множинних пропорцій. В якості вторинного заходу індекс

зазвичай обчислюється для опису варіації, яка в безпосередньому вигляді могла б бути зовсім непомітною. У більш формалізованому варіанті він зазвичай описує відношення між двома величинами, одна з якої взята в якості норми, або очікуваної величини, тоді як інша є вимірюваною величиною. Так, індекс вартості життя порівнює ціни в конкретному році з середніми цінами для «нормального» року. Індекс, рівний 139 в 1995 році, при використанні в якості основного року – 1991, показує, що вартість життя підвищилася на 30% в порівнянні з основним роком, для якого вартість дорівнює 100. Хоча такий, з першого погляду, простий індекс може жваво цитуватися будь-яким журналістом, його внутрішній зміст, що включає охоплення, зважування, метод усереднення спостережень, а також вибір основного періоду, свідчить про його статистичну складність, що проілюстровано у таблиці 4.7. До речі, аналогічно V_q (про який мова йшла вище) порівнює спостережуване число відмінностей ознак заданої множини з максимально можливим числом відмінностей, яке в даному випадку виступає у якості норми .

Таблиця 4.7

Обчислення індексу соціально-економічної класифікації студентів університету та населення регіону

Група	***Університет***		Регіон%	Різниця	Індекс
	Кількість	%%			
Техніки	408	19,3	4,7	14,6	411
Інженери, управлінці	507	24,0	4,7	19,3	511
Бізнесмени	290	13,7	5,0	8,7	274
Клерки	286	13,5	12,8	0,7	105
Фермери	196	9,3	15,9	– 6,6	58
Кваліфіковані робітники	267	12,6	17,0	– 4,4	74
Напівкваліф. робітники	94	4,5	19,9	– 15,4	23
Некваліф. робітники	65	3.1.	20,0	– 16,9	15
Всього:	2,113	100%	100%		

Таблиця 4.7 демонструє процедуру побудови і використання індексу для вимірювання соціальної стратифікації студентів університету, а також те, наскільки представлені в цьому навчальному закладі різні соціальні класи. Логічна основа побудованих в цій таблиці індексів є наступною: дочки заможних батьків складають 19,3% від загального числа дівчат в університеті, тоді як в іншому регіоні найзаможніша група складає 4,7%. Зіставлення процентних відношень між цими парами могло б служити мірою доступності навчання для різних класів. Існує два методи, за допомогою яких можна було б визначити цю міру:

1) за допомогою простої відмінності між процентними відношеннями;

2) за допомогою нормованих індексів.

Що стосується першої альтернативи, то аналіз відмінностей виявляє, що

при переміщенні по соціальній шкалі відмінності зростають. Проте ці відмінності є абсолютними, в силу чого їх неможливо нормувати по величині або в залежності від початку відліку. Щоб унормувати дані відмінності, будується індекс, що складає зміст другої альтернативи. Таким чином, якщо б відвідуваність університету розподілялася випадковим чином між усіма соціальними класами, то можна було б очікувати, що заможна група, яка становить 4,7% всього населення, має дати вклад, рівний 4,7% від загального числа студентів. Очікувана і спостережувана частка студентів із заможних родин були б у даному випадку ідентичними, співвідношення між ними дорівнювало б одиниці. В дійсності, заможна частина студентів університету дорівнює 19,3%, що складає $\frac{19,3}{4,7}\cdot 100 = 41,1\%$ від відповідного відсотка по всьому регіону, або в 4,11 рази перевищує очікуване значення. Таким же чином можна унормувати всі інші соціально-економічні процентні відношення.

Інший підхід до визначення індексу полягає в нормуванні послідовності величин щодо їх середнього. У цьому випадку індекс являє собою просте відношення між даною величиною і середнім значенням послідовності. Наприклад, середня ступінь смертності в ряді міст дорівнює 10,5. Якщо всі фактори, що впливають на ступінь смертності, були б однаковими у всіх містах, тоді всі ступені смертності були б ідентичними і, отже, дорівнювали б середньому значенню послідовності. Оскільки при цьому припущенні середнє є очікуваним або теоретичним значенням, для того, щоб оцінити вагу факторів, які впливають на відмінності, індивідуальні ступені вимірюються по відношенню до середнього. Наприклад, якщо ступінь смертності для міста, яка спостерігається, дорівнює 7, то індекс ми могли б обчислити наступним чином:

$$ІНДЕКС = \frac{спостерігаємий\_ступінь}{очікуваний\_ступінь}\times 100\% = \frac{7}{10,5}\times 100\% = 67\% \ .$$

Це означає, що ступінь смертності в даному місті становить 67% від середнього показника. За допомогою цього методу будь-яке місто можна розташувати на шкалі по відношенню до середнього. Цей прийом нормування є аналогічним до процедури обчислення V, оскільки він висловлює вихідні значення через їх власне середнє. За аналогією обчислюється і, так званий, сезонний індекс: як відношення між місячною та середньорічною мірою. Цей індекс використовується для вимірювання флуктуацій народжуваності, смертності, промислової продукції та деяких інших економічних показників.

Нормування за допомогою підкласифікації. Подібно до того, як немає сенсу у зіставленні окремого абсолютного значення з відповідною нормою, так само і ступінь або процентне відношення практично не має сенсу, якщо виокремлюється з контексту. Вони набувають сенсу у порівнянні з аналогічними показниками. Як от, наприклад, у порівнянні рівня народжуваності двох регіонів і ступеня одруження католиків і християн. Проте не слід робити занадто поспішного висновку про існування причинно-наслідкового співвідношення між такими «спареними» змінними.

Спостережувані варіації ступенів можуть іноді виникати в результаті дії факторів, не врахованих в класифікації. Такі фактори можна назвати прихованими. У багатьох випадках порівняння двох або більшої кількості спостережуваних величин спотворюються в результаті впливу саме прихованих факторів. Так, наприклад, більш висока ступінь народжуваності в одному регіоні може бути наслідком не більшої плідності його населення, а більшої кількості жінок, здатних до дітонародження. Цей випадковий фактор не є класифікованим у наближеному ступені.

Під нормуваннями за допомогою підкласифікації зазвичай розуміється поділ факторів на «зовнішні» і «внутрішні», причому зовнішні фактори не повинні змінюватися в ході дослідження. З усього сказаного витікає те, що не можна розповсюджувати на всі групи результати, отримані для якихось конкретних підгруп. Останні відрізняються не тільки вагою, а й впливовістю інших факторів. Тому необхідною є розробка методу, який дозволив би отримати просту, уточнену, але вільну від впливу ваг, ступінь. Відповідний метод отримав назву «стандартизація», а відповідна ступінь називається стандартизованою.

Було виявлено, наприклад, що підвищений ступінь злочинності населення можна пояснити більш високою часткою молоді в досліджуваній місцевості. Саме цим, а не надлишкової тенденцією до вчинення злочинів, пояснюється підвищений ступінь злочинності місцевого населення. В цілому ж з стандартизовані ступені злочинності для кожної групи населення знаходяться таким чином: обчислюється очікуване число злочинів для місцевого населення за умови, що воно має таку саме вікову структуру, що і інші поселенські групи. Іншими словами, необхідно діяти так, як ніби всі досліджувані групи мають однакове віковий розподіл.

Хоча стандартизація здається повністю формалізованою процедурою, немає ніякої формули, яка допомогла б визначити, наскільки докладної і якою саме має бути підкласифікація. Тому дослідник не звільняється від змістовного аналізу завдання. Так, наприклад, вік можна було б підкласифікувати більш ніж на три інтервали; чим більше число інтервалів − тим більше точність порівнянь. Однак, існують практичні обмеження, за межі яких поширювати підкласифікацію немає сенсу. Іноді достатньо найгрубішого поділу віку на три інтервали, щоб встановити важливість вікового чинника в ступені злочинності.

Застосування стандартизації не обмежується ступенями та відсотками. Будь-який вид середнього арифметичного може бути стандартизований за умови наявності необхідних даних для підкласифікації. Необмежені можливості стандартизації нагадують ще раз, наскільки далека «остаточна істина» від даних, які лежать перед нами і на яких, тим не менш, часто ґрунтуються наші думки та дії. Досить часто наближені ступені, які необхідно перетворити в нормовані, відповідають великим територіям і охоплюють великі інтервали часу. Важко порівнювати статистику різних країн, якщо відсутня загальна узгодженість стандартів. В якості одного з таких стандартів в 1901 р. часто застосовувався англійський «стандартний мільйон». З тієї причини, що розподіл населення за віком є одним з факторів, що дуже

спотворює будь-які показники, в інтерпретації соціальної статистики є таке поняття як «стандартний мільйон», що являє собою віковий розподіл одного мільйона британського населення. Така процедура нормує ступені за віком, тим самим перетворюючи їх у величини, зручні для порівняння.

Перехресні таблиці зазвичай створюються з метою виявлення статистичних асоціацій, однак через присутність прихованих чинників, отримані значення асоціацій не слід розглядати як бездоганно достовірні. Щоб виявити приховані чинники, можна підкласифікувати перехресні таблиці. Для ілюстрації розглянемо дані (фіктивні) для 52 районів, перехресно класифікованих близькістю до певного типу виробництва і за ступенем правопорушень (див. *Табл.4.8*).

Таблиця 4.8.

Наближені коефіцієнти злочинності, промислових міських та сільських районів (віком населення 15 до 75 років)

Тип району	***Коефіцієнт злочинності***					
	Кількість			*Відсоток*		
	Високий	Низький	Сума	*Високий*	Низький	Сума
Промисловий	15	8	23	65	35	100
Сільський	10	19	29	34	66	100
Всього:	25	27	52	48	52	100

Ця таблиця показує, що промислові райони частіше характеризуються високими ступенями правопорушень, ніж райони, де мешкають сільські жителів, вказуючи, немов би то на статистичний зв'язок між місцем проживання та рівнем злочинності. Цей зв'язок виявляється більш чітко в процентному розподілі, яке показує, що 65% всіх промислових районів знаходяться в категорії «високої злочинності», тоді як підкласифіковані аналогічним чином сільські райони становлять в цій категорії лише 34%. З точки зору наближеного ступеня, висновок про зв'язок між місцем проживання і злочинністю здається переконливим. Однак такий висновок є неприйнятним для будь-якого фахівця з соціальної патології міста. Він вказав би, що жителі одних районів зосереджені в регіонах з низьким рівнем життя, тоді як жителі інших районів частіше проживають у більш сприятливих умовах з відносно високим рівнем життя. Розумно тому запитати, чи буде зберігатися різниця між міськими та сільськими районами щодо злочинності, якщо ці райони унормувати за однаковим економічним рівнем. Щоб відповісти на це запитання, необхідно підкласифікувати райони згідно життєвим стандартам і провести порівняння в межах однакових соціально-економічних підкласів (див. *Табл. 4.9).*

Таблиця 4.9.

Коефіцієнти злочинності, віднесені до віку (15 до 75 років) і рівня життя респондентів, що мешкають у промислових та сільських районах

Тип району	***Рівень життя***	
	Високий рівень	*Низький рівень*
	Коефіцієнт злочинності	*Коефіцієнт злочинності*

	Високий	***Низький***	Разом	***Високий***	Низький	Разом
Промисловий	1	6	7	14	2	16
Сільський	3	18	21	7	1	8
Разом:	4	24	28	21	3	24
Процентний розподіл						
Промисловий	14	86	100	88	12	100
Сільський	14	86	100	88	12	100
Разом:	14	86	100	88	12	100

Аналізуючи цю таблицю, можна побачити, що зв'язок між місцем проживання і рівнем злочинності зникає: з 24 економічно гірших районів 21 район або 88% знаходяться в категорії високої злочинності (і це справедливо як для промислових, так і для сільських районів). З 28 районів, більш розвинених економічно, тільки 4 або 14% знаходяться в категорії високої злочинності, незалежно від типу району проживання. В результаті в цій гіпотетичній ілюстрації злочинність повністю залежить від економічного рівня і зовсім не залежить від типу району проживання. Такий зв'язок називається помилковим, тому що він фактично є результатом дії прихованого соціально-економічного чинника, який дає «справжнє» пояснення.

В основному підкласифікація − це процедура уточнення порівнянь, яка, подібно анатомічному розчленуванню, є інструментом для більш повного глибокого статистичного аналізу. Підкласифікація відрізняється від стандартизації тим, що вона чисто є описовою і має стільки ж показників, скільки вибрано підкласів. Стандартизована ступінь, з іншого боку, є скоріше гіпотетичною, ніж описовою; це єдиний, сукупний показник, зважений по ряду частот підкласів, що використовуються як стандарт.