МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

**Є. Т. ВОЛОДАРСЬКИЙ, Л. О. КОШЕВА**

# ТЕОРІЯ ТА ПРАКТИКА ЕКСПЕРИМЕНТАЛЬНИХ ДОСЛІДЖЕНЬ

**Підручник**

Затверджено Вченою радою КПІ ім. Ігоря Сікорського
як підручник для здобувачів ступеня магістра за спеціальністю
175 «Інформаційно-вимірювальні технології»

Електронне мережеве навчальне видання
2-ге видання, доповнене та перероблене

Київ
КПІ ім. Ігоря Сікорського
2024

*Рецензенти* *Романов В. О.,* д-р. техн. наук, проф., завідувач відділу Інституту кібернетики НАН України ім. В.М. Глушкова
*Самойленко О. М.,* д-р. техн. наук, проф., директор Інституту ДП «УКРМЕТРТЕСТСТАНДАРТ»

Відповідальний редактор *Шевченко К. Л.,* д-р техн. наук, професор

*Гриф надано Вченою радою КПІ ім. Ігоря Сікорського (протокол № 6 від 24.06.2024 р.)*

**Володарський Є. Т.**

В 68 Теорія та практика експериментальних досліджень [Електрон. ресурс] : підруч. для здобувачів ступеня магістра спец. 175 «Інформаційно-вимірювальні технології» / Є. Т. Володарський, Л. О. Кошева ; КПІ ім. Ігоря Сікорського. – 2-ге вид., доповн. та переробл. – Електрон. текст. дані (1 файл). – Київ : КПІ ім. Ігоря Сікорського, 2024. – 355 с.

У підручнику стисло і доступно висвітлено питання статистичної оцінки параметрів розподілу даних, розглянуто процедури кореляційного аналізу, та відтворення регресійних залежностей, викладено основи багатовимірного регресійного аналізу та застосування ортогональних поліномів Чебишова. Розглянуто характеристики випадкових процесів, які визначаються за статистичними даними, а також питання побудови моделі об'єкта дослідження, організації й планування експериментів, основи дисперсійного аналізу, який дозволяє встановити зв'язок між якісними величинами. Наводяться приклади побудови моделей, перевірки їх адекватності експериментальним даним.

Призначений для здобувачів ступеня магістра за спеціальністю 175 «Інформаційно-вимірювальні технології». Може бути корисним для здобувачів інших технічних спеціальностей, а також аспірантів і фахівців, що займаються експериментальними дослідженнями та статистичною обробкою експериментальних даних.

УДК 519.24(075.8)

Реєстр. № П 23/24-041. Обсяг 16,75 авт. арк.
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
проспект Берестейський, 37, м. Київ, 03056
https://kpi.ua
Свідоцтво про внесення до Державного реєстру видавців, виготовлювачів і розповсюджувачів видавничої продукції ДК № 5354 від 25.05.2017 р.

# ВСТУП

Головна мета підручника – дати системне та чітке уявлення про статистичні принципи та методи обробки даних експериментальних досліджень, встановлювати зв'язок (статистичні моделі) між кількісними та якісними величинами, допомогти свідомо вибирати для застосування той чи інший підхід. Навчити правильно використовувати статистичні методи обробки даних, окреслити коло експериментальних завдань, для яких ці методи можуть застосовуватися найліпшим чином.

У підручнику відображений багаторічний досвід авторів у викладанні дисциплін, пов'язаних з методологією експериментальних досліджень. Деякі складні положення, засвоєння яких викликає найбільші труднощі, а також які, на думку авторів, є найбільш важливими для практичної діяльності, детально роз'яснені, підтверджені прикладами, що сприяє кращому сприйняттю матеріалу. Ґрунтуючись на логічному зв'язку між етапами обробки експериментальних даних, автори зробили спробу всебічно об'єднати розрізнені розділи.

При написанні підручника використано матеріали навчального посібника Володарського Є. Т. Кошевої Л. О. [1], які було доповнено та адаптовано до спеціальності 175 «Інформаційно-вимірювальні технології».

Зміст, структура та методика викладення матеріалу посібника відповідають вимогам поєднання фундаментальної наукової та практичної направленості навчання, можливостям розвитку самостійної творчої роботи над дисципліною..

Підручник може бути корисним для здобувачів інших технічних спеціальностей, а також аспірантів і фахівців, що займаються експериментальними дослідженнями.

Автори висловлюють глибоку подяку рецензентам д-р. техн. наук, проф. Романову В. О., д-р. техн. наук, проф. Самойленку О. М. за позитивні зауваження та поради, які сприяли поліпшенню якості підручника.

# РОЗДІЛ 1

# ОСНОВНІ ПОЛОЖЕННЯ ТЕОРІЇ ЙМОВІРНОСТЕЙ

## 1.1. ПРЕДМЕТ І ЗАВДАННЯ ТЕОРІЇ ЙМОВІРНОСТЕЙ ТА МАТЕМАТИЧНОЇ СТАТИСТИКИ

Методи статистичної обробки даних дозволяють знаходити цілком закономірний зв'язок між числовими значеннями ознак, що змінюються, і ймовірністю реалізації цих значень у масі проведених спостережень

У наукових дослідженнях, техніці, масовому виробництві ми часто зустрічаємося з дослідами, операціями або явищами, *що багаторазово повторюються в незмінних умовах.*

Відомо, що при повторенні вимірювань параметрів того самого об'єкта, виконуваних за допомогою того самого вимірювального приладу з однаковою старанністю, ми ніколи не одержимо однакових результатів. У цьому випадку говорять, що результати мають *випадкове розсіювання.*

Якщо навіть шляхом застосування спеціальних методів виключити систематичні похибки й промахи, все-таки виявиться, що на результатах вимірювань буде позначатися *вплив численних факторів, які не піддаються контролю й невідомо як змінюються від одного вимірювання до іншого.* До таких факторів можна віднести випадкові вібрації окремих частин приладів, фізіологічні зміни органів почуттів виконавців, різні невраховувані зміни параметрів середовища (температура, оптичні, електричні й магнітні властивості, вологість тощо).

Отже, результат кожного окремого вимірювання за наявності випадкового розсіювання заздалегідь пророчити неможливо, але це ще не означає, що повторні вимірювання не виявляють ніякої закономірності.

**Приклад**. Рухи й зіткнення окремих молекул газу відбуваються хаотично. Кожна з молекул описує складну й заплутану траєкторію, так що не представляється можливим пророкувати, де вона буде перебувати через певний відрізок часу. Проте, виділивши на стінці посудини елементарну площинку, виявимо, що середня кількість молекул газу, що зіштовхнулися зі стінкою в цій площинці в певний відрізок часу й середній імпульс, переданий стінці при зіткненнях, будуть досить стійкою величиною, що визначає пружність і тиск газу.

Цей приклад підтверджує, що в гаданій хаотичності й довільності поводження виникають досить стійкі своєрідні закономірності ймовірнісного типу.

Вивченням закономірностей подібного роду займається *теорія ймовірностей.* Вона вивчає масові випадкові явища та процеси, зв'язує з кожним з можливих результатів (наслідків) дослідів особливу числову міру об'єктивної можливості його появи, яка називається *ймовірністю*.

Теорія ймовірностей розглядає методи визначення ймовірностей складних результатів масових випадкових явищ або процесів за відомими ймовірностями більш простих наслідків. Тим самим відкривається шлях для *аналізу й виявлення імовірнісних закономірностей випадкових явищ***.**

Таким чином, основною передумовою теорії ймовірностей є *відтворюваність умов випробувань і можливість нескінченності їх проведення***.**

На практиці ідеального відтворення умов випробувань реалізувати неможливо, разом з тим, кількість повторних

вимірювань зазвичай буває малим і, саме в цих умовах, особливо важливі методи *математичної статистики,* основою якої є теорія ймовірностей.

Математична статистика розробляє раціональні прийоми обробки *обмеженого обсягу даних,* що належать до масових явищ і відбивають вплив *випадкових факторів*. Методи статистичної обробки даних дозволяють знаходити цілком закономірний *зв'язок* між *числовими значеннями* ознак, що змінюються, *і ймовірністю* реалізації цих значень у масі проведених спостережень. Саме це дає можливість побудувати загальну теорію, що показує, які прийоми обробки спостережень, які середні показники, виведені із зазвичай обмеженого матеріалу спостережень, щонайкраще відповідають специфіці випадкового розсіювання в тому або іншому завданні.

Математична статистика вирішує такі задачі.

• ***Задача опису*** випадкового розсіювання за даними масових явищ**.** Ідеться про питання, пов'язані з виявленням для кожного випадку відповідного закону розподілу, тобто про подання випадкових подій (процесів) у вигляді словесного опису їхнього поводження, графічного зображення, математичної моделі (*описова статистика*)**.**

**Приклад.** Імовірність потрапляння м'яча у ворота під час гри в регбі, коли він може влучити в будь-яке місце площини воріт буде однаковою – це приклад словесного опису. Те саме можна зобразити у вигляді графіка рівномірного закону із границями $x_н$ і $x_в$ – графічне подання (рис. 1.1). Можна застосувати математичну форму – записати ймовірність влучення м'яча у ворота у вигляді формули $p = 1/ (x_в - x_н)$.

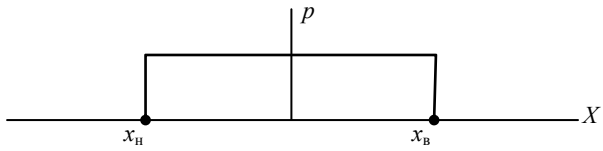



Рис. 1.1. Рівномірний закон розподілу ймовірностей

• ***Задача оцінювання невідомих параметрів*** законів розподілів на підставі наявної кількості спостережень.

Випадкове розсіювання, досліджуване статистичними методами, має важливе практичне значення. Його потрібно враховувати при проектуванні й розрахунку будь-яких пристроїв, коли поряд із контрольованими параметрами доводиться зважати на такі зовсім випадкові, як коливання сили вітру, витрати води, зміни температури навколишнього середовища тощо. У всіх цих випадках тільки знання законів розподілу впливових величин, їх належне врахування можуть забезпечити за допомогою інженерних розрахунків необхідну роботу пристроїв.

• ***Задача перевірки гіпотез***, тобто припущень, що стосуються збігу та розбіжностей параметрів розглянутих розподілів спостережень. В умовах, коли кількість спостережень обмежена й дані про масове явище виявляють значне розсіювання, об'єктивний висновок про переваги того або іншого методу вимірювань чи технологічного процесу, про користь пропонованих ліків тощо можна зробити лише на основі статистичного аналізу й зіставлення даних спостережень, що належать відповідній області.

• ***Задача встановлення наявності залежностей*** між величинами, які змінюються під дією тих самих або різних випадкових факторів (розглядається в *теорії кореляції*).

• ***Задача встановлення виду залежності,*** тобто одержання регресійної моделі (розглядається в *регресійному аналізі*).

• ***Задача прогнозування*** із застосуванням спеціальних методів прогнозування.

На практиці ці задачі пов'язані між собою й виконуються послідовно, від простої до більш складної.

**Приклад**. У книзі бельгійського антрополога Адольфа Кетле (1796 – 1874) «Про соціальну систему й закони, що керують нею» даються приклади використання статистичних спостережень у

медицині. Два відомі професори медицини зі Страсбурзького університету Рамо й Саррю зробили цікаве спостереження щодо швидкості пульсу. Вони помітили, що між зростом і пульсом людини існує залежність. Вік може впливати на пульс тільки при зміні зросту. Частота пульсу перебуває в оберненому відношенні до квадратного кореня зі зросту. Узявши за зріст середньої людини 1,684 м, Рамо й Саррю вважали, що частота її пульсу дорівнює 70 ударів за хвилину. Маючи ці дані, можна обчислити частоту пульсу в людини будь-якого зросту.

**Приклад**. Перші застосування статистичних методів у медицині належать до XVIII століття, коли в Англії було помічено, що відносна частота смертності чоловіків і жінок одного віку, які живуть приблизно в однакових умовах, рік у рік хоч і коливається, але у вузьких межах, що дає змогу з великою точністю прогнозувати частку померлих у тій або іншій категорії населення.

Таким чином, у випадковому явищі – смертності або, навпаки, виживаності людей – було відкрито стійку закономірність: відносна частота для людей однієї статі й близького віку приблизно постійна. Це відкриття мало велику практичну цінність, зокрема лягло в основу сучасного страхування.

**Приклад**. Аналіз виживаності дає змогу проаналізувати неповні або цензуровані дані про виживаність хворих після операції. Однією з важливих характеристик є функція виживаності (імовірності того, що пацієнт проживе $t$ днів після операції (рис. 1.2).

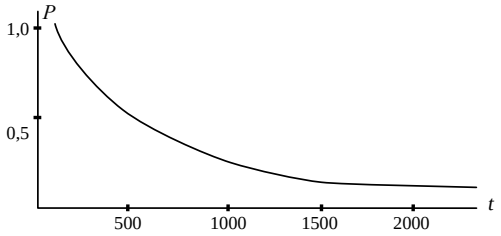



Рис. 1.2. Функція виживаності

Можна порівнювати функції виживаності в різних лікарнях, для різних вікових груп.

**Приклад**. У відомому фремінгхемському дослідженні, виконаному у США, було застосовано статистичний аналіз для оцінювання залежності ризику розвитку ішемічної хвороби серця від семи факторів: віку, кількості холестерину в крові, систолічного тиску, маси тіла, кількості гемоглобіну в крові, кількості викурюваних сигарет за день, показів електрокардіограми. У цьому дослідженні протягом 12 років (так зване лонгітюдне дослідження, тобто тривале дослідження тих самих осіб) було зібрано дані про прояви ішемічної хвороби в 1929 чоловіків і 2540 жінок у віці від 30 до 62 років. На початку обстеження всі пацієнти були здорові. Проведений аналіз дозволив вивчити вплив факторів ризику на розвиток ішемічної хвороби серця.

Уявімо, що знайшовся критик здобутих результатів (нового методу лікування, нових ліків, нового приладу, нового методу діагностування), який зауважує, що ці результати мають виключно випадковий характер і твердження про їхню ефективність сумнівне. У такому разі для доведення вірогідності отриманих результатів користуються ймовірнісними підходами й відповідними *статистичними критеріями.* Ці критерії дають змогу оцінити *силу зв'язку* між досліджуваними змінними. З метою вивчення залежності між цими досліджуваними ознаками використовуються *графіки взаємодії* (візуальний метод). Якщо прямі на графіку перетинаються, то говорять, що ознаки взаємодіють, впливають одна на одну. Якщо прямі паралельні, то говорять, що взаємодії, або залежності між ознаками, немає. Принагідно зазначимо, що в цілому візуальний підхід менш точний, ніж статистичні критерії.

**Приклад**. У галузі засобів телекомунікацій, що бурхливо розвивається, важливо розв'язувати такі задачі: прогнозувати пікові навантаження в мережі, оцінювати тижневі коливання

навантаження, раціонально вибирати місце будівництва нової станції для ефективного розвитку мережі.

Усі ці задачі успішно розв'язуються на основі *регресійних моделей.*

**Приклад**. Цікаві результати для прогнозування доходів телевізійних компаній залежно, приміром, від трьох факторів: кількості продаваних телевізорів, загальної кількості рекламних оголошень і урядових заходів, що обмежують деяку рекламу (наприклад, рекламу сигарет), можна також дістати за допомогою регресійних моделей.

**Приклад**. Успішні дії систем торгівлі «on-line» вимагають від фірм прогнозування поводження індивідуальних покупців. Найбільші фірми, займаючись електронною комерцією, зазнають щорічно величезних збитків через те, що 5-10 % покупців змінюють фірму або переходять у пасивний стан. Системи реєстрації електронної торгівлі дають змогу фіксувати моменти приходу кожного покупця до магазину, суму покупки, кількість товарів та інші параметри. За допомогою системно проведеного статистичного аналізу (описового аналізу даних, складання кореляційної матриці продажу, побудови графіків варіабельності покупок залежно від днів тижня, аналізу споживчого кошика для різних категорій покупців, днів тижня) можна оцінити періоди між покупками й змінити стратегію впливу на покупця, установивши, наприклад, терміни проведення більш активної рекламної кампанії. Знаючи ймовірнісні розподіли, можна легко спрогнозувати кількість покупців у певні проміжки часу.

Ми вжили слово «регресія», що як статистичне поняття має в аналізі даних важливе значення.

*Регресія* (скорочене від «регресія до середнього») – це *тенденція крайніх або незвичайних за значенням параметрів повертатися* (*регресувати*) *до середнього значення.*

Середні результати більш типові, ніж крайні. Так, після настання незвичайної події ситуація схильна вертатися до свого середнього рівня: за екстраординарними випадками, як правило, спостерігаються звичайні явища.

**Приклад**. При дослідженні залежності струму, що протікає через діод, від напруги, яку до нього прикладено, було отримано результати, які включають випадкову похибку вимірювання. Отриману сукупність даних можна апроксимувати деякою залежністю, яка називається вольт-амперною характеристикою діода й показує зміну струму, який протікає через діод, у середньому.

*Важливо запам'ятати, що коли мінливе поводження повертається до нормального стану, найімовірніше спостерігається регрес.* «*Як тільки ви навчитеся розпізнавати регресії, ви будете зауважувати їх усюди*» (*Деніел Канеман*).

У розглянутих прикладах насамперед ішлося про визначення наявності залежностей між факторами в середньому. Другий етап – визначення виду цих залежностей, за якими можна розв'язати задачу прогнозування.

• ***Задача аналізу впливу різних факторів*** на поводження досліджуваної величини (розглядається в *дисперсійному* аналізі).

**Приклад**. Для одержання сертифіката на діагностичний комплекс проводилися сертифікаційні випробування в різних лабораторіях за допомогою різного устаткування різними фахівцями й було отримано різні результати. Потрібно вирішити: чи є ця розбіжність випадковою, тобто зумовленою малою кількістю випробувань і неможливістю виключення впливу випадкових величин, чи причиною є недосконаліть устаткування, або різна кваліфікація фахівців, або методика проведення випробувань, або вплив інших

факторів. Для правильності ухвалення рішення про видачу сертифіката і застосовується дисперсійний аналіз, який дає змогу виявити причину розбіжностей результатів.

Зазначимо, що багато складних задач успішно розв'язуються саме доволі простими статистичними методами. Так, у медицині, зокрема фармакології, здійснюється оцінювання ефективності ліків, класифікація хворих за степеню важкості захворювання, дослідження кардіограм, різні тести, що дають змогу діагностувати пацієнтів на ранньому етапі захворювання, – усе це шлях до доказової медицини, і він лежить через статистичні методи. Так само можна розв'язувати задачі оцінювання технічного стану транспортних засобів, розрахунку податкових пільг для здійснення інвестицій, класифікації об'єктів незавершеного будівництва, класифікації джерел викидів забруднювальних речовин і багато інших, де досі застосовуються емпіричні правила.

Як показує практика, вартість обробки результатів експериментів становить незначну частину вартості експерименту в цілому, але може значно підвищити цінність здобутих результатів. Однак найчастіше цьому питанню не приділяють належної уваги, більше того, нерідко результати громіздких дорогих експериментів не піддаються навіть найпростішій обробці, через що втрачається величезна кількість корисної інформації, а іноді робляться неправильні висновки.

*Необхідно володіти як культурою постановки й проведення експерименту, так і культурою аналізу даних, тобто вмінням обробляти результати досліджень для одержання максимально можливої кількості корисної інформації.*

Таким чином, статистичні методи обробки інформації дають змогу компактно описати дані, зрозуміти їхню структуру, провести класифікацію, побачити закономірності в хаосі випадкових явищ.

**Запитання для самоперевірки**

1. Що є основною передумовою теорії ймовірностей? Яку роль відіграють методи математичної статистики?
2. У чому полягає суть описової статистики? Наведіть приклади словесного, графічного та математичного описів.
3. У чому полягає задача оцінювання невідомих параметрів законів розподілу ймовірностей?
4. Яке практичне значення має випадкове розсіювання параметрів? У чому полягає задача перевірки гіпотез?
5. Які задачі розв'язує теорія кореляції та регресійний аналіз?
6. Сформулюйте задачу прогнозування.
7. За допомогою чого оцінюється сила зв'язку між досліджуваними змінними? Що таке графік взаємодії?
8. Для чого застосовується дисперсійний аналіз?
9. Наведіть приклади застосування статистичних методів.
10. Сформулюйте головну задачу статистичної обробки.

## 1.2. ОСНОВНІ ПОНЯТТЯ ТЕОРІЇ ЙМОВІРНОСТЕЙ

У теорії ймовірностей вивчаються тільки масові випробування, тобто випробування, що відбуваються за незмінних умов неодноразово

Розглянемо основні поняття та визначення теорії ймовірностей як основи для методів статистичної обробки даних.

***Випробування.*** *Випробуванням* називається процес відтворення комплексу умов проведення експериментальних досліджень, котрий можна здійснити як завгодно велику кількість разів.

Це означення передбачає, що можна нескінченну кількість разів повторювати досліди, і умови їх проведення будуть

незмінними. Це можуть бути найрізноманітніші явища, в яких ті самі умови реалізуються багаторазово.

**Приклад**. Контроль при масовому виробництві придатності виробів, визначення розміру виробу.

***Події.*** Явища, що відбуваються в результаті випробувань, називаються *подіями*. Події позначаються великими буквами $A$, $B$, $C$. На рис. 1.3, *а* випадкову подію $A$ подано як потрапляння точки в деяку область $A$, що є частиною квадрата з площею, яка дорівнює одиниці.

**Приклад**. Якщо контроль придатності виробу – це випробування, то поява бракованого виробу при контролі – результат цього випробування, тобто подія; якщо визначення розміру виробу – це випробування, то одержання результату вимірювання є подією.

**Приклад**. При проведенні випробування – підкиданні монети – можливі два результати: герб або цифра.

***Достовірна подія*** – подія, яка в результаті такого випробування неодмінно відбувається.

**Приклад**. Поява бракованого примірника в партії непридатних виробів.

**Приклад**. Поява числа від 1 до 6 при підкиданні грального шестигранного кубика.

***Неможлива подія*** – подія, яка в результаті такого випробування відбутися не може.

**Приклад**. Поява придатного примірника в партії непридатних виробів.

**Приклад**. Поява числа 7 при підкиданні грального кубика.

***Випадкова подія*** – подія, яка в результаті такого випробування може відбутися або не відбутися.

**Приклад**. Поява бракованого примірника в партії виробів, виготовлених при несталому технологічному процесі.

***Несумісні події***. Дві події називаються *несумісними,* якщо при випробуванні поява однієї з них виключає можливість появи іншої.

**Приклад**. Поява герба при підкиданні монети виключає можливість появи цифри.

***Протилежні події.*** Дві події називаються *протилежними* (або взаємно протилежними), якщо поява однієї з них рівносильна непояві іншої. Якщо одну з цих подій позначено через $A$, то протилежну позначають $\overline{A}$ (рис. 1.3, *б*).

***Сумісні події***. Дві події називаються *сумісними*, якщо при випробуванні поява однієї з них не виключає можливості появи іншої.

**Приклад.** При підкиданні грального кубика можливе одночасне випадання парного числа й числа 2.

***Рівноможливі події.*** Якщо при випробуванні можуть з'явитися кілька можливих подій і при цьому немає підстав припускати, що поява одних імовірніша за появу інших, то такі події називаються *рівноможливими***.**

**Приклад**. Рівноможливе випадання герба та цифри при підкиданні монети.

У загальному випадку в результаті випробування залежно від випадкових обставин, що змінюються, може відбутися та або інша подія з множини подій, можливих при даному випробуванні. Така множина називається ***полем подій*.**

Поле подій містить насамперед рівноможливі події. Ці події називаються ***елементарними*.**

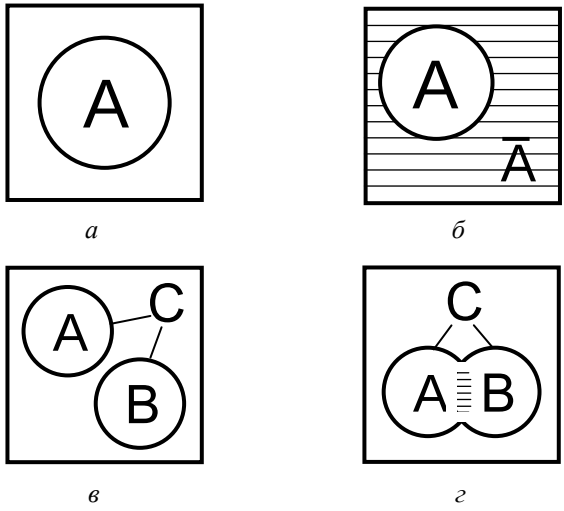
***Незалежні події.*** Події $A$ і $B$ називаються *незалежними*, якщо множини, якими вони подаються, *не перетинаються* (рис. 1.3, *в*), тобто вони не мають спільних точок. Тому ніяка точка, що зображає одну з елементарних подій, не може потрапити одночасно в ці дві області, тобто при повторенні випробування кількість точок, що

потрапляють в область $A + B$, дорівнює сумі кількості точок, що потрапляють в області $A$ або $B$.

**Приклад**. При підкиданні монети не можуть випасти герб і цифра одночасно.

**Приклад**. Парні та непарні цифри грального кубика не можуть випасти одночасно.

***Залежні події***. Події $A$ і $B$ називаються *залежними*, якщо множини, якими вони подаються, *перетинаються*, тобто події частково або повністю збігаються. На рис. 1.3, *г* ілюструється поняття суми залежних подій. Її межу обведено жирною лінією. Кількість точок, які потрапили в область $A + B$, буде меншою, ніж сума точок, що потрапили в області $A$ і $B$ за рахунок можливості потрапляння в їхню спільну частину (область перерізу).




*а* *б* *в* *г*

Рис. 1.3. До поняття суми незалежних та залежних подій

*Для запису суми використовується функція диз'юнкція* ∪ (*логічна операція АБО*)
*Для запису добутку використовується функція кон'юнкція* ∩ (*логічна операція І*).

**Приклад**. У грального кубика випало «непарне число» і «3».

Таким чином, існує деякий взаємний зв'язок між подіями, який подається за допомогою алгебри подій.

**Сумою (об'єднанням)** двох подій $A$ і $B$ називається подія $A + B$, яка настає тоді, коли настає принаймні одна з подій $A$ або $B$.

Сума $n$ подій:

$$\sum_{i=1}^{n} A_i = \left(A_1 + A_2 + ... + A_n\right).$$

**Приклад.** Для трьох подій

$$A_i = A_1 \bigcup A_2 \bigcup A_3 .$$

**Добутком (перетином)** двох подій $A$ і $B$ називається подія $AB$, яка настає тоді, коли настають обидві події $A$ і $B$.

Добуток $n$ подій

$$\prod_{i=1}^{n} A_i = \left(A_1 \cdot A_2 \cdot ... \cdot A_n\right).$$

**Приклад**. Для трьох подій

$$A_i = A_1 \bigcap A_2 \bigcap A_3 .$$

**Відносна частота події.** Відносною частотою (частістю) випадкової події $A$ називається відношення кількості $n_A$ появ цієї події до загальної кількості виконаних випробувань $n$

$$W_n(A) = n_A/n.$$

Частість $n_A/n$ – випадкова величина, її значення залежить від випадкових обставин, що супроводжують випробування, однак у більшості випадків, у довгих повторних серіях, частість має *статистичну стійкість*, тобто зі збільшенням $n$ вона все менше відхиляється від *деякого постійного числа*.

**Приклад**. При багаторазовому підкиданні правильного грального кубика відносна частота випадання кожного числа очок від 1 до 6 коливається біля числа 1/6.

***Стійкість*** відносної частоти відбиває деяку об'єктивну властивість випадкової події, що полягає певною мірою в *можливості її настання.*

Мірою (числовим значенням) об'єктивної можливості настання випадкової події $A$ є її ***ймовірність***, яка позначається символом $P(A)$.

*Конкретний зміст імовірності полягає в тому, що вона визначає середню частість, з якою можна чекати появи події A в довгих серіях випробувань.*

Завдяки стійкості й близькості відносної частоти $W_n(A)$ до ймовірності $P(A)$ *частість* може слугувати *наближеною оцінкою ймовірності*, точність якої зростає зі зростанням кількості випробувань у серії.

**Основні властивості ймовірностей**

**1**. ***Нормування ймовірності.***

З одного боку, кількість $n_A$ поява будь-якої випадкової події $A$ не може бути від'ємною ($n_A > 0$), а з другого боку, якщо випробування повторюються $n$ разів, подія $A$ не може відбутися більш ніж $n$ разів ($n_A \le n$). Тому відносна частота випадкової події завжди задовольняє нерівності:

$$0 \le \frac{n_A}{n} \le 1,$$

а отже, і для ймовірності, яка є граничним значенням для $W_n(A)$, буде виконуватися співвідношення, що нормує її значення: імовірність випадкової події міститься між 0 і 1:

$$0 \le P\left(A\right) \le 1.$$

**2**. ***Імовірність достовірної події дорівнює*** **одиниці**, ***імовірність неможливої події дорівнює*** **нулю**:

$$P(U) = 1,\ P(V) = 0.$$

Як наслідок, сума ймовірностей протилежних подій дорівнює одиниці:

$$P(A) + P(\overline{A}) = 1.$$

На рис. 1.3, *б* цьому випадку відповідає квадрат з нормованою одиничною площею.

**Правило додавання ймовірностей (**властивість адитивності ймовірностей).

Розглянемо спочатку приклад для *несумісних подій* (рис. 1.3, *в*).

**Приклад**. В урні лежать червоні, сині та білі кулі. Подія $A$ полягає у вийманні червоної кулі, подія $B$ – синьої. Подія $C = A + B$ полягає у вийманні будь-якої кольорової кулі. Вийнявши кулю з урни, відзначимо її колір, а потім повернемо кулю знову в урну, щоб повторити випробування в тих самих умовах (за схемою Бернуллі). Тоді при $n$-разовому повторенні випробування кількість вийнятих кольорових куль дорівнює сумі вийнятих червоних і синіх куль:

$$n_C = n_A + n_B,$$

тому для відносних частот завжди виконується співвідношення:

$$n_C/n = n_A/n + n_B/n,$$

що за $n \to \infty$ буде відповідати сумі ймовірностей.

Таким чином, якщо події *несумісні*, імовірність суми цих подій дорівнює сумі їхніх імовірностей:

$$P(A + B) = P(A) + P(B).$$

Правило додавання ймовірностей для двох несумісних подій поширюється на суму будь-якої скінченної кількості випадкових подій: якщо випадкові події $A_1$, $A_2$,…, $A_k$ попарно несумісні, то ймовірність їхньої суми, тобто ймовірність появи хоча б однієї із цих подій, дорівнює сумі їхніх імовірностей:

$$P(A_1 + A_2 + \ldots + A_k) = P(A_1) + P(A_2) + \ldots + P(A_k).$$

Для випадку *сумісних подій*. Як видно з рис. 1.3, *г*, де обведена область відповідає сумі сумісних подій, існує деяка заштрихована область, яка належить одночасно й області події $A$ і області події $B$.

Площа області $C$ буде пропорційною до ймовірності суми подій $A$ і $B$:

$$P(A + B) = P(A) + P(B) – P(AB),$$

де $P(AB)$ – імовірність сумісності подій.

**Приклад.** Два стрільці одночасно стріляють по одній цілі. Подія $C$ – хоча б одне влучення в ціль – дорівнює сумі двох подій $A$ і $B$: $A$ – влучення першого стрільця в ціль, $B$ – влучення другого стрільця в ціль. Цілком можливо, що обидва стрільці влучили в ціль одночасно. Нехай $P(A) = 0{,}8$ і $P(B) = 0{,}9$, тоді, якщо б події були несумісними, то $P(C) = P(A) + P(B) = 1{,}7$, що не відповідає дійсності, бо, як відомо, імовірність не може бути більшою від одиниці. На рис. 2.1, *г* одиничній імовірності відповідає площа квадрата, у межах якого міститься область, що відповідає події $C$. Тоді за формулою для сумісних подій маємо:

$$P(C) = 0{,}8 + 0{,}9 – 0{,}8 \times 0{,}9 = 0{,}98.$$

**Імовірності в повній групі подій**. Із рис. 1.3, *а* випливає, що може відбутися або подія $A$, або подія $A$, й інших наслідків не може бути. У цьому випадку говорять, що події $A$ і $\overline{A}$ утворюють *повну групу подій*, тобто при кожному повторенні випробування може відбутися *хоча б одна з цих подій*.

**Приклад**. Якщо $X$ – кількість очок, що випадає на верхній грані грального кубика, то події $X = 1$, $X = 2$, $X = 3$, $X = 4$, $X = 5$, $X = 6$ утворюють повну групу подій.

**Приклад.** $X$ – парне, $Y$ – непарне утворять повну групу.

Із цих прикладів бачимо, що елементарні події, пов’язані з даним випробуванням, можуть у різний спосіб формувати повні групи подій.

*Сума ймовірностей несумісних подій, що утворюють повну групу, дорівнює одиниці*

$$P(A_1 + A_2 + \ldots + A_k) = P(A_1) + P(A_2) + \ldots + P(A_k) = 1.$$

Здобуте співвідношення часто застосовують для контролю розрахунку ймовірностей.

**Запитання для самоперевірки**

1. Що називається випробуванням? Наведіть приклади.
2. Що таке подія, достовірна та неможлива події? Наведіть приклади.
3. Що називається полем подій? Які події називаються елементарними, рівноможливими, протилежними? Наведіть приклади
4. Які події називаються сумісними і несумісними? Наведіть приклади. Що називається сумою та добутком двох подій?
5. Поясніть, що називається відносною частотою події.
6. У чому полягає конкретний зміст імовірності? Наведіть основні властивості ймовірностей.
7. Наведіть правило додавання ймовірностей для сумісних та несумісних подій.
8. Що таке повна група подій? Чому дорівнює ймовірність у повній групі подій? Чому дорівнює сума ймовірностей протилежних подій?

## 1.3. УМОВНІ ЙМОВІРНОСТІ. ФОРМУЛА БАЄСА

Іноді постає необхідність обчислити ймовірність деякої події В за додаткової умови, яка полягає в тому, що деяка подія А вже відбулася.

**Умовні ймовірності.** Випадкову величину можна визначити як подію, що може відбутися або не відбутися при здійсненні деякої сукупності умов.

Якщо при обчисленні ймовірності події крім зазначеної сукупності умов ніяких інших обмежень не накладається, то така ймовірність називається *безумовною*.

Якщо для здійснення події необхідне виконання деяких додаткових умов, що не входять у зазначену сукупність, то ймовірність виникнення такої події називають *умовною.* Імовірність події $B$, обчислену за припущення, що подія $A$ вже відбулася, позначають $P(B|A)$.

Для з'ясування суті умовної ймовірності насамперед розглянемо поняття *поєднання випадкових подій.* Під поєднанням випадкових подій $A$ і $B$ розуміють випадкову подію, яка полягає в тому, що в результаті випробувань одночасно відбудеться і подія $A$, і подія $B$. Це поєднання позначається через $AB$ (добуток випадкових подій).

**Приклад**. В ящику є $n$ куль, з них $n_A$ – куль кольорових, серед кольорових $n_B$ –куль із зірочкою. Навмання виймається куля і повертається в ящик (для забезпечення однаковості умов проведення дослідів). Уведемо позначення: подія $A$ – виймання кольорової кулі; подія $B$ – виймання кулі із зірочкою, причому подія $B$ не може відбутися без події $A$. Яка ймовірність того, що вийнята куля буде кольоровою і водночас із зірочкою? У результаті проведення $n$ випробувань можна одержати $n_{AB}$ наслідків, що нас цікавлять, тобто відбувається *поєднання подій* (куля і кольорова, і позначена зірочкою). Частота позитивного результату $n_{AB} / n$. Помножимо й поділимо це співвідношення на $n_A$:

$$W_n(AB) = \frac{n_{AB}}{n} \cdot \frac{n_A}{n_A},$$

або

$$W_n(AB) = \frac{n_A}{n} \cdot \frac{n_{AB}}{n_A}. \qquad (1.1)$$

Перший множник у виразі (1.1) говорить про частоту появи події, що витягли кольорову кулю. Другий – про частоту того, що витягли не просто кольорову, а кулю із зірочкою, коли витягали

кольорову, тобто припускається *умова*, що із зірочкою може бути тільки кольорова куля.

Якщо перейти до понять теорії ймовірностей, то перший множник у (1.1) відповідає *безумовній імовірності події A*, а другий – *умовній імовірності* того, що настала подія *B* за умови, що відбулася подія *A*.

Для нескінченної кількості випробувань у граничному випадку $W_n(AB)$ збігається до ймовірності $P(AB)$, тоді:

$$P(A \cdot B) = P(A)P(B \mid A), \qquad (1.2)$$

звідки умовна ймовірність:

$$P(B \mid A) = \frac{P(AB)}{P(A)}.$$

*Умовна ймовірність має певний сенс лише в тих випадках, коли ймовірність умови не дорівнює нулю.*

Аналогічно визначається умовна ймовірність події *A* за умови здійснення події *B*:

$$P(A \mid B) = \frac{P(AB)}{P(B)}.$$

Якщо події несумісні, тобто немає їх перерізу, то $P(B|A) = 0$, а отже, і $P(AB) = 0$.

Вираз (1.2) являє собою *правило множення ймовірностей*: імовірність спільної появи двох подій дорівнює добутку ймовірності однієї з них на умовну ймовірність іншої, обчислену за припущення, що перша подія вже відбулася.

Це правило можна поширити й на більшу кількість подій: імовірність спільної появи кількох подій дорівнює добутку ймовірності однієї з них на умовні ймовірності всіх інших, причому ймовірність кожної наступної події обчислюється за припущення, що всі попередні події вже відбулись.

**Приклад**. Для трьох подій:

$$P(ABC) = P(A)\,P(B|A)\,P(C|AB).$$

У загальному вигляді:

$$P(A_1A_2A_3\ldots A_n) = P(A_1)\,P(A_2|A_1)\,P(A_3|A_1A_2)\,\ldots\,P(A_n|A_1A_2\ldots A_{n-1}),$$

де $P(A_n|A_1A_2\ldots A_{n-1})$ – імовірність події $A_n$, обчислена за припущення, що події $A_1, A_2, \ldots A_{n-1}$ відбулися.

**Незалежні події**. **Правило множення ймовірностей незалежних подій**. Поняття умовної ймовірності використовують для встановлення залежності подій. Подію $B$ називають *незалежною* від події $A$, якщо поява події $A$ не змінює ймовірності події $B$. Інакше кажучи, умовна ймовірність події $B$ дорівнює її безумовній імовірності:

$$P(B|A) = P(B).$$

Із цього випливає, що ймовірність *спільної* появи кількох подій, *незалежних* у сукупності, дорівнює добутку ймовірностей цих подій:

$$P(A_1A_2\ldots A_n) = P(A_1)P(A_2)\ldots P(A_n). \qquad (1.3)$$

**Повна ймовірність.** Імовірність події $A$, що може відбутися за умови появи хоча б однієї з несумісних подій $B_1, B_2, \ldots, B_n$, які утворюють повну групу, дорівнює сумі добутків імовірностей кожної із цих подій на відповідну умовну ймовірність події $A$:

$$P(A) = P(B_1)\,P(A|B_1) + P(B_2)\,P(A|B_2) + \ldots + P(B_n)P(A|B_n). \qquad (1.4)$$

**Приклад**. Є три урни першого типу, в яких дві білі та шість чорних куль, і одна урна другого типу: у ній одна біла й вісім чорних куль. Нехай навмання вибирається одна урна із чотирьох, а звідти виймається куля. Визначити ймовірність того, що буде витягнута біла куля. Уведемо позначення:

- подія $A$ – «вийнята куля – біла», її ймовірність $P(A)$; подія $A$ може настати при вилученні кулі як з урни першого типу, так і з урни другого типу;
- подія $B_1$ – «обрана урна першого типу», $P(B_1) = 3/4$;

- подія $B_2$ – «обрана урна другого типу», $P(B_2) = 1/4$;
- $P(A|B_1) = 1/4$ – ймовірність того, що вийнято білу кулю з урни першого типу;
- $P(A|B_2) = 1/9$ – імовірність того, що вийнято білу кулю з урни другого типу.

Тоді за формулою (1.4) повна ймовірність:

$$P(A) = 3/4 \cdot 1/4 + 1/4 \cdot 1/9 = 0{,}22.$$

**Імовірність гіпотез. Формула Баєса**. Припустимо, що подія $A$ може настати за умови появи однієї з несумісних подій $B_1$, $B_2$ ,…, $B_n$, що утворюють повну групу. Як відомо, імовірність появи події $A$ визначається за формулою повної ймовірності (1.4). Якщо тепер припустити (висунути гіпотезу), що подія $A$ вже відбулася, то можна поставити завдання з'ясувати, яка з умов $B_i$ могла привести до події $A$. Послідовно висуваються гіпотези про те, що відбулася подія $B_i$, яка привела до події $A$. Можливість прояву цих гіпотез визначається умовними ймовірностями:

$$P(B_1|A),\ P(B_2|A),\ \ldots,\ P(B_n|A).$$

Розглянемо можливість появи події $A$ за умови, що відбулась подія $B_1$. За формулою множення ймовірностей знайдемо ймовірність спільної появи подій $A$ та $B_1$:

$$P(AB_1) = P(A)\,P(B_1|A).$$

Оскільки події сумісні, справджуватиметься рівність:

$$P(AB_1) = P(B_1)\,P(A|B_1),$$

звідки

$$P(B_1 \mid A) = \frac{P(B_1)P(A \mid B_1)}{P(A)}.$$

Підставляючи замість імовірності $P(A)$ її значення згідно з формулою повної ймовірності (1.4), маємо:

$$P(B_1 \mid A) = \frac{P(B_1)P(A \mid B_1)}{P(B_1)P(A \mid B_1) + P(B_2)P(A \mid B_2) + \ldots + P(B_n)P(A \mid B_n)}.$$

У загальному випадку умовні ймовірності інших гіпотез $B_i$ $(i=\overline{2,n})$ за умови, що подія $A$ відбулася, можна знайти з формули:

$$P\left(B_i \mid A\right)=\frac{P(B_i)P_i(A \mid B_i)}{P(B_1)P(A \mid B_1)+P(B_2)P(A \mid B_2)+...+P(B_n)P(A \mid B_n)}. \quad (1.5)$$

Вираз (1.5) називається *формулою Баєса*, що показує, яка частка події $B_i$ у можливості появи події $A$ відносно всіх можливих джерел, котрі зумовлюють подію $A$.

Нехай апріорно відомо, що існує кілька джерел – причин $B_1,\ B_2,...,B_m,$ які сприяють появі події $A$. Після проведення випробувань і отримання результату можна визначити, яке із джерел $B_i$ найбільшою мірою в розглянутому випадку може бути причиною появи події $A$. Для цього висувається гіпотеза, що ця подія $A$ відбулася з «провини» деякої події $B_i$, і ймовірність цього визначається за формулою (1.5).

**Приклад**. За умови попереднього прикладу необхідно визначити, яка ймовірність того, що білу кулю вийнято з урни першого типу. За формулою (1.5) маємо:

$$P(B_1 \mid A)=\frac{P(B_1)P(A \mid B_1)}{P(B_1)P(A \mid B_1)+P(B_2)P(A \mid B_2)}\ \frac{3/4\cdot1/4}{3/4\cdot1/4+1/4\cdot1/9}=0{,}87.$$

### Запитання для самоперевірки

1. Що називається умовною і безумовною ймовірностями випадкових подій? Наведіть приклади.
2. Що розуміють під поєднанням випадкових подій? Наведіть приклади.
3. Сформулюйте правило множення ймовірностей.
4. Що таке повна ймовірність? Наведіть приклади.
5. Розкрийте зміст формули Баєса. Що розуміють під імовірністю гіпотез?

# 1.4. ДИСКРЕТНА ВИПАДКОВА ВЕЛИЧИНА. ХАРАКТЕРИСТИКИ ПОЛОЖЕННЯ ТА РОЗСІЮВАННЯ

Для кількісної характеристики, отриманої при експерименті інформації, використовується поняття випадкової величини.

**Випадковою** називається величина, яка в результаті досліду може набувати заздалегідь різних невідомих значень.

Випадкові величини зазвичай позначаються великими латинськими буквами, наприклад випадкова величина $X$ (або $Y$). Значення, яких випадкова величина набуває в результаті досліду, позначаються малими латинськими буквами, наприклад $x_1, x_2, \ldots, x_n$, і називаються *реалізаціями* випадкової величини.

Випадкові величини бувають *дискретними та неперервними*.

**Дискретною** величиною $X$ називається випадкова величина, якщо множина її можливих значень відповідає скінченному або нескінченному ряду чисел, і кожному значенню відповідає певна ймовірність $p_i$.

Для того щоб одержати мінімальну інформацію про поводження випадкової величини, на практиці достатньо знати:

а) деяке *середнє* значення величини, навколо якого групуються значення випадкової величини – *центр групування* (*розсіювання*);

б) як і наскільки розкидані ці значення біля центра групування, тобто точно охарактеризувати за допомогою числового показника *степінь розсіювання*.

Визначати *центр групування* та *числову характеристику розсіювання* можна різними способами. Один із найбільш зручних способів – обчислення *моментів розподілу випадкової величини.*

**Математичне сподівання *MX*.**

Припустимо, що проведено $n$ випробувань і здобуто значення: $x_1 \to n_1;\ x_2 \to n_2\ \ldots\ x_k \to n_k$, де $n_i$ – кількість результатів $x_i$. При цьому:

$$n_1 + n_2 + \ldots + n_k = n.$$

Визначимо середнє значення результату:

$$\overline{x} = \frac{n_1 x_1 + n_2 x_2 + \ldots + n_k x_k}{n} = W_1 x_1 + W_2 x_2 + \ldots + W_k x_k = \sum_{i=1}^{k} W_i x_i ,$$

де $W_i$ – частота появи значень $x_i$.

Звідси випливає, що *середнє значення залежить як від значень випадкової величини $x_i$, так і від частоти $W_i$ їх появи*.

Збільшуючи $n$, у граничному випадку прийдемо до ймовірності появи того чи іншого значення:

$$\lim_{n\to\infty} W_i = p_i .$$

Тоді для середнього значення прийдемо до математичного сподівання, що визначається як сума:

$$MX = \sum_i x_i p_i . \qquad (1.6)$$

**Початкові моменти**. Математичне сподівання випадкової величини $X^k$ називається ***початковим моментом k-го порядку:***

$$\nu_k = MX^k = \sum_i x_i^k p_i . \qquad (1.7)$$

При $k$ = 1 прийдемо до виразу (1.6), тобто *математичне сподівання є першим початковим моментом.*

**Властивості математичного сподівання**

1. Математичне сподівання постійної величини дорівнює самій постійній величині:

$$M(C) = C.$$

2. Математичне сподівання добутку постійної величини на випадкову величину дорівнює добутку постійної величини на математичне сподівання цієї випадкової величини:

$$M(CX) = CMX.$$

3. Математичне сподівання суми постійної та випадкової величин дорівнює сумі постійної величини та математичного сподівання випадкової величини:

$$M(C + X) = C + MX.$$

4. Математичне сподівання добутку двох незалежних випадкових величин дорівнює добутку їхніх математичних сподівань:

$$M(X_1X_2) = (MX_1)\,(MX_2).$$

**Центральні моменти.** При статистичній обробці й вивченні закономірностей окрім математичного сподівання використовуються моменти більш високих порядків. Початкові моменти є незручними для практичної діяльності, і їх найчастіше використовують для обчислення так званих центральних моментів, які позначаються $\mu_k$.

*Початковий момент характеризує випадкову величину відносно початку координат* (*нуля*), *а центральний момент – відносно математичного сподівання MX* (*центра групування*)

Відомо, що для опису випадкової величини недостатньо вказати тільки її центр групування, але й потрібно знати степінь її розсіювання навколо цього центра.

Числовою характеристикою степені розсіювання є **дисперсія.**

Розсіювання випадкової величини *X* характеризує відхилення цієї величини від центра її групування *MX*. Уведемо випадкову

величину $X'$, яка характеризуватиме зміну випадкової величини відносно центра групування, який дорівнює нулю:

$$X' = X - MX,$$

тобто для $X'$ центр координат перенесено в точку *MX,* як показано на рис. 1.4.

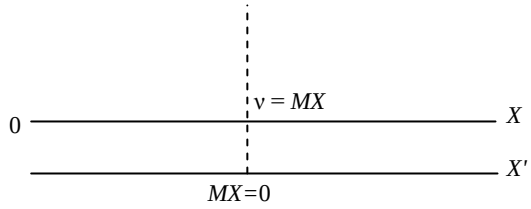



Рис. 1.4. Центрована випадкова величина

Моменти випадкової величини $X'$ називаються *центральними моментами* величини *X*, або моментами *X* щодо центра розподілу ν.

Найбільше поширення знайшов центральний момент другого порядку $\mu_2$, який є математичним сподіванням квадрата відхилення величини *X* і називається *дисперсією.*

Таким чином, за означенням маємо:

$$\mu_2 = M[X']^2 = M[X - MX]^2 = M(X^2 - 2XMX + (MX)^2) =$$
$$= MX^2 - 2(MX)^2 + (MX)^2 = MX^2 - (MX)^2;$$

або

$$\mu_2 = \nu_2 - \nu_1^2.$$

*Центральний момент другого порядку* (*дисперсія*) *дорівнює початковому моменту другого порядку мінус квадрат початкового моменту першого порядку.*

Для дискретної центрованої величини $X' = X - MX$ на підставі (1.7) можна записати:

$$\mu_2 = \sum_i \left(X_i'\right)^2 p_i = \sum_i \left(X_i - MX\right)^2 p_i \,.$$

Таким чином, для обчислення дисперсії потрібно квадрат відхилення кожного значення помножити на ймовірність цього значення і всі добутки такого виду додати.

**Основні властивості дисперсії**

1. Дисперсія постійної величини дорівнює нулю:

$$DC = 0.$$

2. Дисперсія суми постійної та випадкової величин дорівнює дисперсії випадкової величини (тобто не залежить від розміщення центра розсіювання):

$$D(X + C) = DX.$$

3. Дисперсія добутку постійної величини на випадкову величину дорівнює добутку квадрата постійної величини на дисперсію випадкової величини:

$$D(CX) = C^2 DX.$$

4. Дисперсія суми двох незалежних випадкових величин дорівнює сумі дисперсій цих величин:

$$D(X_1 + X_2) = DX_1 + DX_2.$$

Дисперсія рідше трапляється в обробці експериментальних даних через її незручну розмірність. Із цією метою використовується **середнє квадратичне відхилення** (*СКВ, розкид, стандартне відхилення*) – додатний корінь квадратний з дисперсії:

$$\sigma_X = \sqrt{DX} \,.$$

У деяких випадках відхилення випадкової величини характеризується у відносних одиницях у вигляді **коефіцієнта варіації:**

$$\gamma = \frac{\sigma_X}{MX} \cdot 100\% \,,$$

який показує як співвідноситься СКВ і математичне сподівання.

**Запитання для самоперевірки**

1. Яка величина називається випадковою, дискретною випадковою величиною?
2. Для чого використовуються моменти розподілу випадкової величини? Які вам відомі початкові моменти?
3. Які числові характеристики випадкової величини вам відомі?
4. Назвіть властивості математичного сподівання.
5. Що характеризують центральні моменти?
6. Чим початкові моменти відрізняються від центральних?
7. Охарактеризуйте поняття дисперсії. Перелічіть основні властивості дисперсії.
8. Що таке середнє квадратичне відхилення? Як воно обчислюється й що характеризує?
9. Для чого використовується коефіцієнт варіації?

## 1.5. РОЗПОДІЛИ ДИСКРЕТНИХ ВЕЛИЧИН

Якщо відомі всі можливі значення $x_1$, $x_2$,…, $x_n$, яких набула дискретна випадкова величина $X$, і ймовірності $p(x_i)$ для кожного значення, то розподіл цієї величини вважають теоретично заданим.

**Законом розподілу ймовірностей** дискретної випадкової величини $X$ будемо називати будь-яке правило, що дає змогу знаходити ймовірності:

$$P(X = x_i) = p_i \ (i = 1,2...).$$

Таким чином, закон розподілу задає ймовірність $p_i$ як *функцію*, визначену на множині подій $X = x_i$.

Цю функцію можна виразити формулою $p_i = f(x_i)$, тобто подати закон у вигляді аналітичної залежності, графіка, таблиці (якщо є скінченне значення, у таблиці навпроти значення ставиться

ймовірність його появи). Така таблиця називається *таблицею розподілу* (імовірностей) випадкової величини *X*.

*Варто розрізняти*: *розподіл імовірностей – теоретичний розподіл*, *розподіл частостей – емпіричний або статистичний розподіл*.

*P – символьне вираження ймовірності*;

*p – числове значення ймовірності*.

**Біноміальний закон розподілу**. Розглянемо випадок повторення того самого випробування при постійних умовах. Як елементарні події кожного випробування будемо розрізняти лише два результати: поява деякої події $A$ з імовірністю $p$ (позитивний результат – «успіх») або непоява його – $\overline{A}$ з імовірністю $q$ (негативний результат). Результат кожного випробування будемо відзначати, ставлячи букву $A$ або $\overline{A}$ на місці, що відповідає випробуванню.

**Приклад**. При двох випробуваннях можливі такі $2^2 = 4$ результати: $\overline{A}\ \overline{A}$, $\overline{A}A$, $A\overline{A}$, $AA$. При трьох випробуваннях можливі $2^3 = 8$ результатів, де кількість елементарних подій при випробуванні дорівнює 2.

Зважаючи на те, що випробування *незалежні*, ймовірність кожного такого результату визначається перемножуванням імовірностей подій $A$ і $\overline{A}$ у відповідних випробуваннях.

**Приклад**. Для восьми можливих результатів трьох випробувань будемо мати результати, наведені в табл. 1.1.

Таблиця 1.1. Можливі результати трьох випробувань

| Результати | $\overline{A}\ \overline{A}\ \overline{A}$ | $\overline{A}\ \overline{A}\ A$ | $\overline{A}\,A\,\overline{A}$ | $A\,\overline{A}\,\overline{A}$ | $AA\,\overline{A}$ | $A\,\overline{A}\,A$ | $\overline{A}\,AA$ | $AAA$ |
|---|---|---|---|---|---|---|---|---|
| Імовірності | $qqq = q^3$ | $qqp = q^2p$ | $qpq = q^2p$ | $pqq = q^2p$ | $ppq = p^2q$ | $pqp = p^2q$ | $qpp = p^2q$ | $ppp = p^3$ |

Із таблиці випливає, що непоява події $A$ у всіх трьох випробуваннях має ймовірність $p_3(m = 0) = q^3$, що відповідає тільки одному результату $\overline{A}\ \overline{A}\ \overline{A}$. Поява події $A$ один раз протягом трьох випробувань може бути, коли випадає один із трьох варіантів: $\overline{A}\ \overline{A}\ A$, або $\overline{A}\ A\ \overline{A}$, або $A\ \overline{A}\ \overline{A}$, що мають однакову ймовірність $q^2p$. Тому $p_3(1) = P(\overline{A}\ \overline{A}\ A) + P(\overline{A}\ A\ \overline{A}\ ) + P(A\ \overline{A}\ \overline{A}\ ) = 3\,q^2p$.

Як бачимо, імовірність появи події $A$ не залежить від того, в якому саме випробуванні (в якій послідовності здобутих результатів) настане ця подія. Головне – ця подія відбулася.

Оскільки розглянуто повну групу подій (розглянуто всі можливі варіанти результату), у сумі дістанемо:

$$q^3 + 3pq^2 + 3p^2q + p^3 = (q+p)^3 = 1.$$

Цю властивість можна використовувати для контролю правильності обчислень.

Якщо подивитися на останнє співвідношення, можна помітити, що показник степені при $p$ буде $m$, а при $q$ буде $n - m$, де $m$ – кількість позитивних подій $A$ за $n$ випробувань.

У загальному вигляді можна записати ймовірність $p_n(m)$ появи подій $A$ $m$ раз при $n$ випробуваннях:

$$p_n(m) = C_n^m q^{(n-m)} p^m = \frac{n!}{m!(n-m)!} \cdot q^{(n-m)} p^m, \qquad (1.8)$$

де $C_n^m$ – кількість комбінацій із $n$ елементів по $m$.

Після перетворення виразу (1.8) маємо:

$$p_n(m) = \frac{n(n-1)\ldots(n-m+1)}{1 \cdot 2 \ldots m} p^m q^{n-m}.$$

Останній вираз справджується і для крайніх значень $m = 0$ і $m = n$, якщо вважати, як це прийнято, що $0! = 1$ і $C_n^m = 1$. Цей результат відповідає значенням імовірностей, наведеним у табл. 1.1.

Розподіл імовірностей, заданих формулою (1.8) або таблицею, наведеного на с. 36, називається ***біноміальним законом розподілу ймовірностей.***

Ця назва пов'язана з тим, що ймовірності, обчислені за формулою (1.8), збігаються з відповідними членами розкладу бінома Ньютона:

$$(q+p)^n = q^n + npq^{n-1} + \ldots + C_n^m p^m q^{n-m} + \ldots + np^{n-1}q + p^n .$$

Із цього виразу також випливає, що сума всіх імовірностей дорівнює одиниці, зважаючи на те, що $p + q = 1$. Цього й слід очікувати, бо розглядається повна група подій, і один із можливих результатів $m = 0$, або $m = 1$, …, або $m = n$ неодмінно відбудеться.

**Приклад**. Імовірність народження хлопчика дорівнює 0,515. Яка ймовірність того, що з десяти навмання обраних немовлят буде шість хлопчиків?

Для шуканої ймовірності при $n = 10$, $m = 6$ маємо:

$$p(A) = C^6_{10}(0{,}515)^6(0{,}485)^4 = \frac{10!}{6!\cdot 4!}(0{,}515)^6(0{,}485)^4 = 0{,}2167.$$

Біноміальний розподіл зустрічається у випадках, коли відбуваються випробування з поверненням об'єкта, який було обрано на $i$-му $\left(i = \overline{1,n}\right)$ кроці, тобто передбачається, що на кожному кроці число $n$ незмінне. Якщо кількість об'єктів значна ($n > 1000$), то навіть порушення передумови про повернення об'єкта дає змогу з достатньою на практиці точністю використовувати біноміальний розподіл. У такому разі говорять про послідовність незалежних випробувань за схемою Бернуллі або за схемою повторної вибірки.

Таким чином, випадкова величина називається *біноміально розподіленою* з параметрами $n$ і $p$, якщо можливих значень 0,1,…, $n$ вона набуває з імовірностями $p(n, m)$, що задаються формулою (1.8).

Параметри $n$ і $p$ повністю визначають біноміальний розподіл. Якщо кількість $n$ спостережень не дуже велика, для обчислення

ймовірностей $p_n(k)$ можна використати просте співвідношення, що пов'язує два сусідні члени $p_n(m)$ і $p_n(m+1)$:

$$\frac{p_n(m+1)}{p_n(m)} = \frac{(n-m)}{(m+1)} \cdot \frac{p}{q} \ . \qquad (1.9)$$

Якщо відоме значення $p_n(m)$, то за допомогою (1.9) можна розрахувати $p_n(m+1)$. Для біноміального закону складено таблиці для різних значень $n$, $p$ і $m$.

Проаналізуємо вираз (1.9). При зростанні $m$ від 0 до $(n-1)$ відношення в правій його частині змінюється від $np/q$ до $p/q$. Якщо $np > q$ і $nq > p$, то це відношення переходить від значень, більших за одиницю, до значень, менших за одиницю. Якщо $np < q$ або $nq < p$, то ймовірності змінюються монотонно.

Приклади всіх трьох можливих випадків наведено на рис. 1.5.

У всіх випадках найбільш ймовірне значення $X = m_0$ визначається з нерівностей:

$$\frac{n-(m_0-1)}{m_0} \cdot \frac{p}{q} \geq 1; \quad \frac{(n-m_0)}{(m_0+1)} \cdot \frac{p}{q} \leq 1,$$

звідки випливає, що

$$np + p - 1 \leq m_0 \leq np + p \ . \qquad (1.10)$$

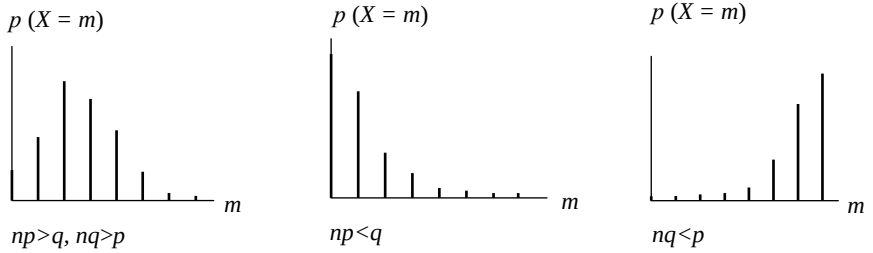



Рис. 1.5. Залежність ймовірності від співвідношення між $np$ та $q$

Для біноміального закону

$$MX = np; \qquad DX = npq; \qquad \sigma = \sqrt{npq} \ .$$

Часто в практичних задачах необхідно обчислити ймовірність того, що подія $A$ настане не більш як $m$ раз у $n$ випробуваннях. Дану умову задовольнятимуть випадки, коли кількість позитивних результатів буде відповідати 0, 1, …, $m$, тобто всі можливі результати, які не перевищуватимуть $m$. Імовірність, що відповідає даній умові, називається ***кумулятивною***, або накопиченою (містить у собі суму ймовірностей усіх можливих результатів, кількість яких не перевищує заданої), імовірністю біноміального розподілу й позначається $p_n(m)$.

За правилом додавання для незалежних подій дістаємо:

$$p_n(m) = p_n(0) + p_n(1) + ... + p_n(m).$$

Підсумовування відбувається за всіма можливими $i$, для яких $x_i < m$. Ця функція має східчастий вигляд (рис. 1.6).

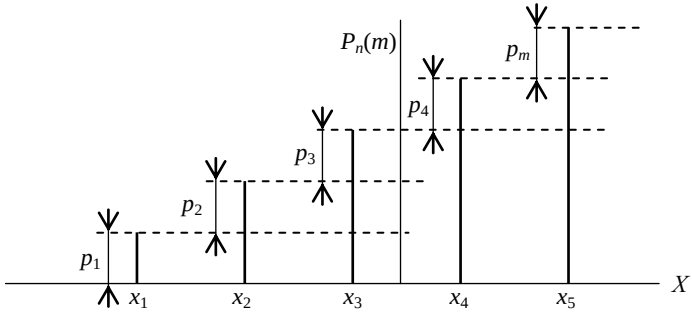



Рис. 1.6. Кумулятивна ймовірність

**Приклад**. Нехай імовірність одержання бракованого виробу дорівнює 0,01. Яка ймовірність того, що серед 100 виробів виявиться не більш як три браковані?

Відповідно до біноміального закону та закону додавання, маємо:

$$P(A) = C_{100}^{0}(0,01)^{0}(0,99)^{100} + C_{100}^{1}(0,01)^{1}(0,99)^{99} + C_{100}^{2}(0,01)^{2}(0,99)^{98} + C_{100}^{3}(0,01)^{3}(0,99)^{97} = 0,9816.$$

**Гіпергеометричний розподіл**. Якщо кількість об'єктів розглянутої сукупності невелика, то на кожному наступному кроці вже буде істотною для вибірки зміна кількості об'єктів, розглянутих на цьому кроці. У разі, коли випробування відбуваються без повернення об'єкта, використовується гіпергеометричний розподіл.

Припустимо, що в партії з $N$ виробів – $M$ стандартних ($M < N$). Із партії навмання беруть $r$ виробів. Випадкова величина $X$ – кількість стандартних виробів серед $r$ відібраних. Очевидно, що $X$ набуває таких можливих значень: 0, 1, 2,..., min ($M$, $r$), а ймовірність появи величини $X = k$ обчислюється за формулою:

$$P(X = k) = \frac{C_M^k C_{N-M}^{r-k}}{C_N^r}.$$

**Розподіл Пуассона**. Розподіл Пуассона часто трапляється в задачах, пов'язаних із *потоком подій*.

**Приклад**. Потік викликів на станції швидкої допомоги, потік заявок у системі масового обслуговування, потік відмов при роботі діагностичного устаткування тощо.

Випадкова величина називається *розподіленою за законом Пуассона*, якщо вона набуває зліченої множини можливих значень 0,1,2... із імовірностями:

$$p_\lambda(m) = \frac{\lambda^m}{m!} e^{-\lambda},\ m = 0,1,2\ldots$$

де $\lambda = A/t$ – середня кількість подій $A$, що настають за час $t$ у потоці.

Імовірність настання тієї або іншої кількості подій за будь-який проміжок часу залежить від тривалості цього проміжку, а не від початку відліку.

Для такого розподілу математичне сподівання та дисперсія збігаються:

$$MX = DX = \lambda; \quad \sigma = \sqrt{\lambda}.$$

Розподіл Пуассона може використовуватись як наближення біноміального розподілу при $n \to \infty$, $p \to 0$ (частинний випадок біноміального закону). Тоді як значення $\lambda$ потрібно брати $np$.

**Приклад**. За умовою останнього прикладу маємо: $n = 100$, $p = = 0{,}01$. Таким чином, $\lambda = 100 \cdot 0{,}01 = 1{,}0$;

$$p(A) = \frac{1^0}{0!}e^{-1} + \frac{1^1}{1!}e^{-1} + \frac{1^2}{2!}e^{-1} + \frac{1^3}{3!}e^{-1} = \frac{1}{e}\left(1 + 1 + \frac{1}{2} + \frac{1}{6}\right) = 0{,}9810,$$

що дає достатньо добрий збіг з точним значенням, але обчислюється набагато простіше.

## Запитання для самоперевірки

1. Як можна задати закон розподілу дискретної випадкової величини?

2. Сформулюйте біноміальний закон розподілу. Чому він має таку назву? У чому полягає суть проведення випробувань за схемою Бернуллі?

3. Якими параметрами характеризується біноміальний розподіл?

4. Як можна визначити ймовірність $(m + 1)$-ї події, знаючи ймовірність настання *m-ї* події?

5. Що таке кумулятивна ймовірність і як вона обчислюється?

6. Що являє собою гіпергеометричний розподіл? В яких випадках стикаємось із розподілом Пуассона?

7. Як визначається ймовірність випадкової події за розподілом Пуассона, і від чого вона залежить? Чому дорівнює математичне сподівання та дисперсія розподілу Пуассона?

# 1.6. НЕПЕРЕРВНІ ВЕЛИЧИНИ. МОМЕНТИ НЕПЕРЕРВНОЇ ВЕЛИЧИНИ

Неперервною називається випадкова величина, якщо вона може набувати нескінченної множини значень на одному або кількох заданих інтервалах.

Раніше було розглянуто, що дискретна випадкова величина може бути задана, якщо є ймовірність появи кожного її значення.

З означення неперервної випадкової величини випливає, що ймовірність того, що неперервна випадкова величина набуде деяке фіксоване значення, дорівнює нулю $P(X = x) = 0$, тому для неперервної випадкової величини, на відміну від дискретної, розглядають імовірність попадання її в нескінченно малий інтервал.

$$P(x < X < x + \Delta x).$$

Зважаючи на те, що ймовірність потрапляння в малий інтервал пропорційна до довжини цього інтервалу, у граничному випадку при $\Delta x \to 0$ маємо:

$$\lim_{\Delta x \to 0} \frac{P(x < X < x + \Delta x)}{\Delta x} = p(x).$$

Функція $p(x)$ називається **щільністю розподілу ймовірностей** неперервної випадкової величини. Так само, як і питома щільність, вона показує, наскільки щільно розподілятимуться імовірності в заданому інтервалі.

Щільність розподілу ймовірностей визначає *закон розподілу ймовірностей неперервної випадкової величини X.* У загальному випадку щільність змінюватиметься зі зміною значення випадкової величини.

Зміна щільності ймовірності неперервної випадкової величини подається у вигляді *кривої розподілу ймовірностей* $p(x)$ *величини X.* На рис. 1.7 наведено криву для нормального закону.

Використовуючи криву розподілу, можна дати графічну інтерпретацію ймовірності $P(x_1 < X < x_2)$. Вона дорівнює площі заштрихованої криволінійної трапеції (див. рис. 1.7).

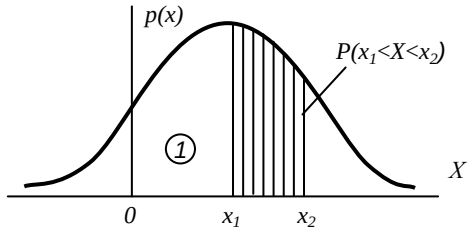



Рис. 1.7. Нормальний закон розподілу ймовірностей неперервної випадкової величини

Таким чином, якщо інтервал скінченний, то ймовірність потрапляння випадкової величини в цей інтервал буде:

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} p(x)dx.$$

**Властивості диференціальної функції**:

— $p(x)$ є *додатною*;

— $p(x)$ має бути *нормована умовою*:

$$\int_{-\infty}^{+\infty} p(x)dx = 1,$$

що відбиває вірогідність події $(-\infty < X < \infty)$, тобто в інтервалі $(-\infty...+\infty)$ випадкова величина набуде принаймні одного зі значень.

Функція $p(x)$ називається також ***диференціальною функцією розподілу ймовірностей***.

Крім диференціальної існує й ***інтегральна функція розподілу*** $F(x)$. Їй відповідає *ймовірність того, що випадкова величина X буде меншою від фіксованого значення x*:

$$F(x) = P(X < x) = \int_{-\infty}^{x} p(x)dx. \qquad (1.11)$$

Графік, що показує, як змінюється $F(x)$ залежно від зміни значення випадкової величини $x$, називається ***інтегральною кривою*** розподілу. На рис. 1.8 наведено інтегральні криві для різних законів розподілу: *а* – рівномірного; *б* – експонентного; *в* – дискретної величини; *г* – нормального.

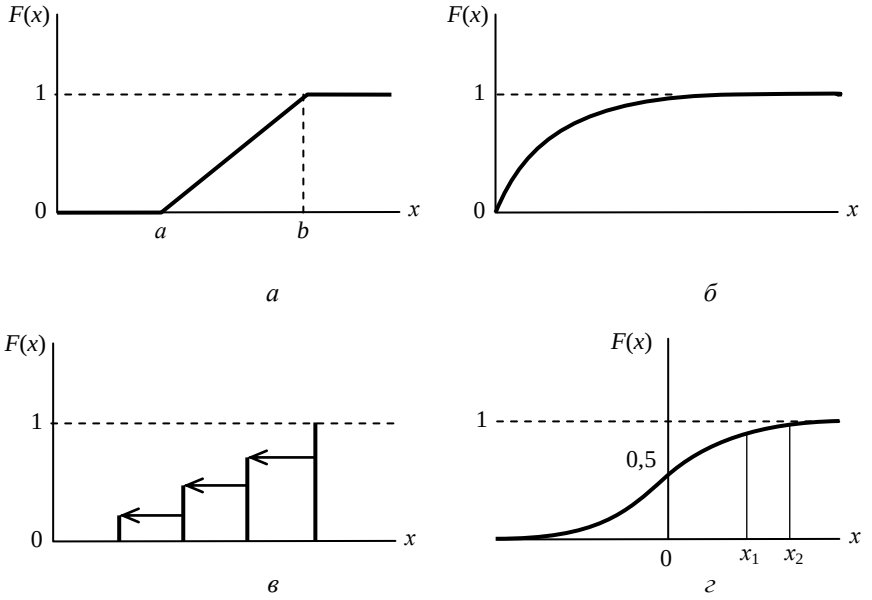



Рис. 1.8. Інтегральні криві розподілу

Інтегральній функції розподілу відповідає площа під кривою розподілу, яка міститься ліворуч від вертикальної прямої, що проходить через точку $X = x$.

**Властивості інтегральної функції розподілу**:

- $F(x_2) \geq F(x_1)$, якщо $x_2 \geq x_1$, тобто $F(x)$ – неспадна функція
- $F(x\ )$ змінюється від значення $F(-\infty) = 0$ до значення $F(+\infty) = 1$.

- $P(x_1 < X < x_2) = F(x_2) - F(x_1)$.
- $P(X \geq x) = 1 - F(x)$.

Якщо взяти похідну від інтегральної функції розподілу, дістанемо щільність розподілу ймовірностей (диференціальну функцію):

$$F' = p(x).$$

***Квантилем***, що відповідає заданому рівню ймовірності $p$, називають таке значення $X = x_p$, при якому функція розподілу набуває значення, що дорівнює $p$, тобто $F(x_p) = p$.

**Приклад**. Для нормального закону розподілу значенню ймовірності $p = 0{,}95$ відповідає квантиль 1,96.

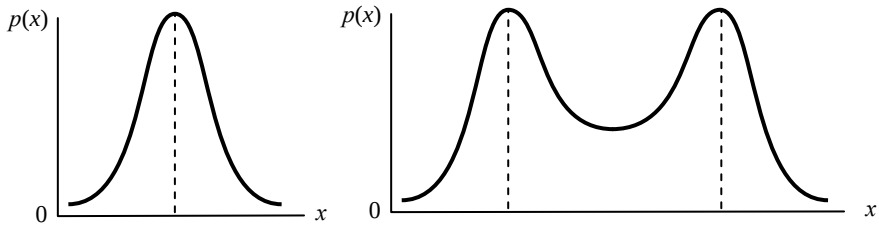
***Медіаною*** розподілу $MeX$ називають квантиль (таке значення $x_{1/2}$), що відповідає значенню ймовірності:

$$P(X < x_{1/2}) = P(X > x_{1/2}) = 1/2,$$

тобто функція розподілу $F(x) = 1/2$.

***Модою*** неперервного розподілу називають значення $x$, при якому щільність розподілу $p(x)$ досягає максимуму.

Розподіл, який має одну моду, називається *унімодальним* (рис. 1.9, *а*). Якщо максимумів функції, наприклад, два (рис. 1.9, *б*), то розподіл називається *двомодальним*. Існують і *багатомодальні* закони розподілу.




*а* *б*

Рис. 1.9. Унімодальний та двомодальний розподіли

**Приклад**. У процесі серійного виробництва виготовлено партію деталей. Для спеціальних цілей було відібрано деталі з малим відхиленням розміру від номінального. Розміри решти деталей можна подати у вигляді двомодального розподілу.

Для симетричних розподілів медіана й мода збігаються із центром розподілу.

**Моменти неперервних розподілів**. Неперервні розподіли відрізняються від дискретних тим, що для них необхідно розглядати ймовірність потрапляння випадкової величини в деякий інтервал.

***Математичне сподівання неперервної випадкової величини***. За аналогією з математичним сподіванням дискретної величини (1.6), для неперервної величини математичне сподівання:

$$MX = \int_{-\infty}^{+\infty} xp(x)dx. \qquad (1.12)$$

*Математичним сподіванням* (центром розподілу) $MX$ неперервної випадкової величини $X$ називається інтеграл від добутку її значень $x$ на щільність розподілу ймовірностей $p(x)$.

*Для неперервної величини у формулах x використовується без індексу, бо йдеться не про конкретне значення $x_i$, а про будь-яке поточне значення x у всьому діапазоні можливих значень X від $-\infty$ до $+\infty$.*

*Операція підсумовування замінюється на операцію інтегрування, тобто підсумовування за елементарними площами p (x) dx.*

***Математичне сподівання функції випадкових величин***. Якщо випадкова величина є *функцією* іншої випадкової величини, $Y = f(X)$, то згідно із загальним виразом для математичного сподівання для випадкової величини $Y$ можна записати:

$$MY = \int_{-\infty}^{+\infty} yp(y)dy\,.$$

З урахуванням функціональної залежності останнє рівняння можна подати у вигляді:

$$MY = \int_{-\infty}^{+\infty} f(x)\,p(x)dx\,.$$

Використовуючи співвідношення $f(X) = X^k$, можна дістати *початкові моменти k*-го порядку:

$$\nu_k = \int_{-\infty}^{+\infty} x^k\,p(x)dx\,. \qquad (1.13)$$

Аналогічно виразу (1.13) можна записати вираз для центральних моментів:

$$\mu_k = \int_{-\infty}^{+\infty} (x-\nu)^k\,p(x)dx \qquad (1.14)$$

Якщо у виразі (1.14) узяти $k = 2$, можна обчислити дисперсію випадкової величини – *другий центральний момент*:

$$DX = \mu_2 = \sigma^2 = \int_{-\infty}^{+\infty} (x-\nu)^2\,p(x)dx\,.$$

Математичне сподівання й дисперсія для неперервної випадкової величини мають ті самі властивості, що й у дискретному випадку.

### Запитання для самоперевірки

1. Яка величина називається неперервною? Наведіть приклади.

2. Що називається щільністю розподілу ймовірностей? Які властивості має функція $p(x)$?

3. Що характеризує крива розподілу ймовірностей?

4. Що являє собою інтегральна функція розподілу? Які властивості має $F(x)$?

5. Що називається квантилем? Наведіть приклад.

6. Що називається медіаною неперервного розподілу? Наведіть приклад.

7. Що є модою неперервного розподілу?

8. Наведіть приклади унімодального, двомодального законів розподілу.

9. Чому дорівнює математичне сподівання неперервної випадкової величини?

10. Чому дорівнює математичне сподівання функції випадкових величин?

11. Що являє собою другий центральний момент?

12. Які властивості мають початкові і центральні моменти неперервної випадкової величини?

## 1.7. ЗАКОНИ РОЗПОДІЛУ НЕПЕРЕРВНИХ ВИПАДКОВИХ ВЕЛИЧИН

Щільність розподілу ймовірностей визначає закон розподілу ймовірностей неперервної випадкової величини X.

**Рівномірний розподіл.** Випадкова величина $X$ розподілена рівномірно у скінченному інтервалі $a$, $b$, якщо всі її можливі значення зосереджені в цьому інтервалі і щільність розподілу її ймовірностей $c$ у цьому інтервалі постійна. Диференціальну форму рівномірного закону розподілу зображено на рис. 1.10.

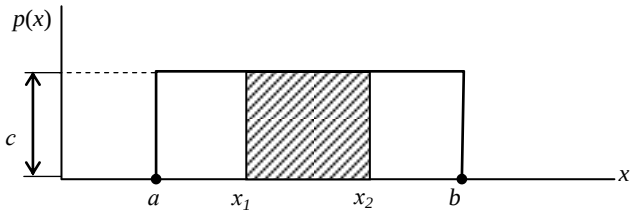


Рис. 1.10. Щільність ймовірностей для рівномірного закону

*Щільність ймовірностей*

$$p(x)=\begin{cases}0 & \text{при} \quad x<a,\ x>b \\ c & \text{при} \quad a\le x\le b\end{cases}.$$

Для випадкової величини $X$, рівномірно розподіленої в інтервалі $(a, b)$, імовірність потрапляння в будь-який інтервал $(x_1, x_2)$, що лежить усередині інтервалу $(a, b)$, пропорційна до довжини цього інтервалу:

$$P(x_1 < X < x_2) = c(x_2 - x_1). \qquad (1.15)$$

Ураховуючи те, що площа прямокутника з основою $(b - a)$ і висотою $c$ дорівнює одиниці (умова нормування), можна визначити щільність рівномірно розподіленої величини:

$$c = p(x) = \frac{1}{b-a}.$$

Підставляючи знайдене значення $c$ в (1.15), знаходимо ймовірність потрапляння значення величини $X$ у інтервал $(x_1, x_2)$, яка дорівнює відношенню довжини цього інтервалу до довжини всього інтервалу $(a, b)$:

$$P(x_1 < X < x_2) = \frac{x_2 - x_1}{b-a}.$$

Крива інтегрального розподілу ймовірностей зображена на рис. 1.8, *а*.

Використовуючи загальний вираз для інтегральної функції (1.11), отримаємо вираз для інтегральної функції рівномірно розподіленої величини:

$$F(x)=\begin{cases} 0 & \text{при} \quad x<a \\ \dfrac{x-a}{b-a} & \text{при} \quad a\le x\le b \\ 1 & \text{при} \quad x>b \end{cases}. \qquad (1.16)$$

Таким чином, усередині інтервалу ($a$; $b$) інтегральна ймовірність буде змінюватися за лінійним законом:

$$F(x)=\int_a^x p(x)dx=\frac{x-a}{b-a}.$$

На інтервалі ($-\infty$, $a$) функція $F(x)=0,$ а на інтервалі ($b$, $+\infty$) маємо $F(x)=1.$

***Центр рівномірного розподілу*** в інтервалі ($a$, $b$) збігається із серединою цього інтервалу:

$$MX=\frac{a+b}{2}.$$

***Дисперсія*** визначається за формулою:

$$DX=\int_a^b\left(x-\frac{a+b}{2}\right)^2 dx=\frac{(b-a)^2}{12}.$$

Звідси випливає, що при рівномірному розподілі в інтервалі ($a$, $b$) ***середнє квадратичне відхилення*** пропорційне до довжини цього інтервалу:

$$\sigma=\frac{b-a}{2\sqrt{3}}.$$

**Нормальний розподіл**. Серед усіх неперервних законів розподілу особливу роль відіграє розподіл імовірностей зі щільністю

$$\varphi(x,a,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-a)^2}{2\sigma^2}}. \qquad (1.17)$$

Такий розподіл називається *нормальним*.

*Важливість нормального закону розподілу визначається тим, що до нього зазвичай приводять задачі, пов'язані з розподілом сум великої кількості випадкових величин, серед яких не можна виділити переважну.*

Про випадкову величину $X$ із таким законом розподілу ймовірностей говорять, що вона *розподілена нормально з параметрами a*, σ. Функція (1.17) являє собою дзвоноподібну криву (диференціальна крива розподілу, рис. 1.11). Параметр $a$ – точка, через яку проходить вісь (центр) симетрії, параметр σ – відстань від осі до точки перегину кривої. При $x \to \pm \infty$ крива розподілу асимптотично наближається до осі абсцис.

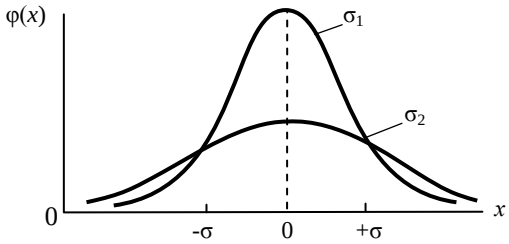



Рис. 1.11. Розсіювання нормально розподілених випадкових при різних значеннях СКВ σ

Якщо σ – мале, крива висока й загострена; якщо σ – велике, вона широка й плоска (на рис. 1.11 $\sigma_1 < \sigma_2$). Це пояснюється тим, що згідно з умовою нормування площа під кривою розподілу дорівнює одиниці і зі зменшенням σ максимум функції буде більший, а випадкова величина «щільніше» розташовуватиметься біля осі симетрії. Таким чином, параметр σ характеризує *розсіювання* випадкової величини $X$.

Розподіл із параметрами $a = 0$, $\sigma = 1$ називається ***стандартним нормальним розподілом***:

$$\varphi(x, a=0, \sigma=1) = \frac{1}{\sqrt{2\pi}} \ell^{\frac{-x^2}{2}}.$$

Крива стандартного нормального розподілу (рис. 1.12) симетрична відносно осі ординат, при $x = 0$ має єдиний максимум, що дорівнює $\frac{1}{\sqrt{2\pi}} \approx 0,40,$ і дві точки перегину при $x = \pm 1$.

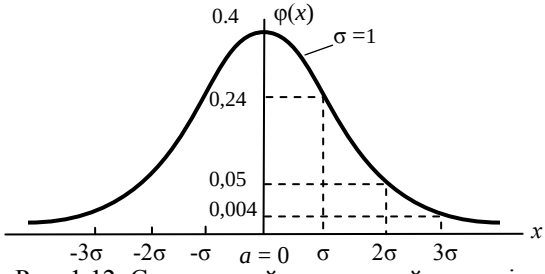


Рис. 1.12. Стандартний нормальний розподіл

*Інтегральну криву* стандартного нормального розподілу зображено на рис. 1.8, *в*.

Інтеграл від щільності $\varphi(x, a, \sigma)$ не виражається через елементарні функції, а крім того, значення інтеграла залежить від $a$ й $\sigma$, тому для розрахунку ймовірностей випадкових величин із нормальним розподілом використовуються таблиці спеціальної функції, яка називається ***інтегралом імовірностей*** (***функцією Лапласа***) $\boldsymbol{\Phi}(z)$.

Існує два різновиди таблиць функції Лапласа, які базуються на виразах:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{\frac{-z^2}{2}} dz \quad ; \quad \Phi_0(z) = \frac{1}{\sqrt{2\pi}} \int_{0}^{z} e^{\frac{-z^2}{2}} dz.$$

Для уніфікації процедури обчислення *x центрують* – початок координат переносять у точку $a$ і *нормують* – значення випадкової величини виражають у частках (одиницях) СКВ. Для цього вводиться нова змінна $z = x - a/\sigma$ – *центрована й нормована* (рис. 1.13).

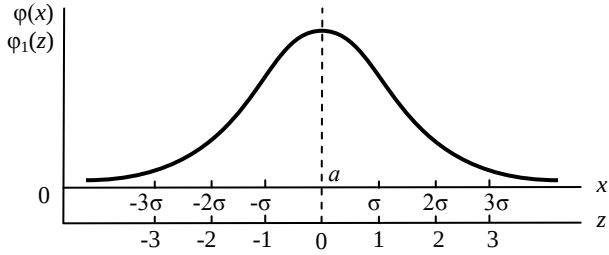


Рис. 1.13. Центрований та нормований нормальний розподіл

Для стандартного нормального закону функція розподілу $F(x)$ пов'язана з інтегралом імовірностей $\Phi(z)$ у такий спосіб. Інтегральна функція розподілу ймовірностей нормально розподіленої величини має вигляд:

$$F(x_1) = P(-\infty < X < x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-a)^2}{2\sigma^2}} dx.$$

Уводимо змінну $z = \frac{x-a}{\sigma}$, тоді, ураховуючи, що $a$ – константа, дістаємо $dz = d\left(\frac{x-a}{\sigma}\right) = \frac{dx}{\sigma}$.

Виходячи з наведених співвідношень, після заміни $dx = dz\sigma$ можна перейти до функції Лапласа:

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz.$$

При $z = 0$ інтегральна функція $\Phi(z) = 0{,}5$, при зміні $z$ від 0 до $+\infty$ функція $\Phi(z)$ зростає від $\Phi(0) = 0{,}5$ до $\Phi(+\infty) = 1$.

Через функцію Лапласа можна виразити ймовірність $P(x_1 < X < x_2)$ у такий спосіб:

$$P(x_1 < X < x_2) = P(X < x_2) - P(X < x_1) =$$
$$= F(x_2) - F(x_1) = \Phi(z_2) - \Phi(z_1).$$

*Існує два види таблиць Лапласа*: *в одних табульовані значення* $\Phi(z)$, *а в інших* $\Phi_0(z)$, *при цьому* $\Phi(z) = 0{,}5 + \Phi_0(z)$. *При обчисленні ймовірностей слід враховувати, що* $\Phi_0(0)=0$; $\Phi_0(-\infty) = \Phi_0(+\infty) = 0{,}5$; $\Phi_0(-z) = -\Phi_0(z)$ (рис. 1.14).

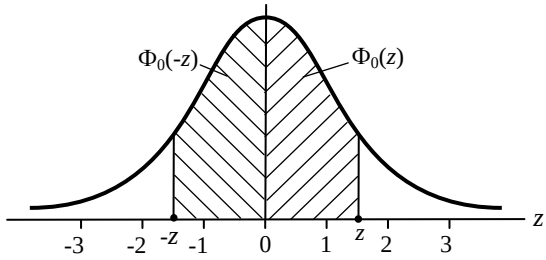



Рис. 1.14. Обчислення ймовірності за таблицями нормального закону

За таблицями нормального закону для $t = 3$ знаходимо $\Phi(3) = 0{,}9973$.

Розглянемо ймовірність того, що випадкова величина буде більшою за $3\sigma$.

$$P(X > 3\sigma) = 1 - P(X < 3\sigma) = 1 - 0{,}997 = 0{,}003,$$

тобто з похибкою 0,3 % можна стверджувати, що всі значення нормально розподіленої випадкової величини містяться в межах $\pm 3\sigma$. Цей висновок на практиці має назву «*правило* $3\sigma$».

*Якщо обчислюється ймовірність потрапляння випадкової величини в інтервал* $(x_1, x_2)$, *то не має значення, якою таблицею ми користуємося, оскільки ця ймовірність буде дорівнювати різниці інтегральних значень.*

Крім математичного сподівання й дисперсії, криві розподілу характеризуються *центральними моментами більш високих порядків* – ***асиметрією*** *A* і ***ексцесом*** *E*.

У тих випадках, коли якісь причини зумовлюють появу значень, більших або, навпаки, менших за середнє, утворюються

*асиметричні розподіли*. При лівосторонній (додатній) асиметрії в розподілі частіше трапляються менші значення (рис. 1.15, *а*), а при правосторонній (від'ємній) асиметрії – більші значення (рис. 1.15, *б*).

Показник *асиметрії* обчислюється за формулою:

$$A = \frac{\mu^3}{\sigma^3} = \frac{M\left(X - \nu\right)^3}{\sigma^3}.$$

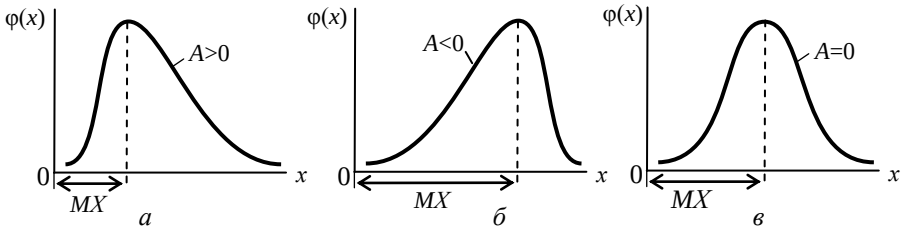



Рис. 1.15. Моментами високих порядків – асиметрія А і ексцес Е

Для симетричних розподілів $A = 0$ (рис. 1.15, *в*).

У тих випадках, коли деякі причини сприяють переважній появі значень, близьких до центра розподілу, говорять про розподіли з *додатним ексцесом* (рис. 1.16, *а*, вершина кривої перебуває вище від вершини кривої нормального розподілу).

Якщо ж у розподілі переважають крайні відносно центра розподілу значення, причому одночасно й малі, і великі, такий розподіл характеризується *від'ємним ексцесом* (рис. 1.16, *б*, вершина кривої перебуває нижче від кривої нормального розподілу), причому в центрі розподілу може утворитися западина, перетворивши його на двомодальний.

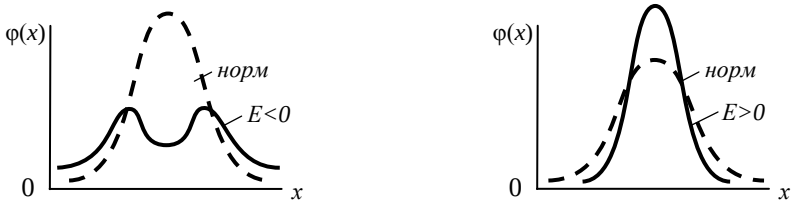



Рис. 1.16. Показник ексцесу

Показник ексцесу $E$ пропорційний до відношення $\frac{\mu^4}{\sigma^4}$. З огляду на те, що для нормального закону $\frac{\mu^4}{\sigma^4} = 3$, показник ексцесу визначається за формулою:

$$E = \frac{\mu^4}{\sigma^4} - 3.$$

Таким чином, для нормального закону $E = 0$, тому нормальний закон є еталоном «вершинності» одномодальних неперервних розподілів.

**Запитання для самоперевірки**

1. Яка випадкова величина вважається рівномірно розподіленою?
2. Який вигляд має диференціальна форма рівномірного розподілу? Чому дорівнює щільність імовірності для рівномірного закону?
3. Який вигляд має інтегральна функція рівномірного розподілу та чому дорівнює?
4. Як обчислюються числові характеристики рівномірного розподілу?
5. Який розподіл називається нормальним? Чому він на практиці відіграє важливу роль?
6. Що собою являє диференціальна форма нормального розподілу?
7. Охарактеризуйте стандартний нормальний розподіл.
8. Що називається функцією Лапласа і для чого вона використовується?
9. Для чого потрібно центрувати та нормувати випадкову величину?
10. Як обчислювати ймовірність за допомогою функції Лапласа?
11. Назвіть центральні моменти високих порядків.

# 1.8. БАГАТОВИМІРНІ ВИПАДКОВІ ВЕЛИЧИНИ

З певними випробуваннями може бути пов'язана не одна, а кілька випадкових величин

Крім одновимірних величин часто доводиться розглядати одночасно системи із двох, трьох і більшої кількості випадкових величин.

**Приклад**. Розміри тієї самої деталі: довжина, висота, ширина; струм завад в електронному підсилювачі з випадковими амплітудою та фазою; зріст та маса тіла людини.

Такі величини називаються *двовимірними*, *тривимірними* тощо відповідно до кількості компонент такої величини.

Для простоти викладу розглядатимемо двовимірний випадок.

Теоретично двовимірну величину (*X*, *Y*) можна розглядати як *випадковий вектор* або точку площини з випадковими координатами (*X*, *Y*). Іноді зручно говорити про *випадковий радіус-вектор* або про систему двох випадкових величин *X* і *Y*.

Закон розподілу ймовірностей неперервної двовимірної випадкової величини (*X*, *Y*) має характеризувати ймовірність потрапляння будь-якої пари її значень (*x*, *y*) у будь-яку область на площині (рис. 1.17).

За аналогією з одновимірними неперервними величинами можна дати означення.

**Неперервною двовимірною випадковою величиною** називається величина (*X*, *Y*), якщо ймовірність потрапляння її значень у будь-яку область *D* на площині (*X*, *Y*) можна подати у вигляді подвійного інтеграла

$$P[(X,Y)\in D]=\iint\limits_{D} p_{XY}(x,y)dx,dy,$$

де $p_{XY}(x, y)$ – щільність двовимірного розподілу ймовірностей або щільність спільного розподілу випадкових величин *X* і *Y*; $p_{XY}(x, y)\,dxdy$

– характеризує ймовірності потрапляння точки ($x$, $y$) у заштрихований прямокутник зі сторонами $dx$ і $dy$.

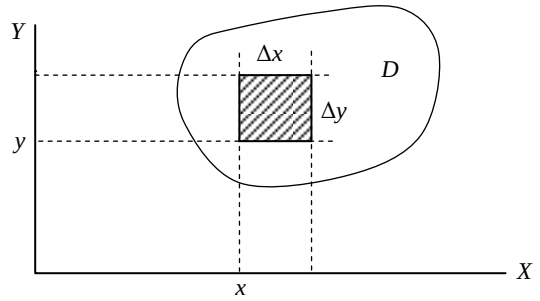



Рис. 1.17. Щільність двовимірної випадкової величини

Для двовимірного, як і для одновимірного, випадку, також буде виконуватися умова нормування:

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} p_{XY}(x,y)dx,dy = 1.$$

Для розглянутого прямокутника імовірність потрапляння в нього координати $X$ дорівнює $p_X(x)dx$, а координати $Y$ – $p_Y(y)\,dy$. Імовірність потрапляння точки в зазначений прямокутник дорівнює $p(x, y)\,dxdy$, оскільки існує *поєднання випадкових подій*

$$x < X < x + dx,\ y < Y < y + dy$$

для незалежних величин.

Таким чином, можна записати:

$$p_{XY}(x,y)dxdy = p_X(x)dx\,p_Y(y)dy,$$

звідки випливає

$$p_{XY}(x,y) = p_X(x)\,p_Y(y). \qquad (1.18)$$

Ця рівність справджується для всіх значень $x$ і $y$ в разі, коли випадкові величини $X$ і $Y$ незалежні.

Таким чином, **незалежними** (**взаємно незалежними**) називають неперервні випадкові величини $X$ і $Y$, якщо щільність їхнього спільного розподілу дорівнює добутку щільностей цих величин.

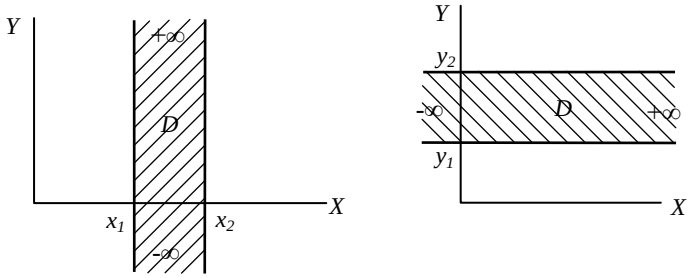
Ця властивість важлива тим, що, знаючи щільності розподілу випадкових величин $X$ і $Y$, можна легко визначити щільність їхнього спільного розподілу, що значно спрощує розрахунки.

Графік функції $z = p_{XY}(x, y)$ називають *поверхнею розподілу ймовірностей.*

***Розподіли координат двовимірної випадкової величини***. Проаналізувавши вираз (1.18), можна дійти висновку, що закон спільного розподілу величин $X$ і $Y$ повністю визначає закони розподілу кожної з величин $X$ і $Y$.

Спочатку розглянемо, як можна визначити щільність $p_X(x)$ випадкової величини $X$. Щоб знайти щільність $p_X(x)$, визначимо ймовірність потрапляння значень величини $X$ у будь-який інтервал $(x_1 < X < x_2)$ на площині $(X, Y)$ (рис. 1.18, *а*).

Влучення значень величини $X$ у інтервал $(x_1, x_2)$ рівносильне влученню точки $(X, Y)$ у вертикальну смугу $D$ на площині. При цьому випадкова величина $Y$ набуде можливі значення від $-\infty$ до $+\infty$ (заштрихована смуга на рис 1.18, *а*).




*а* *б*

Рис. 1.18. Знаходження щільності для двовимірної випадкової величини

Отже,

$$P(x_1 < X < x_2) = \iint_D p_{XY}(x, y)dxdy = \int_{x_1}^{x_2}\left[\int_{-\infty}^{+\infty} p_{XY}(x, y)dy\right]dx. \quad (1.19)$$

Оскільки водночас

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} p_X(x)dx,$$

тоді рівність (1.19) можна записати у вигляді:

$$\int_{x_1}^{x_2} p_X(x)dx = \int_{x_1}^{x_2}\left[\int_{-\infty}^{+\infty} p_{XY}(x,y)dy\right]dx,$$

звідки

$$p_X(x) = \int_{-\infty}^{+\infty} p_{XY}(x,y)dy. \qquad (1.20)$$

*Для того щоб знайти щільність одновимірної величини, знаючи щільність двовимірної величини, необхідно зінтегрувати щільність двовимірної величини за всіма можливими значеннями (від $-\infty$ до $+\infty$) іншої випадкової величини.*

Аналогічно знаходять щільність розподілу координати *Y* (рис. 1.18, *б*).

$$p_Y(y) = \int_{-\infty}^{+\infty} p_{XY}(x,y)dx.$$

**Поняття функції неперервних випадкових величин**. Під функцією $f(X)$ випадкової величини $X$ розуміють таку випадкову величину $Y$, яка набуває значення $y = f(x)$ щоразу, коли величина $X$ набуває значення $x$. При цьому передбачається, що розглянута функція визначена для всіх можливих значень аргументів.

Нехай $y = f(x)$ передбачається монотонно зростаючою неперервною, тобто такою, що має похідну в усіх точках. Для такої функції множина значень випадкової величини $X$, що потрапляє в

інтервал ($x_1$, $x_2$), буде однозначно відображатися множиною значень випадкової величини $Y$ в інтервалі ($y_1, y_2$) (рис. 1.19).

Для функції, яка не має вигинів і розривів, можна записати:

$$P(x < X < x + \Delta x) = P(y < Y < y + \Delta y). \qquad (1.21)$$

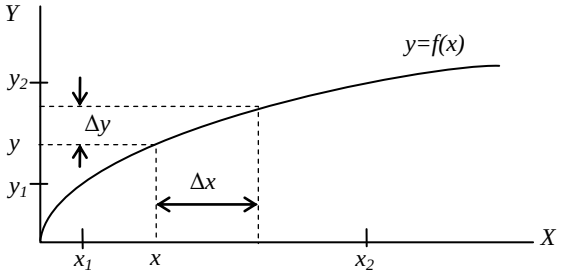



Рис. 1.19. Монотонна функція неперервних випадкових величин

Для неперервних величин при виконанні цієї умови ймовірність потрапляння випадкової величини в елементарний інтервал буде пропорційна до щільності ймовірності в початковій точці інтервалу, помноженої на довжину цього інтервалу. Згідно з виразом (1.21) можна записати, що

$$p_X(x)dx = p_Y(y)dy. \qquad (1.22)$$

Задача полягає у визначенні щільності розподілу ймовірностей $p_Y(y)$ випадкової величини $Y$, функціонально пов'язаної з випадковою величиною $X$, із *використанням* відомого закону розподілу ймовірностей випадкової величини $X$ – щільності цього розподілу $p_X(x)$.

З виразу (1.22) шукана щільність імовірності визначається як

$$p_Y(y) = p_X(x)\frac{dx}{dy}. \qquad (1.23)$$

Необхідно визначити щільність розподілу випадкової величини $Y$, тоді як у вираз (1.23) входить щільність імовірностей величини $X$. Для того щоб виключити з виразу (1.23) значення

аргументу $x$, уведемо обернену функцію $g(y) = x$. Наприклад, якщо $y = Kx$, то $g(y) = x = \frac{y}{K}$.

Другий множник у правій частині виразу (8.6) є похідною від оберненої функції, тобто

$$\frac{dx}{dy} = g'(y) \ .$$

З урахуванням сказаного вираз (1.23) набуде вигляду:

$$p_Y(y) = p_X[g(y)]g'(y). \qquad (1.24)$$

**Приклад**. Випадкова величина $X$ має рівномірну щільність розподілу ймовірностей $p_X(x)$ в інтервалі $(x_1, x_2)$. Знайти щільність розподілу $p_Y(y)$ лінійної функції $Y = ax + b$ $(a \neq 0)$.

Зважаючи на те, що функція $y = ax + b$ монотонна, необхідно:

1) знайти обернену функцію для цієї залежності $g(y) = x = \frac{y-b}{a}$.

2) знайти похідну оберненої функції $g'(y) = \frac{1}{|a|}$ .

3) вирази для $g(y)$ і $g'(y)$ підставити у (1.24).

4) визначити щільність розподілу ймовірностей лінійної функції $Y$:

$$p_Y(y) = p_X\left(\frac{y-b}{a}\right) \cdot \frac{1}{|a|}.$$

Для рівномірного закону $p_X(x) = \frac{1}{x_2 - x_1}$ залишається незмінною на всьому інтервалі, тоді

$$p_X\left(\frac{y-b}{a}\right) = \frac{1}{x_2 - x_1} \ ,$$

при цьому обидві частини виразу не залежать від поточного значення $x$.

Шукана щільність буде дорівнювати

$$p_Y(y) = \frac{1}{x_2 - x_1} \cdot \frac{1}{|a|}$$

і рівномірно розподілиться в інтервалі $[(ax_1+b),(ax_2+b)]$.

*Якщо функція немонотонна, для виключення неоднозначності необхідно розбити її на окремі монотонні ділянки й для кожної з них застосувати формулу* (*1.24*), *а потім скористатися властивістю адитивності ймовірностей.*

**Моменти розподілу двовимірної величини**. Обмежимося визначенням моментів розподілу тільки для двовимірних випадкових величин. Поширення цих понять на випадкові величини розмірності, більшої від двох (або системи двох випадкових величин), не становить труднощів.

***Початкові моменти порядку k + l*** для випадкової величини (*X*, *Y*) визначаються за формулою:

$$\nu_{k,l} = MX^k Y^l,$$

де *k* – степінь величини *X*, *l* – степінь величини *Y*.

Для неперервної величини із щільністю спільного розподілу $p_{XY}(x,y)$

$$\nu_{k,l} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x^k y^l p_{XY}(x,y)dxdy .$$

***Початкові моменти першого порядку***

$$\nu_{1,0} = MX = \nu_X; \quad \nu_{0,1} = MY = \nu_Y$$

визначають точку з координатами (*MX*;*MY*) – центр розподілу двовимірної величини (*X*,*Y*), тобто математичне сподівання ($\nu_X$; $\nu_Y$).

Отже, ***математичним сподіванням випадкового вектора*** $\{X;Y\}$ називається вектор $\{MX;MY\}$.

Математичне сподівання випадкового вектора $\{X;Y\}$ можна виразити через щільність спільного розподілу ймовірностей його координат відповідно до (1.19)

$$MX = \int_{-\infty}^{+\infty} x p_X(x)dx = \int_{-\infty}^{+\infty} x dx \int_{-\infty}^{+\infty} p_{XY}(x,y)dy$$

або

$$MX = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x p_{XY}(x,y)dxdy.$$

Аналогічно одержимо

$$MY = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} y p_{XY}(x,y)dxdy.$$

**Властивості математичного сподівання двовимірної величини:**

- Математичне сподівання суми двох випадкових величин дорівнює сумі математичних сподівань цих величин:
$$M(X+Y) = MX + MY.$$
- Математичне сподівання добутку незалежних випадкових величин дорівнює добутку математичних сподівань цих величин:
$$M(XY) = MX \cdot MY.$$

Ці властивості справджуються для незалежних випадкових величин будь-якої вимірності.

**Центральні моменти порядку *k + l*** визначаються за формулою:

$$\mu_{k,l} = M\left(X - \nu_X\right)^k \left(Y - \nu_Y\right)^l.$$

За означенням обидва центральних моменти першого порядку ($k + l = 1$) дорівнюють нулю:

$$\mu_{1,0} = M\left(X - \nu_X\right) = 0, \quad \mu_{0,1} = M\left(Y - \nu_Y\right) = 0.$$

Із трьох центральних моментів другого порядку ($k + l = 2$) два моменти являють собою дисперсії одновимірних розподілів величин $X$ і $Y$:

$$\mu_{2,0} = M\left(X - \nu_X\right)^2 = DX,$$
$$\mu_{0,2} = M\left(Y - \nu_Y\right)^2 = DY,$$

і лише один момент $\mu_{1,1}$ пов'язаний зі спільним розподілом величин $X$ і $Y$:

$$\mu_{1,1} = M\left(X - \nu_X\right)\left(Y - \nu_Y\right).$$

Для неперервних величин із щільністю спільного розподілу $p_{XY}(x,y)$:

$$\mu_{1,1} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(X - \nu_X\right)\left(Y - \nu_Y\right)p_{XY}(x,y)dxdy.$$

Момент $\mu_{1,1}$ називається ***мішаним моментом другого порядку*** і позначається $\mu_{1,1}$ = cov ($X$, $Y$). Він має також назви *кореляційний момент*, *коваріація*, *момент зв'язку* величин $X$ і $Y$.

Таким чином, **коваріація** є найпростішою характеристикою лінійної стохастичної залежності між випадковими величинами $X$ і $Y$ – *це математичне сподівання добутку відхилень $X$ і $Y$ від їхніх центрів*.

*Між двома випадковими величинами є стохастична залежність тоді, коли існує сукупність випадкових факторів, які впливають на обидві випадкові величини, і є фактори, що діють тільки на одну або тільки на другу випадкову величину, тобто якщо $X = f(Z_1,\ldots,Z_m, X_1,\ldots, X_j)$, $Y = f(Z_1,\ldots, Z_m, Y_1,\ldots, Y_k)$, то $X$ і $Y$ стохастично залежні.*

Зважаючи на те, що коваріація є розмірною величиною, нею безпосередньо не можна скористатися для показника щільності *стохастичної залежності.*

Для того щоб можна було чисельно визначити щільність стохастичної залежності, абстрагуючись від роду й розмірності фізичної величини, вводять коваріацію для *нормованих випадкових* величин, тобто розглядають не $X$ і $Y$, а

$$X^* = \frac{X - \nu_X}{\sigma_X}\ ,\ Y^* = \frac{Y - \nu_Y}{\sigma_Y}.$$

Кожна з них має центром розподілу нуль і дисперсію, рівну одиниці. Тоді

$$\begin{aligned} &\operatorname{cov}\left(X^*, Y^*\right) = M\left(\frac{X - \nu_X}{\sigma_X} \cdot \frac{Y - \nu_Y}{\sigma_Y}\right) = \\ &\frac{M\left[\left(X - \nu_X\right)\left(Y - \nu_Y\right)\right]}{\sigma_X \sigma_Y} = \frac{\operatorname{cov}\left(X, Y\right)}{\sigma_X \sigma_Y} = \rho_{XY}. \end{aligned} \tag{1.25}$$

Коваріація для нормованої величини називається **коефіцієнтом кореляції.**

Таким чином, *коефіцієнт кореляції є безрозмірною величиною й показує щільність стохастичного лінійного зв'язку незалежно від роду фізичної величини.*

Для незалежних величин коефіцієнт кореляції дорівнює нулю, але на підставі рівності нулю коефіцієнта кореляції не можна стверджувати, що величини незалежні.

**Приклад**. Припустимо, що величина $X$ симетрично розподілена відносно початку координат, який і буде центром величини $X$, тобто $MX = 0$. Нехай випадкова величина $X$ функціонально пов'язана з випадковою величиною $Y$ залежністю $Y = X^2$. Згідно з симетрією $X$ відносно центра координат маємо:

$$cov\,(X,Y) = M(X,X^2) = M(X^3) = 0.$$

У результаті, з одного боку, маємо функціонально пов'язані величини, а з другого – кореляція між ними дорівнює нулю. Ця

суперечність пояснюється тим, що рівність нулю коефіцієнта кореляції свідчить про незалежність випадкових величин тільки в разі *лінійного стохастичного зв'язку.*

**Характеристики багатовимірних величин.**

***Дисперсія залежних величин.*** Розглянемо загальний вираз для дисперсії суми випадкових величин:

$$D(X \pm Y) = M\left[(X - \nu_X) \pm (Y - \nu_Y)\right]^2;$$

$$D(X \pm Y) = M\left[(X - \nu_X)\right]^2 + \left[(Y - \nu_Y)\right]^2 \pm 2M\left[(X - \nu_X)(Y - \nu_Y)\right]$$

або

$$D(X \pm Y) = DX + DY \pm 2cov(X, Y).$$

Таким чином, коваріація входить у загальну формулу додавання дисперсій.

***Дисперсія незалежних величин***. Розглянемо вираз для коваріації, подавши його в такий спосіб:

$$\operatorname{cov}(XY) = M\left[(X - \nu_X)(Y - \nu_Y)\right] = M\left[XY - X\nu_Y - Y\nu_X + \nu_X\nu_Y\right] =$$
$$= M(XY) - M(X\nu_Y) - M(Y\nu_X) + M(\nu_X\nu_Y).$$

Зважаючи на те, що $\nu_Y, \nu_X$ – постійні величини і $\nu_Y = MY$, $\nu_X = MX$, дістаємо:

$$\operatorname{cov}(XY) = M(XY) - \nu_Y MX - \nu_X MY + \nu_X\nu_Y =$$
$$= M(XY) - MXMY.$$

Якщо $X$ і $Y$ незалежні, то

$$M(XY) = MXMY, \text{ і } cov(XY) = MXMY - MXMY = 0.$$

Таким чином, дисперсія для суми або різниці незалежних $X$ і $Y$:

$$D(X \pm Y) = DX + DY.$$

Тобто, *дисперсія алгебраїчної суми випадкових незалежних величин дорівнює алгебраїчній сумі їхніх дисперсій.*

Урахувавши те, що $DX = \sigma_x^2$, а $DY = \sigma_y^2$, можна записати:

$$\sigma_{X \pm Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}.$$

***Дисперсія нормованих величин***.

$$D\left(X^*\right)=M\left[\left(\frac{X-\nu_X}{\sigma_X}-0\right)\right]^2=\frac{M\left(X-\nu_X\right)^2}{\sigma_X^2}=\frac{\sigma_X^2}{\sigma_X^2}=1.$$

Аналогічно $D\left(Y^*\right)=1.$

Отже, для *дисперсії суми* або *різниці нормованих величин* можна записати:

$$D\left(X^*\pm Y^*\right)=D\left(X^*\right)+D\left(Y^*\right)\pm 2cov\left(X^*,Y^*\right)$$

або

$$D\left(X^*\pm Y^*\right)=1+1\pm 2\rho_{XY}=2\left(1\pm\rho_{XY}\right).$$

Зважаючи на те, що дисперсія є додатним числом, тобто

$$D\left(X^*\pm Y^*\right)\geq 0,$$

маємо

$$-1\leq\rho_{XY}\leq+1.$$

Таким чином, дістали важливу властивість коефіцієнта кореляції: *за абсолютним значенням коефіцієнт кореляції не перевищує одиниці*.

*Якщо коефіцієнт кореляції за модулем дорівнює одиниці, то стохастична залежність стає лінійною функціональною залежністю. Зміна X неодмінно приведе до певної зміни Y.*

***Усереднення дисперсій***. Висновок про дисперсію суми незалежних величин поширюється й на більшу кількість випадкових величин:

$$D\left(X_1+X_2+...+X_n\right)=DX_1+DX_2+...+DX_n.$$

Припустимо, що випадкові величини $X_i$ мають *однаковий розподіл* (*однакову дисперсію*), тоді останній вираз можна записати у такий спосіб:

$$D\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} DX_i = nDX$$

або

$$\sigma_{\Sigma}^2 = \sum_{i=1}^{n} \sigma_{X_i}^2 .$$

Таким чином, у цьому разі зростання дисперсії суми відбувається пропорційно до кількості доданків, тоді як *середнє квадратичне відхилення* зростає пропорційно до квадратного кореня з кількості доданків:

$$\sigma_{\Sigma} = \sigma_X \sqrt{n} .$$

*Коли намагаються зменшити вплив систематичної похибки приладу на результат дослідження, складають додаткове рівняння «апаратурним шляхом» і розв'язують систему двох рівнянь. При цьому двічі вносять у результат випадкову похибку, тобто дисперсія результату збільшується вдвічі.*

При статистичній обробці результатів вимірювання додаткові рівняння використовують для відшукання середнього значення. При цьому можливі результати $X_i$ характеризуються тим самим законом розподілу.

*Середнє значення* (оцінка математичного сподівання) визначається як

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i . \qquad (1.26)$$

*Дисперсія* середнього

$$D\overline{X} = D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} D\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \sum_{i=1}^{n} DX_i = \frac{1}{n^2} nDX_i = \frac{DX_i}{n} .$$

Отже, коли випадкові величини мають однакові розподіли, дисперсія їхнього середнього буде в $n$ раз меншою за вихідну дисперсію самих випадкових величин.

Для середнього квадратичного відхилення (СКВ)

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}. \qquad (1.27)$$

Таким чином, СКВ середнього результату вимірювань буде в $\sqrt{n}$ разів меншим за вихідне СКВ одиничного результату вимірювань.

*Додаткові досліди дають позитивний ефект у випадку, коли вимірювання того самого значення фізичної величини проводять з метою зменшення впливу випадкової похибки на результат.*

**Запитання для самоперевірки**

1. Дайте визначення двовимірної величини, і наведіть приклади. Чому дорівнює щільність імовірностей двовимірної величини?
2. Як можна знайти щільність імовірностей функції за законом розподілу ймовірностей її аргументу?
3. Дайте означення незалежних неперервних випадкових величин.
4. Як визначаються початкові моменти порядку $k + l$?
5. Чому дорівнюють початкові моменти першого порядку двовимірної величини?
6. Назвіть властивості математичного сподівання двовимірної величини.
7. Як визначаються центральні моменти порядку $k + l$ ?
8. Що таке коваріація і які назви вона ще має?

9. Які величини називаються некорельованими?

10. Які величини вважаються стохастично залежними?

11. Що називається коефіцієнтом кореляції і що він характеризує?

12. Чому дорівнює дисперсія нормованих величин?

13. У чому полягає основна властивість коефіцієнта кореляції?

14. Чому дорівнює дисперсія незалежних величин?

15. Чому дорівнює дисперсія суми незалежних величин?

16. Чому дорівнює дисперсія середнього значення?

17. Яке співвідношення пов'язує СКВ середнього результату й окремих результатів вимірювань?

## 1.9. ПОШИРЕНІ ЗАКОНИ РОЗПОДІЛУ

Найпоширеніші на практиці й найбільш вивчені закони розподілів поділяються на одно- та двопараметричні.

Розглянемо закони, які найчастіше використовуються у практичній діяльності при обробці інформації.

Серед *однопараметричних законів* слід відзначити такі.

- ***Експоненціальний розподіл***. Поширений при розв'язуванні задач, пов'язаних із надійністю, масовим обслуговуванням, оцінкою рідко спостережуваних явищ.

Характеризується параметром $\lambda$, який у теорії надійності розглядається як інтенсивність відказів тих чи інших пристроїв. Функцію щільності ймовірностей наведено на рис. 1.20, *а*.

- $\chi^2$***-розподіл Пірсона.*** Широко використовується у статистичній теорії надійності й прикладному статистичному аналізі при перевірці гіпотез. Якщо $X_1$, $X_2$,..., $X_n$ – незалежні однаково розподілені випадкові величини, кожна з яких має *нормальний* закон

розподілу з нульовим середнім і одиничною дисперсією, то величина $\sum_{i=1}^{n} X_i^2$ має $\chi^2$-розподіл *Пірсона* з $f$ степенями вільності.

Розподіл відповідає розподілу квадрата довжини випадкового вектора в *n-вимірному* просторі, якщо кожна з його проекцій $x_i \in N(0,1)$.

Характеризується кількістю степенів вільності $f$. Функцію щільності ймовірностей наведено на рис. 1.20, *б*.

Окремі випадки розподілу Пірсона – розподіли Максвелла й Релея.

- ***t-розподіл Стьюдента.*** Якщо випадкова величина $X$ має нормальний розподіл, а випадкова величина $Y$ *має* $\chi^2$*-розподіл Пірсона* з $f$ степенями вільності, причому $X$ і $Y$ – незалежні, то випадкова величина

$$T = \frac{X}{\sqrt{Y/n}} = \frac{X_{n+1}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}}$$

має *t-розподіл* Стьюдента. Випадкова величина $T$ відповідає розподілу відношення $(n + 1)$-го результату (серії) вимірювань до середньоарифметичної суми квадратів результатів, отриманих при попередніх $n$ вимірюваннях.

Функцію щільності ймовірностей наведено на рис. 1.20, *в*.

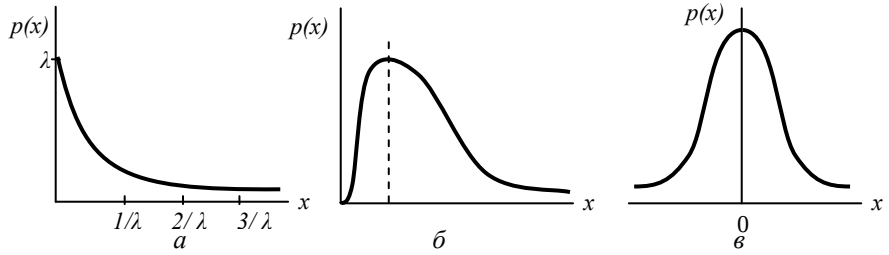



Рис. 1.20. Функція щільності ймовірностей для однопараметричних законів

Серед *двопараметричних законів* варто розглянути такі.

***Рівномірний розподіл.*** Цей розподіл є статистичною моделлю, що описує події, які з однаковою ймовірністю можуть відбутися в заданому інтервалі. Якщо вид розподілу випадкової величини апріорно невідомий, то допускають, що існує рівномірний розподіл.

Характеризується параметрами *b, a*. Функцію щільності ймовірностей наведено на рис. 1.21, *а*.

- ***Нормальний розподіл***. Практично в основу всієї класичної теорії ймовірностей і прикладного статистичного аналізу покладено нормальний розподіл, що є граничною формою численних розподілів.

Характеризується параметрами *a*, σ. Функцію щільності ймовірностей наведено на рис. 1.21, *б*.

- ***Логарифмічно-нормальний розподіл.*** Використовується для опису випадкових величин, логарифм яких розподілений нормально. Широко застосовується в теорії надійності, при апроксимації промислових і атмосферних завад.

Характеризується параметрами *a*, σ. Функцію щільності ймовірностей наведено на рис. 1.21, *в*.

- ***Розподіл Лапласа.*** Описує різницю двох незалежних випадкових величин $X = X_1 - X_2$, які мають експоненціальний розподіл.

Характеризується параметрами *a*, ± λ /2. Функцію щільності ймовірностей наведено на рис. 1.21, *г*.

- ***Гамма-розподіл***. На відміну від розподілу Лапласа, який використовується для відображення різниці незалежних випадкових величин, γ – розподіл описує розподіл суми незалежних випадкових величин, кожна з яких має експоненціальний закон розподілу. Застосовується в теорії масового обслуговування при розв'язуванні задач, пов'язаних з очікуванням у черзі та обслуговуванням клієнтів.

Характеризується параметрами α, β. Функцію щільності ймовірностей наведено на рис. 1.21, *д*.

- ***Екстремальний розподіл***. Застосовується в тому разі, коли необхідно знайти розподіл найменшого або найбільшого елемента вибірки, узятого із сукупності. У загальному випадку розподіл екстремальних значень залежить від обсягу вибірки й характеру початкового розподілу. Використовується в задачах, які стосуються оцінювання довговічності механічних і електромеханічних вузлів та агрегатів.

Характеризується параметрами *a*, σ. Функцію щільності ймовірностей наведено на рис. 1.21, *е*, де *1* – розподіл найбільших значень; *2* – розподіл найменших значень.

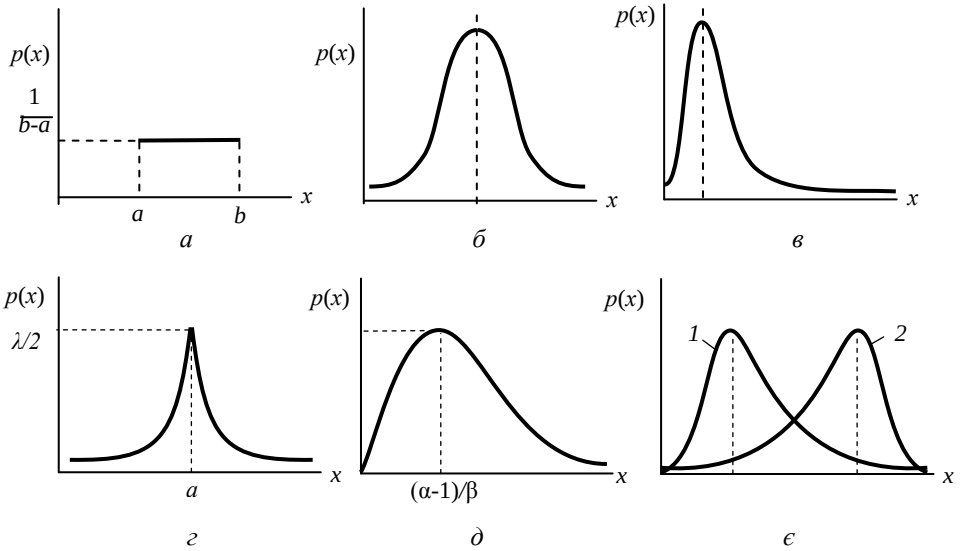
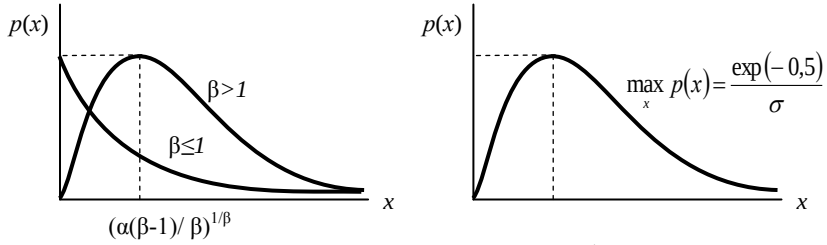



Рис. 1. 21. Двопараметричні закони розподілу

- ***Розподіл Вейбулла.*** Випадкові величини, які не мають певних меж для своїх значень, належать до розподілу Вейбулла. Застосовується для опису розподілу часу непояви раптових відказів деяких елементів радіоапаратури.

Характеризується параметрами α, β. При β ≤ 1 він перетворюється на експоненціальний розподіл. Функцію щільності ймовірностей наведено на рис. 1.22, *а*.

- ***Розподіл Релея***. Частинний випадок розподілу Вейбулла при β = 2. Функцію щільності ймовірностей наведено на рис. 1.22, *б*.




*а* *б*

Рис. 1.22. Розподіли Вейбулла та Релея

Усі перелічені закони розподілу характеризуються *щільністю ймовірностей, інтегральною ймовірністю, початковими та центральними моментами* (математичним сподіванням, дисперсією, коефіцієнтом асиметрії, коефіцієнтом ексцесу).

Оцінки параметрів визначають за *методом моментів, максимальної правдоподібності* або *методу найменших квадратів.*

**Приклад.**

В табл. 1.2 наведені характеристики розподілів для законів, які найбільш зустрічаються на практиці.

Таблиця 1.2. Моментні характеристики розподілів

| Розподіл | Нормальний | Експоненціальний | Максвела | Рівномірний | Лапласа | Екстремальний |
|---|---|---|---|---|---|---|
| Асиметрія А | 0 | 2 | 0,065375 | 0 | 2,12132 | 1,12396 |
| Ексцес Е | 0 | 6 | 1,569972 | 1,2 | 3 | 2,4 |

## Запитання для самоперевірки

1. Які закони розподілу належать до одно- та двопараметричних?

2. Охарактеризуйте експоненціальний розподіл. Які процеси можуть бути ним описані?

3. Охарактеризуйте $\chi^2$-розподіл Пірсона**.** Які процеси можуть бути ним описані?

4. Охарактеризуйте *t*-розподіл Стьюдента. Які процеси можуть бути ним описані?

5. Охарактеризуйте нормальний та логарифмічно-нормальний розподіли. Які процеси можуть бути ними описані?

6. Охарактеризуйте розподіл Лапласа та γ-розподіл. Які процеси можуть бути ними описані?

7. Охарактеризуйте рівномірний розподіл. Які процеси можуть бути ним описані?

8. Охарактеризуйте екстремальний розподіл. Які процеси можуть бути ним описані?

9. Охарактеризуйте розподіл Вейбулла. Які процеси можуть бути ним описані?

10. Якими параметрами характеризуються розподіли?

11. За якими методами обчислюються параметри розподілу?

# РОЗДІЛ 2

# СТАТИСТИЧНІ ОЦІНКИ ПАРАМЕТРІВ РОЗПОДІЛУ

## 2.1. ПЕРВИННИЙ СТАТИСТИЧНИЙ АНАЛІЗ. ГЕНЕРАЛЬНА СУКУПНІСТЬ ТА ВИБІРКА

Первинний аналіз необхідний для того, щоб надалі ефективно й коректно використовувати ті чи інші методи статистичної обробки даних.

При первинному аналізі проводять такі дії.

***Відсіювання грубих похибок вимірювань*** або помилок. Якщо є підозра, що одне або кілька даних є помилковими (значно відрізняються від решти), необхідно зробити перевірку. Із цією метою використовують критерії Грабса, Кохрена та ін. [3-5].

***Перевірка відповідності розподілу результатів вимірювань нормальному закону.*** Існує багато методів перевірки нормальності розподілу: за середнім абсолютним відхиленням, за розмахом варіювання, за показниками асиметрії та ексцесу, за $\chi^2$-критерієм, за критерієм Колмогорова-Смирнова (К-С-критерій) [3-5]. Необхідний певний досвід для вибору найбільш ефективного методу. Наприклад, методика перевірки нормальності розподілу за показниками асиметрії та ексцесу зручна при проведенні експрес-аналізу. Перевірка за $\chi^2$-критерієм використовується для ґрунтовної перевірки на нормальність. Методика перевірки за розмахом варіювання використовується тільки для наближеної перевірки.

**Генеральна сукупність і вибірка**. Завдання математичної статистики полягає в тому, щоб на підставі знання деяких властивостей (виду розподілу, числових характеристик) підмножини

елементів, узятих із деякої множини, висловити деякі твердження про властивості цієї множини – так званої ***генеральної сукупності***. Інакше кажучи, уся множина об'єктів, що підлягає дослідженню й контролю, є генеральною сукупністю. Зазвичай генеральна сукупність містить скінченну множину об'єктів, але вона, як правило, достатньо велика. При теоретичних висновках обсяг генеральної сукупності припускається *нескінченним.*

Повне дослідження генеральної сукупності, як правило, практично неможливе або економічно недоцільне (перепис населення проводиться один раз на 10 років), тому досліджують не всю генеральну сукупність, а тільки її частину, що називається *вибіркою*.

Під ***випадковою вибіркою обсягу n*** розуміють вибір випадковим чином незалежно один від одного $n$ об'єктів із генеральної сукупності. Результатом випадкової вибірки обсягу $n$ є сукупність $(x_1,\ldots, x_n)$ *значень ознаки* (якісної або кількісної). Значення $x_i$ вибірки називають *варіантами*.

Вибірка має достатньо повно відбивати особливості всіх об'єктів генеральної сукупності, щоб оцінки були вірогідними, тобто вона має бути *репрезентативною* (*представницькою*). Це особливо важливо, якщо генеральна сукупність неоднорідна. Представницьку вибірку можна дістати, якщо вибирати об'єкти випадково, тобто гарантувати всім об'єктам генеральної сукупності однакову ймовірність піддатися дослідженню. Існують *поворотна* (*повторна*) і *безповоротна* (*неповторна*) вибірки, коли об'єкт після дослідження можна або не можна повернути. У першому випадку дістаємо більш незалежну й представницьку вибірку, особливо при обмеженому обсязі.

*Якщо після кожного експерименту досліджуваний елемент не повертають до генеральної сукупності, то не можна вважати, що всі експерименти проводяться в однакових умовах і їхні результати незалежні один від одного, оскільки після кожного експерименту склад генеральної сукупності змінюється. У цьому випадку порушуються умови відтворюваності випробувань.*

Не існує загального критерію, який давав би змогу вирішити, коли вибірка вважається *великою*, а коли *малою*. Адже тоді як розподіл однієї функції вибірки вже при $n = 30$ можна наближено замінити асимптотичним розподілом, для іншої функції вибірки навіть і при $n = 100$ неможливе таке наближення.

При дослідженні об'єктів можна фіксувати або вимірювати значення однієї або кількох ознак. Відповідно говорять про *одновимірну*, *двовимірну*, *багатовимірну* вибірку.

*Часто під генеральною сукупністю розуміють власне досліджувану випадкову величину. Сукупність здобутих при випробуваннях значень у цьому разі також називається вибіркою і обробляється статистично.*

**Варіаційний ряд**. Вибір об'єкта з генеральної сукупності й отримання значення ознаки називається *статистичним спостереженням.* На практиці записують усі результати спостережень у порядку їх отримання у вигляді фактичних результатів або у вигляді відхилень від номінального значення. Вибірка буде більш наочною, якщо всі варіанти її впорядкувати за зростанням.

*Варіаційним рядом* $x_1 \le x_2 \le \ldots \le x_n$ називається вибірка, записана в порядку зростання її варіант. Якщо обсяг вибірки великий,

то доцільно скласти варіаційний ряд за інтервалами значень генеральної сукупності. Для цього всю вибірку (зону розсіювання) поділяють на групи.

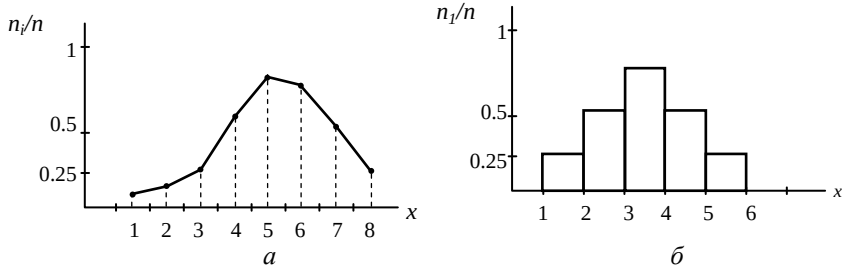
***Зона розсіювання*** (***розмах варіювання***) ***R*** дорівнює різниці між найбільшим і найменшим значеннями величини. Знайдену зону розсіювання поділяють на інтервали, кількість яких рекомендується вибирати в межах від 6 до 14. Як занадто мала, так і занадто велика кількість інтервалів спотворює зовнішній вигляд кривої розсіювання значень.

**Графіки варіаційних рядів.** Використовують два види графіків варіаційних рядів: *полігон* і *гістограму.*

***Полігон*** (***многокутник розподілу***) – це ламана, яку будують, сполучаючи послідовно точки, координатами яких є варіанти $x_i$ та відповідні їм частоти (рис. 2.1, *а*). Вибірки більших обсягів розбивають на рівні інтервали й підраховують для кожного інтервалу *частоту* – кількість спостережень $n_i$, що потрапили в нього. Частоти, віднесені до загальної кількості спостережень $n$, називаються *відносними частотами* (*частостями*). Сума частостей має дорівнювати одиниці:

$$\sum_{i=1}^{m} W_i = \frac{n_i}{n} = 1.$$

***Гістограма*** є графічним поданням розподілу частот за інтервалами. Гістограма будується як множина прямокутників, основою яких є довжина інтервалу, а висота – пропорційна до відповідної частоти (рис. 2.1, *б*).




*а* *б*

Рис. 2.1. Приклад побудови гістограми та полігону частот

За нескінченно великої кількості інтервалів частота $W_i$ є наближенням імовірності потрапляння випадкової величини в $i$-й інтервал:

$$\int_{x_i}^{x_{i+1}} p(x)dx \approx W_i \, .$$

Цю властивість використовують для порівняння теоретичного й емпіричного розподілів. Один із таких способів полягає в графічному зіставленні передбачуваного теоретичного розподілу з емпіричним, тобто *графік щільності* ймовірностей $p(x)$ (*криву розподілу*) зіставляють із *гістограмою.* Якщо передбачуваний теоретичний розподіл добре узгоджується з дослідним, то при досить великому $n$ і вдалому виборі інтервалів гістограма буде близька до кривої розподілу.

*Полігони застосовують зазвичай для характеристики дискретних випадкових величин. Гістограму будують, як правило, для неперервних випадкових величин.*

***Емпірична функція розподілу***. Кожному спостереженню $x_i$ з вибірки обсягом $n$ присвоюють імовірність $p(x_i) = (1/n)$ і дістають розподіл, який називається ***емпіричним***.

Емпіричну функцію цього розподілу $F_n^*(x) \equiv F_n^*(x_1, ..., x_n)$ можна записати у вигляді:

$$F^*(x_i) = \sum_{i=1}^{l} W_i \ ,$$

де $W_i$ – відносна частота варіант $x_i$, які потрапили в інтервал $(-\infty, x)$; $l$ – кількість варіант $x_i$ в даному інтервалі.

За $n \to \infty$ (теорема Глівенка) $F_n^*(x) \to F(x)$.

Емпіричну функцію розподілу можна використовувати як оцінку (наближене значення) функції розподілу $F(x)$ випадкової величини $X$.

Цю властивість можна застосовувати для зіставлення теоретичного розподілу з емпіричним – порівняння функції розподілу $F(x)$, яка для будь-якого $x$ дає ймовірність потрапляння в інтервал $(-\infty, x)$, з емпіричною функцією розподілу функцією $F^*(x)$, яка дорівнює сумі відносних частот тих емпіричних значень $x_i$, що потрапляють у той самий інтервал $(-\infty, x)$.

Функція $F_n^*(x)$ має всі властивості функції розподілу $F(x)$. Аналітичний вигляд цієї функції доволі складний, тому на практиці користуються її графічним поданням.

Графік емпіричної функції розподілу (рис. 2.2) являє собою *східчасту* лінію зі стрибками в точках $x_1$, $x_2,\ldots,x_n$. При великих значеннях $n$ відносні частоти ближчі до відповідних імовірностей, і, отже, буде незначна розбіжність між теоретичною та емпіричною функціями розподілу.

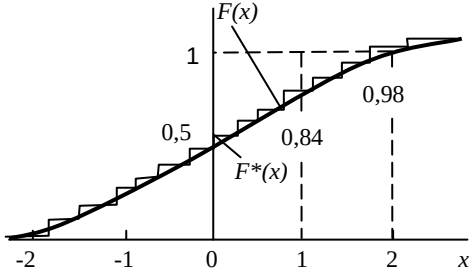



Рис. 2.2. Емпірична функція розподілу

*При виявленні значних розбіжностей між передбачуваним теоретичним і емпіричним розподілами виникає сумнів у правильності передбачення виду розподілу $F(x)$. Для розв'язання питання про припустимість тих чи інших розбіжностей існують критерії згоди, застосування яких потребує великої кількості дослідних даних.*

Характеристики теоретичних розподілів можна розглядати як характеристики, що існують у генеральній сукупності (їх називають ***параметрами***), а характеристики емпіричних розподілів – як вибіркові характеристики (їх називають ***оцінками***, **або *статистиками***).

Теоретичні характеристики (параметри) позначають великими латинськими буквами або буквами грецького алфавіту, вибіркові характеристики (оцінки) – відповідними буквами латинського алфавіту або символами $\hat{\eta}, \hat{\xi}$, іноді із зазначенням обсягу вибірки $\hat{\eta}(i), \hat{\xi}(i)$ (табл. 2.1).

Таблиця 2.1. Теоретичні та вибіркові характеристики

| № з/п | Найменування величини | Теоретичні характеристики (параметри) | Вибіркові характеристики (оцінки) |
|---|---|---|---|
| 1 | Середні значення випадкової величини | $MX$, $M\xi$, $M < x >$, $\mu$, $\nu$ | $\overline{x}$ |
| 2 | Дисперсія | $\sigma^2$, $DX$, $D\xi$, $< \sigma >$ | $S^2$ |
| 3 | Середнє квадратичне відхилення | $\sigma$ | $S$ |
| 4 | Коефіцієнт варіації | $\gamma$ | $v$ |
| 5 | Коефіцієнт кореляції | $\rho$ | $r$ |
| 6 | Залежна змінна | $Y$, $\eta$ | $y$ |
| 7 | Незалежна змінна | $X$, $\xi$ | $x$ |

## Запитання для самоперевірки

1. Що називають генеральною сукупністю й вибіркою?
2. Який обсяг передбачається мати генеральній сукупності, а який – вибірці?
3. Яка вибірка називається репрезентативною?
4. Що таке варіаційний ряд і для чого він будується? Що називається розмахом варіювання?
5. Які ви знаєте графіки варіаційних рядів? Для чого вони будуються?

6. Що називається емпіричною функцією розподілу? Як вона будується?

7. У чому полягає різниця між емпіричною функцією розподілу та функцією розподілу?

8. Чому теоретичні розподіли можуть не збігатися з емпіричними?

9. Яку назву мають теоретичні та вибіркові характеристики розподілу? Як вони позначаються?

## 2.2. ТОЧКОВІ ОЦІНКИ НЕВІДОМИХ ПАРАМЕТРІВ РОЗПОДІЛІВ ТА ЇХ ВЛАСТИВОСТІ

Скориставшись статистичною інформацією, яка міститься у вибірці, потрібно зробити статистичні висновки про справжнє значення невідомого параметра.

Нехай $(X_1,\ldots,X_n)$ – генеральна сукупність випадкової величини. Тоді математичне сподівання випадкової величини $X$ буде

$$MX = \frac{1}{n}\sum_{i=1}^{n} X_i \ .$$

Центр розподілу для конкретної вибірки обсягом $n$ називається ***емпіричним*** (***вибірковим***) ***середнім вибірки*** ($x_1,\ldots, x_n$):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ . \qquad (2.1)$$

Дисперсія випадкової величини може бути визначена як

$$DX = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - MX\right)^2 \ ,$$

для вибірки обчислене значення

$$\tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \qquad (2.2)$$

називається ***емпіричною*** (***вибірковою***) ***дисперсією вибірки*** ($x_1,\ldots, x_n$).

Для ***вибіркового момента порядку k*** справджується рівність:

$$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k \ .$$

Як характеристики положення центра розподілу (групування) на практиці поряд із середнім арифметичним іноді використовуються і інші характеристики, серед яких є *мода* і *медіана*.

***Модою*** називається найбільш імовірне значення випадкової величини й позначається символом *Мо*.

***Медіаною*** називається варіанта, яка поділяє варіаційний ряд на дві частини (серединний елемент ряду) і позначається символом *Ме*.

*Середнє значення, на відміну від моди, може бути спотворене кількома крайніми значеннями. Тому середнє можна брати за оцінку центра розподілу, якщо відомий обсяг вибірки або зазначено якісь межі для цього значення.*

**Приклад**. Розглянемо розподіл доходів у родинах (рис. 2.3). З рис. 2.3 бачимо, що кілька високих доходів роблять середнє значення доходу оманливо високим. Тим часом мода більш об'єктивно характеризує центр групування доходів.

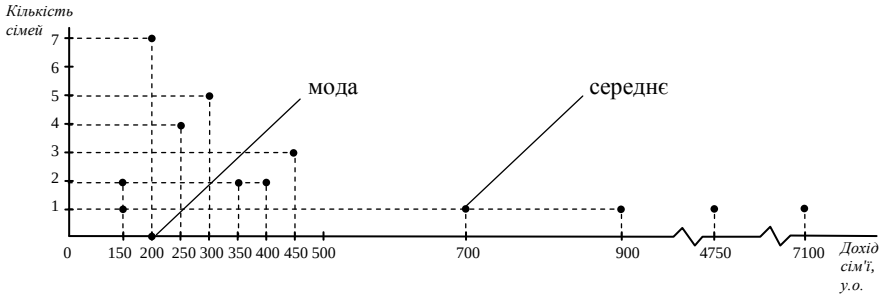



Рис. 2.3. Приклад з застосуванням моди та медіани

**Приклад**. Середня температура по лікарні 36,6. Такий показник шляхом розрахунку середнього значення можна дістати, але він не характеризує стану хворих у лікарні.

**Приклад**. Якби засновник компанії «Майкрософт» Білл Гейтс (статок оцінюється в $54 млрд) жив у місті, де було загалом 54 000 мешканців, кожний з яких не мав би ні цента, то при підрахунку середнього доходу мешканців цього міста з'ясувалося б, що цей дохід дорівнює 1 млн.

**Властивості точкових оцінок**. Як було розглянуто раніше, генеральна сукупність має деякі *постійні числові характеристики* розподілу. За вибіркою можна знайти *оцінки* цих характеристик.

Позначимо невідомий параметр розподілу, тобто числову характеристику генеральної сукупності $X$, через $\theta$, а оцінку невідомого параметра – через $\hat{\theta}$.

Оцінка $\hat{\theta}$ – функція від вибірки, реалізація якої могла б розглядатися як наближення $\theta$. Така функція вибірки називається *точковою оцінкою* $\hat{\theta}$, оскільки вона визначає одну точку на числовій осі.

Оцінки невідомого параметра можна знаходити різними способами. Наприклад, якщо потрібно оцінити істинне значення $\theta = MX$ нормального розподілу, то можна використовувати такі оцінки $\hat{\theta}$:

- перший елемент вибірки $x_1$. На практиці часто так і діють: вимірюють якусь величину тільки один раз і цей результат використовують як оцінку середнього значення цієї величини;
- середнє арифметичне максимального й мінімального елементів вибірки $(x_{max} + x_{min}) / 2$;
- моду *Mo*, що при нормальному розподілі дорівнює середньому значенню $\overline{x}$;

- медіану *Me*, що при нормальному розподілі також дорівнює середньому значенню $\bar{x}$;
- середнє арифметичне вибірки $\bar{x}$.

Для того щоб установити, яка з оцінок краща, потрібно виходити з їхніх основних властивостей: *спроможності*, *незміщеності та ефективності*. Розглянемо ці властивості.

Оцінка $\hat{\theta}$ параметра θ називається ***спроможною***, якщо вона збігається за ймовірністю до параметра θ, тобто для будь-якого як завгодно малого додатного числа ε виконується рівність:

$$\lim_{n\to\infty} P\left(\left|\hat{\theta}-\theta\right|\le\varepsilon\right)=1,$$

що означає: розсіювання випадкової величини *X* в околі *MX* не перевищує величини ε.

**Приклад**. Спроможною оцінкою *MX* є $\bar{x}$, $\tilde{S}^2$ – спроможна оцінка для $DX = \sigma^2$, відносна частота випадкової події $W_i$ – спроможна оцінка її ймовірності $P(x_i)$.

Цінність спроможної оцінки для практики полягає в тому, що, збільшуючи кількість дослідів, можна більш точно оцінити параметр. Спроможність – *неодмінна властивість* використовуваних оцінок.

Оцінка називається ***незміщеною***, якщо її математичне сподівання (середнє значення) дорівнює оцінюваному параметру (відсутня систематична похибка):

$$M\hat{\theta}=\theta.$$

Якщо ця умова не виконується, оцінку називають *зміщеною*, при цьому зсув обчислюється як різниця $M\hat{\theta}-\theta$.

**Приклад**.

$$M\left(\bar{X}\right)=M\left(\frac{1}{n}\sum_i X_i\right)=\frac{1}{n}\sum_{i=1}^{n}MX=\frac{1}{n}nMX=MX$$

є незміщеною оцінкою *MX* (оцінка математичного сподівання дорівнює математичному сподіванню).

**Приклад.** Вибіркова дисперсія $\tilde{S}^2$ є оцінкою дисперсії $\sigma^2$

$$M\tilde{S}^2 = \frac{1}{n} M\left( \sum_{i=1}^{n} X_i - \overline{X} \right)^2 .$$

Можна показати, що

$$M\tilde{S}^2 = \frac{n-1}{n}\sigma^2 ,$$

тобто, вибіркова дисперсія $\tilde{S}^2$ є зміщеною оцінкою для $\sigma^2$. Для того, щоб оцінка була незміщеною, вводять виправлення Шепарда, тобто замість

$$\tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

розглядають

$$S^2 = \frac{n}{n-1}\tilde{S}^2 .$$

У цьому разі

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 . \qquad (2.3)$$

Таким чином, $S^2$ є *незміщеною вибірковою дисперсією* або говорять, *незміщеною оцінкою дисперсії*.

У цій формулі знаменник (*n*-1) називається ***числом степенів вільності.*** Кількість степенів вільності *f* буде дорівнювати різниці між кількістю рівнянь (значень), на підставі яких було обчислено ця оцінку мінус кількість констант, які необхідні для обчислення цієї оцінки, які було визначено на підставі тих самих *n* рівнянь (значень). Наприклад, якщо у виразі (2.3) відомо математичне сподівання (замість $\overline{X}$ у формулі використовується *MX*), то не втрачається число степенів вільності:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - MX\right)^2 .$$

Незміщеність є *бажаною* властивістю. Багато оцінок властивості незміщеності не мають. Зміщення оцінки може призвести до неправильних висновків, оскільки крім впливу випадкових величин може мати місце систематичне відхилення, яке на рис. 2.4 подано у вигляді зсуву.

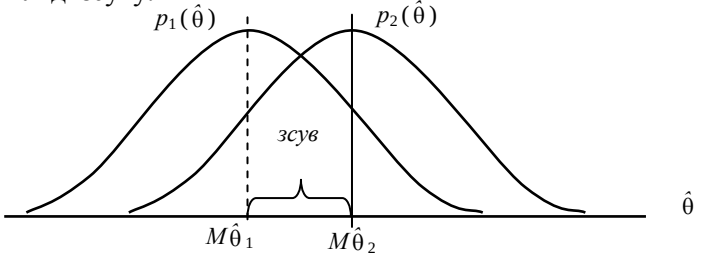



Рис. 2.4. Зсув розподілу

На рис. 2.4 $\hat{\theta}$ – узагальнена статистична оцінка. Для випадку незміщеної оцінки щільність імовірності $p_1(\hat{\theta})$ і $M\hat{\theta}_1 = M\theta$; для випадку з зміщеною оцінкою щільність імовірності $p_2(\hat{\theta})$ і $M\hat{\theta}_2 \neq M\theta$.

$M\hat{\theta}$ – математичне сподівання оцінки (збігається з математичним сподіванням генеральної сукупності).

*Вимога незміщеності особливо важлива за малого обсягу вибірки. При цьому використовують* $S^2$, *а за значного обсягу вибірки –* $\tilde{S}^2$.

Оцінка, дисперсія якої набуває мінімального значення, називається ***ефективною***.

Таким чином, оцінка $\hat{\theta}_1$ параметра $\theta$ називається *більш ефективною*, ніж оцінка $\hat{\theta}_2$, якщо $M(\hat{\theta}_1 - \theta)^2 \leq M(\hat{\theta}_2 - \theta)^2$. Якщо обидві оцінки незміщені, це означає, що

$$D\hat{\theta}_1 \leq D\hat{\theta}_2.$$

Вибіркова оцінка для цієї вибірки є конкретним числом, але це число є одним із можливих значень оцінки, тобто якщо візьмемо нову вибірку, то наявні в ній значення будуть відрізнятися хоча б на один елемент від значень попередньої вибірки, а отже, і обчислена оцінка буде відрізнятися від попередньої. Така ситуація буде спостерігатися й для інших наступних вибірок.

*Таким чином, згідно з вимогами практичної цінності*:
*– оцінка має наближатися до оцінюваного параметра зі збільшенням обсягу вибірки*;
*– оцінка не повинна містити систематичної похибки* (*зсуву*), *це означає, що її математичне сподівання* $M\hat{\theta} = \theta$ *має збігатися з оцінюваним параметром* $\theta$;
*– із усіх спроможних і незміщених оцінок найбільш ефективна та, яка має найменшу дисперсію.*

**Запитання для самоперевірки**

1. Дайте означення вибірковому середньому, вибірковій дисперсії.

2. Що називається модою і медіаною?

3. Що називається точковою оцінкою параметра розподілу? Наведіть приклади точкових оцінок.

4. Які властивості повинні мати точкові оцінки, щоб бути достовірними? 6. Яка оцінка називається спроможною? Наведіть приклади.

5. Яка оцінка називається незміщеною? Наведіть приклад зміщеної та незміщеної оцінок. Це обов’язкова чи бажана вимога?

6. Як визначити найбільш ефективну оцінку?

7. Що називається кількістю степенів вільності?

8. Чому вибіркові характеристики однієї вибірки відрізняються від вибіркових характеристик іншої вибірки тієї самої генеральної сукупності?

# 2.3. ОБЧИСЛЕННЯ ВИБІРКОВИХ ХАРАКТЕРИСТИК

Залежно від обсягу вибірки може бути рекомендована різна методика обчислення параметрів вибірки

Якщо обсяг вибірки невеликий, то для обчислення вибіркового середнього $\bar{x}$ і вибіркової дисперсії $S^2$ користуються формулами (2.1) та (2.2).

Якщо складено варіаційний ряд, для обчислення середнього арифметичного варто використовувати формулу:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{m} x_i n_i \,, \tag{2.4}$$

де $x_i$ – варіанти вибірки; $n_i$ – кількість значень $x_i$; $m$ – кількість варіант; $n$ – обсяг вибірки.

Для подальшого спрощення обчислень застосовують *метод умовних варіант.* Уведемо умовну варіанту

$$\bar{x}^* = \frac{\sum_i (x_i - C)}{n}, \tag{2.5}$$

де $C$ – умовний нуль.

Найбільша простота обчислень досягається тоді, коли за умовний нуль взяти варіанту, яка розміщена приблизно посередині варіаційного ряду, і яка має найбільшу частоту. Легко знайти, що

$$\bar{x} = \bar{x}^* + C \,. \tag{2.6}$$

У такий спосіб простіше обробляти результати, тому що обробці підлягають невеликі значення. Цей спосіб є корисним при обробці масивів даних з великою кількістю значущих цифр. Обробка таких масивів супроводжується похибкою округлення через обмежену розрядність подання результатів. У цьому випадку масив перетворюють до меншої розрядності, виключивши умовний нуль й потім відповідно до виразу (2.5) обчислюють середнє спрощеного

масиву. Після цього істинне середнє визначають на підставі виразу (2.6).

Якщо кількість варіант велика, для спрощення обчислень застосовують *групування емпіричних даних*, тобто варіаційний ряд складається за інтервалами значень. Тоді оцінки обчислюються за формулами:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{m} k_i n_i ,$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{m} n_i k_i^2 - \bar{x}^2 , \qquad (2.7)$$

де $k_i$ – серединне значення $i$-го інтервалу.

***Обчислення моди***. Для дискретних статистичних рядів

$$Mo = x_j, \text{ якщо } n_j = \max_i n_i .$$

Якщо варіаційний ряд складено за інтервалами значень генеральної сукупності, мода обчислюється за такою наближеною формулою:

$$Mo = x_{Mo} + k\frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i - n_{i+1})}, \qquad (2.8)$$

де $x_{Mo}$ – початок модального інтервалу, тобто інтервалу, що має максимальну частоту; $k$ – довжина модального інтервалу; $n_i$ – частота модального інтервалу; $n_{i-1}$ і $n_{i+1}$ – частоти відповідно попереднього та наступного за модальним інтервалів.

***Обчислення медіани***. Якщо варіаційний ряд складено за значеннями генеральної сукупності, при непарному обсязі вибірки медіана – це дійсне значення серединного елемента, а при парному – середнє арифметичне двох серединних елементів, тобто

$$Me = \begin{cases} x_m, & n = 2m-1 \\ \dfrac{x_m + x_{m+1}}{2}, & n = 2m \end{cases}.$$

Якщо варіаційний ряд складено за інтервалами значень, медіана обчислюється за такою наближеною формулою:

$$Me = x_{Me} + k\frac{n/2 - T_{i-1}}{n_i}, \qquad (2.9)$$

де $x_{Me}$ – початок медіанного інтервалу, тобто інтервалу, в якому перебуває серединний елемент; $k$ – довжина медіанного інтервалу; $n$ – обсяг вибірки; $T_{i-1}$ – сума частот інтервалів, що передують медіанному; $n_i$ – частота медіанного інтервалу.

**Оцінка центра розподілу й розсіювання при нерівноточних спостереженнях**. Раніше розглянуті властивості точкових оцінок стосуються *рівноточних* вимірювань і спостережень, тобто вимірювань, які містять тільки випадкову складову.

У практичній діяльності трапляються випадки, коли необхідно спільно обробляти різні вибірки, наприклад для найбільш надійного визначення деякої величини використовують результати вимірювань різного походження, виконані різними приладами й методами, тобто вони можуть мати різні дисперсії.

Припустимо, що значення $x_1$, $x_2$,..., $x_n$ здобуто при дослідженні того самого об'єкта, але з різною точністю, тобто

$$Mx_1 = Mx_2 = ... = Mx_n = a, \text{ але } Dx_1 \neq Dx_2 \neq ... \neq Dx_n.$$

Якщо результати двох груп спостережень мають різні дисперсії, такі спостереження називаються ***нерівноточними***.

Для того щоб знайти центр розподілу нерівноточних спостережень, необхідно використати ***зважену оцінку середнього***

$$\bar{x} = g_1x_1 + ... + g_ix_i + ... + g_nx_n, \qquad (2.10)$$

де $g_i$ , $i=\overline{1,N}$ – ваговий коефіцієнт (визначає вагу кожного з результатів) у визначенні середнього. Для того щоб знайдена оцінка центра розподілу була незміщеною, тобто за будь-якого $a$ $M\bar{x} = a$, необхідно, щоб виконувалася умова:

$$a = ag_1 + ag_2 + ... + ag_n = a(g_1 + g_2 + ... + g_n).$$

Таким чином, умова незміщеності оцінки

$$\sum_{i=1}^{N} g_i = 1. \qquad (2.11)$$

Враховуючи наведене, а також з огляду на те, що вимірювання проводилися незалежно і $g_i$ = const, можна записати вираз для дисперсії середнього:

$$D\bar{x} = D(g_1 x_1) + D(g_2 x_2) + ... + D(g_n x_n) =$$
$$= g_1^2\sigma_1^2 + g_2^2\sigma_2^2 + ... + g_N^2\sigma_N^2 == \sum_{i=1}^{N} g_i^2\sigma_i^2 \quad . \qquad (2.12)$$

З урахуванням виразу (2.11) вираз (2.12) можна переписати у такий спосіб:

$$D\bar{x} = g_1^2\sigma_1^2 + g_2^2\sigma_2^2 + ... + g_{N-1}^2\sigma_{N-1}^2 + (1 - g_1 - g_2 - ... - g_{N-1})^2\sigma_N^2. \quad (2.13)$$

Висунемо умову, щоб оцінка $\bar{x}$ була ефективною, тобто щоб середнє квадратичне відхилення оцінки від її математичного сподівання, знайдене за нерівноточними спостереженнями, було мінімальним.

Для визначення умов, за яких буде забезпечуватися мінімум, розглянемо частинні похідні $\frac{\partial D\bar{x}}{\partial g_i}$, $i = \overline{1, N-1}$ для виразу (2.13) і знайдемо ті значення $\sigma_i$, за яких частинні похідні дорівнюють нулю.

Узявши похідні й прирівнявши їх до нуля, дістанемо:

$$2g_i\sigma_i^2 - 2(1 - g_1 - ... - g_{N-1})\sigma_N^2 = 0.$$

Оскільки вираз у дужках дорівнює $g_N$, маємо:

$$g_i\sigma_i^2 = g_N\sigma_N^2, \quad i = \overline{1, N-1}.$$

Таким чином, доходимо висновку, що *вагові коефіцієнти мають бути обернено пропорційними до дисперсій спостережень*:

$$g_1 : g_2 : \ldots : g_N = \frac{1}{\sigma_1^2} : \frac{1}{\sigma_2^2} : \ldots : \frac{1}{\sigma_N^2}. \qquad (2.14)$$

Для виконання умови нормування, достатньо поділити кожне значення $g_i$ на їхню загальну суму:

$$\frac{g_i}{\sum_{i=1}^{N} g_i},$$

звідки випливає, що

$$\sum_{i=1}^{N} \frac{g_i}{\sum_{i=1}^{N} g_i} = 1,$$

що відповідає умові (2.11).

Таким чином, відповідно до формул (2.10) та (2.13) одержуємо

$$\bar{x} = \frac{\sum_{i=1}^{N} g_i x_i}{\sum_{i=1}^{N} g_i}; \quad \sigma_{\bar{x}}^2 = \frac{\sum_{i=1}^{N} g_i^2 \sigma_i^2}{\left(\sum_{i=1}^{N} g_i\right)^2}. \qquad (2.15)$$

Виходячи зі співвідношення (2.14), доходимо висновку, що

$$g_1\sigma_1^2 = g_2\sigma_2^2 = \ldots = g_N\sigma_N^2 = \sigma^2 ,$$

де $\sigma^2$ – дисперсія деякого «фіктивного» спостереження з ваговим коефіцієнтом, що дорівнює одиниці. Величина $\sigma$ називається ***середнім квадратичним відхиленням усередненого спостереження***.

Підставимо останнє співвідношення у вираз (2.15) для $\sigma_{\bar{x}}^2$:

$$\sigma_{\bar{x}}^2 = \sigma^2 \frac{\sum_{i=1}^{N} g_i}{\left(\sum_{i=1}^{N} g_i\right)^2} = \frac{\sigma^2}{\sum_{i=1}^{N} g_i}. \qquad (2.16)$$

Для оцінки $\sigma^2$ використовується деяка ***середньозважена дисперсія***, обчислена на підставі всіх наявних $N$ результатів:

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N} g_i \left(x_i - \bar{x}\right)^2 . \qquad (2.17)$$

Таким чином, згідно з формулою (2.16) вираз для обчислення оцінки дисперсії середнього, знайденого при нерівноточних спостереженнях, набирає вигляду:

$$s_{\bar{x}}^2 = \frac{s^2}{\sum_{i=1}^{N} g_i} .$$

*Перш ніж спільно обробляти результати кількох вибірок, необхідно перевіряти їх на однорідність, тобто на належність до однієї генеральної сукупності.*

**Запитання для самоперевірки**

1. У яких випадках застосовують групування даних? Як визначається середнє арифметичне значення вибірки, складеної за варіаційним рядом?

2. Як обчислити вибіркове середнє інтервального ряду?

3. Як обчислити вибіркову дисперсію інтервального ряду?

4. Як обчислюється мода для дискретних і інтервальних статистичних рядів?

5. Як обчислюється медіана для дискретних і інтервальних статистичних рядів?

6. Які спостереження називаються нерівноточними?

7. Що таке зважене значення середнього арифметичного? Як вибираються вагові коефіцієнти?

8. Що є оцінкою дисперсії «фіктивного» спостереження?

9. Як мінімізувати дисперсію при нерівноточних спостереженнях?

## 2.4. ІНТЕРВАЛЬНІ ОЦІНКИ

Метою статистичного оцінювання параметрів є знаходження за вибірковими даними найменших інтервалів, які із заданою ймовірністю накривають оцінювані параметри.

Розглянуті раніше точкові оцінки є тільки деякими з можливих значень оцінюваного параметра, й за ними не можна судити про точність і надійність здобутого результату.

Усі точкові оцінки параметрів розподілу генеральної сукупності обчислюють за вибірками, але через випадковість значень, які потрапили у вибірку, оцінки є випадковими величинами, що відрізняються від детермінованого істинного значення параметра $\theta$.

Результат необхідно характеризувати можливим відхиленням оцінки від параметра й імовірністю виконання цієї умови.

Припустимо, необхідно оцінити деякий параметр $\theta$ випадкової величини (математичне сподівання, дисперсію). За результатами спостережень дістаємо оцінку $\hat{\theta}$. Можна завжди вказати ймовірність

$$P\left(\left|\hat{\theta}-\theta\right|<\varepsilon\right)=1-\alpha, \qquad (2.18)$$

з якою відхилення оцінки від її математичного сподівання не буде перевищувати деякого додатного числа $\varepsilon$. Ця ймовірність дорівнює $1-\alpha$, де $\alpha$ – будь-яке число у межах від 0 до 1.

Розглянемо вираз $\left(\left|\hat{\theta}-\theta\right|<\varepsilon\right)$. Його можна подати у вигляді:

$$-\varepsilon<\hat{\theta}-\theta<+\varepsilon$$

або

$$\hat{\theta}-\varepsilon<\theta<\hat{\theta}+\varepsilon,$$

де $\hat{\theta}-\varepsilon$ і $\hat{\theta}+\varepsilon$ – границі інтервалу; $\theta$ – значення параметра.

Інтервал $\hat{\theta}-\varepsilon,\hat{\theta}+\varepsilon$, що із заданою ймовірністю $P_{\text{дов}}=1-\alpha$, *накриває* істинне значення параметра $\theta$, називається ***надійним інтервалом***. Кінці цього інтервалу випадкові, оскільки отримана за результатами спостереження оцінка $\hat{\theta}$ є одним із можливих її значень. Для іншої вибірки оцінка $\hat{\theta}$ буде іншою.

Отже, доходимо висновку, що надійний інтервал матиме випадкові розміри (рис. 2.5), а ймовірність того, що цей інтервал накриє значення параметра $\theta$, дорівнює $1-\alpha$ і називається ***надійною ймовірністю.***

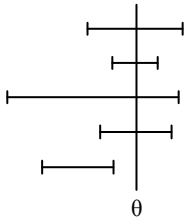



Рис. 2.5. Представлення надійного інтервалу

**Знаходження інтервальних оцінок для центра розподілу при відомому** $\sigma$. Нехай задана генеральна сукупність – випадкова величина *X*, що має математичне сподівання $\nu$ й дисперсію $\sigma^2$. Значення стандартного відхилення випадкової величини $\sigma$ відоме.

Завдання полягає в тому, щоб за результатами прямих спостережень знайти оцінку $\nu$.

Згідно зі співвідношенням (2.18) маємо:

$$P\left(\hat{\theta}-\varepsilon\le\theta\le\hat{\theta}+\varepsilon\right)=1-\alpha\,.$$

За результатами спостережень можна визначити тільки середнє арифметичне:

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i\,.$$

Як було показано у розділі 1, дисперсія середнього арифметичного в $n$ разів менше дисперсії випадкової величини. Оскільки дисперсія випадкової величини вважається відомою, то виходячи із кількості спостережень, результати яких спільно обробляються, можна обчислити дисперсію середнього арифметичного:

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}.$$

У цьому разі можна розглядати нормоване відхилення середнього значення від його математичного сподівання, описуване функцією Лапласа, для якого існують табульовані значення:

$$z = \frac{\bar{x} - \nu}{\sigma / \sqrt{n}}.$$

Таким чином, можна записати:

$$P\left(|z| \le \varepsilon\right) = 1 - \alpha$$

або

$$P\left( -\varepsilon \le \frac{\bar{x} - \nu}{\sigma / \sqrt{n}} \le +\varepsilon \right) = 1 - \alpha \,. \qquad (2.19)$$

Цей запис означає, що нормоване відхилення середнього значення $\bar{x}$ від істинного $\nu$ з імовірністю $1-\alpha$ не перевищуватиме значень $-\varepsilon$; $+\varepsilon$ (рис. 2.6).

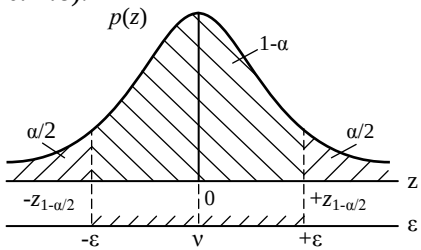



Рис. 2.6. Нормоване відхилення середнього значення від його математичного сподівання

Вираз (2.19) можна подати у вигляді:

$$P\left(-\varepsilon\cdot\frac{\sigma}{\sqrt{n}}\le\bar{x}-\nu\le+\varepsilon\frac{\sigma}{\sqrt{n}}\right)=1-\alpha\,. \qquad (2.20)$$

Для $z$ граничні значення $-\varepsilon$ і $+\varepsilon$ являють собою *квантилі* розподілу Лапласа, визначені для ймовірності $1-\alpha$. З огляду на це вираз (2.19) можна записати так:

$$P\left(-z_{1-\alpha/2}\le z\le z_{1-\alpha/2}\right)=1-\alpha\,,$$

що дає змогу визначити $|\varepsilon|=|z_{1-\alpha/2}|\,\frac{\sigma}{\sqrt{n}}$. Тоді вираз (2.20) можна записати у вигляді:

$$P\left(\bar{x}-z_{1-\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\le\nu\le\bar{x}+z_{1-\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right)=1-\alpha\,. \qquad (2.21)$$

**Інтервальна оцінка центра розподілу при невідомому $\sigma$.**

У цьому разі нормоване відхилення середнього від математичного сподівання не можна подати функцією Лапласа. Замість невідомого $\sigma$ використовують його незсунену оцінку $s$, обчислену на підставі наявних результатів спостережень:

$$s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}\,.$$

Тоді статистика

$$t=\frac{\bar{x}-\nu}{s/\sqrt{n-1}} \qquad (2.22)$$

підпорядковується закону розподілу Стьюдента ($t$-розподілу). Величина $t$ визначається за таблицями розподілу Стьюдента залежно від $\alpha$ та кількості степенів вільності $n-1$.

*У деяких таблицях наводяться значення для процентних точок $q$, %* ($\alpha = q/100$), *тому потрібно бути уважним при користуванні таблицями.*

Розподілом Стьюдента (рис. 2.7, крива *1*) для знаходження інтервальних оцінок користуються в разі, коли $n < 30...50$. За $n > 50$ розподіл Стьюдента асимптотично наближається до нормального (рис. 2.7, крива *2*), тобто за $n > 50$ можна користуватися таблицями Лапласа. Зазвичай кількість спостережень обмежують, оскільки впродовж тривалого експерименту змінюються умови його проведення.

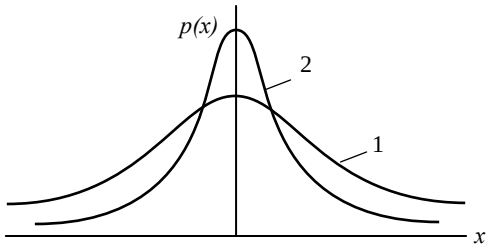



Рис. 2.7. Представлення розподілу Стьюдента

Визначивши значення $t$ за таблицями розподілу Стьюдента, можна записати

$$P\left(-t_{\alpha,n-1} \le \frac{\bar{x}-\nu}{s/\sqrt{n-1}} \le t_{\alpha,n-1}\right) = 1-\alpha,$$

звідки $$P\left(\bar{x} - t_{\alpha,n-1}\cdot\frac{s}{\sqrt{n-1}} \le \nu \le \bar{x} + t_{\alpha,n-1}\cdot\frac{s}{\sqrt{n-1}}\right) = 1-\alpha.$$

У лівій і правій частинах нерівності записано граничні значення надійного інтервалу.

**Надійний інтервал для дисперсії нормального розподілу**. Розглянемо суму квадратів нормально розподілених випадкових величин із нульовим математичним сподіванням і з одиничною дисперсією (центрованих і нормованих):

$$\xi^2 = \sum_{i=1}^{n}\left(\frac{X_i - \bar{X}_i}{\sigma_i}\right)^2. \qquad (2.23)$$

Величина $\xi^2$ матиме $\chi^2$-розподіл Пірсона (рис. 2.8). Для розподілу $\chi^2$ у табличному вигляді задаються критичні значення $\chi^2_{\text{кр}}$, для яких виконується співвідношення:

$$P\left(\chi^2 > \chi^2_{\text{кр}}\right) = \alpha \ . \qquad (2.24)$$

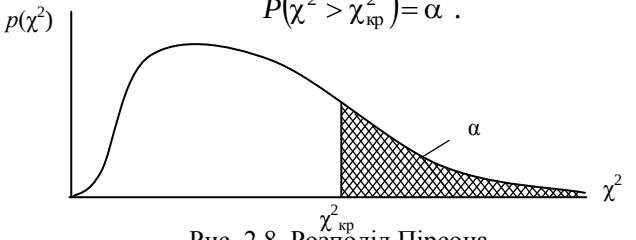



Рис. 2.8. Розподіл Пірсона

Для вибірки значень $x_i$ об'ємом $n$ можна знайти незміщену оцінку СКВ:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)} \, .$$

З останнього виразу випливає, що

$$\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = (n-1)s^2 \, .$$

Підставимо це співвідношення у (2.23) і дістанемо величину

$$\xi^2 = \frac{(n-1)s^2}{\sigma^2} \ , \qquad (2.25)$$

яка має $\chi^2$-розподіл Пірсона з кількістю степенів вільності $n - 1$. Цю статистику використовують для знаходження інтервальних оцінок для дисперсії.

Згідно з означенням надійного інтервалу для величини $\frac{(n-1)s^2}{\sigma^2}$ можна записати $P\left(\chi_1^2 \le \chi^2 \le \chi_2^2\right) = 1-\alpha$ або з урахуванням (2.25):

$$P\left(\chi_1^2 \le \frac{(n-1)s^2}{\sigma^2} \le \chi_2^2\right) = 1-\alpha \, . \qquad (2.26)$$

Для визначення надійного інтервалу для $\sigma^2$ виконаємо обернене до (2.26) перетворення, потім помножимо на $(n-1)s^2$ і, узявши до уваги, що той дріб менший, в якого знаменник більший, дістанемо:

$$P\left(\frac{(n-1)s^2}{\chi_2^2} \le \sigma^2 \le \frac{(n-1)s^2}{\chi_1^2}\right) = 1-\alpha, \qquad (2.27)$$

де в дужках задано межі надійного інтервалу для дисперсії $\sigma^2$. Невідомими тут є $\chi_1^2$ і $\chi_2^2$. За таблицями $\chi^2$-розподілу можна знайти такі два числа $\chi_1^2$ і $\chi_2^2$, які задовольняють умову (2.27). Таких пар чисел існує нескінченна множина. Щоб зафіксувати одну таку пару $\chi_1^2$ і $\chi_2^2$ введемо додаткову умову – симетричність інтервалу за ймовірністю (рис. 2.9).

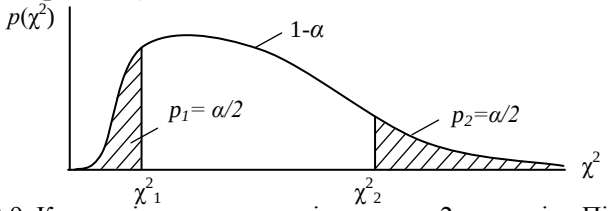



Рис. 2.9. Квантилі симетричного інтервалу χ2-розподілу Пірсона

Зважаючи на те, що $\chi^2$-розподіл Пірсона несиметричний, існують два критичні значення

$$\chi_{\text{кр}}^2 = \chi_1^2 \quad \text{і} \quad \chi_{\text{кр}}^2 = \chi_2^2,$$

для яких критична область визначається згідно з умовою:

$$P\left(\chi^2 \le \chi_1^2\right) = P\left(\chi^2 > \chi_2^2\right) = \frac{\alpha}{2}.$$

Із таблиці $\chi^2$-розподілу Пірсона безпосередньо визначають значення однієї критичної точки $\chi_2^2$. Зважаючи на те, що в таблицях подається значення $\chi_{\text{кр}}^2$, для якого $P\left(\chi^2 \ge \chi_{\text{кр}}^2\right)$, необхідно скористатися умовою для визначення $\chi_1^2$:

$$P\left(\chi^2 \le \chi_1^2\right) = 1 - P\left(\chi^2 > \chi_1^2\right) = 1 - \left(1 - \frac{\alpha}{2}\right) = \frac{\alpha}{2}.$$

Знаючи $\chi_1^2$ і $\chi_2^2$, можна визначити межі надійного інтервалу для СКВ:

$$P\left(\frac{\sqrt{n-1}s}{\chi_2} \le \sigma \le \frac{\sqrt{n-1}s}{\chi_1}\right) = P_{\text{дов}}.$$

**Визначення обсягу вибірки**. Досі ми розглядали обробку вибірок довільного обсягу *n*. Який мінімальний обсяг повинна мати вибірка, щоб отримана оцінка параметра розподілу відрізнялася від його істинного значення на деяке задане $\varepsilon$ з відповідною надійною ймовірністю? За законом великих чисел перевагу варто надати вибіркам із більшим обсягом, але зазвичай більший обсяг вибірки потребує й більших витрат для її одержання та обробки. Іншими словами, варто визначити, яким має бути мінімальний обсяг вибірки, щоб

$$P\left(\left|\hat{\theta} - \theta\right| \le \varepsilon\right) = P_{\text{дов}}. \qquad (2.28)$$

Розглянемо випадок, коли як $\theta$ береться середнє арифметичне значення.

Середньоквадратичне значення $\sigma$ для генеральної сукупності відоме і визначене при попередніх дослідженнях. За результатами поточних вимірювань $x_i$ знаходять середнє значення $\bar{x}$. Згідно з центральною граничною теоремою середнє значення матиме нормальний закон розподілу незалежно від вихідного розподілу генеральної сукупності. Запишемо вираз (2.28) для нормованої випадкової величини для випадку $\theta = MX$ і $\hat{\theta} = \bar{x}$. Тоді

$$P\left(\left|\frac{\bar{x} - MX}{\sigma_{\bar{x}}}\right| \le z_{1-\alpha/2}\right) = 1 - \alpha. \qquad (2.29)$$

Зіставивши (2.28) і (2.29) можна записати, що припустиме відхилення середнього від математичного сподівання з імовірністю 1– α не перевищує

$$\varepsilon = \sigma z_{1-\alpha/2}.$$

Узявши до уваги, що дисперсія середнього в *n* раз менша за вихідну дисперсію, можна записати:

$$\varepsilon = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Тоді мінімальний обсяг вибірки, тобто мінімальна кількість спостережень $n_{\min}$, за якої отримане середнє буде відрізнятися від істинного значення не більше ніж на $\varepsilon$ з імовірністю $1-\alpha$, набирає вигляду:

$$n_{min} = \frac{z_{1-\alpha/2}^2\sigma^2}{\varepsilon^2}.$$

**Запитання для самоперевірки**

1. Що таке довірчий інтервал?
2. Що таке довірча ймовірність?
3. Як знайти інтервальну оцінку для центра розподілу при відомому $\sigma$?
4. Як знаходять квантилі нормального розподілу?
5. Як знайти інтервальну оцінку для центра розподілу при невідомому $\sigma$.
6. Як користуватися таблицями *t*-розподілу?
7. Як знайти довірчий інтервал для дисперсії нормального розподілу?
8. У якому випадку користуються *z*, *t* та $\chi^2$ - статистиками?
9. Як за таблицями $\chi^2$-розподілу Пірсона визначати критичні значення $\chi^2$?
10. Що потрібно знати, щоб обчислити необхідний мінімальний об’єм вибірки?

# РОЗДІЛ 3

# СТАТИСТИЧНА ПЕРЕВІРКА ГІПОТЕЗ

## 3.1. ПОНЯТТЯ СТАТИСТИЧНОЇ ГІПОТЕЗИ

При опрацюванні випадкових величин на відміну від детермінованих, можна зробити тільки припущення щодо їхнього походження, тобто висунути гіпотезу.

**Статистичною гіпотезою** називається будь-яке припущення щодо виду або властивостей розподілу спостережуваних в експерименті випадкових величин.

Сутність перевірки статистичної гіпотези полягає у тому, щоб встановити, узгоджуються чи ні результати спостережень та висунута гіпотеза, чи можна розбіжність між гіпотезою та результатом вибіркових спостережень віднести за рахунок похибки, зумовленої механізмом випадкового відбору.

Гіпотези бувають прості і складні.

***Проста гіпотеза*** має тільки одне твердження.

**Приклад.** Параметр $\theta$ має одне конкретне значення ($\theta = \theta_0$).

***Складна гіпотеза*** складається з безлічі простих гіпотез.

**Приклад**. Параметр $\theta$ має деяке значення із сукупності

$$(\theta < \theta_0;\ \theta > \theta_0;\ \theta \neq \theta_0.).$$

Гіпотезу, що висувають, називають ***основною гіпотезою.*** У зв'язку з тим, що вона найчастіше полягає у передбаченні відсутності систематичної розбіжності (нульової розбіжності) між невідомим параметром генеральної сукупності й заданою величиною, то її також називають ***нульовою гіпотезою*** і позначають $\boldsymbol{H_0}$**.** Зміст гіпотези записують після двокрапки. Так, у розглянутому прикладі $H_0$: $\theta = \theta_0$.

Зазвичай формулюють ще й ***альтернативну*** (***конкуруючу***) ***гіпотезу*** $\boldsymbol{H_1}$. У результаті перевірки можна приймати тільки одну з гіпотез $\boldsymbol{H_0}$ або $\boldsymbol{H_1}$, відхиляючи водночас іншу.

Гіпотезу перевіряють статистичними методами на підставі *вибірки*, отриманої з генеральної сукупності. Через випадковість відхилень значень у вибірці в результаті перевірки можуть виникати помилки й прийматися неправильне рішення. Коли рішення не відповідає справжньому стану, можуть виникнути помилки двох видів.

**Помилка першого роду** відбувається тоді, коли відхиляється правильна гіпотеза $H_0$.

**Помилка другого роду** полягає в тому, що приймається гіпотеза $H_0$, коли правильною є альтернативна гіпотеза.

При перевірці гіпотез може виникнути одна із чотирьох ситуацій:

- гіпотеза $H_0$ правильна й вона приймається;
- правдива гіпотеза $H_0$, однак приймається неправильна гіпотеза $H_1$ (помилка першого роду);
- правдива гіпотеза $H_1$, однак приймається неправильна гіпотеза $H_0$ (помилка другого роду);
- гіпотеза $H_1$ правильна й вона приймається.

Здебільшого наслідки зазначених помилок нерівнозначні. Одна з помилок приводить до більш обережного, консервативного рішення, інша – до невиправданих дій. Що краще, що гірше – залежить від конкретної постановки задачі і змісту нульової гіпотези. Природним є прагнення зменшити втрати від обох помилок одночасно. Але, як ми побачимо далі, вони є конкуруючими, і зменшення ймовірності появи однієї з них призводить до збільшення ймовірності появи іншої. Таким чином, необхідно вибирати компромісне рішення. Ухвалення рішення ґрунтується на деякому *статистичному критерії.*

**Статистичним критерієм** називається певне *правило обробки* статистичного матеріалу (результатів дослідження), на підставі якого одна з гіпотез приймається, а всі інші відхиляються.

Кожен критерій має свою числову характеристику, що називається ***потужністю***. ***Функцією потужності критерію*** називається *ймовірність* того, що основна гіпотеза $H_0$ відхиляється, тоді як альтернативна гіпотеза приймається. Чим більша потужність критерію, тим менша ймовірність здійснення помилки 2-го роду.

Рішення приймається за значенням деякої *функції вибірки*, яка називається ***статистикою***, або ***статистичною характеристикою*** (z, *t*, $\chi^2$).

Для того щоб прийняти або відхилити передбачувану гіпотезу, потрібно вибрати межу *припустимих* при висунутій гіпотезі відхилень від математичного сподівання, тобто призначити таке *критичне відхилення*, перевищення якого при висунутій гіпотезі настільки малоймовірне, що його можна вважати практично неможливим. Якщо воно фактично спостерігалося, то це вказує на несумісність висунутої гіпотези з наявними спостереженнями, або якщо фактичне відхилення менше за критичну межу, є підстави вважати, що результати досліду не суперечать висунутій гіпотезі, а спостережене відхилення від центра розподілу можна пояснити впливом випадкових величин.

Множину значень обраної статистики можна поділити на дві неперетинні підмножини:

- підмножина значень статистики, за якою *гіпотеза* $H_0$ *приймається* (*не відхиляється*), називається **областю прийняття гіпотези (припустимою областю)**;
- підмножина значень статистики, за якою $H_0$ *відхиляється* і *приймається гіпотеза* $H_1$, називається **критичною областю.**

**Критичними точками** $\theta_{кр}$ називаються точки, які відокремлюють критичну область від припустимої. Критичні точки визначають за таблицями розподілу статистики.

Кількісно помилки оцінюють за ймовірністю їх виникнення.

***Припустима ймовірність помилки першого роду*** *позначається* **α** (імовірність потрапляння статистичної характеристики в критичну область, тобто ймовірність практично неможливих відхилень) і називається ***рівнем статистичної значущості.***

*Значення* α *зазвичай вибирається малим* (*0,05*; *0,02*; *0,01 – для технічних завдань*; *для завдань, пов'язаних із життям людини – 0,001*). *Чим менший рівень значущості, тим менша ймовірність зробити помилку першого роду* α, *але тим вища ймовірність зробити помилку другого роду.*

Для *визначення критичної області* статистики рівень значущості α вибирають з урахуванням виду альтернативної гіпотези $H_1$.

**Приклад**. Основна гіпотеза $H_0$: $\theta = \theta_0$. Альтернативна гіпотеза $H_1$ може при цьому мати такий вигляд: $H_1$: $\theta < \theta_0$; $H_1$ : $\theta > \theta_0$ або $H_1$ : $\theta \neq \theta_0$. Відповідно можна дістати ***лівобічну*** (рис. 3.1, *а*), ***правобічну*** (рис. 3.1, *б*) або ***двобічну критичну область*** (рис. 3.1, *в*).

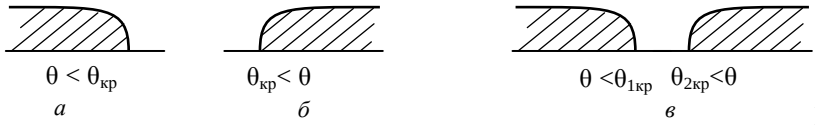



Рис. 3.1. Критичні області при перевірці гіпотез

**Загальна схема перевірки статистичної гіпотези** складається з таких етапів:

1. Формулюють гіпотези $H_0$ і $H_1$.

2. Вибирають статистику, за значенням якої приймається рішення про правильність гіпотези. Необхідно, щоб статистика мала відомий закон розподілу.

3. Задаються рівнем статистичної значущості α.

4. Визначають критичні значення за таблицями α і вид альтернативної гіпотези $H_1$.

5. Обчислюють за вибіркою значення обраної статистики.

6. Порівнюють значення обчисленої статистики з її критичним значенням.

7. Приймають рішення: якщо значення обчисленої статистики не потрапляє у критичну область, приймається гіпотеза $H_0$ і відхиляється гіпотеза $H_1$ і навпаки.

## Запитання для самоперевірки

1. Дайте означення статистичної гіпотези. Яка гіпотеза називається нульовою, а яка – конкуруючою?

2. Яка гіпотеза називається простою, а яка – складною?

3. У чому полягають помилки першого та другого роду?

4. Що називається статистичним критерієм?

5. Що таке критична точка і як вона обирається?

6. Які значення величини містить критична область, а які – область прийняття гіпотези?

7. В яких випадках використовується лівобічна, правобічна та двобічна критичні області?

8. Який порядок перевірки гіпотез?

## 3.2. ГІПОТЕЗИ ЩОДО ПАРАМЕТРІВ РОЗПОДІЛУ. ВИЗНАЧЕННЯ ОБ'ЄМУ ВИПРОБУВАНЬ

Критерії бувають параметричні і непараметричні. Параметричні критерії передбачають нормальний розподіл і пов'язані з обчисленням оцінок параметрів.

**Гіпотеза про середнє значення нормального розподілу при відомому СКВ.** Припустимо, що генеральна сукупність має нормальний розподіл, СКВ якого відоме.

При рівні значущості α потрібно перевірити гіпотезу $H_0$: $\nu = \nu_0$. Альтернативна гіпотеза $H_1$ може бути

$$H_1: \nu < \nu_0;\ H_1: \nu > \nu_0 \text{ або } H_1: \nu \neq \nu_0.$$

Як статистику використовують випадкову величину

$$Z = \frac{\overline{X} - \nu_0}{\sigma} \sqrt{n}\ ,$$

що має нормований нормальний розподіл.

Критичну область визначають за допомогою таблиці нормального розподілу. Якщо альтернативна гіпотеза має вигляд $H_1$: $\nu < \nu_0$, використовують лівобічну критичну область, що задовольняє умову:

$$P(Z < -z_{кр}) = \Phi(-z_{кр}) = \alpha.$$

Звідси випливає, що критична область – це множина таких $Z$, для яких $Z < -z_{кр}$ (рис. 3.2, *а*).

Якщо альтернативна гіпотеза має вигляд $H_1$: $\nu > \nu_0$, використовують правобічну критичну область, що задовольняє умову

$$P(Z > z_{кр}) = \alpha.$$

Із таблиці знаходять значення $z_{кр}$ з огляду на те, що

$$P(Z < z_{кр}) = \Phi(z_{кр}) = 1 - \alpha.$$

Знаходять критичну область $Z > z_{кр}$ (рис. 3.2, *б*).

За альтернативної гіпотези $H_1$: $\nu \neq \nu_0$ використовують двосторонню критичну область, що задовольняє умові:

$$P\left(|Z| > z_{кр}\right) = \alpha,$$

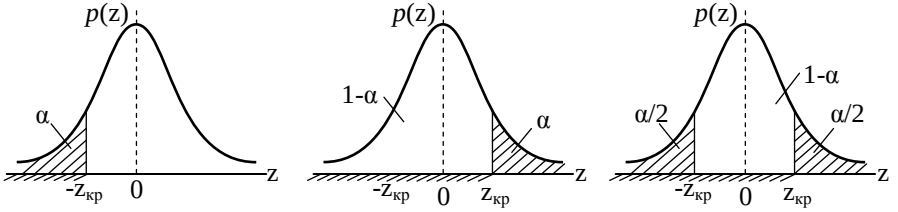
або

$$P(Z < z_{кр}) = P(Z > z_{кр}) = \alpha/2.$$

За таблицями знаходять

$$\Phi(z_{кр}) = 1 - \alpha/2.$$

У цьому разі критична область $|Z| > z_{кр}$ (рис. 3.2, *в*).




*а* *б* *в*

Рис. 3.2. Приклад вибору критичних областей при нормальному законі розподілі

**Гіпотеза про середнє значення нормального розподілу при невідомому СКВ**. Припущення аналогічні розглянутим раніше. У разі, коли СКВ невідоме, використовують випадкову величину (статистику *T*)

$$T = \frac{\overline{X} - \nu_0}{S}\sqrt{n-1},$$

яка має *t*-розподіл Стьюдента із кількістю степенів вільності $n - 1$.

Критичні області визначають так само, як і в попередньому випадку. Використання таблиць *t*-розподілу Стьюдента простіше, оскільки вони складені саме для визначення критичних областей.

**Гіпотеза про дисперсії нормального розподілу.** Припустимо, що генеральна сукупність має нормальний розподіл, де СКВ невідоме.

Дисперсію визначають за результатами експерименту, що складається з *n* дослідів.

Висувають гіпотезу $H_0 : \sigma^2 = \sigma_0^2$. Як статистику використовують випадкову величину

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}, \qquad (3.1)$$

яка має $\chi^2$-розподіл Пірсона із числом степенів вільності *n*-1.

Вибір критичної області визначають залежно від обраної альтернативної гіпотези $H_1$ та рівня статистичної значущості α, для чого за таблицями $\chi^2$ визначають критичні значення обраної статистики.

Якщо альтернативна гіпотеза має вигляд $H_1 : \sigma^2 < \sigma_0^2$, можна записати:

$$\frac{1}{\sigma^2} > \frac{1}{\sigma_0^2}$$

або з використанням виразу (3.1)

$$\frac{(n-1)S^2}{\sigma^2} > \frac{(n-1)S^2}{\sigma_0^2}.$$

Для перевірки гіпотези $P(\chi^2 > \chi^2_{кр}) = \alpha$ варто використовувати *правобічну* критичну область (рис. 3.3, *а*).

За альтернативної гіпотези $H_1 : \sigma^2 > \sigma_0^2$, міркуючи аналогічно, виходять з умови:

$$P(\chi^2 < \chi^2_{кр}) = \alpha.$$

Ураховуючи особливості складання таблиці $\chi^2$-розподілу, значення $\chi^2_{кр}$ знаходять згідно з умовою:

$$P\left(\chi^2 < \chi^2_{кр}\right) = 1 - P\left(\chi^2 > \chi^2_{кр}\right) = 1 - \alpha.$$

У цьому разі використовують *лівобічну* критичну область (рис. 3.3, *б*.)

За альтернативної гіпотези $H_1 : \sigma^2 \neq \sigma_0^2$ знаходять *двобічну* (рис. 3.3, *в*) критичну область згідно з умовою:

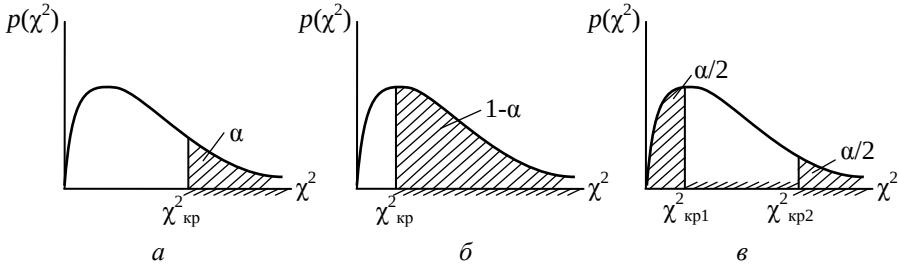
$$P\left\{\left(\chi^2 < \chi^2_{кр1}\right) \cup \left(\chi^2 > \chi^2_{кр2}\right)\right\} = \alpha .$$

Зазвичай приймають симетричну за ймовірністю критичну область, що задовольняє умову:

$$P(\chi^2 < \chi^2_{кр1}) = P(\chi^2 > \chi^2_{кр2}) = \alpha/2.$$

Відповідно до цієї умови з таблиці можна безпосередньо знайти $\chi^2_{кр2}$, а для визначення $\chi^2_{кр1}$ необхідно, як у попередньому випадку, використати умову:

$$P(\chi^2 > \chi^2_{кр1}) = 1- \alpha/2.$$




*а* *б* *в*

Рис. 3.3. Вибір критичних областей для χ2-розподілу Пірсона

**Перевірка гіпотези щодо належності вибірки до генеральної сукупності**. Як уже було розглянуто, помилкове рішення залежить від розмірів критичної області. З одного боку, чим менший розмір критичної області, тим частіше буде прийматися висунута гіпотеза $H_0$, але з другого боку, критерій втрачатиме чутливість стосовно альтернативної гіпотези.

Припустимо, на вхід приладу може бути подано одне з двох значень величин $\nu_0$ або $\nu_1$. При багаторазовому вимірюванні вихідної величини отримано значення $x_i$ ($i$ = 1,...,$n$) та обчислено середнє $\bar{x}$ (рис. 3.4). Необхідно визначити, чи є $\bar{x}$ оцінкою випадкової величини із центром $\nu_0$ або $\nu_1$. Висуваємо гіпотезу $H_0 : \bar{x} \in \nu_0$ за альтернативної гіпотези $H_1 : \bar{x} \in \nu_1$.

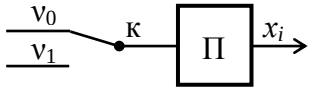



Рис. 3.4. Структурна реалізація перевірки гіпотези щодо належності вибірки до генеральної сукупності

Оскільки обсяг експериментальних даних обмежений, можливі обчислювані середні значення для цієї вибірки будуть відповідно розсіяні навколо $\nu_0$ і $\nu_1$ (рис. 3.5).

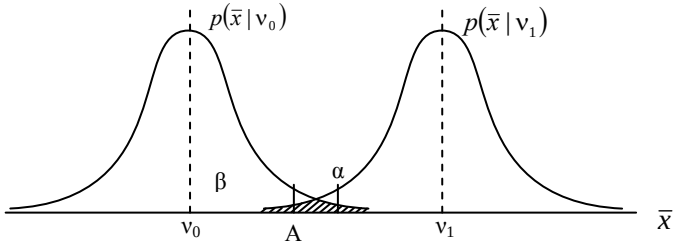



Рис. 3.5. Вибір граничного значення, що відокремлює альтеративні гіпотези

Зважаючи на те, що середні значення мають нормальний розподіл, в інтервалі $\nu_0, \nu_1$ спостерігатиметься перекриття законів розподілу, а отже, і неоднозначність прийнятих рішень, що виражається в їхній можливій помилковості. Граничним значенням $\bar{x}$, що відокремлює гіпотези $H_0$, $H_1$, є $A$. Якщо $\bar{x}$ буде ліворуч від $A$, то приймають гіпотезу $H_0$, а якщо праворуч, – гіпотезу $H_1$. Через перекриття законів розподілу для гіпотези $H_0$ буде мати місце помилка першого роду, тобто буде прийматися гіпотеза $H_1$, тоді як правдива гіпотеза $H_0$. Таким чином, відхилення $\bar{x}$ у більший бік від $A$ може бути викликане впливом випадкових факторів та обмеженим обсягом вибірки, а не належністю до $\nu_1$.

Імовірність виникнення помилки першого роду:

$$\alpha = P\left[\left(\bar{x} - \nu_0\right) > A\right];$$

де $A = z_{1-\alpha}\sigma_{\bar{x}}$, або

$$\alpha = \frac{1}{\sigma_{\bar{x}}\sqrt{2\pi}} \int_{A}^{+\infty} e^{-\frac{(\bar{x}-\nu_0)^2}{2\sigma_{\bar{x}}^2}} d\bar{x}.$$

Крім помилки першого роду може виникнути й помилка другого роду, яка полягає в тому, що справджується гіпотеза $H_1$, а обчислене середнє значення виявилося ліворуч $A$, тобто буде хибно прийматися гіпотеза $H_0$.

Імовірність виникнення помилки другого роду:

$$\beta = P\left[\left(\bar{x} - \nu_1\right) < A\right]; \beta = \frac{1}{\sqrt{2\pi}\sigma_{\bar{x}}} \int_{-\infty}^{A} e^{-\frac{(\bar{x}-\nu_1)^2}{2\sigma_{\bar{x}}^2}} d\bar{x}.$$

**Визначення необхідного обсягу випробувань**. Помилки першого та другого роду при фіксованих значеннях $\nu_0, \nu_1$ визначаються обсягом випробувань. Чим більший обсяг випробувань, тим ближче будуть групуватися можливі значення $\bar{x}$ до $\nu_0$ або $\nu_1$, а отже, зменшуватиметься область перекриття законів розподілів, тобто імовірність помилкових рішень (див. рис. 3.5).

Виходячи з оптимальності процедури перевірки гіпотез, необхідно визначити *мінімальний обсяг випробувань, який забезпечує помилки першого й другого роду, що не перевищують заданих значень.*

Нехай є два невідомі $\alpha$ і $\beta$ (див. рис. 3.5), які є визначальними для встановлення мінімального обсягу. Крім того, $\alpha$ і $\beta$ залежать від вихідного розсіювання $\sigma$, визначеного на підставі заздалегідь проведених досліджень, відстані між центрами розподілів $\nu_0$ і $\nu_1$ та вибору граничного значення $A$ між ознаками (статистичними характеристиками).

Припустимо, що $\nu_0, \nu_1$; $A$; $\sigma, \beta$ – задані. Оскільки є два невідомі значення $\alpha$ і $\beta$, а обсяг мінімальної кількості випробувань одночасно залежить від цих ймовірностей, необхідно скласти систему з двох рівнянь.

Для гіпотези $H_0$ (рис. 3.6) можна записати:

$$P(\bar{x} \le A \mid \nu_0) = 1 - \alpha .$$

де $A$ – фіксоване значення, при якому необхідно забезпечити одночасно значення $\alpha$ і $\beta$, менші від заданих.

У цьому разі застосовують статистику:

$$\frac{A - \nu_0}{\sigma / \sqrt{n}} = z_{1-\alpha} \ . \qquad (3.2)$$

На рис. 3.6 наведена графічна інтерпретація співвідношення (3.2).

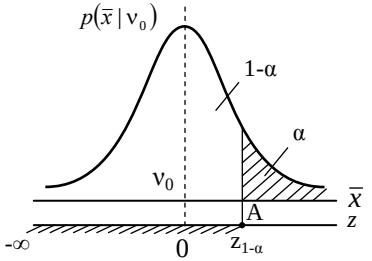



Рис. 3.6. Пояснення до визначення необхідного обсягу випробувань

Для гіпотези $H_1$ (рис. 3.6) має виконуватися співвідношення:

$$P(\bar{x} \le A \mid \nu_1) = \beta .$$

У цьому разі використовують статистику:

$$\frac{A - \nu_1}{\sigma / \sqrt{n}} = z_{\beta} . \qquad (3.3)$$

Графічну інтерпретацію співвідношення (3.3) наведено на рис. 3.7.

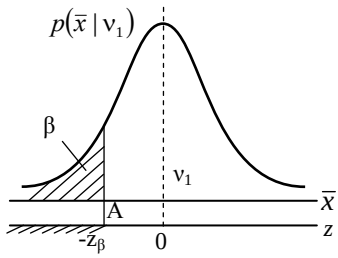



Рис. 3.7. Вибір граничного значення для нормованого нормального розподілу

Із виразу (3.3) знаходимо:

$$A = z_\beta \, \sigma \Big/ \sqrt{n} + \nu_1 \, .$$

Підставляючи знайдене значення $A$ у вираз (3.2), дістаємо:

$$\frac{z_\beta \dfrac{\sigma}{\sqrt{n}} + \nu_1 - \nu_0}{\sigma \Big/ \sqrt{n}} = z_{1-\alpha} \, .$$

Після перетворення маємо:

$$\sqrt{n}\left(\nu_1 - \nu_0\right) = z_{1-\alpha}\sigma - z_\beta\sigma \, ,$$

звідки кількість випробувань:

$$n_{min} = \left[ \frac{\sigma\left(z_{1-\alpha} - z_\beta\right)}{\nu_1 - \nu_0} \right]^2 . \qquad (3.4)$$

*Якщо границя A задана, обчислену кількість випробувань n округлюють до цілого числа в більшу сторону.*
*Можна, використавши ту саму систему рівнянь, визначити границю A, якщо заданий обсяг вибірки.*

**Запитання для самоперевірки**

1. Які критерії називаються параметричними?
2. Як перевіряється гіпотеза про середнє значення з відомим СКВ?
3. Як перевіряється гіпотеза про середнє значення з невідомим СКВ?
4. Як перевіряється гіпотеза про дисперсії?
5. Як перевіряється гіпотеза про наявність зсуву?
6. Як обчислюється ймовірність помилки першого роду?
7. Як обчислюється ймовірність помилки другого роду?
8. Що характеризує параметр $A$ при визначенні необхідного обсягу випробувань?
9. Як обчислюється необхідний обсяг випробувань?

## 3.3. ПАРАМЕТРИЧНІ КРИТЕРІЇ

Параметричні критерії можна застосовувати при визначенні можливої розбіжності параметрів розподілу сукупностей.

**Порівняння двох середніх**. Завдання такого плану трапляються тоді, коли необхідно переконатися в тому, чим викликана розбіжність середніх, отриманих на підставі наявних двох вибірок з однієї генеральної сукупності випадкових величин $X$ і $Y$ обсягами $n_1$ і $n_2$. Іншими словами, чи зумовлена ця розбіжність впливом випадкових величин і обмеженим обсягом вибірки або ж справді існує розбіжність між центрами розподілів (систематичними похибками приладів, продуктивністю праці, довговічністю, надійністю тощо).

***Для випадку з відомими СКВ генеральних сукупностей***.

Відомі $\sigma_x$ і $\sigma_y$ величин $X$ і $Y$. Висуваємо гіпотезу $H_0$: $\nu_x = \nu_y$ при альтернативній $H_1$: $\nu_x \neq \nu_y$. Необхідно визначити, чи істотна розбіжність між середніми $\bar{x}$ і $\bar{y}$, отриманими відповідно з вибірки

обсягом $n_1$ для випадкової величини $X$ і обсягом $n_2$ для випадкової величини $Y$.

Позначимо:

$$\bar{x} - \bar{y} = u;$$

$$\nu_x - \nu_y = Mu.$$

Розглянемо різницю $u$ як параметр $Z$, тобто у вигляді нормованої різниці:

$$Z = \frac{u - Mu}{\sigma_u} = \frac{(\bar{x} - \bar{y}) - (\nu_x - \nu_y)}{\sigma_{(\bar{x}-\bar{y})}},$$

де $\sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{n_1}$, $\sigma_{\bar{y}}^2 = \frac{\sigma_y^2}{n_2}$, а за висунутої гіпотези $\nu_x - \nu_y = 0$.

Величини $X$ і $Y$ – незалежні величини, для яких $\sigma^2(x \pm y) = \sigma^2(x) + \sigma^2(y)$, тому можна записати:

$$\sigma_{\bar{x}-y}^2 = \frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2}. \qquad (3.5)$$

Для перевірки правильності гіпотези використовують двобічну критичну область. З урахуванням виразу (3.5) значення статистичної характеристики:

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2}}}.$$

Значення $z$ порівнюють з критичним значенням $z_{кр} = z_{1-\alpha/2}$.

Якщо $|z| \le z_{кр}$, приймають гіпотезу $H_0$.

***Для випадку з невідомими СКВ генеральних сукупностей***. У цьому разі визначають емпіричне значення $s$ і використовують статистичну характеристику:

$$t = \frac{(\bar{x} - \bar{y}) - (\nu_x - \nu_y)}{s_{\bar{x}-\bar{y}}},$$

розподілену за законом Стьюдента.

Відомо, що для незалежних випадкових величин:

$$s_{\bar{x}-\bar{y}} = s_{\bar{x}}^2 + s_{\bar{y}}^2 .$$

За умови, що вибірки взяті з однієї генеральної сукупності, тобто $s_x^2 = s_y^2 = s$, можна записати:

$$s_{\bar{x}-\bar{y}}^2 = \frac{s^2}{n_1} + \frac{s^2}{n_2},$$

або

$$s_{\bar{x}-\bar{y}} = s\sqrt{\frac{n_1 + n_2}{n_1 n_2}} . \qquad (3.6)$$

Зважаючи на те, що вибірки взято з однієї генеральної сукупності, доцільно скористатися всіма наявними експериментальними даними, тобто розглядати вибірку обсягом $n_1 + n_2$. Це дає змогу підвищити статистичну надійність знайденої оцінки СКВ генеральної сукупності. Тоді:

$$s = \frac{1}{\sqrt{(n_1 + n_2) - 2}} \cdot \left[ \sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{i=1}^{n_2} (y_i - \bar{y})^2 \right] \cdot \qquad (3.7)$$

Виходячи з виразу для дисперсії можна записати:

$$\sum_{i=1}^{n_1} (x_i - \bar{x})^2 = s_x^2 (n_1 - 1) .$$

Аналогічно маємо:

$$\sum_{i=1}^{n_2} (y_i - \bar{y})^2 = s_y^2 (n_2 - 1) .$$

Підставляють отримані значення сум у вираз (3.7) і дістають:

$$s = \frac{1}{\sqrt{(n_1 + n_2) - 2}} \left[ s_x^2 (n_1 - 1) + s_y^2 (n_2 - 1) \right]. \qquad (3.8)$$

Підставляють вираз (3.8) у вираз (3.6) і отримують:

$$s_{\bar{x}-\bar{y}} = \frac{1}{\sqrt{(n_1+n_2)-2}}\left[s_x^2(n_1-1)+s_y^2(n_2-1)\right]\sqrt{\frac{n_1+n_2}{n_1 n_2}}. \quad (3.9)$$

З урахуванням виразу (3.9) розрахункове значення коефіцієнта Стьюдента буде:

$$t_{\text{р}} = \frac{(\bar{x}-\bar{y})-(v_x-v_y)}{s_{\bar{x}-\bar{y}}} = \frac{(\bar{x}-\bar{y})}{\sqrt{s_x^2(n_1-1)+s_y^2(n_2-1)}}\sqrt{\frac{(n_1+n_2-2)n_1 n_2}{n_1+n_2}}.$$

Для перевірки гіпотези необхідно знайти критичне значення коефіцієнта Стьюдента $t_{\text{кр}}$ для заданого рівня статистичної значущості $\alpha$ і кількості степенів вільності $n_1+n_2-2$. Якщо розрахункове значення $|t_{\text{р}}| \le t_{\text{кр}}$, приймають нульову гіпотезу.

У випадку, коли вибірки взято не з однієї генеральної сукупності, тобто $\sigma_x^2 \neq \sigma_y^2$, для визначення суттєвості розбіжності обчислених середніх $\bar{x}$ і $\bar{y}$ коефіцієнт Стьюдента обчислюють за формулою:

$$t'_{\text{р}} = \left|(\bar{x}-\bar{y})\right|\left(\sqrt{\frac{s_{\bar{x}}^2}{n_1}+\frac{s_{\bar{y}}^2}{n_2}}\right)^{-1}.$$

Обчислене значення $t'_{\text{р}}$ порівнюють із критичним значенням коефіцієнта Стьюдента, яке залежно від $\alpha$ і кількості степенів вільності обчислюють за формулою:

$$f = E\,|\,f'+1|,$$

де

$$f' = \frac{(n_1-1)(n_2-1)\left(\frac{s_{\bar{X}}^2}{n_1}+\frac{s_{\bar{Y}}^2}{n_2}\right)^2}{\frac{(n_1-1)s_{\bar{X}}^4}{n_1^2}+\frac{(n_2-1)s_{\bar{Y}}^4}{n_2^2}}.$$

**Порівняння двох дисперсій**. Другою важливою ознакою, за якою можуть порівнюватися дві сукупності, є дисперсії в кожній із них. Гіпотези про дисперсії відіграють важливу роль у техніці, оскільки

саме дисперсія характеризує такі важливі конструкторські й технологічні показники, як точність приладу, похибку результату, точність технологічного процесу тощо.

При спільній обробці результатів необхідно, насамперед, переконатися в тому, що умови проведення дослідів, при яких отримані результати, були однаковими. З цією метою можна використати оцінку розбіжності дисперсій.

***F-критерій*** (***розподіл Фішера***). Припустимо, що задано дві генеральні сукупності $X$ та $Y$ з нормальним розподілом $X \in N(\mu_1, \sigma_1)$ *і* $Y \in N\ (\mu_2, \sigma_2)$. Із цих генеральних сукупностей зроблено незалежні вибірки з параметрами відповідно $n_1, s_1^2,\ n_2,\ s_2^2$. Потрібно при рівні значущості α перевірити гіпотезу $H_0 : \sigma_1^2 = \sigma_2^2$ при альтернативній гіпотезі $H_1 : \sigma_1^2 > \sigma_2^2$.

Математик-статистик Р. Фішер установив, що відношення незсунених оцінок двох дисперсій підпорядковується закономірності, що залежить від кількості степенів вільності цих дисперсій. Припускаючи, що $s_1^2 > s_2^2$, приймають як статистику величину $F$, котра задовольняє *F-розподіл*, для якого $F = \dfrac{s_1^2}{s_2^2}$ з кількістю степенів вільності, що дорівнює $n_1 - 1$ і $n_2 - 1$ відповідно.

Критична область буде тільки *правобічна* і визначається умовою $P(F_р > F_{кр}) = \alpha$.

Значення $F_{кр}$ знаходять із таблиць $F$-розподілу, яке залежить від трьох величин: рівня значущості α і двох чисел, якими виражаються степені вільності $f_{чис} = k_1,\ f_{знам} = k_2$. Таблиці складені окремо для кожного значення α (тривимірні таблиці). Задаючись рівнем статистичної значущості $\alpha$, вибравши в таблиці колонку $f_{чис}$ і рядок $f_{знам}$, на їх перетині знаходять критичне значення коефіцієнта $F_{кр}$.

Якщо $F_р < F_{кр}$, можна стверджувати на підставі наявних експериментальних даних, що при рівні статистичної значущості $\alpha$ вибіркові дисперсії будуть однорідні.

Іноді перед обробкою даних необхідно переконатися в однорідності дисперсій, отриманих на підставі даних двох експериментальних вибірок. У цьому випадку також використовується критерій Фішера.

Відмінністю є те, що альтернативна гіпотеза $H_1 : \sigma_1^2 \neq \sigma_2^2$. Для перевірки гіпотези використовують двобічну критичну область (рис. 3.8).

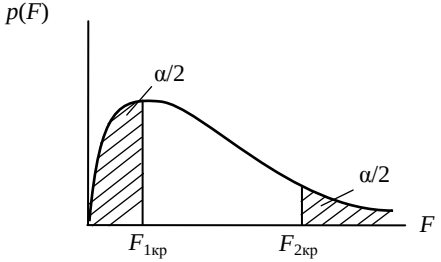



Рис. 3.8. Перевірка гіпотези про однорідність дисперсій двох експериментальних вибірок

Передбачається, що ймовірність того, що розрахункове значення $F_р$ буде меншим за критичне значення $F_{кр1}$ або більшим за $F_{кр2}$ не перевищує значення:

$$P\left(F < F_{кр1}\right) = P\left(F > F_{кр2}\right) = \frac{\alpha}{2}.$$

Оскільки в таблиці є тільки значення $P(F > F_{кр})$, то значення $F_{кр2}$ можна знайти безпосередньо з таблиць. Для визначення $F_{кр1}$, яке відповідає умові $P(F < F_{кр1})$, вводять коефіцієнт

$$F' = \frac{1}{F}.$$

Тоді $$P(F < F_{кр1}) = P(F' > \frac{1}{F_{кр1}}).$$

Останнє співвідношення показує, що $F_{кр1}$ можна знайти виходячи з умови $P\left(F' > \frac{1}{F}\right)$. Надалі перевірка залишається такою самою, як і для однобічної критичної області, тобто нуль-гіпотеза, приймається в тому випадку, коли розрахункові значення $F_р$ містяться між критичними $F_{кр1}$ й $F_{кр2}$.

***Перевірка гіпотези про однорідність ряду дисперсій G-критерій***). Якщо потрібно побудувати аналітичну залежність (аналітичну модель) на основі експериментальних даних, то насамперед необхідно переконатися в однорідності вибіркових дисперсій. Для цього серед наявних вибіркових дисперсій вибирають максимальну, а потім розглядають відношення максимальної дисперсії до суми всіх вибіркових дисперсій. Для випадку, коли обсяги вибірок однакові, тобто в кожній вибірці дисперсія обчислювалася на підставі однакової кількості даних *m*, застосовують *G-критерій* (критерій Кохрена). Статистика *G* вказує, яку частку має максимальна дисперсія в загальній дисперсії, і обчислюється за формулою:

$$G_р = \frac{s^2\{y_i\}_{max}}{\sum_{i=1}^{N} s^2\{y_i\}}, \qquad (3.10)$$

де $s^2\{y_i\} = \frac{1}{m-1}\sum_{u=1}^{m}(y_{iu} - \bar{y}_i)^2$ – дисперсія для кожної $i$-ої вибірки, $i = 1,\overline{N}$.

Висувають гіпотезу $H_0$: дисперсії однорідні. Для перевірки нульової гіпотези знайдене розрахункове значення $G_р$ порівнюють із критичним значенням коефіцієнта $G_{кр}$, яке знаходять із таблиць критичних значень *G*-критерію на перетині

стовпця $f_{\text{чис}} = m - 1$ та рядка $f_{\text{знам}} = N$ для заданого рівня статистичної значущості α.

Якщо $G_{\text{р}} < G_{\text{кр}}$, на підставі наявних даних можна стверджувати, що вибіркові дисперсії будуть однорідні, тобто гіпотеза $H_0$ приймається.

Якщо гіпотеза про рівність дисперсій відхиляється, то збільшують обсяг вибірок і знову перевіряють гіпотезу. Якщо ж гіпотеза знову не підтверджується, такі вибіркові дані спільно обробляти не можна.

## Запитання для самоперевірки

1. Що називається параметричним критерієм?
2. Записати критерій перевірки гіпотези про рівність двох середніх із відомим СКВ.
3. Записати критерій перевірки гіпотези про рівність двох середніх з невідомим СКВ.
4. За яким критерієм перевіряють гіпотези про рівність дисперсій двох вибірок?
5. Як перевіряють гіпотези про рівність дисперсій двох вибірок при правобічній критичній області?
6. Як перевіряють гіпотези про рівність дисперсій двох вибірок при двобічній критичній області?
7. За яким критерієм та як перевіряють гіпотези про рівність дисперсій кількох вибірок?
8. Як користуватися таблицями *F*-розподілу та *G*-розподілу?

# 3.4. КРИТЕРІЇ ЗГОДИ

Критерії згоди використовують для перевірки узгодженості наявних експериментальних даних з теоретичним законом, що припускається.

Розглянуті раніше методи перевірки гіпотез припускали, що функціональна форма закону розподілу відома, й визначалися лише значення параметрів цього закону.

Однак у деяких випадках сам вид закону розподілу потребує статистичної перевірки. Раніше було розглянуто, як, зіставляючи ймовірності потрапляння значень в інтервали з відповідними частостями, отриманими зі спостережень, або проводячи графічне порівняння полігонів і гістограм із кривою розподілу, можна скласти первинне уявлення про близькість теоретичного та емпіричного розподілів.

Постає питання про критерії перевірки гіпотези про те, що величина $X$ відповідає конкретному закону розподілу зі щільністю ймовірності $p(x)$.

Подібні критерії зазвичай називають *критеріями згоди* (*відповідності*). Вони ґрунтуються на виборі певної міри розбіжності між теоретичним та емпіричним розподілами. Якщо така міра розбіжності для розглянутого випадку перевищує встановлену межу, гіпотеза відхиляється й навпаки.

Таким чином, перевірити таку гіпотезу – означає переконатися в тому, що наявні дані справді взято з генеральної сукупності, яка має передбачуваний закон розподілу.

Розглянемо застосування одного з найбільш уживаних критеріїв $\chi^2$-*критерію Пірсона*.

***Критерій*** $\chi^2$ припускає, що наявні експериментальні дані розбивають на $l$ елементарних інтервалів, кожний з яких містить не менш ніж 8-10 значень. За відомим правилом будують гістограму, за

видом якої, якщо немає додаткових джерел, можна зробити припущення про можливий закон розподілу. Підбирають закон розподілу, якому щонайбільше відповідають наявні дані.

Для кожного елементарного інтервалу можна визначити $\hat{m}_i$ – кількість значень, що потрапили в $i$-й інтервал, $i = 1,..., l$.

Висувають гіпотезу $H_0$: емпіричний розподіл належить до передбачуваного теоретичного. Для того щоб перевірити висунуту гіпотезу, необхідно оцінити розбіжності між емпіричною $\hat{m}_i$ і теоретичною $m_i$ кількістю потраплянь за елементарний інтервал за припущення, що *випадкова* величина $X$ має передбачуваний закон розподілу.

Установлено, що величина

$$\frac{\hat{m}_i - m_i}{\sqrt{np_i}} = \xi_i \qquad (3.11)$$

розподілена нормально з нульовим математичним сподіванням і одиничною дисперсією. Відомо, що сума квадратів нормованих нормально розподілених величин підпорядковується $\chi 2$-розподілу Пірсона:

$$\chi^2 = \sum_{i=1}^{l} \xi_i^2 \quad .$$

Виходячи з виразу (3.11) можна обчислити *розрахункове значення* $\chi^2$:

$$\chi_{\text{р}}{}^2 = \sum_{i=1}^{l} \frac{\left(\hat{m}_i - m_i\right)^2}{np_i} . \qquad (3.12)$$

Вираз (3.12) використовують як статистику для перевірки гіпотези про передбачуваний закон розподілу.

Для знаходження теоретичної кількості потраплянь $m_i$ розглянемо співвідношення $\hat{m}_i \dfrac{n}{n}$ і запишемо його як $n\dfrac{\hat{m}_i}{n}$.

Відношення $\hat{m}_i / n$ є частотою (частістю) потрапляння результатів у *i*-й елементарний інтервал. Відомо, що при збільшенні загальної кількості спостережень *n* частота наближається до теоретичної ймовірності $p_i$ потрапляння випадкової величини в *i*-й елементарний інтервал при передбачуваному законі розподілу. Виходячи із цього, можна записати:

$$\lim_{n\to\infty} \hat{m}_i = n \lim_{n\to\infty} \frac{\hat{m}_i}{n}.$$

Пам'ятаючи про те, що границею, до якої прямуватиме $\hat{m}_i$ при збільшенні *n*, є теоретичне значення $m_i$ кількості потраплянь в інтервал, можна записати $m_i = np_i$.

Таким чином, можна обчислити теоретичні значення $m_i$, а потім знайти різницю $\hat{m}_i - m_i$, що характеризує розбіжність кількості потраплянь у кожний *i*-й інтервал.

Потрапляння результатів в *i*-й інтервал характеризується біноміальним законом розподілу, для якого

$$\sigma = \sqrt{npq} = \sqrt{np(1-p)} = \sqrt{np - np^2} \approx \sqrt{np}.$$

Для знаходження $p_i$ необхідно користуватися співвідношенням:

$$p_i = \int_{x_i}^{x_{i+1}} p(x)dx,$$

де $x_i$ й $x_{i+1}$ – кінці інтервалу, а $p(x)$, наприклад, для нормального закону має вигляд:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-Mx)^2}{2\sigma^2}}.$$

У випадку, коли $Mx$ і $\sigma$ невідомі, необхідно обчислити їхні оцінки $\bar{x}$ і $s$ на підставі дослідних даних і увести їх у вираз для

закону розподілу. Так, для нормального закону оцінка щільності ймовірності:

$$\hat{p}(x) = \frac{1}{\sqrt{2\pi}s} e^{-\frac{(x-\bar{x})^2}{2s^2}}.$$

*Оскільки на підставі тих самих експериментальних даних додатково обчислюють $\bar{x}$ і $s$, то втрачаються два степені вільності, тобто $f = l - 1 - 2$.*

Отримане значення $\chi^2_{\text{р}}$ порівнюють із критичним $\chi^2_{\text{кр}}$, узятим з таблиці $\chi^2$-розподілу для обраного рівня статистичної значущості $\alpha$ і кількості степенів вільності $f = l - 3$.

Якщо $\chi^2_{\text{р}} < \chi^2_{\text{кр}}$, можна стверджувати, що з обраним рівнем статистичної значущості $\alpha$ приймається гіпотеза $H_0$, тобто наявні дані відповідають передбачуваному закону розподілу. Гіпотеза $H_0$ відхиляється при значеннях $\chi^2_{\text{р}}$, більших за $\chi^2_{\text{кр}}$, тому використовується тільки правобічна критична область (див. рис. 3.3, *б*).

*Взагалі у практиці перевірки гіпотез про відповідність закону розподілу намагаються застосовувати два методи перевірки, що підвищує вірогідність прийняття рішення. Особливо це важливо, коли характерні риси закону розподілу перебувають на його «хвостах», що не завжди можна виявити при малому обсязі випробувань.*

**Критерій $\omega^2$ (омега-квадрат)**. Критерій $\chi^2$, незважаючи на свою простоту, не завжди забезпечує надійну перевірку гіпотези про закон розподілу, оскільки частина вихідної інформації втрачається при здійсненні процедури групування даних.

Розглянемо критерій згоди при простій гіпотезі, що повністю фіксує закон розподілу генеральної сукупності, з якої отримано вибірку. Цей критерій, що дістав назву $\omega^2$, на відміну від $\chi^2$, ґрунтується *безпосередньо на спостережених* (*незгрупованих*) значеннях величини *X*.

Нехай гіпотеза полягає в припущенні, що величина *X* розподілена відповідно до деякого неперервного закону розподілу з інтегральною функцією *F*(*x*), яка вважається відомою.

Розглянемо ряд вибіркових значень $x_1, x_2, ..., x_n$ випадкової величини *X*. Позначимо $n_x$ кількість значень, що перебуває ліворуч від вибраного значення. Тоді відношення $\frac{n_x}{n} = W_n(x)$ і є емпіричною функцією розподілу. При великому обсязі вибірки емпірична функція розподілу $W_n(x)$ є апроксимацією теоретичної функції розподілу $F(x)$:

$$F(x) = \int_{-\infty}^{x} p(x)dx.$$

Різниця величин $W_n(x)$ і $F(x)$ може бути використана як *міра розбіжності* між вибірковими даними та передбачуваним законом розподілу генеральної сукупності. За міру цієї розбіжності беруть середній квадрат відхилень за всіма можливими значеннями аргументу:

$$\omega^2 = \int_{-\infty}^{+\infty} [W_n(x) - F(x)]^2 dF(x),$$

де $dF(x) = F'(x)dx$.

З наявних вибіркових даних будується варіаційний ряд $x_1 < x_2 < ... < x_n$ . Теоретично рівність будь-яких двох членів у цьому ряді через неперервність функції $F(x)$ практично неможлива, тобто

має ймовірність, що дорівнює нулю. Вважаємо, що для $x < x_1$ емпірична функція розподілу дорівнює нулю, а при $x > x_n$, дорівнює одиниці. Відповідно до такого визначення функції $W_n(x)$ маємо:

$$\begin{cases} W_n(x) = 0 & \text{при} \quad x < x_1 \\ W_n(x) = \dfrac{k}{n} & \text{при } x_1 < x < x_k, \text{ де} \quad k = 1,2,..n-1 \\ W_n(x) = 1 & \text{при} \quad x > x_n \end{cases} \qquad (3.13)$$

Емпірична функція стрибком змінює свої значення в точках $x = x_к$ на $\dfrac{1}{n}$, зберігаючи це значення протягом усього інтервалу $x_к$ - $x_{к+1}$ (рис. 3.9).

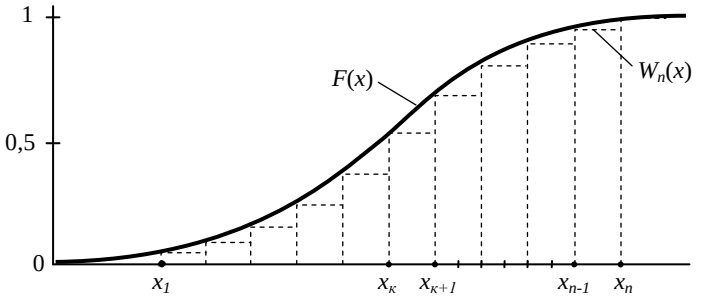



Рис. 3.9. Розбіжності між вибірковими даними та передбачуваним законом розподілу генеральної сукупності

Виходячи зі співвідношення (3.13) можна представити вираз для критерію $\omega^2$ у вигляді окремих доданків:

$$\omega^2 = \int_{-\infty}^{x_1} [0 - F(x)]^2 dF(x) + \sum_{k=1}^{n-1} \int_{x_k}^{x_{k+1}} \left[ \frac{k}{n} - F(x) \right]^2 dF(x) + \int_{x_n}^{+\infty} [1 - F(x)]^2 dF(x). \qquad (3.14)$$

Розглянемо першу складову інтеграла (3.14):

$$\int_{-\infty}^{x_i} [0 - F(x)]^2 dF(x) = -\frac{1}{3} [0 - F(x)]^3 \Big|_{-\infty}^{x_i} = \frac{1}{3} F^3(x_1). \qquad (3.15)$$

Розглянемо другу складову інтеграла (3.14):

$$\int_{x_k}^{x_{k+1}} \left[\frac{k}{n} - F(x)\right]^2 = -\frac{1}{3}\left[\frac{k}{n} - F(x)\right]^3 \Big|_{x_k}^{x_{k+1}} =$$

$$= \frac{1}{3}\left[F(x_{k+1}) - \frac{k}{n}\right]^3 - \frac{1}{3}\left[F(x_k) - \frac{k}{n}\right]^3. \qquad (3.16)$$

Розглянемо третю складову інтеграла (3.14):

$$\int_{x_n}^{\infty} \left[1 - F(x)\right]^2 dF(x) = -\frac{1}{3}\left[1 - F(x)\right]^3 \Big|_{x_n}^{+\infty} = +\frac{1}{3}\left[1 - F(x_n)\right]^3. \qquad (3.17)$$

Підставляємо (3.15), (3.16), (3.17) у вираз (3.14), дістаємо:

$$\omega^2 = \frac{1}{3}F^3(x_1) + \sum_{k=1}^{n-1}\left\{\frac{1}{3}\left[F(x_{k+1}) - \frac{k}{n}\right]^3 - \frac{1}{3}\left[F(x_k) - \frac{k}{n}\right]^3\right\} + \frac{1}{3}\left[1 - F(x_n)\right]^3.$$

Після спрощень останній вираз набирає вигляду:

$$\omega^2 = \frac{1}{12n^2} + \frac{1}{n}\sum_{k=1}^{k=n}\left[F(x_k) - \frac{2k-1}{2n}\right]^2. \qquad (3.18)$$

Ця рівність показує, в який спосіб $\omega^2$ залежить від індивідуальних членів варіаційного ряду.

Точний розподіл $\omega^2$ дуже складний, але дослідження показують, що вже при $n > 40$ розподіл добутку $n\omega^2$ близький до деякого граничного розподілу, для якого складено таблиці. Фрагмент такої таблиці наведено у табл. 3.1.

Таблиця 3.1. Залежність $\omega^2$ від індивідуальних членів варіаційного ряду

| α | 0,5 | 0,4 | 0,3 | 0,2 | 0,1 |
|---|---|---|---|---|---|
| $(n\omega^2)_{кр}$ | 0. 1184 | 0. 1467 | 0. 1843 | 0. 2412 | 0. 3473 |
| α | 0,05 | 0,03 | 0,02 | 0,01 | 0.001 |
| $(n\omega^2)_{кр}$ | 0. 4614 | 05489 | 0. 6198 | 0. 7435 | 1. 6979 |

За таблицями визначають критичні значення для величини $n\omega^2$ при $n > 40$:

$$P\left[(n\omega^2)_{р} > (n\omega^2)_{кр}\right] = \alpha.$$

Якщо розрахункове значення критерію буде меншим від критичного, приймають гіпотезу про передбачуваний закон розподілу.

**Запитання для самоперевірки**

1. Які критерії називаються критеріями згоди?
2. Як за критерієм $\chi^2$ перевіряють гіпотезу про закон розподілу?
3. Яка міра розбіжності між емпіричним та теоретичним розподілом застосовується для критерію $\chi^2$?
4. Яка величина підлягає розподілу $\chi^2$?
5. Як визначається число степенів вільності для розподілу $\chi^2$?
6. Яка величина застосовується як статистика для перевірки гіпотези про закони розподілу?
7. Чому використовується тільки правостороння критична область?
8. Які відмінні особливості має критерій $\omega^2$?
9. Яка міра розбіжності між емпіричним та теоретичним розподілом застосовується для критерію $\omega^2$?
10. Як знаходиться критичне значення для критерію $\omega^2$?

## 3.5. НЕПАРАМЕТРИЧНІ КРИТЕРІЇ

У непараметричних критеріях не робиться припущень щодо розподілу генеральної сукупності. Вони застосовуються при будь-яких законах як для кількісних, так і якісних ознак.

В експериментальних дослідженнях часто потрібно порівняти два паралельних ряди спостережень над об’єктами, що належать до різних різновидів, типів, сортів.

При цьому виконують $N$ дослідів при систематичній і планомірній зміні від досліду до досліду рівня якого-небудь фактора, розбіжність у дії якого на кожний із двох різновидів об'єктів мають з'ясувати. У кожному такому досліді над парою об'єктів, виконаному, природно, у незмінних умовах, мають два значення порівнюваної ознаки двох об'єктів.

*Непараметричні критерії* використовуються для перевірки того, чи належать дві вибірки до тієї самої генеральної сукупності, тобто чи однорідні ці вибірки (значення результатів випробувань мають однакові функції розподілів). Такі критерії ґрунтуються на вивченні послідовностей реалізацій випадкової величини. Відповідні обчислювальні процедури є ефективними навіть при обмежених обсягах вибірок.

**Критерій Уїлкоксона (критерій серій).** Цей критерій дає змогу на основі наявних вибірок, використовуючи інтегральну функцію ймовірностей, перевірити гіпотезу про те, що дві вибірки мають однаковий закон розподілу.

Таким чином, гіпотеза $H_0 : F_X(x) \equiv F_Y(x)$ перевіряється за допомогою однієї вибірки ($x_1$,..., $x_{n1}$) з $X$ і однієї вибірки ($y_1$,…, $y_{n2}$) із $Y$. Щодо розподілів $X$ і $Y$ ніяких припущень не робиться.

Значення $x_1$,..., $x_{n1}$ та $y_1$,…, $y_{n2}$ обох вибірок упорядковуються у спільний варіаційний ряд. Коли в цьому ряді елемент однієї вибірки більший від елемента іншої вибірки, говорять, що пари значень ($x_i$, $y_j$) утворюють *інверсію*.

Під ***інверсією*** розуміють перехід від елемента однієї вибірки до елемента іншої вибірки при послідовному просуванні вздовж варіаційного ряду, причому інверсія не передбачається для першого елемента варіаційного ряду. Як контрольна величина береться повна кількість *інверсій и*.

**Приклад**. Для ряду $x_1$, $y_1$, $y_2$, $x_2$, $y_3$, $y_4$, $y_5$, $x_3$ перед $y_1$ є тільки один елемент $x$, а отже, кількість інверсій дорівнює одиниці.

Елементу $y_2$ передує один елемент $x$, кількість інверсій також дорівнює одиниці. Елементу $y_3$ передують два елементи першої групи, а отже, кількість інверсій дорівнює двом. Аналогічно для $y_4$ і $y_5$. Повна кількість інверсій $u = 1+1+2+2+2 = 8$.

Для перевірки нульової гіпотези можна застосовувати два способи.

1. Якщо гіпотеза правильна, обчислене значення $u$ не повинне відхилятися від свого математичного сподівання:

$$Mu = \frac{mn}{2}, \tag{3.19}$$

де $m$ і $n$ – обсяги вибірок.

Від гіпотези відмовляються, якщо $\left|u - \frac{mn}{2}\right|$ більше від певного критичного значення $u_\alpha$. Критичне значення $u_\alpha$ беруть для заданого рівня значущості α з таблиці критичних значень Уїлкоксона.

2. У загальному випадку, а також у випадку великих $m$ і $n$, для яких $u_\alpha$ не можна взяти з таблиці, справджується формула:

$$u_\alpha = z_\alpha \sqrt{\frac{mn(m+n+1)}{12}},$$

де $z_\alpha$ – визначається з таблиць Лапласа.

***Критерій знаків*** (***медіанний критерій***). Його використовують тільки в разі, коли вибірки $X$ та $Y$ однакові за обсягом, тобто значення можна розглядати у вигляді масиву пар чисел $(x_i, y_i)$, $i = 1…n$.

При застосуванні цього критерію будується спільний варіаційний ряд. Спочатку визначається медіана для цього ряду. Якщо кількість значень непарна, то як медіана використовується середній елемент. Якщо кількість парна, медіана перебуває між $x_{\frac{m}{2}}$ і $x_{\frac{m}{2}+1}$ та обчислюється за формулою:

$$Me = \frac{x_{\frac{m}{2}} + x_{\frac{m}{2}+1}}{2}.$$

Після обчислення медіани визначається знак відхилення поточного значення $x_i$, $i = \overline{1,m}$, від медіани. У такий спосіб одержують послідовність знаків, що розпадається на окремі *серії* знаків одного виду, укладених між знаками іншого виду.

**Приклад.** + + – – – + – – – – + + + – . Кількість серій $r = 6$.

За спеціальними таблицями для відомих значень кількості спостережень у вибірках $N_1$ і $N_2$ визначається для певного рівня статистичної значущості α $r_{\text{кр. н}}$ $(u_{0,025})$ і $r_{\text{кр. в}}$ $(u_{0,975})$. Якщо виконується співвідношення

$$r_{\text{кр. н}} < r < r_{\text{кр. в}},$$

гіпотеза про належність вибірок до однієї генеральної сукупності, для якої $F_X(x) \equiv F_Y(x)$, приймається, у протилежному разі – відхиляється.

Критерій знаків набув поширення в дослідницьких роботах завдяки тому, що процедура його застосування винятково проста, вимірювання ознаки, що враховується, можуть виконуватися грубими засобами, тому що має значення лише знак різниці результатів вимірювань, а в основу критерію покладені прості положення, яких майже завжди дотримуються (наприклад, не припускають нормального розподілу досліджуваних величин).

### Запитання для самоперевірки

1. Які критерії називаються непараметричними?
2. Яка гіпотеза перевіряється за допомогою критерію Уїлкоксона?
3. Що таке інверсія і як вона визначається?

4. Як двома способами за допомогою критерію Уїлкоксона перевірити правильність гіпотези?

5. Як визначається критичне значення й критична область для критерію Уїлкоксона?

6. Що таке медіана? Як вона визначається для вибіркових значень?

7. Що таке серія знаків та як вона визначається ?

8. Як визначаються критичне значення та критична область для критерію знаків?

## 3.6. ПЕРЕВІРКА ГІПОТЕЗ ВІДНОСНО ЧАСТКИ ОЗНАКИ ПОРІВНЯННЯ ДВОХ ВИБІРОК

Критерії цієї групи дають змогу використовувати частку ознак і приймати рішення про властивість генеральної сукупності та відповідність нормам.

Будемо розглядати задачі: порівняння частки ознаки з нормативним значенням (стандартом) та порівняння частки ознаки у двох сукупностях.

**Порівняння частки ознаки з нормативним значенням.**

Нехай потрібно перевірити гіпотезу, що частка $p$ деякої ознаки у генеральній сукупності дорівнює деякому нормативному значенню $a$, тобто висувається нуль-гіпотеза $H_0$: $p = a$ за альтернативної гіпотези $H_1$: $p \neq a$. Для перевірки гіпотези $H_0$ застосовують двобічний критерій, оскільки порушення гіпотези $H_0$ може бути як у разі $p > a$, так і в разі $p < a$. Як критерій застосовуємо статистику

$$W = m/n,$$

де $n$ – загальна кількість випробувань; $m$ – кількість позитивних результатів, $W$ – частота.

Ця статистика при будь-якому $n$ розподілена за біноміальним (для вибірки з поверненням) або за гіпергеометричним (для вибірки

без повернення) законом розподілу. Однак при достатньо великому $n$ при розрахунках можна скористатися асимптотичними розподілами, найчастіше – нормальним. Виходячи з нормального закону розподілу при заданому рівні значущості α значення $z$ знаходять із таблиць нормального розподілу згідно з рівністю:

$$P\left\{\left|\frac{m}{n}-a\right| \le z\sigma\left(m/n\right)\right\} = 1-\alpha\,.$$

Узявши до уваги, що для біноміального закону $\sigma^2\left(m/n\right) = \frac{pq}{n}$, а також, що $q = 1 - p$ і для розглянутого випадку вихідним припущенням є $p = a$, дістанемо вираз:

$$\sigma\left(m/n\right) = \sqrt{\frac{a(1-a)}{n}}\,,$$

який надалі можна використати для знаходження надійних меж для $W$. Критичні точки в цьому випадку:

$$W_{\text{кр}1} = a - z_{\alpha/2}\sqrt{\frac{a(1-a)}{n}}\,; \quad W_{\text{кр}2} = a + z_{\alpha/2}\sqrt{\frac{a(1-a)}{n}}\,.$$

Якщо вибіркова частість $W$ буде в межах $[W_1, W_2]$, гіпотеза $H_0$ приймається.

У випадку, коли перевіряють альтернативну гіпотезу $H_1$: $p > a$, використовують однобічну критичну область із граничним значенням $z$ згідно з рівнянням:

$$P\left\{\frac{m}{n} > W_{\text{кр}2}\right\} = \alpha\,.$$

**Приклад**. Перевіряється внесок деякого компонента до складу продукції.

1. Перевірити відповідність вкладу складової 10 %-вому стандартному значенню, тобто перевірити гіпотезу $H_0$: $p = 0{,}1$.

2. Перевірити, що наявність цього компонента у продукції не перевищує 10 %, тобто перевірити гіпотезу $H_1$: $p > a$.

**Порівняння частки ознаки у двох сукупностях.** Припустимо, що маємо $m_1 / n_1$ та $m_2 / n_2$ частки однієї ознаки у двох сукупностях з $n_1$ та $n_2$ одиниць. Висувається гіпотеза $H_0$: розбіжність між $m_1 / n_1$ та $m_2 / n_2$ є результатом впливу випадкових факторів та обмеженого обсягу вибірок.

Розглянемо випадок *великих вибірок*. Якщо $n_1$ та $n_2$ більші за 30, то розподіл вибіркових частостей при виконанні припущення про нульову розбіжність буде близьким до нормального з параметрами:

$$M\left(\frac{m_1}{n_1}\right) = M\left(\frac{m_2}{n_2}\right) = p;$$

$$\sigma^2\left(\frac{m_1}{n_1}\right) = \frac{p(1-p)}{n_1}; \qquad \sigma^2\left(\frac{m_2}{n_2}\right) = \frac{p(1-p)}{n_2}.$$

Для перевірки гіпотези застосовують статистику:

$$W = \frac{m_1}{n_1} - \frac{m_2}{n_2}.$$

Статистика $W$ також може бути подана за допомогою нормального закону з параметрами:

$$M(W) = M\left(\frac{m_1}{n_1} - \frac{m_2}{n_2}\right) = M\left(\frac{m_1}{n_1}\right) - M\left(\frac{m_2}{n_2}\right) = p - p = 0;$$

$$\sigma^2(W) = \sigma^2\left(\frac{m_1}{n_1} - \frac{m_2}{n_2}\right) = \sigma^2\left(\frac{m_1}{n_1}\right) + \sigma^2\left(\frac{m_2}{n_2}\right) = p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

Для перевірки гіпотези $H_0$ будемо використовувати двобічний критерій. Задаючись рівнем значущості $\alpha$, знаходимо $z$ згідно з рівнянням:

$$P\{|W| \le z\sigma(W)\} = 1 - \alpha,$$

а потім визначимо критичні точки:

$$W_{\text{кр}1} = -z\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}; \qquad W_{\text{кр}2} = z\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},$$

де $\hat{p} = (m_1 + m_2)/(n_1 + n_2)$– оцінка $p$, яку отримують на підставі наявних даних двох вибірок.

Якщо вибіркове значення $W$ знаходиться в інтервалі $[W_{\text{кр}1}, W_{\text{кр}2}]$, то гіпотеза про несуттєвість розбіжності приймається.

Розглянемо випадок *малих вибірок*. Якщо $n_1$ та $n_2$ – малі числа, то використання нормального розподілу для статистики $W = \left(\frac{m_1}{n_1} - \frac{m_2}{n_2}\right)$ є невірним. У цьому випадку необхідно використовувати критерій $\chi^2$, за допомогою якого, як було показано раніше, при ідентифікації законів розподілу можна визначити розбіжність між теоретичними і вибірковими частками.

Для розглянутого випадку $\chi^2$ обчислюється в такий спосіб. Припустимо, що нас цікавить деяка ознака $A$. Беруть дві сукупності обсягами $n_1$ та $n_2$ і результати для позитивних $A$ і негативних $\bar{A}$ наслідків заносять у табл. 3.2.

У табл. 3.2 $\bar{m}_1$ і $\bar{m}_2$ – кількість елементів у кожній вибірці, які не мають ознаки $A$. Виходячи з припущення, що вибірки взято з тієї самої генеральної сукупності з часткою ознаки $p$, можна визначити теоретичні частоти, які відповідають фактичним частотам $pn_1$, $(1-p)n_1$, $pn_2$, $(1-p)n_2$.

Таблиця 3.2. Співставлення фактичних і теоретичних частот

| Сукупність | Фактичні частоти | | | Оцінки теоретичних частот | |
|---|---|---|---|---|---|
| | $A$ | $\bar{A}$ | Усього | $A$ | $\bar{A}$ |
| Вибірка 1 | $m_1$ | $\bar{m}_1$ | $n_1$ | $pn_1$ | $(1-p)n_1$ |
| Вибірка 2 | $m_2$ | $\bar{m}_2$ | $n_2$ | $pn_2$ | $(1-p)n_2$ |
| Усього | $m_1 + m_2$ | $\bar{m}_1 + \bar{m}_2$ | $n_1 + n_2$ | – | – |

У останніх двох рядках табл. 3.2 наведені оцінки теоретичних частот, де замість $p$ використовується $\hat{p} = (m_1 + m_2)/(n_1 + n_2)$.

На підставі даних, наведених у табл. 3.2, можна обчислити $\chi^2$ за формулою:

$$\chi^2 = \frac{(m_1 - \hat{p}n_1)^2}{\hat{p}n_1} + \frac{[\overline{m}_1 - (1-\hat{p})n_1]^2}{(1-\hat{p})n_1} + \frac{(m_2 - n_2\hat{p})^2}{\hat{p}n_2} + \frac{[\overline{m}_2 - (1-\hat{p})n_2]^2}{(1-\hat{p})n_2},$$

де у знаменниках записано оцінки відповідних дисперсій.

Беручи до уваги, що між чотирма теоретичними частотами існує три незалежні співвідношення, у розподілі $\chi^2$ необхідно враховувати тільки одну степінь вільності.

Якщо нульова гіпотеза, відповідно до якої обидві сукупності є вибірками з однієї генеральної сукупності, правильна, то розбіжність між теоретичними та дослідними частотами можна віднести тільки на рахунок випадкового відбору. Тому, визначивши для рівня значущості α значення $\chi^2$, приймемо рішення про відхилення гіпотези $H_0$, якщо $\chi^2 > \chi^2_{\text{кр}}$, або про незначущість розбіжності при $\chi^2 \le \chi^2_{\text{кр}}$.

**Приклад**. Проводились випробування нового методу лікування. Одна група (експериментальна) – з 50 осіб ($n_1 = 50$) лікувалася за новим методом, а друга («традиційна»), яка складалася з 30 осіб ($n_2 = 30$), – за традиційним методом. Після завершення лікування у першій групі залишилося 9 хворих ($m_1 = 9$), а в другій – 7 ($m_2 = 7$). Необхідно перевірити суттєвість ефективності нового методу.

Обчислимо оцінку теоретичної частоти хворих після лікування:

$$\hat{p} = \frac{9+7}{50+30} = 0{,}2\,.$$

Позначимо позитивний результат для хворого – «вилікувалися за зазначений період» як подію $A$, тоді – «залишилися хворими» буде $\overline{A}$. Вихідні дані та отримані результати наведено в табл. 3.3.

Таблиця 3.3. Дані для порівняння методів лікування

| Групи, що обстежувалися | Результати дослідження | | Усього | Теоретичні результати | |
|---|---|---|---|---|---|
| | $A$ | $\overline{A}$ | | $A$ | $\overline{A}$ |
| Експериментальна | 9 | 41 | 50 | 10 | 40 |
| «Традиційна» | 7 | 23 | 30 | 6 | 24 |
| Усього: | 16 | 64 | 80 | | |

Розрахунок теоретичної кількості позитивних результатів будемо проводити відповідно за виразами $\hat{p}n_1$ та $\hat{p}n_2$. Внесемо у табл. 3.3 оцінку $\hat{p}$ теоретичної кількості $A$ та $\overline{A}$. Виходячи з даних, наведених у табл. 3.3, обчислимо значення критерію:

$$\chi^2 = \frac{(9-10)^2}{10} + \frac{(41-40)^2}{40} + \frac{(7-6)^2}{6} + \frac{(23-24)^2}{24} = 0{,}33.$$

Скориставшись таблицями розподілу $\chi^2$, для $\alpha = 0{,}05$ та степені вільності $f = 1$, знайдемо критичне значення $\chi^2_{\text{кр}} = 3{,}8.$ Таким чином, $\chi^2 < \chi^2_{\text{кр}}$, тобто робимо висновок, що розбіжність частки хворих, які залишилися в обох групах після закінчення терміну лікування, не значуща, а отже, новий метод лікування дає такий самий ефект, як і традиційний.

**Порівняння двох залежних вибірок (парні зіставлення).** Часто трапляються випадки, коли дві вибірки, які порівнюються, не можуть розглядатися як незалежні.

**Приклад.** Перевірка ефективності нової технології за результатами роботи тієї самої бригади до і після впровадження нової технології.

**Приклад**. Оцінка стану хворих до і після прийняття нових ліків.

Проводячи реєстрацію по кожному об'єкту спостережень до нововведення – $x$ і після нього – $y$, дістаємо два ряди спостережень (табл. 3.4):

Таблиця 3.4. Порівняння двох залежних вибірок

| $x_1$ | $x_2$ | … | $x_i$ | … | $x_n$ |
|---|---|---|---|---|---|
| $y_1$ | $y_2$ | … | $y_i$ | … | $y_n$ |

Таким чином, ідеться про *парні спостереження*, тобто про $n$ зв'язаних пар $(x_i, y_i)$. Якщо досліджуваний фактор впливає тільки на одну з ознак $x$ або $y$, то між цими парами спостережень фіксуватиметься суттєва розбіжність. Завдання полягає в тому, щоб визначити, коли розбіжність між парами спостережень можна віднести на рахунок випадкових відхилень, а коли вона суттєва і її потрібно пов'язувати з впливом якогось фактора. Нехай різниця між спостереженнями в кожній парі становить $d_i = x_i - y_i$. Тоді узагальненою величиною розбіжності пар спостережень може бути середня різниця:

$$\bar{d} = \sum_{i=1}^{n} d_i / n .$$

Чим менша різниця $d$, тим більш правдоподібне припущення щодо несуттєвості розбіжності між рядами спостережень. Таким чином, перевірці підлягає гіпотеза $H_0$: $d = 0$. Критерієм для перевірки може бути статистика $t = \bar{d} / S(\bar{d})$, де

$$S(\bar{d}) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2} .$$

За нормального розподілу різниць $(x_i - y_i)$ $t$-статистика має розподіл Стьюдента з кількістю степенів вільності $n - 1$. Подальший механізм перевірки не відрізняється від перевірки розбіжності середніх.

*Перевірку гіпотез широко використовують у дисперсійному аналізі та в теорії кореляції, де перевіряють гіпотези про значущість відповідних параметрів, наявність стохастичного зв'язку, суттєвість впливу випадкових величин тощо.*

**Запитання для самоперевірки**

1. Чому при порівнянні частки ознаки з нормативним значенням використовують двобічний критерій?
2. Коли застосовують критерій $\chi^2$ при порівнянні частки ознак у двох сукупностях?
3. Який критерій застосовують при порівнянні двох залежних вибірок?
4. Чому при порівнянні частки ознак у двох сукупностях потрібно вводити чотири складові?
5. Як визначають оцінку теоретичної частки?
6. Коли застосовують парні спостереження?

## РОЗДІЛ 4

# ОСНОВИ ТЕОРІЇ КОРЕЛЯЦІЙНОГО ТА РЕГРЕСІЙНОГО АНАЛІЗУ

## 4.1. УМОВНІ ЗАКОНИ РОЗПОДІЛУ

Щоб досліджувати вплив однієї випадкової величини на зміну іншої, необхідно розглядати умовні закони розподілу ймовірностей однієї величини при фіксованих значеннях іншої величини.

Статистичні дані про двовимірні величини подають у вигляді таблиці, яку теоретично можна розглядати як вектор або точку з випадковими координатами ($X$, $Y$) (п. 1.8).

Розглянемо систему двох дискретних випадкових величин $X$ і $Y$, поданих таблицею, де як заголовки рядків та стовпців використовують можливі значення випадкових величин $X$ і $Y$, а на їх перетині, тобто в комірках таблиці, містяться ймовірності одночасної появи відповідних комбінацій цих випадкових величин.

Таким чином, у комірці з координатами ($x_i$, $y_j$) указано ймовірність $p(x_i, y_j)$ того, що випадкова величина $X$ набуде значення $x_i$, тобто $X = x_i$, тоді як друга випадкова величина $Y$ набуде значення $y_j$, тобто $Y = y_j$.

У табл. 4.1 наведено всі можливі сполучення ($x_i$, $y_j$), $i = \overline{1,n}$; $j = \overline{1,m}$, які утворюють повну групу подій, тому можна записати:

$$\sum_i \sum_j p(x_i, y_j) = 1.$$

Знайдемо для $i$-го стовпця таблиці 4.1 суму ймовірностей усіх комірок:

$$\sum_j p(x_i, y_j) = p(x_i, y_1) + p(x_i, y_2) + \ldots + p(x_i, y_j) + p(x_i, y_m). \quad (4.1)$$

Події $X = x_i$,, $Y = y_j$ є несумісними, тому вираз (4.1) можна подати як

$$\sum_j p(x_i, y_j) = p(x_i)p(y_1) + p(x_i)p(y_2) + \ldots + p(x_i)p(y_j) + \ldots +$$
$$+ p(x_i)p(y_m) = p(x_i)\sum_j p(y_j).$$

Беручи до уваги, що сума ймовірностей усіх можливих значень дискретної випадкової величини дорівнює одиниці, можна записати:

$$\sum_j p(x_i, y_j) = p(x_i). \qquad (4.2)$$

Таблиця 4.1. Можливі сполучення ($x_i$, $y_j$), які утворюють повну групу подій

| $Y$ | $X$ | | | | | | |
|---|---|---|---|---|---|---|---|
| | $x_1$ | $x_2$ | … | $x_i$ | … | $x_n$ | $\sum_i$ |
| $y_1$ | $p(x_1y_1)$ | $p(x_2y_1)$ | … | $p(x_iy_1)$ | … | $p(x_ny_1)$ | $p(y_1)$ |
| $y_2$ | $p(x_1y_2)$ | $p(x_2y_2)$ | … | $p(x_iy_2)$ | … | $p(x_ny_2)$ | $p(y_2)$ |
| … | … | … | … | … | … | … | … |
| $y_j$ | $p(x_1y_j)$ | $p(x_2y_j)$ | … | $p(x_iy_j)$ | … | $p(x_ny_j)$ | $p(y_j)$ |
| … | … | … | … | … | … | … | … |
| $y_m$ | $p(x_1y_m)$ | $p(x_2y_m)$ | … | $p(x_iy_m)$ | … | $p(x_ny_m)$ | $p(y_m)$ |
| $\sum_j$ | $p(x_1)$ | $p(x_2)$ | … | $p(x_i)$ | … | $p(x_n)$ | 1 |

Аналогічно для рядка таблиці, де фіксованим є значення випадкової величини $Y = y_j$, а випадкова величина $X$ набуває всіх можливих значень $x_i$, маємо:

$$\sum_i p(x_i, y_j) = p(y_j)\sum_i (x_i) = p(y_j).$$

Отже, якщо відомий закон розподілу двовимірної величини $X$ і $Y$, то можуть бути окремо визначені одновимірні закони розподілу випадкової величини $X$ або $Y$.

Таким чином, двовимірні закони показують деякий зв'язок між випадковими величинами $X$ і $Y$. Раніше було встановлено (п. 1.8), що залежність між двома випадковими подіями проявляється в тому, що умовна ймовірність появи однієї події при появі другої відрізняється від безумовної ймовірності першої події.

Для двовимірної величини маємо аналогічну ситуацію. Різниця полягає лише в тому, що фіксованому значенню однієї величини, наприклад $X = x_i$, відповідає сукупність можливих значень другої випадкової величини $Y$. Сукупність можливих значень $y_j$ випадкової величини характеризується відповідним законом розподілу. Беручи до уваги, що випадкова величина $X$ набуває значень $X = x_i$ з імовірністю $p(x_i)$, для характеристики двовимірних розподілів широко використовують *умовні розподіли*, що дають змогу аналізувати, як зміна значення однієї випадкової величини впливає на зміну розподілу можливих значень іншої випадкової величини.

*Згадаємо! Умовною ймовірністю $P(B|A)$ називають імовірність події $B$, обчислену за припущення, що подія $A$ вже відбулась.*

Умовна ймовірність події $Y = y_j$, якщо мала місце подія $X = x_i$ з ймовірністю $p(x_i)$, буде дорівнювати

$$p(y_j \mid x_i) = \frac{p(x_i, y_j)}{p(x_i)}. \qquad (4.3)$$

При розгляді випадкових величин $X$ і $Y$ існує *сукупність умовних імовірностей* $p(y_1|x_i)$, $p(y_2|x_i)$,…,$p(y_j|x_i)$,…, яка відповідає тій самій умові $X = x_i$. Сукупність умовних імовірностей утворює ***умовний розподіл*** $Y$ *при* $X = x_i$, для якого забезпечується властивість імовірностей – ***властивість нормування***:

$$\sum_j p\left(y_j \mid x_i\right) = \sum_j \frac{p\left(x_i, y_j\right)}{p(x_i)} = \frac{\sum_j p\left(x_i, y_j\right)}{p\left(x_i\right)} = \frac{p\left(x_i\right)}{p\left(x_i\right)} = 1.$$

Аналогічно можна записати:

$$\sum_i p(x_i \mid y_j) = \frac{\sum_i p(x_i, y_j)}{p\left(y_j\right)} = 1.$$

**Приклад**. При виготовленні деталі можливі відхилення від номінальних розмірів як за довжиною, так і за товщиною, які мають випадковий характер. У табл. 4.2 наведено можливі комбінації цих відхилень та відповідні їм ймовірності

Таблиця 4.2. Можливі відхилення від номінальних розмірів як за довжиною, так і за товщиною

| $X$ | $Y$ | | | |
|---|---|---|---|---|
| | –1 | 0 | 1 | $p(x_i)$ |
| –2 | 0,15 | 0,35 | 0,05 | 0,55 |
| 3 | 0,10 | 0,25 | 0,10 | 0,45 |
| $p(y_j)$ | 0,25 | 0,60 | 0,15 | $\sum_{i=1}^{2}\sum_{j=1}^{3} p\left(x_i, y_j\right) = 1{,}0$ |

Виходячи з наведених даних, можна обчислити умовні ймовірності $y_i$ для фіксованих значень $X = x_1$ і $X = x_2$. Припустимо, що необхідно знайти умовні ймовірності $p\left(y_j \mid x_1\right)$, тобто як змінюється ймовірність появи значень випадкової величини $Y$ при фіксованому значенні величини $X = x_1$. Для цього необхідно розглянути перший рядок табл. 20.2 і скористатися формулою (4.2).

Таким чином, дістанемо значення, які занесено в табл. 4.3.

Таблиця 4 3. Чисельні значення умовних ймовірностей

| $Y$ | –1 | 0 | 1 | $\sum_{j=1}^{3} p\left(y_j \middle\| x_1\right)$ |
|---|---|---|---|---|
| $p(y_j \mid x_1)$ | 0,275 | 0,636 | 0,091 | 1,0 |

Отже, ***умовні закони розподілу ймовірностей*** відображають вплив однієї випадкової величини на зміну іншої.

Двовимірна величина може характеризуватися законами розподілів – спільними й умовними. Найбільш інформативним законом розподілу є *умовний закон розподілу*, який, як уже відомо, можна отримати із спільного.

Основні властивості умовних розподілів відображаються, як і для одновимірних розподілів, за допомогою числових характеристик: *умовного математичного сподівання*, *умовної дисперсії* та *моментів більш високих порядків.*

Найбільш важливою характеристикою, яка часто використовується на практиці, є *умовне математичне сподівання* $M(Y|x)$ величини $Y$ при фіксованому значенні $X = x$. Таким чином, для кожного з можливих значень $x$ існує «своє» математичне сподівання іншої випадкової величини $Y$ (для величини $Y$ розглядають усю сукупність можливих значень), яке залежить від значення величини $X$. З урахуванням виразу (4.3) можна записати:

$$M(Y \mid x) = \sum_j y_j p(y_j \mid x) = \sum_j \frac{y_j p(x, y_j)}{p(x)}. \qquad (4.4)$$

Іноді замість $M(Y \mid x)$ використовують позначення $\bar{y}(x)$. Беручи до уваги вирази (4.2) і (4.3), вираз (4.4) можна записати як

$$\bar{y}(x) = \frac{\sum_j y_j p(x, y_j)}{\sum_j p(x, y_j)}.$$

Останній вираз дає змогу розглядати $M(Y|x)$ як центр мас $p(x_i, y_j)$, розміщених на вертикальній прямій $X = x = \text{const}$ (рис. 4.1).

Зі зміною $x$, тобто при переході від одного стовпця табл. 4.1 до іншого, у загальному випадку змінюється й значення умовного математичного сподівання $M(Y|x)$.

Сукупність умовних математичних сподівань $M(Y|x)$, визначених для значень $x_1, x_2, ..., x_i, ..., x_N$, називається ***регресією***

випадкової величини $Y$ відносно випадкової величини $X$, тобто *регресія є функцією умовних математичних сподівань*.

Хоча для кожного значення $X = x$ величина $Y$ залишається випадковою з розсіюванням своїх значень, залежність $Y$ від $X$ проявляється у зміні центрів умовних розподілів $\bar{y}(x)$ при переході від одного значення $x$ до іншого. Цю залежність у середньому описує ***крива регресії*** $\bar{y}(x)$ (рис. 4.1).

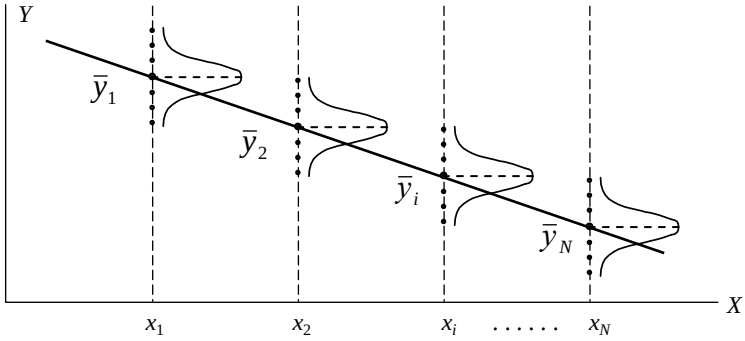



Рис. 4.1. Побудова лінії регресії

Беручи до уваги, що величини $X$ і $Y$ випадкові, можна розглядати умовні закони розподілу ймовірностей величини $X$ за фіксованих значень $Y = y_j$, що обумовлені сукупністю умовних ймовірностей $p(x_i|y_j)$. Тобто, можна для фіксованих значень $Y = y_j$, розглядати *регресію величини X на Y:*

$$\bar{x}(y) = M(X \mid y) = \sum_i x_i\, p(x_i \mid y) = \frac{\sum\limits_i x_i\, p(x_i, y)}{p(y)} = \frac{\sum\limits_i x_i\, p(x_i, y)}{\sum\limits_i p(x_i, y)}.$$

Таким чином, маємо два умовні математичні сподівання: одне з них –

$$\bar{y}(x) = M(Y \mid x) \qquad (4.5)$$

є функцією від $x$ і називається *регресією величини Y на величину X*, а друге –

$$\bar{x}(y) = M(X \mid y) \tag{4.6}$$

називається *регресією величини X на величину Y* або ***оберненою регресією.***

Криві в площині *X, Y*, обумовлені рівняннями (4.5), (4.6), називаються ***лініями регресії*** *Y* відносно *X* та відповідно *X* відносно *Y*.

Графіки цих функцій наведено на рис. 4.2, крива *1 – лінія регресії Y* відносно *X* показує, як у *середньому* змінюється величина *Y* при зміні величини *X*; крива *2* – графік *оберненої регресії* .

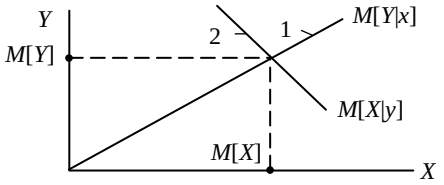



Рис. 4.2. Пряма та обернена лінії регресії

Варто зазначити, що функції регресії $M(Y \mid x)$ і $M(X \mid y)$ не є взаємно оберненими.

Найбільш простий випадок, коли обидві функції регресії – *лінійні*, тобто обидві лінії регресії будуть прямими. У цьому випадку вони називаються ***прямими регресії***, а випадкові величини *X* та *Y* називаються *лінійно корельованими*.

У випадку неперервних величин *X* та *Y* математичне сподівання величини *Y* за умови, що величина *X* набула конкретного значення *x*, набирає вигляду:

$$M(Y \mid X = x) = \int_{-\infty}^{\infty} y p_{Y|x}(y) dy .$$

Аналогічно для $M(X|Y = y)$:

$$M(X \mid Y = y) = \int_{-\infty}^{\infty} x p_{X|y}(x) dx .$$

*Тут і надалі при розгляді неперервних величин операція* $\Sigma$ *замінюється на операцію* $\int$ *з межами інтегрування від* $-\infty$ *до* $+\infty$*, тобто за всіма можливими значеннями випадкової величини на елементарних інтервалах dx. Це означає, що для неперервної випадкової величини можна використовувати вирази, отримані для дискретної випадкової величини.*

У випадку неперервного розподілу величин $X$ і $Y$ їхній спільний розподіл задається за допомогою спільної щільності ймовірностей – функції $p_{XY}(x, y)$.

За аналогією з дискретними випадковими величинами, можна подати умовні математичні сподівання:

$$\bar{y}(x) = M(Y \mid x) = \int_{-\infty}^{+\infty} y p_{XY}(y \mid x)dy = \frac{\int_{-\infty}^{+\infty} y p_{XY}(x,y)dy}{\int_{-\infty}^{+\infty} p_{XY}(x,y)dy};$$

$$\bar{x}(y) = M(X \mid y) = \int_{-\infty}^{+\infty} x p_{XY}(x \mid y)dx = \frac{\int_{-\infty}^{+\infty} x p_{XY}(x,y)dx}{\int_{-\infty}^{+\infty} p_{XY}(x,y)dx}.$$

*Таким чином, розсіювання* $y_j$ *буде групуватися навколо деякого центра, яким є умовне математичне сподівання.*

*Геометричне місце точок центрів розподілів утворює лінію регресії, яка показує, яке в середньому (очікуване) значення набуде випадкова величина Y для можливих значень випадкової величини X. Залежність випадкової величини Y від випадкової величини X проявляється у зміні* $\bar{y}(x_i)$ *за* $i = \overline{1,n}$.

Усі ці поняття можуть бути легко узагальнені на будь-яку кількість випадкових величин.

**Приклад**. У випадку тривимірної величини $(X,Y,Z)$ уводиться тривимірна щільність $p_{XYZ}(x, y, z)$, виходячи з якої можна обчислити три двовимірні щільності:

$$p_{XY}(x, y),\ p_{XZ}(x, z) \text{ і } p_{YZ}(y, z)$$

і три одновимірні щільності

$$p(x)\ p(y)\ p(z).$$

*Регресія посідає важливе місце в статистичній обробці даних. На її основі будуються аналітичні залежності (математичні моделі), які дають змогу не тільки якісно, але й кількісно оцінити стохастичний взаємозв'язок між величинами.*

### Запитання для самоперевірки

1. Що називається умовною ймовірністю? Що називається комбінацією випадкових подій?

2. Для чого вводиться в розгляд поняття умовного закону розподілу ймовірностей? Що називається умовним розподілом випадкової величини?

3. Якою властивістю ймовірностей характеризується умовний розподіл?

4. Які числові характеристики використовуються для опису умовних законів розподілів?

5. Що таке регресія? Яке значення на практиці вона має?

6. Чому дорівнює $\bar{x}(y)$? Чому дорівнює $\bar{y}(x)$?

7. Що таке рівняння регресії, лінійна регресія, лінія регресії, пряма регресія, обернена регресія?

8. Поясніть, що таке умовне математичне сподівання випадкової величини і який його зв'язок з лінією регресії?

## 4.2. ПОНЯТТЯ ПРО КОРЕЛЯЦІЙНИЙ АНАЛІЗ

Теорію кореляції фактично можна назвати теорією прогнозування, коли вказуються межі, в яких з наперед заданою надійністю перебуватиме випадкова величина *Y*, якщо інші пов'язані з нею величини *X* набувають певних значень.

У будь-якій галузі науки і техніки постає потреба у вивченні залежності між різними ознаками. Зміст статистичної обробки полягає в тому, щоб *установити залежність* (*зв'язок*) *між ознаками*.

Ці залежності, їхній якісний прояв намагаються, зазвичай, подати в кількісній формі. Найбільш простим видом зв'язку є ***функціональна*** залежність, коли кожному значенню однієї величини відповідає певне значення іншої. З функціональною залежністю часто стикаються в техніці та природознавстві.

**Приклад**. Залежність між висотою місцевості та тиском атмосфери.

**Приклад**. Значення спаду напруги на активному опорі при заданому значенні струму, що проходить через нього.

Однак часто трапляються і такі зв'язки між величинами, які не можна віднести до функціональних залежностей.

**Приклад.** Зв'язок між врожаєм та кількістю опадів, між зростом батьків та дітей.

У наведених прикладах кожному значенню однієї величини відповідає множина можливих значень іншої, тобто зміна однієї величини викликає зміну *розподілу* іншої величини. Випадковий розкид цих можливих значень пояснюється впливом великої кількості додаткових факторів, які не враховують, вивчаючи зв'язок між даними величинами. Такі залежності називаються ***стохастичними*** або ***статистичними*** (п. 1.8).

*Окремим випадком статистичної залежності є кореляційна залежність, коли середнє значення однієї випадкової величини функціонально залежить від значень іншої.*

Наявність додаткових впливових факторів призводить до неможливості зробити однозначні висновки про залежність, яка нас цікавить.

При встановленні виду фактичної залежності зазвичай результати вимірювань або спостережень фіксують у *таблиці спостережень* (*кореляційній таблиці*):

Таблиця 4.4. Кореляційна таблиця

| X | $x_1$ | $x_2$ | … | $x_n$ |
|---|---|---|---|---|
| Y | $y_1$ | $y_2$ | … | $y_n$ |

або зображують на координатній площині (***кореляційному полі***) у вигляді точок, координатами яких є значення ознак $X$ і $Y$ одного об'єкта $(x_i, y_i)$, $i = \overline{1,n}$. У такий спосіб будується ***діаграма розсіювання*** (рис. 4.3).

З рис. 4.3 бачимо, що у випадку, наведеному на рис. 4.3 *а*, варто шукати нелінійну залежність, у випадку рис. 4.3, *б* – лінійну залежність, а у випадку на рис. 4.3, *в* на перший погляд, навряд чи якась залежність існує.

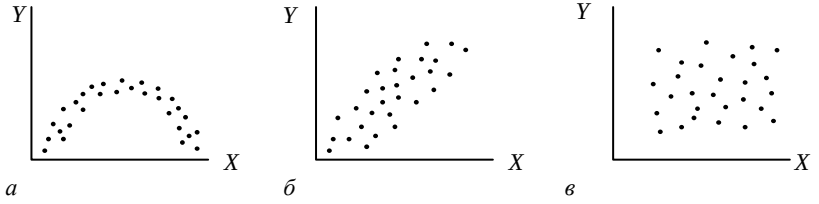



Рис. 4.3. Проклади поля кореляції

**Коефіцієнт кореляції**. Щільність стохастичного *лінійного* зв'язку між величинами $X$ і $Y$ незалежно від їхнього роду характеризує коефіцієнт кореляції, який визначається за формулою (1.20).

Графічну інтерпретацію коефіцієнта кореляції наведено на рис. 4.4. Якщо $\rho > 0$, маємо додатну кореляцію, яка означає, що при зростанні однієї величини інша також зростає. У випадку $\rho < 0$ існує від'ємна кореляція, коли при зростанні однієї величини інша має тенденцію в середньому спадати. За відсутності статистичного зв'язку між величинами $X$ та $Y$ значення $\rho = 0$.

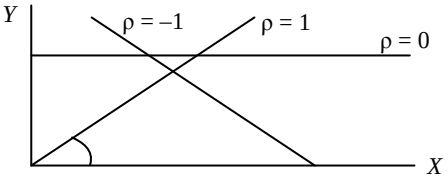



Рис. 4.4. Відображення коефіцієнтів кореляції

У будь-якому випадку значення коефіцієнта кореляції чисельно дорівнює косинусу кута нахилу прямої регресії.

**Оцінка коефіцієнта кореляції**. Для оцінки коефіцієнта кореляції між двома випадковими величинами $X$ і $Y$ виконують ряд незалежних випробувань, результатом кожного з яких є пара величин $(x_i, y_i)$. Двовимірні величини $(x_1, y_1)$, $(x_2, y_2)$,…, $(x_n, y_n)$ незалежні й відповідають тому самому закону розподілу, що й двовимірна величина $(X, Y)$.

Будується діаграма розсіювання й робиться припущення щодо виду залежності між випадковими величинами $X$ і $Y$.

Використовуючи вираз для теоретичного коефіцієнта кореляції (п.. 1.8):

$$\rho = \frac{cov(X,Y)}{\sigma_x \sigma_y},$$

знаходять емпіричний коефіцієнт кореляції:

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}},$$

де $\bar{x} = \frac{1}{N}\sum_{i=1}^{N}(x_i),\ \bar{y} = \frac{1}{N}\sum_{i=1}^{N}(y_i).$

У цьому виразі математичні сподівання замінені на відповідні середні арифметичні, а дисперсії – на вибіркові дисперсії, що є незміщеними і спроможними оцінками центрів розподілів і дисперсій.

*Зазвичай прийнято вважати, що при лінійній кореляції значення r > 0,7 свідчать про високий коефіцієнт зв'язку між величинами;*
*0,7 < r < 0,3 – про середній коефіцієнт зв'язку;*
*r < 0,3 – про слабкий зв'язок.*

**Надійний інтервал для коефіцієнта кореляції**. Емпіричний коефіцієнт кореляції $r$ є спроможною точковою оцінкою коефіцієнта кореляції $\rho$. Для того щоб говорити про статистичну надійність цієї оцінки, тобто вірогідність отриманого результату, необхідно використовувати *інтервальну оцінку*, тобто знаходити *надійний інтервал* для $\rho$.

Якщо випадкові величини *X* і *Y* близькі до нормального закону розподілу, то для великих вибірок $(n > 30\ldots50)$ можна знайти оцінку середньоквадратичного відхилення коефіцієнта кореляції *r*, обчислену відповідно до виразу:

$$S_r = \frac{1-r^2}{\sqrt{n}},$$

і вважати, що *r* наближено має нормальний закон розподілу з параметрами $(\rho, \sigma_r)$, де як СКВ генеральної сукупності $\sigma_r$ використовується $S_r$, тобто $\sigma_r = S_r$.

Тоді можна, скориставшись таблицями нормального розподілу, визначити $z$, що дає змогу обчислити межі надійного інтервалу для коефіцієнта кореляції ρ:

$$r - z\frac{1-r^2}{\sqrt{n}} \le \rho \le r + z\frac{1-r^2}{\sqrt{n}}. \qquad (4.7)$$

Цей вираз означає, що з імовірністю $1-\alpha$ знайдений інтервал накриє істинне значення коефіцієнта кореляції ρ. Даними виразами можна користуватися для великих вибірок

На практиці зазвичай обсяги спостережень невеликі. У випадку, коли $n < 30$, необхідно користуватися статистичною характеристикою, яку запропонував Р. Фішер. Ним було доведено, що величина

$$Z = \frac{1}{2} ln\left(\frac{1+r}{1-r}\right) \qquad (4.8)$$

приблизно розподілена за нормальним законом з математичним сподіванням

$$M[Z] \approx \frac{1}{2} ln\left(\frac{1+\rho}{1-\rho}\right),$$

і дисперсією

$$\sigma^2(Z) \approx \frac{1}{n-3}. \qquad (4.9)$$

Три степені вільності втрачаються на обчислення $S_x$, $S_y$ і $r$.

Надійний інтервал для вибіркового значення коефіцієнта кореляції визначається за формулою:

$$Z - z\sigma(Z) \le \frac{1}{2} ln\frac{1+\rho}{1-\rho} \le Z + z\sigma(Z).$$

**Перевірка гіпотези про значущість коефіцієнта кореляції.** Беручи до уваги, що оцінки коефіцієнту кореляції $r$ обчислювалися за обмеженого обсягу даних, навіть за відсутності стохастичного зв'язку коефіцієнт кореляції $r$ може не дорівнювати нулю. З огляду

на це необхідно встановити, чи є знайдені оцінки *r статистично значущими*. Висувається гіпотеза $H_0$: $\rho = 0$ за альтернативної гіпотези $H_1$: $\rho \neq 0$. Для перевірки гіпотези використовують статистичний коефіцієнт кореляції *r*.

За великих вибірок, користуючись співвідношенням (4.7), будують область з критичними точками $\pm z\sigma_r$ за рівня значущості α. Якщо отримане за даними вибірки значення *r* виявиться у критичній області, тобто $|r| > z\sigma_r$, гіпотезу $H_0$ відхиляють.

У випадку малого обсягу вибірки можна скористатися такою властивістю, що величина

$$t_{n-2}\left(n-2+t^2{}_{n-2}\right)^{-1/2}$$

за $\rho = 0$ має такий самий розподіл, що й *r*. Тоді отримати границю довірчого інтервалу для $|r|$ можна, скориставшись таблицями розподілу Стьюдента із заданим рівнем статистичної значущості α для степені вільності *n*-2.

Існує інший варіант критерію значущості *r*, який оснований на використанні апроксимації, запропонованої Р. Фішером, у відповідності до якого знаходиться *r*, а потім за формулою (4.8) обчислюється Z. Якщо $\rho = \rho_0$, то статистика

$$U = \left[ Z - \frac{1}{2}\ln\frac{1+\rho}{1-\rho} \right](n-3)^{1/2} \approx \frac{Z - M(Z)}{\sigma(Z)}$$

має розподіл, близький до нормального, з математичним сподіванням, що дорівнює нулю, і дисперсією, що дорівнює одиниці.

Для цієї величини $U$ застосовується критерій значущості як для нормального розподілу, тобто за таблицями знаходять критичні значення $U_{кр}$ і далі діють за відомим алгоритмом.

*Якщо знайдений надійний інтервал для* $\rho$ *включає число нуль, то гіпотеза* $H_0$ *приймається.*

**Кореляційне відношення**. У разі нелінійної залежності величин $X$ і $Y$ оцінка щільності зв'язку за допомогою коефіцієнта кореляції може призвести до помилкових рішень.

**Приклад**. Нехай величина $X$ симетрично розподілена біля початку координат, який є центром розподілу величини $X$, тобто $MX = 0$. Випадкова величина $X$ функціонально зв'язана з випадковою величиною $Y$ залежністю $Y = X^2$. Згідно з симетрією $X$ відносно до початку координат:

$$cov\,(X,Y) = M(X,X^2) = M(X^3) = 0,$$

оскільки математичне сподівання непарної степені центрованої випадкової величини дорівнює нулю. У результаті, з одного боку, маємо функціонально зв'язані величини, а з другого – кореляція між ними дорівнює нулю.

Тому критеріями щільності нелінійного зв'язку є показники, які характеризують концентрацію дослідних точок біля кривих регресії.

Найбільш поширеним показником щільності нелінійного зв'язку є ***кореляційне відношення*** $\eta_{Y|x}$ чи $\eta_{X|y}$, яке запропонував К. Пірсон.

Вихідним є подання дисперсії змінної величини $Y$ у вигляді двох складових:

$$\sigma_Y^2 = M\left(Y - \nu_Y\right)^2 = M\left[\left(Y - \bar{y}\left(X\right)\right) + \left(\bar{y}\left(X\right) - \nu_Y\right)\right]^2. \qquad (4.10)$$

Перша складова правої частини виразу (4.10)

$$M\left(Y - \bar{y}\left(X\right)\right)^2 = \sigma_{Y|x}^2$$

є середньою умовною дисперсією і характеризує степінь розсіювання випадкової змінної $Y$ стосовно кривої регресії.

Друга складова виразу – є *систематичною складовою* загальної дисперсії величини *Y*, яка характеризує відхилення кривої регресії, що апроксимує нелінійну залежність, від математичного сподівання.

Таким чином:

$$\sigma_Y^2 = \sigma_{Y|x}^2 + M\left[\bar{y}(X) - \nu_Y\right]^2.$$

Кореляційне відношення визначається як

$$\eta_{Y|x}^2 = \frac{M\left[\bar{y}(X) - \nu_Y\right]^2}{\sigma_Y^2} \qquad (4.11)$$

і показує, яку частину у загальній дисперсії займає систематична її складова.

Враховуючи, що

$$M\left[\bar{y}(X) - \nu_Y\right]^2 = \sigma_Y^2 - \sigma_{Y|x}^2,$$

остаточно отримаємо

$$\eta_{Y|x}^2 = 1 - \frac{\sigma_{Y|x}^2}{\sigma_Y^2}.$$

*Властивості кореляційного відношення*:

- значення кореляційного відношення завжди лежить у межах від 0 до 1;
- якщо $\eta_{Y|x} = \eta_{X|y} = 1,$ то випадкові складові обох дисперсій дорівнюють нулю, тобто залежність між *X* і *Y* є функціональною;
- якщо $\eta_{Y|x} = 0,$ то величини *X* і *Y* – некорельовані;
- у загальному випадку $\eta_{Y|x} \neq \eta_{X|y},$ тобто міра нелінійного зв'язку несиметрична;
- співвідношення між коефіцієнтом кореляції *r* і кореляційними відношеннями $\eta_{Y|x}$ та $\eta_{X|y}$ визначаються нерівностями:

$$|r| \le \eta_{Y|x}; \qquad |r| \le \eta_{X|y};$$

– тому через рівність нулю будь-якого з показників $\eta_{Y|x}$ чи $\eta_{X|y}$ випливає, що $r = 0$;

– близькість величин кореляційного відношення до коефіцієнта кореляції свідчить про те, що залежність наближено може вважатися лінійною.

Емпірично кореляційне відношення оцінюється, виходячи з виразу:

$$\hat{\eta}_{Y|x} = \frac{\sum_{i=1}^{N} m_i (\bar{y}_i - \bar{y})^2}{\sum_{i=1}^{N} \sum_{j=1}^{m_i} (\tilde{y}_{ij} - \bar{y})^2},$$

де $\bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} \tilde{y}_{ij}$ – умовне середнє, отримане для значень $x_i$ за $j = \overline{1, m_i}$;

$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} \bar{y}_i$ – загальне середнє за всіма експериментальними точками;

$\tilde{y}_{ij}$ – результати спостережень у $i$-ій експериментальній точці.

**Множинна кореляція**. Процес, який характеризується залежностями між трьома, чотирма та більш змінними, вивчають за допомогою методів множинної кореляції. Розглядаючи одну зі змінних ($y$) як функцію, а решту ($u, v, \ldots, t$) як аргументи, можна визначити середні значення $y$ для будь-якої сукупності значень $u, v, \ldots, t$ і скласти рівняння:

$$\bar{y} = \bar{y}(u, v, \ldots, t).$$

У разі кореляції між трьома змінними рівняння регресії геометрично зображується у вигляді деякої поверхні, навколо якої розсіяні дослідні точки.

***Множинне кореляційне відношення***. За аналогією з парною кореляцією загальну дисперсію змінної *Y*, що розглядається як функція, можна подати у вигляді суми систематичної $\sigma^2_{\bar{y}(u,v,\ldots,t)}$ і випадкової $\sigma^2_{y|u,v,\ldots,t}$ складових:

$$\sigma^2_Y = \sigma^2_{\bar{y}(u,v,\ldots t)} + \sigma^2_{y|u,v,\ldots t}\,.$$

Тоді відношення

$$\eta_{Y|u,v,\ldots t} = \frac{\sigma^2_{\bar{y}(u,v,\ldots t)}}{\sigma_Y}$$

називається *множинним кореляційним відношенням y* за *u, v, …, t*.

Як і в разі парної кореляції, множинне кореляційне відношення міститься в межах від 0 до 1. При $\eta = 0$ спостерігається відсутність кореляції між *Y* та *u, v, …, t*. Стосовно кореляції трьох змінних поверхня регресії в цьому випадку буде площиною з рівнянням $y = C$, яка паралельна координатній площині (*u*, v).

***Сукупний коефіцієнт кореляції***. В особливо важливих для практики випадках, коли рівняння множинної регресії є *лінійним*, множинне кореляційне відношення $\eta_{Y|u,v,\ldots,t}$ перетворюється на множинний, або сукупний, коефіцієнт кореляції $R_{Y|u,v,\ldots,t}$. Сукупний коефіцієнт кореляції пов'язаний з коефіцієнтами кореляції для кожної пари змінних певними співвідношеннями. Зокрема, для лінійної кореляції величини *y* з величинами *u, v* таке співвідношення має вигляд:

$$R_{Y|u,v} = \sqrt{\frac{r^2_{yu} + r^2_{yv} - 2r_{yu}r_{yv}r_{uv}}{1 - r^2_{uv}}}\,.$$

Сукупний коефіцієнт кореляції $R_{Y|u,v}$ та коефіцієнти парної кореляції між *y* і кожною з величин *u, v* пов'язані співвідношеннями:

$$R_{Y|u,v,\ldots,t.} \geq |r_Y|\,; \quad R_{Y|u,v,\ldots,t.} \geq |r_Y|.$$

Аналогічні співвідношення спроваджуються і для кореляції будь-якої кількості змінних.

***Частинні коефіцієнти множинної кореляції***. Якщо при вивченні кореляції трьох змінних знайдено коефіцієнти кореляції між кожною парою з них, то з'являється можливість виключити одну з цих змінних, тобто визначити кореляцію двох інших змінних за умови, що ця величина залишається сталою. Коефіцієнти кореляції, що характеризують отримані при цьому залежності, називаються *частинними коефіцієнтами кореляції*. Якщо занумерувати три змінні числами 1, 2, 3, то загальну формулу для визначення частинних коефіцієнтів кореляції можна подати у вигляді:

$$r_{12(3)} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{\left(1 - r_{13}^2\right)\left(1 - r_{23}^2\right)}},$$

де $r_{12(3)}$ – частинний коефіцієнт кореляції між змінними 1 і 2 при виключенні змінної 3.

Якщо кількість змінних множинної кореляції перевищує три, за допомогою останньої формули можуть бути послідовно виключені дві і більше змінних.

Частинний коефіцієнт кореляції може не тільки значно відрізнятися за значенням від відповідного загального коефіцієнта кореляції, а й мати інший знак.

**Рангова кореляція**. Існують ознаки, які не піддаються безпосередньому кількісному оцінюванню, однак мають ряд якісних градацій, які дають змогу порівнювати між собою окремі об'єкти за степеню вільності цієї ознаки. Такі ознаки називаються *якісними*.

**Приклад**. Дисперсність розчину, інтенсивність болю.

У тих випадках, коли ознакам явищ, які спостерігаються, не вдається однозначно приписати ті або інші абсолютні значення, користуються методами *рангової кореляції*. За допомогою цих методів

вдається значно розширити можливості статистичного аналізу. Останнім часом при розв'язуванні задач діагностики та прогнозуванні широко застосовуються експертні системи, для яких методи рангової кореляції є майже не єдиним шляхом узагальнення експертних оцінок.

Послідовність об'єктів, які розміщено відповідно до деякої ознаки, називається *впорядкованою*. Сам процес такого впорядкування називається *ранжуванням*, а кожному елементу *ранжованого ряду* присвоюється *ранг*. Найчастіше ранги позначають порядковими числами 1, 2, ..., *n*. Таким чином, якщо після впорядкування об'єкт займає п'яту позицію в ранжованому ряді, то він має ранг 5. Припустимо, що при ранжуванні деякої сукупності за ознакою *A* ранг об'єкта дорівнює 5, тоді як при ранжуванні за ознакою *B* його ранг становить 8. Таким чином, при ранжуванні за ознакою *B* кількість вищих за рангом об'єктів буде на 3 більшою, ніж при ранжуванні за ознакою *A*. Можна зробити висновок, що цифра 3 вже являє собою не порядковий, а *кількісний* показник.

Кількісну характеристику, яка може змінювати своє значення при переході від одного елемента сукупності до другого, вважають випадковою величиною. Порівнюючи для кожного об'єкта сукупності рангів за першою та другою ознаками, можна встановити показники зв'язку між ними, які називаються ***коефіцієнтами рангової кореляції.***

**Приклад.** Значення зросту, об'єму легень тощо, які можуть бути виміряні, являють собою в межах відповідної шкали вимірювання випадкові величини.

Таку сукупність завжди можна впорядкувати, керуючись місцем, яке займає на шкалі вимірювання кожний об'єкт. Після цього можна казати, що значення випадкової величини подано відповідними рангами.

Якщо у групі є об'єкти, які не різняться за даною ознакою, то кожній парі, трійці тощо таких об'єктів призначається середній ранг,

який дорівнює середньому арифметичному з тих рангів, які б мали ці об'єкти, якщо їх можна було б розрізнити.

Таким чином, існує можливість розглядати процес упорядкування як не зовсім точний спосіб подання порядкових відношень між елементами, бо він не дозволяє судити про те, наскільки близько один до одного розміщені на шкалі вимірювання різні елементи розглядуваної сукупності. Але процес ранжування, програючи в точності, виграє у загальності підходу.

*Методи рангової кореляції доцільно застосовувати за невеликої кількості спостережень і значній кількості якісних градацій ознак.*

***Коефіцієнт рангової кореляції Кендалла.***

**Приклад.** Припустимо, що групу учнів розташовано згідно з їхніми здібностями, які вони виявили на уроках музики і математики. Можна їх розташувати, наприклад, за зростом і масою тіла, тобто за кількісними ознаками. Позначимо учнів літерами від *A* до *J* і запишемо наведені далі дві послідовності рангів, вважаючи, що ознаки, які беруться для дослідження, – α і β.

Таблиця 4.5. Дані для обчислення коефіцієнта Кендалла

| | Учні | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Ознака | *A* | *B* | *C* | *D* | *E* | *F* | *G* | *H* | *I* | *J* |
| Математика α | 7 | 4 | 3 | 10 | 6 | 2 | 9 | 8 | 1 | 5 |
| Музика β | 5 | 7 | 3 | 10 | 1 | 9 | 6 | 2 | 8 | 4 |

Постає питання: чи існує залежність між ознаками α і β (музикальними та математичними здібностями, зростом та масою тіла)? Наявність чи відсутність зв'язку між цими ознаками буде більш наочною, якщо ми розмістимо елементи першого ряду в порядку зростання у послідовності натуральних чисел.

Таблиця 4.6. Модифікація Таблиці

| | Учні | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Ознака | *A* | *B* | *C* | *D* | *E* | *F* | *G* | *H* | *I* | *J* |
| Математика α | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Музика β | 2 | 9 | 3 | 7 | 4 | 1 | 5 | 2 | 6 | 10 |

Необхідно визначити степінь відповідності між цими двома послідовностями порядкових оцінок або, іншими словами, *щільність рангової кореляції*. Коефіцієнт кореляції Кендалла позначимо τ. Він має такі властивості:

- якщо між послідовностями порядкових оцінок існує повна відповідність, тобто кожний елемент займає одне й те саме місце в обох рядках, то $\tau = +1$;
- якщо існує від'ємна залежність, тобто якщо в першій послідовності оцінки розміщені у зворотному порядку порівняно з другою, то $\tau = -1$;
- у всіх інших випадках $-1 < \tau < +1$.

Візьмемо в першій послідовності пару рангів, наприклад *A* і *B*. Їх значення відповідають 7 і 4. За першою ознакою утворюють зворотний порядок величин. Параметрам, які утворюють зворотний порядок, будемо приписувати значення «мінус»1. За другою ознакою відповідні ранги 5 і 7 утворюють прямий порядок, і тому парі *A* і *B* приписуємо значення +1. Перемножуючи значення, які приписані цим парам за першою і другою ознакою, отримаємо «мінус» 1, тобто ця пара між собою *не узгоджена*.

Аналогічні обчислення проводять для кожної пари й отримують *P* результатів +1 і *Q* результатів «мінус» 1.

Значення τ обчислюється за формулою:

$$\tau = \frac{\textit{Справжня сума приписаних значень}}{\textit{Максимально можлива сума цих значень}} = \frac{P-Q}{P+Q}.$$

Розглянемо цю задачу в загальній постановці. Нехай є дві послідовності рангів, кожна з яких має *n* членів. Кількість пар, які можуть бути порівняні, дорівнює кількості способів, за допомогою

яких можна вибрати два елементи з набору з $n$ елементів, тобто кількість комбінацій:

$$C_n^2 = \frac{n(n-1)}{2!},$$

що характеризує найбільш можливу суму приписаних значень, які можна було б отримати в разі, коли порядок рангів в обох послідовностях повністю збігається. Тоді:

$$\tau = \frac{S}{C_n^2} = \frac{(P-Q)}{n(n-1)/2}. \qquad (4.12)$$

Введемо у чисельник ± $P$ або ± $Q$ і, взявши до уваги, що

$$P + Q = C_n^2$$

можна переписати вираз (4.12) як

$$\tau = 1 - \frac{2Q}{n(n-1)/2} = \frac{2P}{n(n-1)/2}.$$

Значущість коефіцієнта рангової кореляції Кендалла оцінюється за допомогою $t$-критерію за формулою:

$$t = \frac{6(S-1)}{\sqrt{2n(n-1)(2n+5)}},$$

за кількості степенів вільності $n - 2$.

За $n \le 10$ оцінка значущості коефіцієнта рангової кореляції $\tau$ оцінюється за допомогою спеціальних таблиць для критичних значень коефіцієнта кореляції рангів Кендалла.

***Коефіцієнт рангової кореляції Спірмена.***

**Приклад**. Розглянемо таблицю ознак α і β і обчислимо для них різницю рангів. Позначимо $d$ різницю між відповідними парами рангів ознак першої та другої послідовності і занесемо в табл. 4.7.

Таблиця 4.7. Коефіцієнт рангової кореляції Спірмена

| Ознака α | 7 | 4 | 3 | 10 | 6 | 2 | 9 | 8 | 1 | 5 |
|---|---|---|---|---|---|---|---|---|---|---|
| Ознака β | 5 | 7 | 3 | 10 | 1 | 9 | 6 | 2 | 8 | 4 |
| Різниця ознак $d$ | 2 | –3 | 0 | 0 | 5 | – 7 | 3 | 6 | – 7 | 1 |
| $d^2$ | 4 | 9 | 0 | 0 | 25 | 49 | 9 | 36 | 49 | 1 |

Коефіцієнт Спірмена обчислюють як

$$\rho_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)},$$

де $n$ – кількість пар спостережень.

Якщо коефіцієнт рангової кореляції Спірмена $\rho_S$ визначений для кількісних показників, які розподілені за нормальним законом, то статистична оцінка для істинного коефіцієнта кореляції $r$ може бути знайдена за допомогою співвідношення:

$$r = 2\sin\frac{\pi}{6}\rho_S .$$

Коефіцієнт рангової кореляції Спірмена має такі властивості:

- $|\rho_S| \le 1$;
- якщо $\rho_S = 1$, існує повна узгодженість величин $X$ і $Y$;
- якщо $\rho_S = -1$, існує протилежне впорядкування послідовностей рангів (від'ємна кореляція);
- якщо $\rho_S = 0$, кореляція відсутня.

Незважаючи на те, що коефіцієнт Спірмена $\rho_S$ обчислити легше, ніж коефіцієнт Кендалла $\tau$, однак $\tau$ використовують частіше. Це зумовлено тим, що додавання нових елементів у вихідну послідовність потребує заново обчислювати $\rho_S$, бо цей коефіцієнт базується на використанні всієї сукупності даних, тоді як для коефіцієнта Кендалла додавання до послідовності нових елементів не потребує повного перерахунку даних.

*Коефіцієнт Кендалла може бути рекомендований для застосування у поглиблених дослідженнях, оскільки критерій значущості його більш обґрунтований порівняно з коефіцієнтом рангової кореляції Спірмена.*

Значущість отриманого коефіцієнта кореляції $\rho_S$ оцінюється за допомогою *t*-критерію за формулою

$$t = \rho_S \sqrt{\frac{n-2}{1-\rho_s^2}}$$

на підставі перевірки гіпотези $H_0$: $\rho_S = 0$ при кількості степенів вільності $n - 2$.

При $n \leq 30$ значущість рангової кореляції Спірмена оцінюється за допомогою спеціальних таблиць для критичних значень коефіцієнта кореляції рангів Спірмена.

*При великій кількості спостережень* (*за винятком значень, близьких до одиниці*) *коефіцієнт рангової кореляції Кендалла приблизно у 1,5 рази менший за коефіцієнт Спірмена.*

***Тетрахоричний коефіцієнт кореляції.*** Для класу задач, в яких масиви даних *X* і (або) *Y* формуються шляхом *дихотомізації* («так-ні», «добре-погано», згоден-незгоден»), тобто наявністю або відсутністю цієї ознаки в об'єкта, кореляційний зв'язок визначається *тетрахоричним коефіцієнтом кореляції*.

У цьому разі загальну чисельність усіх об'єктів у вибірці можна розбити на чотири частини:

*a* – кількість об'єктів, які мають обидві ознаки (+ +);

*b* – кількість об'єктів, які мають першу ознаки, але не мають другої (+ –);

*c* – кількість об'єктів, які не мають першої ознаки, але мають другу (– +);

*d* – кількість об'єктів, які не мають обох ознак (– –).

Знаючи значення *a*, *b*, *c*, *d*, можна обчислити тетрахоричний коефіцієнт кореляції за формулою:

$$r_t = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}.$$

Оцінка значущості $r_t$ виконується за допомогою розподілу $\chi^2$.

Обчислюється розрахунковий коефіцієнт Пірсона

$$\chi_{\text{P}}{}^2 = n\, r_t$$

за кількості степенів вільності $f = 1$ та порівнюється з критичним значенням коефіцієнта Пірсона $\chi^2_{\text{кр}}$, узятим за таблицями розподілу $\chi^2$ для заданого рівня статистичної значущості α.

**Хибна кореляція**. Головне обмеження при використанні результатів кореляційного аналізу полягає у проблемі хибної кореляції.

Серед причин появи хибної кореляції розглядають такі:

- кількість значень змінних залежить від оператора або впливу якихось випадкових величин;
- у досліджуваних рядах спостерігаються деякі тенденції, а також періодичні та сезонні хвилі;
- наявність *стратифікації* – неоднорідності досліджуваного матеріалу;
- похибки реєстрації тощо.

Щоб у кожному конкретному випадку дати відповідь на запитання про те, що слід вважати хибною кореляцією, необхідно попередньо вивчити сутність завдання, з'ясувати, у якій формі має бути подано дані, що підлягають дослідженню, а також якою мірою впливові величини можуть призвести до спотворення досліджуваного зв'язку.

**Особливості кореляційного аналізу**. При вивченні залежностей явищ стикаються з двома різними типами передумов. У першому випадку експериментатор задає певні значення залежної змінної $X = x$, для яких і спостерігаються значення змінної $Y$. Тобто величини $x$ – не випадкові (фіксовані), причому кожному значенню $x$

відповідає деякий генеральний розподіл $Y$ з дисперсією $\sigma^2$. Спостережені значення $y$ розглядають як вибіркові з цих розподілів. У цьому випадку здобуту модель називають ***регресійною***.

У другому випадку спостережені значення $y$ і $x$ можуть являти собою вибірки з двовимірного нормального розподілу, тобто на відміну від попереднього випадку значення $X$ уже не є фіксованими або контрольованими змінними. Таку модель називають ***кореляційною***.

Якщо досліджувана модель є регресійною, то коефіцієнт кореляції характеризує *якість добору рівняння регресії*. У разі кореляційної моделі коефіцієнт кореляції є мірою щільності лінійного зв'язку між випадковими величинами.

*Слід зазначити, що на практиці не завжди можна чітко визначитися з видом моделі, що може призвести до хибного тлумачення здобутих результатів.*

*Основною передумовою* кореляційного аналізу є *незалежність* випадкових аргументів і можливість їх контролю або керування ними. *Усі* вхідні та вихідні випадкові змінні мають бути підпорядковані *нормальному* закону розподілу ймовірностей.

При проведенні кореляційного аналізу реалізується двоетапна процедура.

*Перший етап* – будується *діаграма розсіювання* (*поле кореляції*), тобто графічне відображення точок ($x_i$, $y_i$) на площині ($X$, $Y$), (див. рис. 4.3). Проаналізувавши діаграму розсіювання, можна з'ясувати, чи припустимим є припущення про залежність між $X$ та $Y$, і прийняти рішення щодо форми кривої залежності.

*Другий етап.* Якщо припускається наявність лінійної кореляції, то обчислюється значення коефіцієнта кореляції $r$ та перевіряється його статистична значущість.

За умови, що коефіцієнт кореляції значущий, існує твердження про лінійний регресійний зв'язок між величинами $X$ і $Y$.

Якщо ідентифікується нелінійна регресійна залежність, обчислюється кореляційне відношення $\eta_{Y|x}$ і перевіряється його статистична значущість. Вид залежності можна уточнити накладанням на неї типових кривих.

Проводячи статистичний аналіз, перевіряють гіпотези про форму функції регресії. Найбільш потужним є критерій Фішера.

*Американський математик Тьюккі висловив цікаву думку про те, що використовувати коефіцієнти кореляції є сенс у двох випадках: або коли вони є коефіцієнтами регресії, або коли неможливо виміряти одну чи обидві змінні в якомусь певному масштабі.*

Перша частина цього твердження стосується випадку, коли величини $X$ і $Y$ мають спільний двовимірний нормальний закон розподілу, тоді за рівних дисперсій величин $X$ і $Y$ та наявній лінійній залежності коефіцієнт кореляції відбиває вплив $X$ на $Y$.

Що ж до другого випадку, то можна, скажімо, навести таку сферу застосування статистичних методів, як суспільні науки, медицина тощо, де коефіцієнт кореляції широко використовується, але застосувати детерміновані шкали для вимірювань не завжди можна.

### Запитання для самоперевірки

1. У чому полягає різниця між функціональною та кореляційною залежностями?

2. Що являє собою кореляційне поле; діаграма розсіювання?

3. Що характеризує коефіцієнт кореляції і яка область його можливих значень?

4. Як визначається емпіричний коефіцієнт кореляції?

5. Чому обчислюється надійний інтервал для коефіцієнта кореляції?

6. Як визначається надійний інтервал для коефіцієнта кореляції в разі малої та в разі великої вибірки?

7. Як перевіряється гіпотеза про значущість коефіцієнта кореляції для малої (великої) вибірки?

8. Що характеризує кореляційне відношення? Як обчислюється ця характеристика?

9. Які властивості характерні для кореляційного відношення?

10. Якими показниками характеризується множинна лінійна кореляція?

11. В яких випадках використовуються методи рангової кореляції?

12. Який порядок обчислень коефіцієнтів Спірмена і Кендалла?

13. Як визначаються критичні значення коефіцієнтів Спірмена і Кендалла?

14. В яких випадках застосовується тетрахоричний коефіцієнт кореляції?

15. Чим відрізняється регресійна модель від кореляційної?

## 4.3. ПОНЯТТЯ ПРО РЕГРЕСІЙНИЙ АНАЛІЗ

Якщо кореляційний аналіз довів наявність залежності досліджуваних змінних, то форма залежності може бути уточнена методами регресійного аналізу.

***Основні передумови регресійного аналізу*** деякою мірою збігаються з передумовами кореляційного аналізу. До них належать такі положення:

1. Вхідні величини не є випадковими і задаються (вимірюються) з малою похибкою.

2. Значення контрольованих вхідних величин незалежні між собою.

3. Вихідна величина *Y* при фіксованому значенні $X$ має нормальний закон розподілу.

4. Дисперсія вихідної величини *Y* залишається постійною зі зміною значень вхідної величини $X$.

Наявність рівняння регресії дає змогу прогнозувати поводження об'єкта. Якщо задачу оцінювання параметрів буде розв'язано, можна з деякою надійністю судити про поводження величини *Y* залежно від значення аргументу *X*. Інакше кажучи, можна з певною надійністю робити прогноз щодо можливих значень $\hat{y}$.

Лінія (рівняння) регресії показує, як у середньому буде змінюватися вихідна величина *Y* при незмінному значенні вхідної величини *X*. Центр розсіювання вихідної величини $y_i$, $i=\overline{1,N}$ ($N$ – кількість точок, в яких проводяться спостереження) міститиметься на *теоретичній* лінії регресії.

*Задачею регресійного аналізу є добір такої залежності, яка повною мірою відповідала б теоретичній лінії регресії.*

При проведенні досліджень із застосуванням регресійного аналізу може розв'язуватись одна з двох узагальнених задач:

1. Вид залежності відомий, необхідно якомога точніше знайти оцінки коефіцієнтів регресії.

2. Вид моделі невідомий. У цьому разі висувається гіпотеза про вид передбачуваної залежності. Побудована математична модель перевіряється на адекватність об'єкта. Якщо модель не адекватна, то змінюється гіпотеза і знову проводиться статистична обробка.

Конкретний вид функціональної залежності між величинами *X* і *Y*, установлений за двовимірною вибіркою, називають

***емпіричною залежністю***, яка відображається ***емпіричною формулою***. Якщо побудувати графік емпіричної залежності на кореляційному полі, то він не обов'язково має пройти через усі точки $(x_i, y_i)$ вибірки, а *буде найкращим наближенням до цих точок*. Найчастіше перевага надається лінійній залежності

$$Y = a_0 + a_1 X \,. \tag{4.13}$$

**Знаходження лінійної емпіричної формули**. Для одержання лінійної емпіричної формули $y = a_0 + a_1 x$ існує кілька методів: *метод* «*натягнутої нитки*», *метод сум*, *метод найменших квадратів*.

У ***методі «натягнутої нитки»*** всі результати вимірювань зображують у вигляді точок на кореляційному полі. При цьому варто подумки натягнути між цими точками нитку в такий спосіб, щоб по обидва боки залишилася приблизно однакова кількість точок. З цією метою візьмемо на прямій, що збігається з напрямком нитки, дві точки з координатами ($x_1$, $y_1$) і ($x_2$, $y_2$), які не обов'язково мають бути присутніми у вибірці, але мають бути достатньо віддаленими одна від одної.(рис. 4.5).

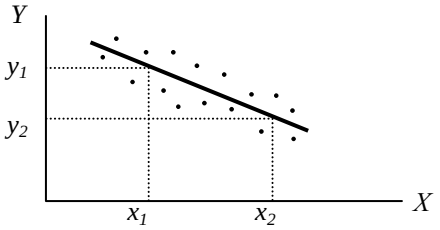



Рис. 4.5. Методі «натягнутої нитки»

Підставивши ці координати у формулу (4.13), дістанемо систему лінійних рівнянь:

$$\begin{cases} \hat{y}_1 = \hat{a}_0 + \hat{a}_1 x_1 \\ \hat{y}_2 = \hat{a}_0 + \hat{a}_2 x_2 \end{cases}, \tag{4.14}$$

де невідомими є коефіцієнти $\hat{a}_0$ й $\hat{a}_1$. Розв'язуючи систему (4.14), дістаємо емпіричну формулу (4.13).

*Через те, що експериментальні дані отримані з похибкою, у результаті статистичної обробки визначають тільки оцінки коефіцієнтів* $\hat{a}_s$.

У ***методі сум*** міркують у такий спосіб. Нехай є двовимірна вибірка. Якщо підставити в емпіричну формулу (4.13) значення величини $X$ із вибірки, можна записати відповідні значення випадкової величини $\hat{y}$ у вигляді

$$\hat{y} = \hat{a}_0 + \hat{a}_1 x\,,\ \ i = 1\ ,..., N.$$

Знайдемо відхилення між вимірюваними $y_i$ й обчисленими значеннями $\hat{y}_i$, яке називається ***відхилом:***

$$\Delta_i = y_i - \hat{y}_i = y_i - \hat{a}_1 x_i - \hat{a}_0\,. \qquad (4.15)$$

Необхідна вимога – для всієї вибірки *сума цих відхилів має дорівнювати нулю*:

$$\sum_{i=1}^{N} \Delta_i = 0\,. \qquad (4.16)$$

У такий спосіб побудовано одне рівняння для визначення коефіцієнтів $\hat{a}_0$ й $\hat{a}_1$. Щоб мати два рівняння, поділимо таблицю спостережень на дві частини. Нехай у першій із них $k$ спостережень, а у другій частині $N - k$ спостережень. В обох частинах рівнянь має виконуватися умова (4.16). У результаті дістанемо таку систему лінійних рівнянь:

$$\begin{cases} \sum\limits_{i=1}^{k} \left(y_i - \hat{a}_0 - \hat{a}_1 x_1\right) = 0 \\ \sum\limits_{i=k+1}^{N} \left(y_i - \hat{a}_0 - \hat{a}_1 x_2\right) = 0 \end{cases}. \qquad (4.17)$$

Якщо розкрити дужки й підсумувати подібні члени, отримаємо:

$$\begin{cases} \sum_{i=1}^{k} y_i - \hat{a}_1 \sum_{i=1}^{k} x_i - k\hat{a}_0 = 0 \\ \sum_{i=k+1}^{N} y_i - \hat{a}_1 \sum_{i=k+1}^{N} x_i - (N-k)\hat{a}_0 = 0 \end{cases}. \qquad (4.18)$$

З виразу (4.18) випливає, що в обох частинах таблиці спостережень потрібно підсумувати значення $x_i$ і $y_i$, а далі розв'язати систему (4.18) відносно параметрів $\hat{a}_0$ й $\hat{a}_1$.

***Метод найменших квадратів***. Зазвичай емпіричну формулу, тобто коефіцієнти рівняння регресії знаходять за допомогою *більш точного методу – методу найменших квадратів* (*МНК*). Ця процедура ще називається *лінійним парним регресійним аналізом.*

МНК ґрунтується на такій теоремі.

*Якщо f(X) є функцією регресії величини Y щодо величини X, то середній квадрат відхилення величини Y від функції f(X) менший, ніж від будь-якої іншої функції h(X), тобто:*

$$M[Y - f(X)]^2 \le M[Y - h(X)]^2,$$

причому рівність досягається лише для таких функцій $h(X)$, для яких

$$M[f(X) - h(X)]^2 = 0.$$

З цієї теореми випливає, що коли вид функції регресії відомий, а невідомі тільки її параметри, то їх можна знаходити з умови мінімізації середнього квадратичного відхилення:

$$M[Y - f(X)]^2.$$

Припустимо, що зв'язок між випадковими величинами має лінійний характер $f(x) = a_0 + a_1 X$. Необхідно знайти таку пряму у площині $(X, Y)$, щоб виконувалася умова:

$$Q = M[(Y - a_0 - a_1 X)]^2 = \min. \qquad (4.19)$$

Внесемо під дужки виразу (4.19) $\pm MY$ та $\pm a_1 MX$ і розіб'ємо цей вираз на окремі складові:

$$Q = M\left[(Y - MY) - a_1(X - MX) + (MY - a_0 - a_1 MX)\right]^2.$$

Після перетворення дістанемо:

$$Q = M[(Y - MY)^2 - 2a_1(X - MX)(Y - MY) - 2a_1\left(X - MX\right)\times$$
$$\times\left(MY - a_0 - a_1 MX\right) + a_1^2(X - MX)^2 + (MY - a_0 - a_1 MX)^2] \quad .$$

Візьмемо до уваги, що

$$M\left[(Y - MY)\right]^2 = \sigma_y{}^2,\ M\left[(X - MX)\right]^2 = \sigma_x{}^2$$

і $$M\left[(X - MX)(Y - MY)\right] = \mathrm{cov}(X, Y),$$

а $$M\left[(X - MX)\right] = 0,$$

виконавши математичні перетворення, перейдемо до виразу:

$$Q = \sigma_y^2 - 2a_1\mathrm{cov}(X, Y) + a_1^2\sigma_x^2 + (MY - a_0 - a_1 MX)^2. \qquad (4.20)$$

Для забезпечення мінімуму виразу (4.19) знаходимо частинні похідні та прирівняємо їх до нуля. У результаті дістаємо систему рівнянь:

$$\begin{cases} \dfrac{\partial Q}{\partial a_0} = -2(MY - a_0 - a_1 MX) = 0 \\ \dfrac{\partial Q}{\partial a_1} = 2a_1\sigma_x{}^2 - 2(MY - a_0 - a_1 MX)MX - 2\mathrm{cov}(X, Y) = 0 \end{cases} . \qquad (4.21)$$

Урахувавши перше рівняння системи (4.21), перетворимо друге її рівняння і дістанемо:

$$\begin{cases} MY - a_0 - a_1 MX = 0 \\ a_1\sigma_x^2 - \mathrm{cov}(X, Y) = 0 \end{cases}, \qquad (4.22)$$

розв'язок якої дає значення коефіцієнтів $a_0$ і $a_1$, за яких вираз (4.19) приймає мінімальне значення. З другого рівняння системи (4.22) знаходимо:

$$a_1 = \frac{cov(X,Y)}{\sigma_x^2}$$

або
$$a_1 = \rho \frac{\sigma_y}{\sigma_x}, \qquad (4.23)$$

де $\rho = \frac{cov(X,Y)}{\sigma_x \sigma_y}$ – нормований коефіцієнт кореляції.

Підставимо знайдене значення $a_1$ у перше рівняння системи (4.22), дістанемо:

$$a_0 = MY - \rho \frac{\sigma_y}{\sigma_x} MX \ . \qquad (4.24)$$

*Коефіцієнти рівняння регресії, отримані на підставі критерію мінімізації Q, називаються оцінками найменшого квадрата або МНК-оцінками.*

З урахуванням (4.23) і (4.24) рівняння прямої регресії *Y* щодо *X* (4.19), яке мінімізує *Q*, матиме вигляд:

$$\hat{Y} = MY + \rho \frac{\sigma_y}{\sigma_x}(X - MX), \qquad (4.25)$$

де $\hat{Y} = M[Y \mid X = x]$.

Визначимо значення $Q_{\min}$, яке можна дістати підставлянням заданих значень $a_0$ і $a_1$ з виразу (4.19). Після алгебраїчних перетворень маємо:

$$Q_{\min} = M[Y - f(X)]^2 = \sigma_y^2(1 - \rho^2).$$

З останнього виразу випливає: чим більша щільність лінійного кореляційного зв'язку, тим менша середньоквадратична розбіжність між теоретичною лінійною залежністю величин *X* і *Y* та

лінійною регресією. Коли існує лінійна функціональна залежність, то $\rho = 1$ і $M[Y - f(X)] = 0$.

Таким чином, *задача* методу найменших квадратів МНК полягає в тому, щоб, знаючи положення точок на площині, так провести лінію регресії, щоб *сума квадратів відхилень Q уздовж осі Oy* (*ординати*) *цих точок від проведеної прямої була б мінімальною.*

*МНК застосовується тільки для рівнянь, лінійних за параметрами або таких, що припускають можливість лінеаризації.*

**Властивості МНК-оцінок**. Оцінки МНК мають низку цінних властивостей:

- ***незміщеність***, тобто математичне сподівання параметра дорівнює істинному його значенню, зокрема для парної регресії:

$$M(a_0) = \alpha \text{ і } M(a_1) = \beta.$$

Незміщеність означає, що вибіркові оцінки параметрів концентруються біля істинних невідомих значень параметрів.

Незміщеність означає, що вибіркові оцінки параметрів концентруються біля істинних невідомих значень параметрів.

- ***спроможність***, коли дисперсія оцінки прямує до нуля зі зростанням *n*. Для парної регресії можна записати:

$$\lim_{n\to\infty} \sigma^2_{a_0} = 0 \text{ і } \lim_{n\to\infty} \sigma^2_{a_1} = 0 .$$

Властивість спроможності означає, що при збільшенні обсягу спостережень оцінки параметрів стають більш надійними, тобто щільніше концентруються біля істинних невідомих значень параметрів.

- ***ефективність***, тобто оцінки мають мінімальну дисперсію порівняно з будь-якими іншими лінійними і незміщеними оцінками цього параметра.

**Геометрична інтерпретація коефіцієнтів регресії.** Коефіцієнт $a_0$ (вільний член рівняння регресії) геометрично являє собою відстань від початку координат до точки перетину лінії регресії з ординатою або, інакше кажучи, це відрізок, що відтинається на ординаті лінією регресії.

Коефіцієнт $a_1$ являє собою тангенс кута нахилу лінії регресії до осі абсцис і пропорційний до коефіцієнта кореляції. На рис. 4.6 наведено *графіки емпіричної регресії* залежно від значень коефіцієнтів регресії: $a_0 = 0$, $a_1 > 0$ (рис. 4.6, *а*); $a_0 > 0$, $a_1 > 0$ (рис. 4.6, *б*); $a_0 < 0$, $a_1 > 0$ (рис. 4.6, *в*); $a_0 > 0$, $a_1 < 0$ (рис. 4.6, *г*).

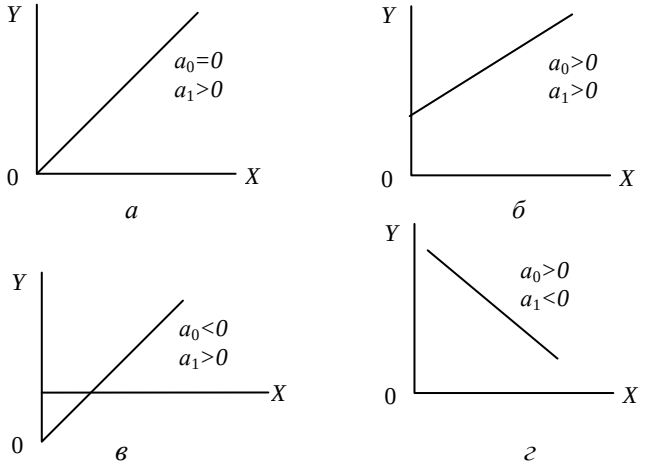



Рис. 4.6. Графіки емпіричної регресії

**Емпірична регресія**. На практиці мають справу не з *теоретичною* лінією регресії, а з *емпіричною*, яка будується на підставі обмеженого обсягу експериментальних даних.

Емпірична лінія регресії дає змогу визначити, яке в середньому буде значення $Y$, якщо при фіксованому значенні $x_i$, $i = \overline{1, N}$, дослід повторюватиметься багато разів.

*Для Y і X у літературі трапляються такі найменування*: *X – вхідна змінна, незалежна змінна, аргумент, фактор, регресор, предиктор; Y – вихідна змінна, залежна змінна, ордината функція відгуку.*

**Запитання для самоперевірки**

1. Сформулюйте задачі та основні передумови кореляційного аналізу.
2. Які ви знаєте методи побудови емпіричної формули? Який із методів найточніший?
3. Що таке відхил?
4. Для чого використовуються кореляційні таблиці?
5. На якій теоремі ґрунтується метод найменших квадратів?
6. У чому полягає сутність методу найменших квадратів?
7. Як пов'язані між собою коефіцієнти рівняння регресії та коефіцієнт кореляції?
8. Що характеризують коефіцієнти рівняння регресії?
9. Наведіть геометричну інтерпретацію коефіцієнтів рівняння регресії.
10. Сформулюйте поняття емпіричної регресії.
11. Які назви можуть мати величини *X* і *Y*?

## 4.4. ЗНАХОДЖЕННЯ КОЕФІЦІЄНТІВ ЛІНІЙНОЇ ПАРНОЇ РЕГРЕСІЇ

Найпростішою формою оцінки стохастичної залежності є одновимірний регресійний аналіз, згідно з якими формуються обчислювальні процедури відтворення лінії регресії.

У регресійному аналізі припускається, що можна безпосередньо або опосередковано контролювати одну або кілька незалежних змінних. Причому ці змінні не є випадковими величинами.

У загальному випадку об’єкт розглядають як «чорний ящик», на вхід якого подаються незалежні змінні (рис. 4.7). Крім того, на результат вимірювання відгуку впливає випадкова похибка засобу вимірювання ε. Її можна інтерпретувати як заваду або шум.

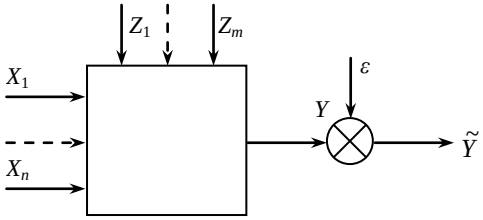



Рис. 4.7. Представлення об’єкта як «чорний ящик»

Вхідні змінні (фактори) можуть бути розбиті на два класи: ***контрольовані*** $X_i$ (вимірювані) і ***неконтрольовані*** $Z_j$. У свою чергу, контрольовані змінні не є випадковими і можуть бути керованими або некерованими. ***Керовані*** – такі фактори, цілеспрямована зміна яких у ході експерименту можлива. Фактори, для яких цілеспрямована зміна неможлива, називаються ***некерованими***.

Регресійний аналіз дозволяє встановити зв’язок між вхідною $X$ і вихідною $Y$ величинами у вигляді *математичної моделі.* Кінцевою метою експериментального дослідження є встановлення математичної залежності, що адекватно описує поводження об’єкта. Коефіцієнти цієї моделі дають змогу оцінити вплив незалежних величин $X_i$ на вихідну величину $Y$.

**Знаходження оцінок коефіцієнтів парної лінійної регресії**. Зазвичай залежності оцінюють за підсумками спостережень, результати яких сформовано у вигляді масиву даних.

Нехай при дослідженні залежності між величинами $X$ та $Y$ було проведено вимірювання при $N$ дискретних значеннях $x_i$. Для кожного значення $x_i$ було отримано $m$ значень вихідної величини $\tilde{y}_{ij}$, тобто задано масив даних:

$$\left\{x_i, \tilde{y}_{ij}\ i=\overline{1,N}, j=\overline{1,m}\right\},$$

де $N$ – кількість дослідів у кожній точці факторного простору; $m$ – кратність спостережень в $i$-й точці.

Для $N$ наявних пар значень будують поле кореляції, за зовнішнім виглядом якого можна припустити, що залежність між випадковими величинами лінійна.

На підставі наявних даних можна побудувати множину прямих, при цьому можуть бути використані різноманітні критерії. Як було встановлено (п. 4.3), найбільш ефективним при апроксимації наявних даних є *критерій мінімізації суми квадратів відхилень* експериментальних даних від передбачуваної залежності, тобто для знаходження оцінок коефіцієнтів регресії $\hat{a}_0$ й $\hat{a}_1$ необхідно застосовувати метод найменших квадратів, що дає змогу мінімізувати суму квадратів різниці відхилень експериментальних даних $\tilde{y}_i$ і розрахункових значень $\hat{y}$, тобто

$$Q = \sum_{i=1}^{N} (\tilde{y}_i - \hat{y}_i)^2 \to min. \tag{4.26}$$

Для формалізації процедури обчислення оцінок коефіцієнтів помножимо оцінку $\hat{a}_0$ на «фіктивну змінну» $x_{i0} \equiv 1$.

Перепишемо (4.26) у вигляді:

$$Q = \sum_{i=1}^{N} [\tilde{y}_i - (\hat{a}_0 x_{i0} + \hat{a}_1 x_{i1})]^2 \to min,$$

і знайдемо оцінки коефіцієнтів $\hat{a}_0$ й $\hat{a}_1$.

Таким чином,

$$\begin{cases} \dfrac{\partial Q}{\partial \hat{a}_0} = -2\sum\limits_{i=1}^{N} [\tilde{y}_i - (\hat{a}_0 x_{i0} + \hat{a}_1 x_{i1})] x_{i0} = 0 \\ \dfrac{\partial Q}{\partial \hat{a}_1} = 2\sum\limits_{i=1}^{N} [\tilde{y}_i - (\hat{a}_0 x_{i0} + \hat{a}_1 x_{i1})] x_{i1} = 0 \end{cases} \tag{4.27}$$

Виконавши відповідні перетворення, систему (4.27) можна переписати в такий спосіб:

$$\begin{cases} \hat{a}_0 \sum_{i=1}^{N} x_{i0}^{\,2} + \hat{a}_1 \sum_{i=1}^{N} x_{i0} x_{i1} = \sum_{i=1}^{N} x_{i0} \tilde{y}_i \\ \hat{a}_0 \sum_{i=1}^{N} x_{i0} x_{i1} + \hat{a}_1 \sum_{i=1}^{N} x_{i1}^{\,2} = \sum_{i=1}^{N} x_{i1} \tilde{y}_i \end{cases}. \qquad (4.28)$$

Як бачимо, систему (4.28) можна дістати й без знаходження частинних похідних, а скориставшись відомим правилом побудови *нормальної системи рівнянь*: якщо послідовно множити базове рівняння регресії:

$$\hat{a}_0 \sum_{i=1}^{N} x_{i0} + \hat{a}_1 \sum_{i=1}^{N} x_{i1} = \sum_{i=1}^{N} \tilde{y}_i \,. \qquad (4.29)$$

на відповідний фактор, дістанемо систему нормальних рівнянь (4.28).

*Тобто якщо всі члени рівняння* (*4.29*) *помножити на* $x_{i0}$, *то дістанемо перше рівняння системи* (*4.28*), *а якщо всі члени помножити на* $x_{i1}$, *то дістанемо друге рівняння цієї системи.*

Подана в такому вигляді система нормальних рівнянь має особливості:

- у правій частині системи під знаком суми містяться добутки, отримані в результаті послідовного множення сукупності значень вихідних величин на сукупність відповідних значень вхідної величини;
- на діагоналі лівої частини системи рівнянь під знаком суми послідовно розміщено квадрати незалежних змінних;
- у лівій частині спостерігається симетрія відносно діагоналі виразів, записаних під знаком суми.

Оцінки коефіцієнтів рівняння регресії, що задовольняють метод найменших квадратів, можна знайти за допомогою визначників:

$$\hat{a}_0 = \frac{A_0}{A}; \quad \hat{a}_1 = \frac{A_1}{A},$$

де $A$ – головний визначник системи, за умови, що $x_{i0} = 1$; $A_0$, $A_1$ – алгебраїчні доповнення:

$$A = \begin{vmatrix} \sum_{i=1}^{N} x_{i0}^2 & \sum_{i=1}^{N} x_{i0}x_{i1} \\ \sum_{i=1}^{N} x_{i0}x_{i1} & \sum_{i=1}^{N} x_{i1}^2 \end{vmatrix} = \begin{vmatrix} N & \sum_{i=1}^{N} x_{i1} \\ \sum_{i=1}^{N} x_{i1} & \sum_{i=1}^{N} x_{i1}^2 \end{vmatrix} = N\sum_{i=1}^{N} x_{i1}^2 - \left(\sum_{i-1}^{N} x_{i1}\right)^2;$$

$$A_0 = \begin{vmatrix} \sum_{i=1}^{N} x_{i0}\tilde{y}_i & \sum_{i=1}^{N} x_{i0}x_{i1} \\ \sum_{i=1}^{N} x_{i1}\tilde{y}_i & \sum_{i=1}^{N} x_{i1}^2 \end{vmatrix} = \sum_{i=1}^{N} \tilde{y}_i \sum x_{i1}^2 - \sum_{i=1}^{N} x_{i1}\tilde{y}_i \sum_{i=1}^{N} x_{i1};$$

$$A_1 = \begin{vmatrix} \sum_{i=1}^{N} x_{i0}^2 & \sum_{i=1}^{N} x_{i0}\tilde{y}_i \\ \sum_{i=1}^{N} x_{i0}x_{i1} & \sum_{i=1}^{N} x_{i1}\tilde{y}_i \end{vmatrix} = \begin{vmatrix} N & \sum_{i=1}^{N} \tilde{y}_i \\ \sum_{i=1}^{N} x_{i1} & \sum_{i=1}^{N} \tilde{y}x_{i1} \end{vmatrix} = N\sum_{i=1}^{N} \tilde{y}_i x_{i1} - \sum_{i=1}^{N} x_{i1}\sum_{i=1}^{N} \tilde{y}_i .$$

Отже,

$$\hat{a}_0 = \frac{\sum_{i=1}^{N} \tilde{y}_i \sum_{i=1}^{N} x_{i1}^2 - \sum_{i=1}^{N} x_{i1}\sum_{i=1}^{N} x_{i1}\tilde{y}_i}{N\sum_{i=1}^{N} x_{i1}^2 - (\sum_{i=1}^{N} x_{i1})^2}; \qquad (4.30)$$

$$\hat{a}_1 = \frac{N\sum_{i=1}^{N} \tilde{y}_i x_{i1} - \sum_{i=1}^{N} x_{i1}\sum_{i=1}^{N} \tilde{y}_i}{N\sum_{i=1}^{N} x_{i_1}^2 - (\sum_{i=1}^{N} x_{i1})^2}. \qquad (4.31)$$

Знайдені в такий спосіб оцінки коефіцієнтів $\hat{a}_0$, $\hat{a}_1$ дозволяють побудувати пряму регресії, що мінімізує суму квадратів відхилень розрахункових значень від експериментальних даних.

*Знак перед коефіцієнтами показує, як зміна факторів приводить до зміни відгуку. Якщо стоїть знак «–», то збільшення x приведе до зменшення y.*

**Метод максимальної правдоподібності**. Знаходження оцінок методом найменших квадратів базується на припущенні, що умовне математичне сподівання *Y* лінійно залежить від поточного значення *x*. Похибка вимірювання вихідної величини при цьому є незалежною і має нормальний розподіл. Таким чином, отримані значення вихідної величини $\tilde{y}$ також будуть випадковими і матимуть нормальний розподіл.

Розглянемо задачу визначення оцінок коефіцієнтів лінійної парної регресії з іншого боку, виходячи з передумови, що найкращою оцінкою параметра випадкової величини є та оцінка, яка найбільш імовірна за результатами проведення експерименту.

Для цього скористаємось *методом максимальної правдоподібності*, який полягає у виборі такої гіпотези, згідно з якою ймовірність отримання у процесі вимірювання величин, що спостерігаються, була б *максимальною* (*правдоподібною*).

Якщо при кожному значенні *x* значення *Y* розподілені за нормальним законом із середнім, що містяться на прямій регресії $M[Y \mid x_i] = \bar{y}_i = a_0 + a_1 x_1$, то умовна щільність імовірності *Y* визначається за формулою:

$$p(\tilde{y}_i \mid x_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(\tilde{y}_i - \bar{y}_i)^2}{2\sigma^2}}.$$

У разі наявності $N$ вибіркових точок, в яких проводився експеримент, знаходиться *функція правдоподібності* $L$ (від *Likelihood*-правдоподібність), яка дорівнює добутку умовних імовірностей вихідних величин, параметри розподілу яких відповідають максимуму ймовірності появи наявного значення випадкової величини:

$$L = \prod_{i=1}^{N} p(\tilde{y}_i \mid x_i) \to \max.$$

З урахуванням наведених припущень дістанемо:

$$L = \frac{1}{(\sqrt{2\pi})^N \sigma^N} e^{\sum_{i=1}^{N} \frac{[-(\tilde{y}_i - \bar{y}_i)]}{2\sigma^2}},$$

Перш ніж проводити дослідження функції правдоподібності на максимум, прологарифмуємо її:

$$\ln L = ln\left[\frac{1}{\sqrt{(2\pi)^N}\sigma^N}\right] - \frac{\sum_{i=1}^{N}(\tilde{y}_i - \bar{y}_i)^2}{2\sigma^2} \cdot ln\,\ell = ln1 - N\,ln\sqrt{2\pi} - N\,ln\sigma - \quad (4.32)$$

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(\tilde{y}_i - \bar{y}_i)^2 = -N\,ln\sqrt{2\pi} - ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(\tilde{y}_i - \bar{y}_i)^2.$$

Завдання полягає у визначенні оцінок коефіцієнтів $a_{0\text{МП}}, a_{1\text{ММ}}$ (індекс «МП» – вказує на те, що оцінки знайдені за методом максимальної правдоподібності), при яких функція правдоподібності $L$ приймає максимальне значення. Для розглянутого випадку $\ell nL$ буде приймати максимальне значення тоді й тільки тоді, коли третій доданок правої частини рівняння (4.32) набуде мінімальне значення, тому що перший і другий доданки – константи. Таким чином, необхідно досліджувати умови, що забезпечують мінімізацію

$$\sum_{i=1}^{N}(\tilde{y}_i - a_0 - a_1 x_i)^2 .$$

Порівнявши умову, при якій третій доданок забезпечує максимум функції правдоподібності, до умови одержання оцінок коефіцієнтів регресії з використанням методу найменших квадратів, можна дійти висновку, що *метод найменших квадратів є окремим випадком методу максимальної правдоподібності* при нормальному розподілі випадкових величин.

*Оцінки коефіцієнтів регресії, знайдені на основі методу найменших квадратів, матимуть також і властивості максимальної правдоподібності, тобто будуть найбільш імовірними.*

*Чим більша правдоподібність, тим краще модель узгоджується з вибірковими даними.*

**Коефіцієнт детермінації**. Цей коефіцієнт дає змогу відповісти на питання щодо якості опису залежності за допомогою рівняння регресії. Чим ближче спостереження розміщені до лінії регресії, тим краще регресія описує відповідну залежність змінних і з більшою надійністю може бути застосована для розв'язання практичних задач.

Для того щоб пояснити зміст коефіцієнта детермінації, розглянемо в полі кореляції для парної залежності лінію регресії, що задовольняє метод найменших квадратів (рис. 4.8).

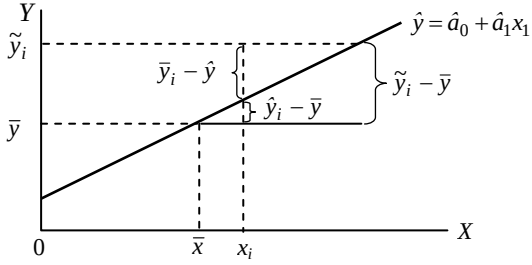



Рис. 4.8. Знаходження коефіцієнта детермінації

Візьмемо довільну точку з координатами $(x_i, \tilde{y}_i)$ і розглянемо відхилення значення $\tilde{y}_i$ вихідної величини від середнього значення $\bar{y}_i$ (центра ваги). Згідно з рис. 23.2 можна записати:

$$(\tilde{y}_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (\tilde{y}_i - \hat{y}_i). \qquad (4.33)$$

Аналогічну рівність можна записати для будь-якої точки $i = \overline{1, N}$ поля кореляції. Піднесемо до другої степені обидві частини рівняння (4.33) і запишемо для всієї сукупності точок поля кореляції співвідношення:

$$\sum_{i=1}^{N}(\tilde{y}_i - \bar{y})^2 = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{N}(\hat{y}_i - \bar{y})(\tilde{y}_i - \hat{y}_i) + \sum_{i=1}^{N}(\tilde{y}_i - \hat{y}_i)^2 .$$

За відсутності кореляції між $(\hat{y}_i - \bar{y})$ і $(\tilde{y}_i - \hat{y}_i)$ друга складова у правій частині буде дорівнювати нулю і тоді

$$\sum_{i=1}^{N}(\tilde{y}_i - \bar{y})^2 = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{N}(\tilde{y}_i - \hat{y}_i)^2 . \qquad (4.34)$$

Цей вираз характеризує повне розсіювання вихідної величини (точок кореляційного поля) відносно середнього значення $\bar{y}$, яке складається із суми розсіювань умовних математичних сподівань відносно середнього значення і суми середніх умовних розсіювань. Таким чином, перший добуток у правій частині рівняння (4.34) характеризує розсіювання за рахунок впливу на вихідну величину вхідної величини (фактора) і зумовлений регресійним зв'язком. Відхилення $(\tilde{y}_i - \hat{y}_i)$ відображають вплив випадкових величин.

Коефіцієнт детермінації регресії визначається як відношення суми квадратів відхилень, які можна пояснити впливом досліджуваного фактора, до всієї (повної) суми квадратів відхилень:

$$R^2_{Y|x} = \frac{\sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N} (\tilde{y}_i - \bar{y})^2}$$

або з урахуванням (4.34)

$$R^2_{Y|x} = 1 - \frac{\sum_{i=1}^{N} (\tilde{y}_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (\tilde{y}_i - \bar{y})^2}.$$

З останнього виразу випливає, що $R^2_{Y|x}$ буде наближатися до одиниці, коли $\sum_{i=1}^{N} (\tilde{y}_i - \hat{y}_i)^2$ буде наближатися до нуля, тобто коли

$$\sum_{i=1}^{N} (\tilde{y}_i - \hat{y}_i)^2 = \sum_{i=1}^{N} (\tilde{y}_i - \bar{y})^2$$

і розсіювання навколо лінії регресії дорівнює розсіюванню біля загального середнього $\bar{y}$ при будь-якому *x*. З цього випливає, що між досліджуваними величинами існує *функціональний зв'язок*.

Взаємозв'язок між коефіцієнтами кореляції і детермінації можна виразити як

$$r^2_{X|Y} \le R^2_{Y|x}.$$

Рівність між ними виконуватиметься лише в тому випадку, коли матиме місце строга лінійна залежність.

Чим менша різниця $\sum_{i=1}^{N} (\tilde{y}_i - \hat{y}_i)^2$ та $\sum_{i=1}^{N} (\tilde{y}_i - \bar{y})^2$, тим більше значення $R^2_{Y|x}$ і тим «тісніше» розташовуються окремі спостереження до лінії регресії. У випадку, коли $R^2_{Y|x} = 1$, центри розподілу емпіричних точок лежать на лінії регресії.

Згідно зі сказаним $R^2_{Y|x}$ може розглядатися як міра якості опису залежності ознак *x* і *y* за допомогою рівняння регресії.

**Знаходження надійних інтервалів для коефіцієнтів регресії.**

Якщо провести дослідження оцінок параметрів і врахувати, що вони отримані за методом максимальної правдоподібності, отримані оцінки будуть *незміщеними*, *спроможними* й *ефективними*.

Зважаючи на те, що оцінки коефіцієнтів було знайдено на підставі вибіркових значень випадкових величин *X* і *Y*, вони, по суті, є *точковими*, і при проведенні іншого аналогічного експерименту можна одержати дані, які приведуть до інших значень цих коефіцієнтів. Тому найбільш надійною оцінкою параметрів є оцінка у вигляді *надійного інтервалу*. При побудові надійних інтервалів вважають, що кожна оцінка має *нормальний розподіл.*

***Надійний інтервал для коефіцієнта регресії*** $\hat{a}_1$. Розглянемо вираз для оцінки коефіцієнта $\hat{a}_1$ (4.31). Щоб позбутися $x_i$ (поточного значення) і одержати середнє арифметичне, котре є конкретним числом, поділимо чисельник і знаменник на $N$ і виконаємо перетворення з огляду на те, що $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_{i1}$, а другий доданок знаменника отримано шляхом його множення на $\frac{N}{N}$, тобто

$$\frac{N}{N}\cdot\frac{1}{N}\left(\sum_{i=1}^{N} x_{i1}\right)^2 = N\left(\frac{\sum_{i=1}^{N} x_{i1}}{N}\right)^2 = N\overline{x}_1^2,$$

тоді

$$\hat{a}_1 = \frac{\sum_{i=1}^{N} \tilde{y}_i x_{i1} - \bar{x} \sum_{i=1}^{N} \tilde{y}_i}{\sum_{i=1}^{N} x_{i1}{}^2 - N\bar{x}^2}. \qquad (4.35)$$

Розглянемо чисельник виразу (4.35)

$$\sum_{i=1}^{N} \tilde{y}_i x_{i1} - \bar{x} \sum_{i=1}^{N} \tilde{y}_i = \sum_{i=1}^{N} (x_{i1} - \bar{x}) \tilde{y}_i . \qquad (4.36)$$

Розглянемо знаменник виразу 4.35)

$$\sum_{i=1}^{N} x_{i1}^2 - N\bar{x}^2 .$$

Доведемо, що

$$\sum_{i=1}^{N} x_{i1}^2 - N\bar{x}^2 = \sum_{i=1}^{N} \left(x_{i1} - \bar{x}\right)^2 .$$

Доказ проведемо у зворотному порядку. Розглянемо вираз:

$$\sum_{i=1}^{N} \left(x_{i1} - \bar{x}\ \right)^2 = \sum_{i=1}^{N} \left(x^2{}_{i1} - 2x_{i1}\bar{x} + \bar{x}^2\right) =$$

$$= \sum_{i=1}^{N} x_{i1}^2 - 2\frac{N}{N} \sum_{i=1}^{N} x_{i1}\bar{x} + \sum_{i=1}^{N} \bar{x}^2 = \sum_{i=1}^{N} x_{i1}^2 - 2N\bar{x}^2 + N\bar{x}^2 = \sum_{i=1}^{N} x_{i1}^2 - N\bar{x}^2 .$$

Таким чином, можна замінити знаменник

$$\sum_{i=1}^{N} x_{i1}^2 - N\bar{x}^2$$

на вираз

$$\sum_{i=1}^{N} \left(x_{i1} - \bar{x}\right)^2 , \qquad (4.37)$$

що й було доведено.

Підставимо отримані співвідношення (4.36) і (4.37) у вираз (4.35) для $\hat{a}_1$:

$$\hat{a}_1 = \frac{\sum_{i=1}^{N}(x_{i1} - \bar{x})\tilde{y}_i}{\sum_{i=1}^{N}(x_{i1} - \bar{x})^2}.$$

Позначимо

$$k_i = \frac{(x_{i1} - \bar{x})}{\sum_{i=1}^{N}(x_{i1} - \bar{x})^2}, \qquad (4.38)$$

де $k$ – детермінована величина.

Тоді $$\hat{a}_1 = \sum_{i=1}^{N} k_i \tilde{y}_i . \qquad (4.39)$$

Крім знайденої оцінки коефіцієнта $\hat{a}_1$ для знаходження надійного інтервалу необхідно обчислити його дисперсію:

$$D\{\hat{a}_1\} = D\left\{\sum_{i=1}^{N} k_i \tilde{y}_i\right\} = \sum_{i=1}^{N}(k_i)^2 D\{\tilde{y}_i\} = \sum_{i=1}^{N}(k_i)^2 \sigma^2, \qquad (4.40)$$

де $k_i$ = *const*, виноситься за знак дисперсії, $D\{\tilde{y}_i\}$ – дисперсія результату вимірювання $\sigma^2$.

Підставимо у вираз (4.40) замість $k_i$ його значення з виразу (4.38).

Тоді

$$D\{\hat{a}_1\} = \frac{\sum_{i=1}^{N}(x_{i1} - \bar{x})^2}{\left[\sum_{i=1}^{N}(x_{i1} - \bar{x})^2\right]^2}\sigma^2 = \frac{\sigma^2}{\sum_{i=1}^{N}(x_{i1} - \bar{x})^2}. \qquad (4.41)$$

Ураховуючи те, що дисперсія $\sigma^2$ невідома, замість неї скористаємося незміщеною оцінкою $S^2$.

Розрахувавши дисперсію коефіцієнта $\hat{a}_1$ можна записати границі довірчого інтервалу:

$$\hat{a}_1 - t\frac{S}{\sqrt{\sum_{i=1}^{N}(x_{i1}-\bar{x})^2}} < a_1 < \hat{a}_1 + t\frac{S}{\sqrt{\sum_{i=1}^{N}(x_{i1}-\bar{x})^2}}, \qquad (4.42)$$

де $t$ – коефіцієнт Стьюдента, який визначають за таблицями $t$-розподілу Стьюдента для кількості степенів вільності $N$–2 та рівня статистичної значущості α/2; $S = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N}(\tilde{y}_i - \hat{a}_0 - \hat{a}_1 x_i)^2}$ – незміщена оцінка дисперсії.

***Надійний інтервал для коефіцієнта регресії*** $\hat{a}_0$. Нехай $x_1 = \bar{x},$ тоді рівняння регресії для точки $(\bar{x}, \bar{y})$ можна записати:

$$\bar{y} = \hat{a}_0 + \hat{a}_1\bar{x}. \qquad (4.43)$$

З рівняння (4.43) випливає, що

$$\hat{a}_0 = \bar{y} - \hat{a}_1\bar{x}.$$

Дисперсія оцінки коефіцієнта $\hat{a}_0$ дорівнює сумі дисперсій:

$$D\{\hat{a}_0\} = D\{\bar{y}\} + D\{\hat{a}_1\bar{x}\}. \qquad (4.44)$$

Дисперсія першого доданка виразу (4.44) – це дисперсія середнього арифметичного результату вимірювання, що в $N$ разів менша за дисперсію окремого результату вимірювання:

$$D\{\bar{y}\} = \frac{\sigma^2}{N}. \qquad (4.45)$$

Винесемо за знак дисперсії другого доданку виразу (4.44) константу $\bar{x}$

$$D\{\hat{a}_1\bar{x}\} = \bar{x}^2 D\{\hat{a}_1\}. \qquad (4.46)$$

Підставимо всі складові, тобто вирази (4.45), (4.46), (4.41) у рівняння (4.44) і отримаємо:

$$D\{\hat{a}_0\} = \frac{\sigma^2}{N} + \bar{x}^2\frac{\sigma^2}{\sum_{i=1}^{N}(x_{i1}-\bar{x})^2},$$

або у випадку невідомої дисперсії:

$$D\{\hat{a}_0\} = S^2\left[\frac{1}{N} + \frac{\bar{x}^2}{\sum_{i=1}^{N}(x_{i1} - \bar{x})^2}\right]. \qquad (4.47)$$

Розрахувавши вираз для дисперсії оцінки коефіцієнта $\hat{a}_0$, можна записати межі надійного інтервалу:

$$\hat{a}_0 - tS\left[\frac{1}{N} + \frac{\bar{x}^2}{\sum_{i=1}^{N}(x_{1i} - \bar{x})^2}\right]^{1/2} < a_0 < \hat{a}_0 + tS\left[\frac{1}{N} + \frac{\bar{x}^2}{\sum_{i=1}^{N}(x_{1i} - \bar{x})^2}\right]^{1/2}.$$

*Отже, при проведенні регресійного аналізу будують рівняння регресії, яке пов'язує залежну змінну з незалежними. Якщо рівняння лінійне відносно параметрів, говорять про лінійну регресію, якщо ні – регресія нелінійна.*

*Рівняння наближеної лінії регресії* $\hat{Y} = \hat{a}_0 + \hat{a}_1 X$ *мінімізує відхил, тобто коефіцієнти цього рівняння знайдені з використанням методу найменших квадратів.*

*Зважаючи на те, що оцінки коефіцієнтів були знайдені на підставі вибіркових значень величин* $X$ *і* $Y$ *, вони є точковими й при проведенні іншого експерименту можуть бути обчислені інші їхні значення. Тому параметри оцінюють у вигляді надійного інтервалу.*

**Запитання для самоперевірки**

1. Назвіть передумови регресійного аналізу. У чому полягає задача регресійного аналізу?

2. Які величини називають контрольованими, неконтрольованими, керованими?

3. У чому полягає правило побудови системи нормальних рівнянь? Які особливості має система нормальних рівнянь?

4. Якому критерію має відповідати пряма регресії?

5. Для чого застосовується метод максимальної правдоподібності? У чому полягає суть методу максимальної правдоподібності? Що характеризує функція правдоподібності?

6. Чому коефіцієнти регресії оцінюють у вигляді надійного інтервалу?

7. Як обчислюється дисперсія коефіцієнта $a_1$?

8. Запишіть надійний інтервал для коефіцієнта $a_1$.

9. Як обчислюється дисперсія коефіцієнта $a_0$?

10. Запишіть надійний інтервал для коефіцієнта $a_0$.

## 4.5. ПЕРЕВІРКА СТАТИСТИЧНОЇ ЗНАЧУЩОСТІ КОЕФІЦІЄНТІВ РІВНЯННЯ РЕГРЕСІЇ. ПЕРЕВІРКА АДЕКВАТНОСТІ МОДЕЛІ

Необхідно встановити: зміна вихідної величини зумовлена насправді впливом факторів чи це прояв впливу випадкових величин.

**Перевірка статистичної значущості коефіцієнтів рівняння регресії**. Вплив факторів оцінюється відповідними коефіцієнтами рівняння регресії, тому необхідно перевіряти *статистичну значущість* отриманих оцінок коефіцієнтів, бо навіть за відсутності впливу факторів спостерігатиметься розкид значень, зумовлений впливом випадкових величин. Оцінка статистичної значущості здійснюється за *t*-критерієм Стьюдента.

Відомо, що для вибірки обмеженого обсягу відношення відхилення оцінки випадкової величини від її математичного сподівання до середньоквадратичного відхилення цієї оцінки має *t*-розподіл:

$$t_j = \frac{|(\hat{a}_j - a_j)|}{S\{\hat{a}_j\}}, \quad j = \overline{0,1},$$

де $t_j$ – коефіцієнт Стьюдента; $S\{\hat{a}_j\}$ – оцінка середньоквадратичного відхилення коефіцієнта $\hat{a}_j$; $a_j$ – математичне сподівання.

Висувається гіпотеза $H_0$: оцінка коефіцієнта $\hat{a}_j$ є статистично незначущою. Перевіряючи цю гіпотезу, припускають, що математичне сподівання зазначеної оцінки дорівнює нулю, тобто $a_j = 0,$ тому розрахункові значення $t_j$ знаходять за формулою:

$$t_j = \frac{|\hat{a}_j|}{S\{\hat{a}_j\}}.$$

Коефіцієнт $t_j$ обчислюють для кожного коефіцієнта $\hat{a}_j$ і порівнюють із критичним значенням $t_{кр}$, знайденим за таблицями Стьюдента при $f = N - 2$ і $\frac{\alpha}{2}$.

*При обчисленні оцінки середньоквадратичного відхилення $S\{\hat{a}_0\}$ коефіцієнта $\hat{a}_0$ слід використовувати формулу* (*4.46*), *а для оцінки середньоквадратичного відхилення $S\{\hat{a}_j\}$ інших коефіцієнтів $\hat{a}_j$ слід скористатися формулою* (*4.41*).

Якщо $t_j < t_{кр}$, гіпотеза $H_0$ приймається, тобто за результатами порівняння приймається рішення, що за такого обсягу вибірки

розглядуваний коефіцієнт є статистично незначущим і його слід виключити з рівняння регресії.

Чим більше $t_j$ порівняно з $t_{кр}$, *тим істотніший вплив* $j$-го фактора.

*Варто спочатку переконатися, що сама по собі вибіркова дисперсія* $S^2$ *не перевищує діапазону зміни фактора при проведеному експерименті. Для цього необхідно провести додаткові досліди* (*збільшити обсяг вибірки*) *і, якщо висновок повториться, то справді коефіцієнт статистично незначущий.*

Зважаючи на те, що оцінки коефіцієнтів пов'язані між собою, тобто обчислювалися на підставі того самого визначника, то з виключенням статистично незначущого коефіцієнта з моделі постає необхідність перерахування решти коефіцієнтів. Це недолік, якого потрібно позбуватися.

**Перевірка адекватності лінійної моделі.** Переконавшись, що рівняння регресії містить тільки статистично значущі коефіцієнти, необхідно перевірити *адекватність* отриманої моделі, тобто наскільки модель відповідає експериментальним даним. Вимога адекватності є однією з найважливіших при побудові моделі. Адекватність означає правильний аналітичний опис об'єкта за вибраними характеристиками з достатньою точністю. Розглянемо два випадки.

***1. Експерименти проводяться одноразово*** в кожній точці $x_i$ факторного простору. Як правило, така ситуація може спостерігатися при проведенні вимірювань характеристик динамічних процесів або в інших випадках, в яких неможливо забезпечити відтворюваність вимірювань. У цьому разі (див. рисунок) результатам дослідження можуть відповідати:

– горизонтальна лінія, коли $\hat{a}_0 = \bar{y} = \frac{1}{N}\sum_{i=1}^{N} \tilde{y}_i$, тобто $\hat{y} = \bar{y}$;

– похила лінія, отримана з використанням МНК, тобто $\hat{y} = \hat{a}_0 + \hat{a}_1 x_1$.

Якщо результатам відповідає горизонтальна лінія, це говорить про відсутність впливу фактора *X* на функцію відгуку *Y* (розкид значень зумовлений впливом випадкових величин і (або) випадковою похибкою засобу вимірювання). Лінія з нахилом свідчить про вплив досліджуваного фактора.

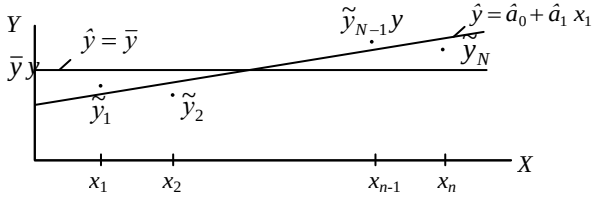



Рис. 4.9. – Оцінка суттєвості впливу фактора

Для прийняття рішення про вплив фактора *X* на функцію відгуку висувається гіпотеза $H_0 : \hat{y} = \bar{y}$ (розкид значень зумовлений тільки впливом випадкових величин) при альтернативній гіпотезі $H_1$: $\hat{y} = \hat{a}_0 + \hat{a}_1 x_1$ (фактор впливає лінійно).

Для перевірки гіпотези $H_0$ використовують статистику Фішера. Розрахункове значення коефіцієнта Фішера обчислюють як відношення більшої дисперсії до меншої. При цьому розглядають ***дисперсію адекватності*** (***залишкову дисперсію***), що характеризує відхилення дослідних даних $\tilde{y}_i$ від значень $\hat{y}_i$, знайдених з лінійної регресії $\hat{y} = \hat{a}_0 + \hat{a}_1 x_1$, і обчислюється за формулою:

$$S^2_{\text{ад}} = \frac{1}{N-2}\sum_{i=1}^{N} (\tilde{y}_i - \hat{y}_i)^2 .$$

У випадку одноразових вимірювань для врахування сукупного впливу всіх випадкових величин використовують оцінку

дисперсії, що характеризує відхилення експериментальних значень відносно середнього $\bar{y}$:

$$S^2\{\tilde{y}\} = \frac{1}{N-1}\sum_{i=1}^{N}(\tilde{y}_i - \bar{y})^2 .$$

Оскільки знайдене з використанням методу найменших квадратів рівняння регресії $\hat{y} = \hat{a}_0 + \hat{a}_1 x_1$ за означенням мінімізує суму квадратів відхилень дослідних даних від лінії регресії, то $S^2_{\text{ад}}$, обчислена через різницю $(\tilde{y}_i - \hat{y})$, буде меншою, ніж $S^2\{\tilde{y}\}$, обчислена через різницю $(\tilde{y}_i - \bar{y})$. Тому розрахунковий коефіцієнт Фішера $F_\text{р}$ визначають за формулою:

$$F_\text{Р} = \frac{S^2\{\tilde{y}\}}{S_{\text{ад}}^{\ 2}}. \qquad (4.48)$$

Розрахункове значення $F_\text{р}$ порівнюють із критичним $F_\text{кр}$, що визначають із таблиць розподілу Фішера за обраним рівнем статистичної значущості α, кількістю степенів вільності чисельника $f_\text{чис} = N - 1$ і кількістю степенів вільності знаменника $f_\text{знам} = N - 2$ .

Якщо $F_\text{р} < F_\text{кр}$, приймається гіпотеза $H_0$, тобто можна стверджувати, що за наявними даними із обраним рівнем статистичної значущості α, розсіювання значень зумовлене тільки впливом випадкових величин, а не впливом фактора $X$ (дані більше відповідають залежності $\hat{y} = \bar{y}$ ).

Якщо $F_\text{р} > F_\text{кр}$, то лінійна залежність $\hat{y} = \hat{a}_0 + \hat{a}_1 x_1$ більше відповідає наявним даним і фактор $X$ справді впливає на функцію відгуку.

**Приклад**. Для α = 0,05 значення $F_\text{кр} = 8{,}2$, а $F_\text{р} = 6{,}7$. З огляду на те, що $F_\text{р} < F_\text{кр}$, не можна стверджувати, що знайдене рівняння регресії $\hat{y} = \hat{a}_0 + \hat{a}_1 x_1$ з імовірністю 0,95 адекватне дослідним даним. Але для α = 0,1 значення $F_\text{кр} = 6{,}2$, тобто $F_\text{кр} < F_\text{р}$. Це говорить про те,

що з імовірністю $P = 1 - \alpha = 0{,}9$ можна на підставі наявних даних стверджувати, що рівняння регресії відповідає дослідним даним.

*Якщо модель неадекватна, необхідно спочатку збільшити обсяг даних, а якщо модель знову виявиться неадекватною, то змінити гіпотезу.*

***2. Експерименти проводяться багаторазово***. Для перевірки, чи справді виконується *лінійна залежність*, необхідно мати додаткову інформацію про дисперсію засобу вимірювання, за допомогою якого визначають значення вихідної величини.

Для врахування впливу похибки вимірювання для кожного фіксованого значення $x_{i1}, i = \overline{1,N},$ проводять $n_i$ вимірювань і дістають сукупність значень $\tilde{y}_{ij},\ j = \overline{1,n_i}.$

Для кожного фіксованого значення $x_{i1}$ (точки факторного простору) знаходять середні значення вихідної величини:

$$\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\tilde{y}_{ij}\ .$$

*Для кожної точки факторного простору можна обчислити* точкові дисперсії, які характеризують розкид значень відносно умовного середнього:

$$S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(\tilde{y}_{ij} - \bar{y}_i)^2, \quad i = \overline{1,N}\ .$$

Якщо умови проведення дослідів однорідні, то точкові дисперсії $S_i^2$ є вибірковими значеннями дисперсії генеральної сукупності, що характеризує випадкову похибку засобу вимірювань. Тому для більш надійної оцінки цієї дисперсії використовують *усереднену дисперсію* (знайдену за всіма $x_i$ точками), яка називається ***дисперсією***

***відтворюваності*** $S_{в}^{2}$ і характеризує *середньозважену точність одного вимірювання*:

$$S_{в}^{2} = \frac{1}{N}\sum_{i=1}^{N} S_i^{2} = \frac{1}{N(n_i - 1)}\sum_{i=1}^{N}\sum_{j=1}^{n_i}(\tilde{y}_{ij} - \bar{y}_i)^2.$$

Беручи до уваги, що за означенням дисперсія адекватності характеризує відхилення експериментальних даних від розрахункових, а в розглядуваному випадку для розрахунків було використано середні значення, доходимо висновку, що дисперсія адекватності набирає вигляду:

$$S^{2}_{ад} = \frac{1}{N-2}\sum_{i=1}^{N} n_i(\bar{y}_i - \hat{y}_i)^2.$$

Відмінність від раніше розглянутого полягає лише в тому, що замість $\tilde{y}_i$ використовується $\bar{y}_i$. Тоді умова мінімізації квадрата розбіжностей між експериментальними та розрахунковими даними запишеться як

$$\sum_{i=1}^{N}(\bar{y}_i - \hat{y}_i)^2 \to min.$$

Оцінки коефіцієнтів $\hat{a}_0, \hat{a}_1$ для розрахунку $\hat{y}_i$ знаходять згідно з цією умовою.

Таким чином, $S^{2}_{ад}$ враховує як вплив похибки засобу вимірювань, так і можливі відхилення через неадекватність моделі. У свою чергу, $S_{в}^{2}$ залежить тільки від похибки засобу вимірювань. Таким чином, у випадку багаторазового вимірювання при кожному фіксованому $x_i$ значення $S^{2}_{ад}$ більше, ніж $S^{2}_{в}$. При цьому

$$F_{р} = \frac{S^{2}{}_{ад}}{S^{2}{}_{в}}. \qquad (4.49)$$

Отримане значення $F_{р}$ порівнюють із $F_{кр}$, отриманим із таблиць Фішера для рівня статистичної значущості α, кількості

степенів вільності чисельника $f_{\text{чис}} = f_{\text{ад}} = N - 2$ і кількості степенів вільності знаменника:

$$f_{\text{знам}} = f_{\text{в}} = N\sum_{i=1}^{N} n_i - N = N\left(\sum_{i=1}^{N} n_i - 1\right).$$

Якщо $F_{\text{р}} < F_{\text{кр}}$, то з обраним рівнем статистичної значущості $\alpha$ можна стверджувати, що досліджувана залежність між $X$ і $Y$ є лінійною, тобто лінійна апроксимація відповідає наявним дослідним даним. Якщо ж ця умова не виконується, то лінійна модель не адекватна дослідним даним, тобто розбіжність між $\hat{y}_i$ і $\overline{y}_i$ зумовлюється впливом похибки вимірювання та обмеженим обсягом експериментальних даних.

## Запитання для самоперевірки

1. Для чого необхідно перевіряти коефіцієнти регресії на статистичну значущість?
2. За яким критерієм перевіряються коефіцієнти регресії на статистичну значущість і яка при цьому висувається основна гіпотеза?
3. За якою формулою обчислюють значення розрахункового коефіцієнту Стьюдента?
4. Як знаходиться критичне значення коефіцієнта Стьюдента для випадку перевірки статистичної значущості коефіцієнтів регресії?
5. За якої умови приймається основна гіпотеза?
6. Що потрібно робити зі статистично незначущими коефіцієнтами?
7. Для чого необхідно перевіряти модель на адекватність експериментальним даним?
8. За яким критерієм відбувається перевірка моделі на адекватність і яка при цьому висувається гіпотеза?

9. Як визначити критичне значення коефіцієнту Фішера?

10. Як обчислюється розрахункове значення коефіцієнту Фішера?

11. Що характеризує дисперсія адекватності?

12. Що характеризує дисперсія відтворюваності? Чому вона називається середньозваженою?

13. У якому випадку дисперсія відтворюваності буде більша: при одноразовому чи багаторазовому експерименті?

14. Що потрібно робити у випадку неадекватної моделі?

## 4.6. МНОЖИННА ЛІНІЙНА РЕГРЕСІЯ

Множинна лінійна регресія характеризує зв'язок між трьома і більше змінними.

У випадку кореляції між кількома змінними рівняння регресії геометрично подають у вигляді певної поверхні, у просторі якої розсіяні дослідні точки.

*Задача* множинного регресійного аналізу полягає у виборі такої прямої в *n*-вимірному просторі, квадрат відхилення результатів спостережень від якої був би мінімальним.

Таким чином, якщо на об'єкт впливає не один, а кілька факторів *n*, рівняння лінійної регресії в загальному випадку матиме вигляд:

$$y = a_0 + a_1 x_1 + \ldots + a_n x_n .$$

Для визначення $n + 1$ коефіцієнта рівняння регресії необхідно мати $n + 1$ рівняння. Вихідним буде рівняння:

$$\hat{a}_0 \sum_{i=1}^{N} x_{i0} + \hat{a}_1 \sum_{i=1}^{N} x_{i1} + \ldots + \hat{a}_n \sum_{i=1}^{N} x_{in} = \sum_{i=1}^{N} \tilde{y}_i ,$$

з якого за відомим правилом (п. 4.4) можна скласти систему нормальних рівнянь і визначити оцінки коефіцієнтів $\hat{a}_j$, $j = \overline{0,n}$:

$$\begin{cases} \hat{a}_0 \sum_{i=1}^{N} x_{i0}^{\,2} + \hat{a}_1 \sum_{i=1}^{N} x_{i0} x_{i1} + \ldots + \hat{a}_n \sum_{i=1}^{N} x_{in} x_{i0} = \sum_{i=1}^{N} \tilde{y}_i x_{i0} \\ \hat{a}_0 \sum_{i=1}^{N} x_{i0} x_{i1} + \hat{a}_1 \sum_{i=1}^{N} x_{i1}^{\,2} + \ldots + \hat{a}_n \sum_{i=1}^{N} x_{in} x_{i1} = \sum_{i=1}^{N} \tilde{y}_i x_{i1} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \hat{a}_0 \sum_{i=1}^{N} x_{i0} x_{in} + \hat{a}_1 \sum_{i=1}^{N} x_{i1} x_{in} + \ldots + \hat{a}_n \sum_{i=1}^{N} x_{in}^{\,2} = \sum_{i=1}^{N} \tilde{y}_i x_{in} \end{cases}. \qquad (4.50)$$

Якщо в кожній точці факторного простору спостереження виконували $n_i$ раз, то замість миттєвих значень $\tilde{y}_i$ необхідно використовувати середні значення $\overline{y}_i$.

*Чим більша кількість впливових факторів, тим більш громіздкою буде система нормальних рівнянь. З огляду на це для зручності обчислень використовують матричну форму подання вхідних величин.*

Вхідні величини досліджуваного об’єкта можна подати у вигляді матриці, доповненої вектором-стовпцем «фіктивної» змінної $x_{i0} \equiv 1$, в якій рядок показує поєднання значень факторів для *i*-го досліду (*i*-ї точки факторного простору), $i = \overline{1, N}$.

Розмір матриці $(n + 1) \times N$, де $N$ – кількість дослідів; $n$ – кількість факторів:

$$\mathbf{X} = \begin{vmatrix} x_{10} & x_{11} & x_{12} & \ldots & x_{1n} \\ x_{20} & x_{21} & x_{22} & \ldots & x_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ x_{N0} & x_{N1} & x_{N2} & \ldots & x_{Nn} \end{vmatrix},$$

Кожному поєднанню факторів відповідають відгуки, які також можна подати у вигляді вектора-рядка**:**

$$\tilde{\mathbf{Y}}^T = \left| \tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_N \right| .$$

Необхідно знайти вектори-рядки оцінок коефіцієнтів:

$$\hat{\mathbf{A}}^T = \left|\hat{a}_0, \hat{a}_1, \hat{a}_2, ..., \hat{a}_n\right|.$$

Рівняння регресії у векторній формі набирає вигляду:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{A}}.$$

З огляду на те, що наявними є тільки експериментальні дані $\tilde{Y}$, задача полягає у визначенні вектора-стовпця $\hat{\mathbf{A}}$, такого, що $\left|\hat{\mathbf{Y}} - \tilde{\mathbf{Y}}\right|^2 \to \min$, що відповідає умові знаходження МНК-оцінок. Тому надалі будемо виходити з рівняння:

$$\tilde{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{A}}. \qquad (4.51)$$

Для безпосереднього знаходження вектора-стовпця $\hat{\mathbf{A}}$ необхідно матричне рівняння (4.51) помножити зліва на обернену матрицю $\mathbf{X}^{-1}$. Однак у загальному випадку матриця може не бути квадратною ($(n + 1) \neq N$), тому попередньо рівняння (4.50) помножують зліва на транспоновану матрицю $\mathbf{X}^T$. У результаті дістаємо матрицю:

$$\mathbf{X}^T\tilde{\mathbf{Y}} = \mathbf{X}^T\mathbf{X}\hat{\mathbf{A}}. \qquad (4.52)$$

Дістанемо квадратну *інформаційну* матрицю $\mathbf{C} = \mathbf{X}^T\mathbf{X}$, яка має вид:

$$\mathbf{C} = \begin{vmatrix} \sum_{i=1}^{N} x^2_{i0} & \sum_{i=1}^{N} x_{i0}x_{i1} & \dots & \sum_{i=1}^{N} x_{i0}x_{in} \\ \sum_{i=1}^{N} x_{i0}x_{i1} & \sum_{i=1}^{N} x^2_{i1} & \dots & \sum_{i=1}^{N} x_{i1}x_{in} \\ \dots & \dots & \dots & \dots \\ \sum_{i=1}^{N} x_{i0}x_{in} & \sum_{i=1}^{N} x_{i1}x_{in} & \dots & \sum_{i=1}^{N} x^2_{in} \end{vmatrix}.$$

*Обернена матриця $\mathbf{X}^{-1}$ існує для квадратної матриці. Для отримання квадратної матриці необхідно матрицю вхідних величин помножити зліва на транспоновану матрицю.*

Зважаючи на те, що матриця **C** квадратна (розмір матриці $(n + 1) \times (n + 1)$), для обчислення вектора-стовпця $\hat{\mathbf{A}}$ можна помножити обидві частини матричного рівняння (4.52) на обернену матрицю $\mathbf{C}^{-1}$.

*Інформаційна матриця* **C** *ще називається дисперсійною, або коваріаційною, оскільки квадратичні діагональні елементи її пропорційні до дисперсії, а решта елементів пропорційна до коваріації.*

У результаті маємо:

$$\mathbf{C^{-1}X}^T\tilde{\mathbf{Y}} = \mathbf{C^{-1}C}\hat{\mathbf{A}}.$$

З огляду на те, що $\mathbf{C}^{-1}\mathbf{C} = 1$, дістаємо:

$$\hat{\mathbf{A}} = \mathbf{C^{-1}X}^T\tilde{\mathbf{Y}}. \qquad (4.53)$$

Звідси оцінки коефіцієнтів можна визначити як

$$\hat{a}_j = \sum_{j=1}^{n} c_{ij}^{-1} \sum_{i=1}^{N} x_{ij}\tilde{y}_i \ . \qquad (4.54)$$

де $c_{ij}^{-1}$ – відповідний елемент матриці $\mathbf{C^{-1}}$.

Знайдені оцінки коефіцієнтів $\hat{a}_j, j = \overline{0,n},$ при цьому будуть *взаємозалежними*, що випливає з вигляду матриці **C**. Наявність взаємозалежності коефіцієнтів незручна в разі потреби їх перерахування. Це можливо у випадках, коли було прийнято хибне припущення про значущість впливових факторів (наприклад, насправді, не впливає або мало впливає фактор $x_k$, який введено до моделі, тобто можна вважати $a_k = 0$), або довелося ввести в модель додатковий фактор, раніше не врахований. У цих випадках доводиться перераховувати *всі* коефіцієнти. Отже, після уточнення рівняння регресії необхідно знову транспонувати матрицю факторів, знаходити коваріаційну матрицю, потім відповідно до (4.54) визначати $\hat{a}_j$.

Знаходити незалежно оцінки коефіцієнтів можна було б, якби матриця **C** була *діагональною*, тобто всі її елементи, крім діагональних, дорівнювали нулю. Для перетворення матриці **C** на діагональну необхідно виконати умову ***ортогональності***:

$$\sum_{k=1}^{N} x_{iu} x_{ik} = 0, u \neq k, \tag{4.55}$$

де $i$ – номер рядка; $j$, $k$ – номери стовпців.

Коваріаційна матриця $\mathbf{C^{-1}}$, як і матриця **C,** також буде діагональною і матиме вигляд:

$$\mathbf{C^{-1}} = \begin{vmatrix} 1/\sum_{i=1}^{N} x^2_{i0} & 0 & \dots & 0 \\ 0 & 1/\sum_{i=1}^{N} x^2_{i1} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1/\sum_{i=1}^{N} x^2_{in} \end{vmatrix}.$$

З урахуванням викладеного вираз (4.54) можна переписати так:

$$\begin{vmatrix} \hat{a}_0 \\ \hat{a}_1 \\ \hat{a}_n \end{vmatrix} = \begin{vmatrix} \sum_{i=1}^{N} x_{i0} \tilde{y}_i / \sum_{i=1}^{N} x^2_{i0} \\ \sum_{i=1}^{N} x_{i1} \tilde{y}_i / \sum_{i=1}^{N} x^2_{i1} \\ \dots \\ \sum_{i=1}^{N} x_{in} \tilde{y}_i / \sum_{i=1}^{N} x^2_{in} \end{vmatrix}.$$

У результаті матричне рівняння (4.54) розпадається на $n$ +1 незалежне рівняння:

$$\hat{a}_j = \frac{\sum_{i=1}^{N} x_{ij} \tilde{y}_i}{\sum_{i=1}^{N} x^2_{ij}}; j = \overline{0,n}. \tag{4.56}$$

Таким чином, вираз (4.56), отриманий при виконанні умови ортогональності, дозволяє *незалежно* знаходити оцінки коефіцієнтів багатофакторної регресійної лінійної моделі.

*Незалежність знаходження оцінок коефіцієнтів дає змогу уточнювати апроксимуючий поліном* (*математичну модель*) *з метою досягнення його адекватності, не перераховуючи інших коефіцієнтів моделі.*

**Запитання для самоперевірки**

1. Що називається лінійною множинною регресією?
2. В якому випадку використовується матричне задання вхідної величини?
3. У чому полягає задача множинного регресійного аналізу?
4. Наведіть рівняння регресії у векторній формі.
5. Що собою являє інформаційна матриця? Яка матриця називається коваріаційною (дисперсійною?)
6. У чому полягає умова ортогональності? Чому необхідно виконувати умову ортогональності?
7. Як незалежно обчислюються коефіцієнти моделі?

## 4.7. НЕЛІНІЙНА ПАРНА РЕГРЕСІЯ

У процесі ідентифікації кореляційного поля виявляється, що в багатьох випадках потрібно відтворювати нелінійну регресійну залежність.

У тому випадку, коли при графічному зображенні точок нелінійність явно проглядається або гіпотеза лінійності не підтверджується, доцільно відтворювати за експериментальними

залежність, яка включає в себе фактори у другому та вищих степенях, а також взаємодії факторів, тобто необхідно розглядати парну залежність, *нелінійну за аргументами* або *нелінійну відносно параметрів.*

**Приклад**. Нелінійна залежність відносно параметрів – вольтамперна характеристика діода.

**Приклад**. Нелінійна модель за факторами: $E = \kappa I^2$.

*Відтворювати нелінійну залежність доцільно тоді, коли розраховують на те, що нелінійна модель дасть меншу дисперсію адекватності, тобто краще відображатиме експериментальні дані.*

Якщо вихідна залежність ***нелінійна відносно параметрів***, то спочатку необхідно здійснити її *лінеаризацію*, а потім застосовувати традиційну процедуру знаходження коефіцієнтів регресії з використанням МНК.

Для нелінійних функцій, які приводять до лінійної форми, виконують різного роду перетворення (логарифмування, гіперболічне перетворення, заміну змінних тощо).

Можна класифікувати типи кривих регресії, які приводять до лінійного вигляду, у такий спосіб.

**1**. ***Поліноміальні залежності:***

$$y = a_0 + \sum_{i=1}^{N} a_i x_i + \sum_{i=1}^{N} b_{ij} x_i x_j + \sum_{i=1}^{N} c_i x_i^2$$

шляхом *введення заміни* $z_i = x_i x_j$, $u_i = x_i^2$ зводять до лінійного вигляду:

$$y = a_0 + \sum_{i=1}^{N} a_i x_i + \sum_{i=1}^{N} b_{ij} z_i + \sum_{i=1}^{N} c_i u_i \, .$$

**2**. ***Залежності, які зводять до лінійного вигляду:***

$$y = a_0 + \sum_{i=1}^{N} a_i z_i$$

шляхом перетворення:

– *зворотного* $z_i = \frac{1}{x_i}$, для $y = a_0 + \sum_{i=1}^{N} a_i \frac{1}{x_i}$;

– *логарифмічного* $z_i = \ln x_i$, для $y = a_0 + \sum_{i=1}^{N} a_i \ln x_i$;

– *степеневого* $z_i = x_i^{\,n}$, для $y = a_0 + \sum_{i=1}^{N} a_i x^{n}{}_{i}$.

***3. Мультиплікативні залежності***:

$$y = a_0 \prod_{i=1}^{N} x_i^{a_i}, \qquad y = a_0 \prod_{i=1}^{N} a_i^{x_i},$$

які шляхом логарифмування

$$\ln y = \ln a_0 + \sum_{i=1}^{N} a_i \ln x_i; \qquad \ln y = \ln a_0 + \sum_{i=1}^{N} x_i \ln a_i$$

приводяться до відповідного лінійного виду.

***4. Експонентні залежності:***

$$y = exp\left( a_0 + \sum_{i=1}^{N} a_i x_i \right),$$

які зводяться до вигляду:

$$\ln y = a_0 + \sum_{i=1}^{N} a_i x_i .$$

Для лінеаризованого рівняння за відомим правилом (розд. 4.4) записують систему нормальних рівнянь.

**Приклад**. Модель має вигляд:

$$y = a_0 + a_1 x_1 + a_{11} x^2{}_1.$$

Зробимо заміну $a_{11}x_1^{\,2} = a_2x_2$ і складемо систему нормальних рівнянь

$$\begin{cases} \hat{a}_0\sum\limits_{i=1}^{N}x_{i0}^2 + \hat{a}_1\sum\limits_{i=1}^{N}x_{i0}x_{i1} + \hat{a}_2\sum\limits_{i=1}^{N}x_{i0}x_{i2} = \sum\limits_{i=1}^{N}x_{i0}\tilde{y}_i \\ \hat{a}_0\sum\limits_{i=1}^{N}x_{i0}x_{i1} + \hat{a}_1\sum\limits_{i=1}^{N}x_{i1}^2 + \hat{a}_2\sum\limits_{i=1}^{N}x_{i1}x_{i2} = \sum\limits_{i=1}^{N}x_{i1}\tilde{y}_i \\ \hat{a}_0\sum\limits_{i=1}^{N}x_{i0}x_{i2} + \hat{a}_1\sum\limits_{i=1}^{N}x_{i1}x_{i2} + \hat{a}_2\sum\limits_{i=1}^{N}x_{i2}^2 = \sum\limits_{i=1}^{N}x_{i2}\tilde{y}_i \end{cases}.$$

Розв'язавши цю систему рівнянь, дістанемо оцінки коефіцієнтів парної квадратичної моделі $\hat{a}_0$, $\hat{a}_1$, $\hat{a}_{11}$.

Скориставшись правилом побудови системи нормальних рівнянь, можна записати систему рівнянь, яка дає змогу дістати оцінки коефіцієнтів парної залежності будь-якого порядку. Однак практично при виконанні матричних операцій та операцій піднесення до степені за допомогою комп'ютера для одержання формул порядку вищого за четвертий, похибки округлення настільки великі, що зводиться нанівець увесь виграш від підвищення порядку регресії. Тому з деякого моменту при підвищенні порядку рівняння регресії залишкова дисперсія (дисперсія адекватності) замість того, щоб зменшуватися, може збільшуватися.

**Приклад.** *Застосування логарифмування.*

Функціональна залежність має вигляд:

$$y = a_0x_1^{\,a1}.$$

Необхідно знайти $a_0$ й $a_1$.

Оскільки МНК придатний тільки для рівняння виду $y = a_0 + a_1x_1$, то щоб позбутися степені та замінити добуток додаванням, застосовують логарифмування наведеного виразу:

$$\ln y = \ln a_0 + a_1 \ln x_1.$$

Зробимо заміну:

$$y' = \ln y;\ a'_0 = \ln a_0;\ a'_1 = a_1;\ x'_1 = \ln x_1. \qquad (4.57)$$

З урахуванням введених замін отримаємо рівняння, лінійне за параметрами:

$$y' = a'_0 + a'_1 x'_1.$$

Для лінеаризованого рівняння за відомим правилом записують систему нормальних рівнянь із двома невідомими $a'_0$ і $a'_1$. Знаходять оцінки коефіцієнтів $\hat{a}'_0$ і $\hat{a}'_1$. Ураховуючи співвідношення (4.57), знаходять значення коефіцієнтів $a_0$ і $a_1$.

**Приклад**. *Застосування оберненого перетворення.*

Модель має вигляд: $y = \dfrac{1}{a_0 + a_1 x_1}$.

Уведемо позначення $y' = \dfrac{1}{y}$.

Виконавши обернене перетворення, дістають лінійне рівняння $y' = a_0 + a_1 x_1$. Надалі діють за відомим алгоритмом.

**Приклад**. *Застосування логарифмічного та оберненого перетворень*.

Модель має вигляд :

$$y = \frac{1}{1 + exp(a_0 + a_1 x_1)}.$$

Виконують обернене перетворення:

$$y^{-1} = 1 + exp(a_0 + a_1 x_1) \quad \text{або} \quad y^{-1} - 1 = exp(a_0 + a_1 x_1),$$

а далі логарифмують обидві частини цього рівняння і дістають:

$$ln\left(y^{-1} - 1\right) = a_0 + a_1 x_1.$$

Позначивши

$$y' = ln\left(y^{-1} - 1\right),$$

дістають лінійне за параметрами рівняння

$$y' = a_0 + a_1 x_1.$$

*Строго кажучи, ортодоксальної теорії нелінійної регресії не існує, однак зведення до лінійної форми щодо шуканих параметрів дає змогу реалізовувати статистичні критерії лінійної регресії.*

Перетворення регресійної залежності, нелінійної за аргументами, до лінійної форми має свої недоліки. Вони полягають у тому, що оцінки параметрів $\hat{a}_j$, які ввійшли в лінеаризовану модель і були знайдені за допомогою методу найменших квадратів, насправді мінімізують не суму квадратів відхилень значень $\tilde{y}_i$ отриманих дослідним шляхом, від регресійної кривої $\hat{y}_i$ а мінімізують суму квадратів відхилень перетворених значень вихідної величини $\tilde{y}_i'$ від відповідних значень лінеаризованої регресійної моделі $\hat{y}_i'$

$$\sum_{i=1}^{N}(\tilde{y}_i' - \hat{y}_i')^2 .$$

Тому при оберненому перетворенні і переході до вихідної моделі матиме місце зсув МНК-оцінок коефіцієнтів нелінійної регресійної моделі. Виходячи з цього потрібно вводити коригувальні коефіцієнти або застосовувати спеціальні прийоми, які забезпечують найкраще наближення нелінійної регресійної залежності до експериментальних даних. Із цією метою, наприклад, для квазілінійних залежностей застосовують різноманітні *ітераційні процедури.*

Слід зазначити, що перспективними алгоритмами одержання стійких оцінок є *робасні процедури* [4]. Реалізувавши робасні процедури, можна зняти обмеження на застосування методу найменших квадратів завдяки тому, що припускається порушення початкової умови – вимоги про нормальний закон розподілу випадкової величини.

**Запитання для самоперевірки**

1. Які залежності належать до нелінійних відносно параметрів?
2. До залежностей якого виду застосовується МНК?
3. Наведіть класифікацію типів кривих регресії, які зводяться до лінійного вигляду.
4. Сформулюйте правило, за яким отримується система нормальних рівнянь.
5. Які ви знаєте методи лінеаризації? Наведіть приклади.
6. Які процедури застосовують для квазілінійних залежностей?

## 4.8. ОРТОГОНАЛЬНІ ПОЛІНОМИ ЧЕБИШОВА. ОРТОГОНАЛІЗАЦІЯ РІВНЯННЯ РЕГРЕСІЇ

Найпростіша обчислювальна схема побудови поліноміальної регресії ґрунтується на ортогональних поліномах Чебишова.

**Поняття про апроксимацію та інтерполяцію.** Кінцевою метою експериментальних досліджень є добір такої аналітичної залежності – функції $f(x)$, яка б за вибраним критерієм щонайкраще відповідала наявним даним (рис. 4.10).

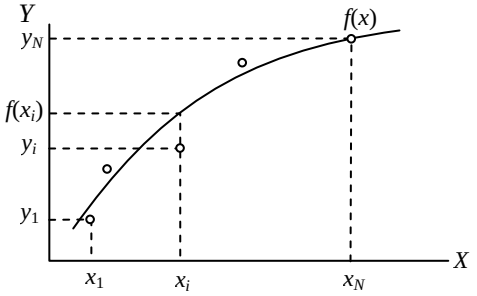



Рис. 4.10. Апроксимація експериментальних даних функціональною залежністю

Експериментальні дані, на базі яких потрібно відтворити функціональну залежність $f(x)$, мають розсіювання, зумовлене або наявністю випадкової похибки засобів вимірювань, або впливом неврахованих випадкових величин. Тому немає сенсу вимагати, щоб функція $f(x)$ проходила через усі вузлові точки. У цьому випадку більш ефективним є вибір залежності, яка б *згладжувала* (усереднювала) вплив розсіювання експериментальних даних, тобто перехід до процедури ***апроксимації***, що дає змогу невідому реальну залежність між величинами $X$ та $Y$ фактично замінити іншою залежністю, причому крива, яка її відображає, має проходити між парами спостережень $x_i$, $y_i$ згідно з вибраним критерієм оптимальності. Найчастіше таким критерієм є мінімізація $\sum_{i=0}^{N}[\tilde{y}_i - f(x_i)]^2$, тобто застосовується *квадратична апроксимація*. Іноді застосовують *лінійну апроксимацію*, коли

$$\sum_{i=0}^{N}|\tilde{y}_i - f(x_i)| = min.$$

У будь-якому випадку при проведенні експериментальних досліджень добирають таку аналітичну залежність, яка найкраще відповідає отриманим експериментальним даним.

Апроксимуюча функція дає змогу визначити значення функції $f(x)$ у *проміжних* точках $x_j$ при відомих значеннях її в *опорних*, або вузлових, точках $x_i$. У цьому випадку існує процедура, яка називається ***інтерполяцією***, за якої значення $f(x)$ знаходять у вигляді узагальненого багаточлена:

$$f(x) = \sum_{i=0}^{N} k_i f_i^*(x),$$

де $f_i^*(x)$ – деяка система функцій, лінійно незалежна на інтервалі експериментальних даних; $k_i$ – дійсні коефіцієнти; $i = \overline{1, N}$ – вузлові точки, які відповідають парам наявних даних $x_i$, $y_i$.

Побудова конкретної інтерполяційної функції $f(x)$ зводиться до вибору вузлових точок та пошуку $k_i$ з тим, щоб різниця

$$\sum_{j=0}^{N} k_i f_i^*(x_j) - f(x_j) \qquad (4.58)$$

відповідала вибраному критерію.

*Коли за функцією $f(x)$, визначеною у вузлових точках, визначаються (прогнозуються) її значення для аргументів $x_j$, які містяться поза наявним полем кореляції, відбувається екстраполяція. Варто зазначити, що екстраполяція має обмежене застосування. Наприклад, для функцій із перегинами її не можна застосовувати.*

На практиці найчастіше застосовують інтерполяцію з використанням *алгебраїчних багаточленів* (*параболічна інтерполяція*), тобто багаточленів, побудованих за системою функцій 1, $x$, $x^2$,..., $x^n$. Такий спосіб наближення ґрунтується на гіпотезі, що на невеликих відрізках зміни $x$ функцією $f_i^*(x)$ можна достатньо добре відобразити наявні експериментальні дані за допомогою параболи деякого порядку. Із цією метою можна використати інтерполяційні багаточлени Лагранжа, Ньютона.

Проте процедура визначення функції $f(x)$ за інтерполяційними багаточленами потребує великої кількості обчислювальних операцій. Крім того, якщо знайдено на базі $i = \overline{1, N}$ вузлових точок функцію $f_i^*(x)$, то для ($N$+1)-ї вузлової точки всі обчислення потрібно виконувати знову.

Якщо поліном третього порядку не може достатньо відобразити існуючу залежність, потрібно переходити до багаточлена більш високого порядку. У цьому випадку краще переходити до інтерполяції *кусково-аналітичними* функціями – ***сплайнами***.

Найчастіше використовуються кубічні параболи, які в усіх випадках точно проходять через дві опорні точки. При $N$ опорних точках *сплайнова функція* складається з $N - 1$ окремого полінома (рис. 4.11). *Кубічний сплайн* являє собою групу кубічних багаточленів виду:

$$f_i^*(x) = a_i x + b_i x^2 + c_i x^3 + d\,,$$

для яких у місцях поєднання (стиків) перша і друга похідна неперервні і за значеннями збігаються.

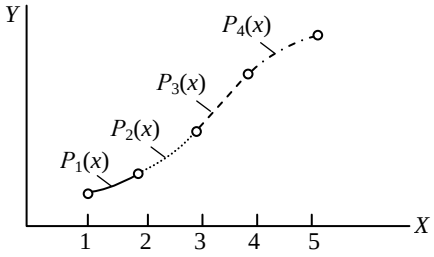



Рис. 4.11. Представлення експериментальних даних кубічним сплайном

Сплайн-інтерполяцію можна застосовувати для *нерівновіддалених вузлів* (залежність може мати кілька екстремумів), проте вона потребує значних обчислювальних витрат для знаходження коефіцієнтів сплайнів.

При апроксимації, на відміну від інтерполяції, не висуваються вимоги, щоб шукана функціональна залежність та крива, що її відображає, *точно проходили через вузлові точки*, тобто *відсутня умова* $y_i = f(x_i)$.

У загальному випадку апроксимуючий поліном

$$P(x) = a + bx + cx^2 + dx^3.$$

Максимальний степінь полінома залежить від кількості пар значень $x_i$, $y_i$, $i = \overline{1, N}$, за якими будується апроксимуюча залежність $f(x)$, і дорівнює $N$–1.

*Без застосування спеціальних методів використовувати поліноми більш високих порядків для апроксимації не рекомендується, бо похибка обчислень і округлень практично перекреслює ефект, якого вдається досягти збільшенням порядку полінома.*

**Апроксимація Чебишова**. При обробці експериментальних даних, особливо для випадку широкого діапазону зміни аргументу $X$, і наявності кількох екстремумів, може виявитися, що побудована передбачувана залежність $f(x)$ не відповідає повною мірою наявним даним. У такому разі необхідно змінювати вид моделі, ускладнюючи її, але при цьому, якщо не застосовувати особливих заходів, необхідно заново обробляти всі дані та знаходити нові коефіцієнти цієї моделі.

Як зазначалося раніше, апроксимація залежності нелінійним поліномом має обмеження через вплив похибки округлення при збільшенні його степені. Крім того, при кожному підвищенні степені полінома доводиться обчислювати не тільки коефіцієнт нововведеного доданка, але й перераховувати решту коефіцієнтів.

Уникнути цих недоліків дає змогу апроксимація залежності *поліномами Чебишова*. Основна ідея такої апроксимації полягає в тому, що багаточлен знаходиться не просто у вигляді суми степенів $x$, а у вигляді комбінації багаточленів, які вибираються у спеціальний спосіб.

При цьому шуканий багаточлен $y = f(x)$ для парної регресії запишеться як

$$y = \alpha_0 P_0(x) + \alpha_1 P_1(x) + \ldots + \alpha_m P_m(x),$$

де $P_j(x)$ – поліном степені $j$, $j = \overline{0,m}$ ; $\alpha_j$ – коефіцієнти регресії.

*Коефіцієнти $\alpha_j$ відрізняються від коефіцієнтів $a_j$ звичайної множинної регресії* (*п. 4.6*). *Коефіцієнти $\alpha_j$ не відображають впливу фактора, а відповідають вкладу полінома $P_j(x)$ у результуючу модель залежності.*

Припустивши, що визначено вид апроксимуючого полінома, можна знайти значення коефіцієнтів $\alpha_0$, $\alpha_1$,..., $\alpha_m$ цього полінома згідно з принципом Лежандра, тобто знайти мінімум функції:

$$U = \sum_{i=1}^{N} \{\tilde{y}_i - [\hat{\alpha}_0 P_0(x_i) + \hat{\alpha}_1 P(x_i) + ... + \hat{\alpha}_m P_m(x_i)]\}^2 .$$

Обчислення частинних похідних функції $U$ за всіма коефіцієнтами $\alpha_0$, $\alpha_1$,...,$\alpha_m$ дає можливість побудувати систему нормальних рівнянь за методом найменших квадратів. Розв'язавши цю систему, можна знайти оцінки коефіцієнтів полінома.

Однак знайдені безпосередньо з цієї системи оцінки коефіцієнтів будуть взаємозв'язаними, а процедура їх визначення буде трудомісткою. Цього можна уникнути шляхом відповідного вибору складових багаточлена $P_0(x)$, $P_1(x)$,..., $P_m(x)$, тобто забезпеченням виконання *умови їхньої ортогональності*:

$$\sum_{i=1}^{N} P_j(x_i) P_k(x_i) = 0, j \neq k, \quad , \quad j,k = \overline{0,(N-1)}. \qquad (4.59)$$

Це дає змогу *незалежно визначити оцінки коефіцієнтів* $\hat{\alpha}_j$ і оцінити внесок кожної складової у точність апроксимації. При цьому введення нових доданків у модельне рівняння не змінює обчислених раніше коефіцієнтів. Додаючи у такий спосіб доданок за доданком до полінома, можна спостерігати, як зменшується залишкова дисперсія.

На відміну від апроксимації у вигляді суми доданків степенів $x$, поліноми Чебишова дозволяють апроксимувати залежності, які мають *точки перегину.*

*Основною умовою* застосування поліномів Чебишова є рівновіддаленість вузлових точок.

Вплив похибки округлення при використанні поліномів Чебишова неістотний, бо складові вищої степені уточнюють апроксимуючу функцію і похибка округлення у цьому випадку буде величиною другого порядку мализни.

Таким чином, ортогональні поліноми Чебишова мають позитивні властивості, притаманні інтерполяції параболічними багаточленами (знаходять коефіцієнти за системою нормальних рівнянь) і апроксимації кусково-аналітичними функціями (апроксимують залежності, які мають точки перегину).

Варто зазначити, що при ***пасивному експерименті*** ортогональні поліноми Чебишова – це практично єдиний «інструмент», який дає змогу при статистичній обробці експериментальних даних незалежно оцінювати коефіцієнти моделі і проводити її уточнення, вводячи додаткові складові**.**

*Пасивним називається експеримент, при якому в хід функціонування об'єкта не втручаються, а дані одержують при робочому, а не тестовому функціонуванні досліджуваного об'єкта.*

**Приклад**. У результаті дослідження у п'яти точках $x$={2; 5; 8; 11; 14}, отримані відповідні результати $y$={2,0; 7,8; 6,2; 5,8; 8,2}, яким, з точки зору експериментатора, може відповідати залежність, представлена на рис. 4.12. Необхідно отримати аналітичну функцію, яка б щонайкраще відповідала наявним даним.

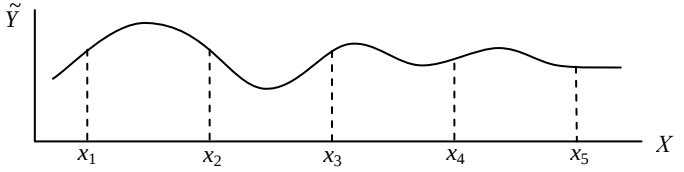



Рис. 4.12. Залежність з точками перегину

Як бачимо з рис. 4.12, залежність має точки перегину, а $X$ змінюється у широкому діапазоні. З математичного аналізу відомо, що через точки на площині можна провести криву $(N-1)$-го порядку

(проведення кола через три точки). Для наведеної залежності діапазон зміни аргументу розбитий на чотири рівні частини з вузловими точками $X_i$, $i = 1,...,5$, яким відповідають значення відгуків $\tilde{Y}_i$. Виходячи з викладеного, на підставі дослідних даних можна побудувати модель, яка містить поліном четвертої степені:

$$y(x) = \alpha_0 P_0(x) + \alpha_1 P_1(x) + \alpha_2 P_2(x) + \alpha_3 P_3(x) + \alpha_4 P_4(x), \quad (4.60)$$

де $P_j$ – ортогональний поліном; $j$ – показник степені полінома, $j = \overline{0,4}$; $\alpha_j$ – коефіцієнти моделі, які підлягають визначенню за даними експерименту; *x – кодована змінна*.

Для того щоб формалізувати обробку експериментальних даних, на практиці здійснюють ***кодування*** незалежних змінних, тобто подання сукупності наявних значень у вигляді *цілих безрозмірних чисел*. Це дає змогу при обробці й побудові моделей не прив'язуватися до якоїсь конкретної фізичної величини й динамічного діапазону її змін. За допомогою кодування можна перейти до відносної шкали для всіх вхідних величин. У відносних (кодованих) одиницях фізичну величину позначають $x_i$, а відліки беруть через *рівні* інтервали

$$d = x_N - x_{N-1}.$$

Для створення кодованих шкал розглядають випадки, коли кількість відліків $N$ парна і непарна.

Для випадку, коли $N$ – парне, формула кодування набирає вигляду:

$$x_i = \frac{X_i - \overline{X}}{d/2},$$

де $X_i$ – дослідні розмірні дані.

У цьому випадку $x_i$ набуває значень ± 1; ± 3; …; ± ($N$ – 1), які містяться симетрично відносно $\overline{x}$ (рис. 4.13).

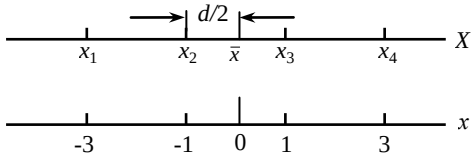



Рис.4.13. Створення кодованих шкал

**Приклад**. Подання значень $X_1$ і $X_2$ у кодованому вигляді при парній кількості відліків $N$:

$$x_1 = \frac{X_1 - \left(X_1 + 3/2\, d\right)}{d/2} = -3; \quad x_2 = \frac{X_2 - \left(X_2 + d/2\right)}{d/2} = -1.$$

Для випадку, коли $N$ – непарне, формула кодування така:

$$x_i = \frac{X_i - \overline{X}}{d},$$

при цьому $x_j$ набуває значень 0; ±1; ±2; …; ±($N$−1)/2... (рис. 4.14).

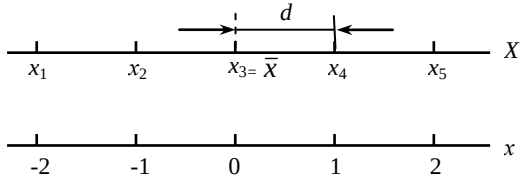



Рис. 4.14. Створення кодованої шкали для випадку, коли $N$ – непарне

**Приклад**. Подання значень $X_1$ і $X_2$ у кодованому вигляді при непарній кількості відліків $N$:

$$x_1 = \frac{X_1 - \left(X_1 + 2d\right)}{d} = -2; \quad x_2 = \frac{X_2 - \left(X_2 + d\right)}{d} = -1.$$

В обох випадках кодована змінна набуває симетричних щодо нуля цілочислових значень.

Для поліномів Чебишова існує рекурентне співвідношення, яке дає змогу здійснити їх обчислення:

$$P_{j+1}(x)=P_1(x)P_j(x)-\frac{j^2\left(N^2-j^2\right)}{4\left(4j^2-1\right)}P_{j-1}(x).$$

З урахуванням формул кодування, а також цього співвідношення перші п'ять складових полінома Чебишова з точністю до постійного множника $\lambda_j$ визначають у такий спосіб:

$$P_0(x)=1;$$

$$P_1(x)=\lambda_1\frac{X-\overline{X}}{d};$$

$$P_2(x)=\lambda_2\left[\left(\frac{X-\overline{X}}{d}\right)^2-\frac{N^2-1}{12}\right]; \qquad (4.61)$$

$$P_3(x)=\lambda_3\left[\left(\frac{X-\overline{X}}{d}\right)^3-\left(\frac{X-\overline{X}}{d}\right)\left(\frac{3N^2-7}{20}\right)\right];$$

$$P_4(x)=\lambda_4\left[\left(\frac{X-\overline{X}}{d}\right)^4-\left(\frac{X-\overline{X}}{d}\right)^2\left(\frac{3N^2-13}{14}\right)\right]+\frac{3\left(N^2-1\right)\left(N^2-9\right)}{560},$$

де $X$ – будь-яке поточне значення фактора, $\overline{X}$ – середнє значення фактора; $\lambda_j$ – множники.

Множники $\lambda_j$ залежать від кількості експериментальних точок $N$, а також степені складової полінома $j$ і вводяться для того, щоб у точках $i=\overline{1,N}$ поліноми $P_j(x_i)$ набували цілочислових значень. Так, для полінома першої степені $\lambda_1 = 1$ при непарних $N$ і $\lambda_2 = 2$ при парних $N$.

Значення ортогональних поліномів $P_j(x_i)$ залежать від кількості експериментальних точок $N$ і від степені полінома $j$. Деякі кодовані значення ортогональних поліномів $P_j(x_i)$ наведено в табл. 4.7.

Згідно з основними передумовами побудови ортогональних поліномів Чебишова мають виконуватися умови ортогональності:

$$\sum_{i=1}^{N} P_j P_k = 0,$$

а також симетрії:

$$\sum_{i=1}^{N} P_j = 0.$$

Як випливає з табл. 4.7, ці умови виконуються.

Таблиця 4.7. Кодові значення ортогональних складових

| Номер досліду, $i$ | **N = *3*** | | **N = *4*** | | | **N = *5*** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $P_1$ | $P_2$ | $P_1$ | $P_2$ | $P_3$ | $P_1$ | $P_2$ | $P_3$ | $P_4$ |
| 1 | -1 | +1 | -3 | +1 | -1 | -2 | +2 | -1 | +1 |
| 2 | 0 | -2 | -1 | -1 | +3 | -1 | -1 | +2 | -4 |
| 3 | +1 | +1 | +1 | -1 | -3 | 0 | -2 | 0 | +6 |
| 4 | | | +3 | +1 | +1 | +1 | -1 | -2 | -4 |
| 5 | | | | | | +2 | +2 | +1 | +1 |
| $\sum_{i=1}^{N} P_j^2(x)$ | 2 | 6 | 20 | 4 | 20 | 10 | 14 | 10 | 70 |
| $\lambda_j$ | 1 | 3 | 2 | 1 | $\frac{10}{3}$ | 1 | 1 | $\frac{5}{6}$ | $\frac{35}{12}$ |

**Обчислення коефіцієнтів моделі $\alpha_j$.** З огляду на те, що поліноми Чебишова за означенням є ортогональними, для знаходження оцінок коефіцієнтів $\hat{\alpha}_j$ можна за аналогією з (4.53) і (4.54) записати матричне рівняння:

$$\hat{\boldsymbol{\alpha}}_{\mathbf{j}} = \mathbf{C}^{-1}\mathbf{P}^{\mathbf{t}}\overline{\mathbf{Y}}, \qquad (4.62)$$

яке розпадається на $N + 1$ незалежних рівнянь. У виразі (4.62) $\overline{\mathbf{Y}}$ – вектор-стовпець середніх значень вихідних величин (припускається, що вимірювання в кожній точці проводились кілька разів); $\hat{\boldsymbol{\alpha}}_{\mathbf{j}}$ – вектор-стовпець шуканих оцінок коефіцієнтів моделі; $\mathbf{C}^{-1} = \left(\mathbf{P}^t \cdot \mathbf{P}\right)^{-1}$; $\mathbf{P}^t$ – матриця, транспонована щодо матриці

$$\mathbf{P}=\begin{vmatrix} P_0(x_1) & P_1(x_1) & P_2(x_1) & \dots & P_j(x_1) & \dots & P_n(x_1) \\ P_0(x_2) & P_1(x_2) & P_2(x_2) & \dots & P_j(x_2) & \dots & P_n(x_2) \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ P_0(x_i) & P_1(x_i) & P_2(x_i) & \dots & P_j(x_i) & \dots & P_n(x_i) \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ P_0(x_N) & P_1(x_N) & P_2(x_N) & \dots & P_j(x_N) & \dots & P_n(x_N) \end{vmatrix},$$

де елементами матриці є кодовані значення поліномів $P_j(x_i)$ степені $j$ у точці $i$. Оцінку кожного з коефіцієнтів кодованих змінних $\hat{\alpha}_j$ визначають незалежно за формулою:

$$\hat{\alpha}_j=\frac{\sum_{i=1}^{N}\tilde{y}_i P_j(x_i)}{\sum_{i=1}^{N}P_j^2(x_i)},\ j=\overline{1,N-1}. \qquad (4.63)$$

*Вираз* (*4.63*) *подібний до виразу* (*4.54*). *Відмінність полягає в тому, що замість кодованого значення* $x_{ij}$ *використовують кодоване значення поліномів* $P_j(x_i)$.

Обчислимо значення коефіцієнтів за виразом (4.63), для результатів $y_i$ , наведених у попередньому числовому прикладі**.** Отримаємо такі значення коефіцієнтів моделі:

$$\alpha_0=\frac{1}{5}\cdot(2,0+7,8+6,2+5,8+8,2)=6,0;$$

$$\alpha_1=\frac{1}{10}\cdot\big((-2)\cdot 2,0+(-1)\cdot 7,8+(0)\cdot 6,2+(+1)\cdot 5,8+(+2)\cdot 8,2\big)=1,02;$$

$$\alpha_2=\frac{1}{14}\cdot\big((+2)\cdot 2,0+(-1)\cdot 7,8+(-2)\cdot 6,2+(-1)\cdot 5,8+(+2)\cdot 8,2\big)=-0,47;$$

$$\alpha_3=\frac{1}{10}\cdot\big((-1)\cdot 2,0+(+2)\cdot 7,8+(0)\cdot 6,2+(-2)\cdot 5,8+(+1)\cdot 8,2\big)=1,02;$$

$$\alpha_4=\frac{1}{70}\cdot\big((+1)\cdot 2,0+(-4)\cdot 7,8+(+6)\cdot 6,2+(-4)\cdot 5,8+(+1)\cdot 8,2\big)=-0,1.$$

**Перевірка значущості коефіцієнтів моделі та її адекватності при відомій дисперсії**. Знайдені коефіцієнти згідно з процедурою статистичної обробки результатів необхідно перевірити на ***статистичну значущість***. Припускається, що раніше було проведено додаткові дослідження і це дає змогу вважати дисперсію вимірювання $\sigma_y^2$ відомою. У цьому разі істотність відхилень розглядається відносно

$$\alpha_{j\,кр} = z \cdot S\{\hat{\alpha}_j\}, \tag{4.64}$$

де $z$ – квантиль нормального розподілу; $S^2\{\hat{\alpha}_j\}$ – дисперсії знайдених оцінок коефіцієнтів, зумовлені впливом випадкових величин або обмеженим обсягом вибірки, які визначаються за формулою:

$$S^2\{\hat{\alpha}_j\} = \frac{S_в^2}{\sum_{i=1}^{N} P_j^2(x_i) m}, \tag{4.65}$$

де $m$ – кількість паралельних спостережень в $i$-й точці; $S_в^2$ – дисперсія відтворюваності, яка характеризує середньозважену точність проведення одного вимірювання.

У випадку, коли дисперсія вимірювання відома, можна вважати $S_в^2 = \sigma_y^2$, і в кожній точці проводити тільки по одному вимірюванню, тобто $m = 1$.

Зважаючи на те, що в знаменнику виразу (4.65) для кожного $j$-го коефіцієнта, на відміну від лінійної регресії, будуть знаходитись різні суми значень $P_j^2(x_i)$, то й дисперсії $S^2\{\hat{\alpha}_j\}$ знайдених коефіцієнтів $\hat{\alpha}_j$ також будуть різними.

Розсіювання значень оцінок коефіцієнтів $\hat{\alpha}_j$ представляється у вигляді нормального розподілу з центром $\alpha_j$ і дисперсією $S^2\{\hat{\alpha}_j\}$.

Знайдене значення $\left|\hat{\alpha}_j\right|$ порівнюється з $\alpha_{j\ кр.}$. Чим більше $\left|\hat{\alpha}_j\right|$ за $\alpha_{j\,кр}$, тим суттєвіший вплив складової багаточлена $x_j$.

Якщо $\left|\hat{\alpha}_j\right| < \alpha_{j\,кр}$, гіпотеза $H_0$ приймається, тобто можна стверджувати з вибраним рівнем статистичної значущості, що $j$-й коефіцієнт буде статистично незначущим, і його слід виключити з математичної моделі. Беручи до уваги, що поліноми, які входять до моделі, ортогональні, це не призведе до необхідності перерахунку інших коефіцієнтів моделі.

Для числового прикладу, що розглядається, перевіримо отримані оцінки коефіцієнтів на статистичну значність. Візьмемо $S^2_{в} = 0{,}2$; $m = 1$, та обчислимо значення дисперсії знайдених оцінок коефіцієнтів. Значення знаменників береться з табл. 4.7 за $N$=5. Тоді:

$$S^2(a_0) = \frac{0{,}2}{5} = 0{,}201; \quad S^2(a_2) = \frac{0{,}2}{14} = 0{,}120; \quad S^2(a_4) = \frac{0{,}2}{70} = 0{,}054.$$
$$S^2(a_1) = \frac{0{,}2}{10} = 0{,}142; \quad S^2(a_3) = \frac{0{,}2}{10} = 0{,}142;$$

Виходячи з обчислених значень дисперсії знайдених оцінок коефіцієнтів, визначимо критичні значення для цих коефіцієнтів. Для нормального закону розподілу за $P = 0{,}95$; $z_{кр} = 1{,}96$. Тоді:

$$\alpha_{кр\_0} = 1{,}96 \cdot 0{,}201 = 0{,}39 \qquad \alpha_{кр\_1} = 1{,}96 \cdot 0{,}142 = 0{,}28$$
$$\alpha_{кр\_2} = 1{,}96 \cdot 0{,}120 = 0{,}24 \qquad \alpha_{кр\_3} = 1{,}96 \cdot 0{,}142 = 0{,}28$$
$$\alpha_{кр\_4} = 1{,}96 \cdot 0{,}054 = 0{,}105$$

Порівняємо розраховані значення коефіцієнтів із критичними значеннями:

$$|a_0| = 6{,}0 > 0{,}39; \quad |a_2| = 0{,}47 > 0{,}24; \quad |a_4| = 0{,}1 < 0{,}105.$$
$$|a_1| = 1{,}02 > 0{,}28; \quad |a_3| = 1{,}02 > 0{,}28;$$

Як бачимо, усі оцінки коефіцієнтів окрім $\alpha_4$ перевищують відповідні критичні значення, тому вважаються статистично значущими. У цьому разі модель (4.60) набуде вигляду:

$$y(x) = 6.0 \cdot P_0(x) + 1.02 \cdot P_1(x) - 0.47 \cdot P_2(x) + 1.02 \cdot P_3(x). \qquad (4.66)$$

Модель зі статистично значущими коефіцієнтами необхідно перевірити на ***адекватність*** об'єкту, використовуючи для цього наявні експериментальні дані. За аналогією з раніше розглянутим (п. 4.6) необхідно обчислити коефіцієнт Фішера за формулою (4.42).

У цій формулі:

$$S^2_{\text{ад}} = \frac{1}{N-l} \sum_{i=1}^{N} \left(\tilde{y}_i - \hat{y}_i\right)^2, \qquad (4.67)$$

де $l$ – кількість статистично значущих коефіцієнтів моделі; $\tilde{y}_i$ – експериментальні значення; $\hat{y}_i$ – значення, обчислені на підставі математичної моделі для $i$-ї точки.

Розрахункове значення $F_{\text{р}}$ порівнюють із критичним значенням коефіцієнта Фішера $F_{\text{кр}}$ при рівні статистичної значущості $\alpha$ і кількості степенів вільності чисельника $f_{\text{ад}} = N - l$ і знаменника $f_{\text{в}} = \infty$ (значення $\sigma_y^2$ відоме, тобто вважається, що воно визначене при нескінченній кількості вимірювань).

Якщо $F_{\text{р}} < F_{\text{кр}}$, то з обраним рівнем статистичної значущості можна стверджувати, що знайдена за експериментальними даними модель адекватна об'єкту.

Якщо $F_{\text{р}} > F_{\text{кр}}$, то модель неадекватна і для уточнення її необхідно включити доданок $\alpha_{m+1} P_{m+1}(x)$, тобто складову більш високого порядку.

Оскільки виконується умова ортогональності, оцінка коефіцієнта для додаткової складової моделі $\hat{\alpha}_{m+1}$ буде визначатися незалежно від інших коефіцієнтів, тобто раніше знайдені оцінки коефіцієнтів не треба перераховувати.

*Дану процедуру можна здійснити, якщо дозволяє кількість пар експериментальних даних N, оскільки степінь полінома має не перевищувати N–1.*

**Перевірка значущості коефіцієнтів моделі та її адекватності при невідомій дисперсії**. У цьому випадку, на відміну від раніше розглянутого, необхідно обчислювати дисперсію відтворюваності $S_{\text{в}}^2$ на підставі дослідних даних як середнє значення *точкових дисперсій*:

$$S_{\text{в}}^2 = \frac{1}{N}\sum_{i=1}^{N} S^2\{\tilde{y}_i\}.$$

Оскільки умови проведення експерименту дозволяють проводити в кожній точці $m$ спостережень, то для кожного значення $X_i$ при $m \geq 2$ можна обчислити дисперсію точкового результату:

$$S^2\{\tilde{y}_i\} = \frac{1}{m-1}\sum_{u=1}^{m}(\tilde{y}_{iu} - \bar{y}_i)^2 ,$$

де $\bar{y}_i = \frac{1}{m}\sum_{u=1}^{m} y_{iu}$ – середнє значення для $i$-ї точки; $u$ – поточне спостереження в $i$-й точці.

Статистичну значущість у цьому разі визначають із використанням критерію Стьюдента. Висувається нуль-гіпотеза стосовно значення, що перевіряється, і обчислюється значення

$$t_{j\text{p}} = \frac{|\hat{\alpha}_j|}{S\{\hat{\alpha}_j\}},$$

яке порівнюють з критичним значенням для числа степенів вільності $N(m\text{-}1)$ і рівня статистичної значущості $\alpha$.

яке порівнюють із критичним значенням для кількості степенів вільності $N(m\text{-}1)$ і рівня статистичної значущості $\alpha$.

Деякі відмінності будуть і в перевірці адекватності моделі для цього випадку. Тут для визначення $F_{\text{кр}}$ необхідно враховувати, що кількість степенів вільності для знаменника $N(m\text{-}1)$, а при обчисленні $S_{\text{ад}}^2$ у виразі (4.67) у чисельнику має враховуватися $m$, відмінне від одиниці.

Для адекватної моделі записують її вираз у природній (фізичній) системі координат, для чого $x_j$ замінюють співвідношеннями (4.61) і розв'язують рівняння відносно будь-якого $X$.

Для розглядуваного числового прикладу переходимо до нормальної системи координат за середнього арифметичного значення вхідної величини $\overline{X} = 8$ та $d = 3$. За співвідношеннями (4.61) знаходимо:

$$P_1(x) = 1 \cdot \frac{X-8}{3},$$

$$P_2(x) = 1 \cdot \left[ \left( \frac{X-8}{3} \right)^2 - \frac{5^2-1}{12} \right],$$

$$P_3(x) = 1 \cdot \left[ \left( \frac{X-8}{3} \right)^2 - \left( \frac{X-8}{3} \right) \cdot \left( \frac{3 \cdot 5^2 - 7}{20} \right) \right].$$

Підставивши ці значення у модельне рівняння (4.66) отримаємо:

$$y(x) = 6,0 + 1,02 \cdot \frac{X-8}{3} - 0,47 \cdot \left[ \left( \frac{X-8}{3} \right)^2 - \frac{5^2-1}{12} \right] +$$

$$+1,02 \cdot \frac{5}{6} \cdot \left[ \left( \frac{X-8}{3} \right)^2 - \left( \frac{X-8}{3} \right) \cdot \left( \frac{3 \cdot 5^2 - 7}{20} \right) \right].$$

Спростивши вираз, отримаємо регресійну модель, яка відповідає експериментальним даним:

$$y(x) = -7,47 + 5,99 \cdot x - 0,81 \cdot x^2 + 0,03 \cdot x^3.$$

***Ддвофакторний експеримент.*** Даний експеримент дозволяє досліджувати поверхню результатів вимірювань за широкого діапазону зміни аргументу.

Розглянемо побудову поверхні відгуку на прикладі варійованих незалежних змінних $X_1$ та $X_2$ . Фактор $X_1$ змінюється на двох рівнях $X_{11} = 1$ та $X_{12} = 5$ з кроком $d_1 = 4$. Фактор $X_2$ в даному

прикладі набуває п’ять рівновіддалених значень із кроком $d_2 = 3$, а саме $X_{21} = 2$; $X_{22} = 5$; $X_{23} = 8$; $X_{24} = 11$; $X_{25} = 14$.

На рис. 4.15 наведена поверхня відгуку для двофакторного експерименту

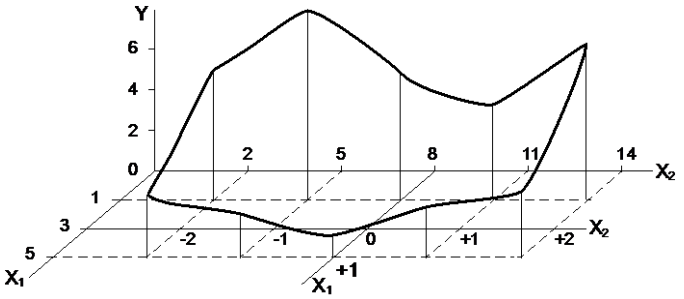



Рис. 4.15. Поверхня відгуку для двофакторного експерименту

Як і раніше, для спрощення процесу аналізу та оброблення експериментальних даних, необхідно перейти до кодованої системи координат:

$$X_1 = \frac{X_1 - \overline{X_1}}{d_1 / 2} = \frac{X_1 - 3}{2}; \quad X_2 = \frac{X_2 - \overline{X_2}}{d_2} = \frac{X_1 - 8}{3}. \qquad (4.68)$$

Оскільки, перший фактор варіюється на двох рівнях та, за умови що $\lambda_1 = 1$, отримаємо поліном $P_{11} = P_1(x_1) = x_1$.

Другий фактор варіюється на п’яти рівнях, відповідно у модельне рівняння можуть входити поліноми від першого до четвертого степенів, а саме:

$$P_1(x_2) = P_{12} = x_2;$$
$$P_2(x_2) = P_{22} = x_2^2 - 2;$$
$$P_2(x_3) = P_{32} = \frac{5}{6}x_2^2 - \frac{17}{6}x_2;$$
$$P_2(x_4) = P_{42} = \frac{35}{12}x_2^4 - \frac{155}{12}x_2^2 + 6.$$

Для двофакторного експерименту план повинен містити $N$ вузлових точок, тобто для розглядуваного випадку $N = N_1 \cdot N_2 = 10$ та містити не більше 10 невідомих елементів.

Модельне рівняння, з загальному випадку, окрім перерахованих елементів, також, може містити і ефекти взаємодії, що визначаються співвідношенням $P_1(x_1) \cdot P_s(x_2)$, де $s = 1 \ldots 4$.

У такий спосіб, модель може бути представлена рівнянням:

$$\begin{aligned}\tilde{y}(x_1, x_2) = \alpha_0 \cdot P_0 + \alpha_{11} \cdot P_1(x_1) + \alpha_{12} \cdot P_1(x_2) + \\ + \alpha_{22} \cdot P_2(x_2) + \alpha_{32} \cdot P_3(x_2) + \alpha_{42} \cdot P_4(x_2) + \\ + \alpha^*_{11} \cdot P_1(x_1) \cdot P_1(x_2) + + \alpha^*_{12} \cdot P_1(x_1) \cdot P_2(x_2) + \\ + \alpha^*_{13} \cdot P_1(x_1) \cdot P_3(x_2) + \alpha^*_{14} \cdot P_1(x_1) \cdot P_4(x_2),\end{aligned} \tag{4.69}$$

де $\alpha^*_{1s}$ показує вплив ефекту взаємодії $P_1(x_1)$ та $P_s(x_2)$ за $s = 1 \ldots 4$.

Складемо структурну матрицю розміром 10×10 для цього експерименту, яка подана у табл. 4.8. У цій таблиці 3-й та 4-й стовпці є матрицею плану двофакторного експерименту, а з 5-го по 7-й – взяті із таблиці 4.7 для $N = 5$.

Побудовані у такий спосіб вектор-стовпці мають властивості симетрії та ортогональності, що дозволяє використовувати МНК, згідно з яким усі коефіцієнти моделі визначаються шляхом послідовного множення елементів відповідного вектор-стовпця на елементи вектор-стовпця вихідної величини $\tilde{y}$ з подальшим їх підсумовуванням та діленням на суму квадратів елементів цього вектор-стовпця.

При цьому, випадкові величини $\tilde{y}_i$ мають бути статистично незалежними, мати одну і ту ж дисперсію $\sigma^2_y$ у всіх точках плану $i = \overline{1, N}$ та відповідати нормальному закону розподілу.

Таблиця 4.8. Структурна матриця для двофакторного експерименту

| $i$ | $P_{0i}$ | $P_{11i} = x_{1i}$ | $P_{12i} = x_{2i}$ | $P_{22i} = P_2(x_{2i})$ | $P_{32i} = P_3(x_{2i})$ | $P_{42i} = P_4(x_{2i})$ | $P_{11i} \cdot P_{12i}$ | $P_{11i} \cdot P_{22i}$ | $P_{11i} \cdot P_{32i}$ | $P_{1i} \cdot P_{42i}$ | $y_i$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
| 1 | +1 | -1 | -2 | +2 | -1 | +1 | +2 | -2 | +1 | -1 | 7 |
| 2 | +1 | -1 | -1 | -1 | +2 | -4 | +1 | +1 | -2 | +4 | 10 |
| 3 | +1 | -1 | 0 | -2 | 0 | +6 | 0 | +2 | 0 | -6 | 6 |
| 4 | +1 | -1 | +1 | -1 | -2 | -4 | -1 | +1 | +2 | +4 | 5 |
| 5 | +1 | -1 | +2 | +2 | +1 | +1 | -2 | -2 | -1 | -1 | 8 |
| | | | | | | | | | | | |
| 6 | +1 | +1 | -2 | +2 | -1 | +1 | -2 | +2 | -1 | +1 | 3 |
| 7 | +1 | +1 | -1 | -1 | +2 | -4 | -1 | -1 | +2 | -4 | 2 |
| 8 | +1 | +1 | 0 | -2 | 0 | +6 | 0 | -2 | 0 | +6 | 1 |
| 9 | +1 | +1 | +1 | -1 | -2 | -4 | +1 | -1 | -2 | -4 | 3 |
| 10 | +1 | +1 | +2 | +2 | +1 | +1 | +2 | +2 | +1 | +1 | 4 |
| | | | | | | | | | | | |
| $\Sigma$ | 10 | 10 | 20 | 28 | 20 | 140 | 20 | 28 | 20 | 140 | |
| $\lambda_i$ | 4.9 | -2.3 | 0 | 0.36 | 0.5 | -0.11 | 0.3 | 0.14 | 0.6 | 0.011 | |
| | 0.063 | 0.063 | | 0.045 | 0.045 | 0.017 | 0.046 | 0.038 | 0.045 | 0.017 | |

**Запитання для самоперевірки**

1. Розкрийте зміст поняття апроксимації. Яка апроксимація називається лінійною; квадратичною?
2. Яка різниця між інтерполяцією і екстраполяцією?
3. Для чого використовується сплайнова функція?
4. Розкрийте поняття апроксимуючого полінома.
5. У чому полягає суть умови ортогональності?
6. Який експеримент називається пасивним?
7. Які поліноми називаються ортогональними поліномами Чебишова?
8. Що таке кодування і для чого воно виконується?
9. Як утворюються кодовані шкали для парної та непарної кількості відліків?
10. Як обчислюють поліноми Чебишова?
11. Від чого залежать значення ортогональних поліномів?
12. Наведіть матричне рівняння для обчислення оцінок коефіцієнтів моделі.
13. Як перевірити коефіцієнти моделі на статистичну значущість та модель на адекватність у випадку з відомою дисперсією?
14. Як перевірити коефіцієнти моделі на статистичну значущість та модель на адекватність у випадку з невідомою дисперсією?
15. Як перевести модель із кодованої до природної системи координат?

## 4.9. РЕГРЕСІЙНІ МОДЕЛІ З ВИКОРИСТАННЯМ ТРИГОНОМЕТРИЧНИХ ПОЛІНОМІВ

Коли досліджувана залежність змінюється періодично, то для побудови математичної моделі доцільно використовувати тригонометричну регресію, що базується на ряді Фур'є.

У випадку графічного подання результатів експериментальних досліджень або при дослідженні аналогового сигналу доцільно неперервну залежність зображати у вигляді окремих значень, узятих

через рівні проміжки часу (рис. 4.16). Завдання полягає у визначенні аналітичної залежності у вигляді полінома на підставі наявних даних $\tilde{y}(t_i)$.

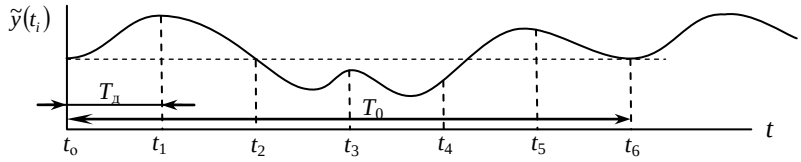



Рис. 4.16. Представлення неперервної залежності дискретними значеннями

Подамо неперервну величину у дискретній формі. На рис. 4.14 $T_0$ – період, що відповідає частоті $f_0 = \dfrac{1}{T_0}$ першої гармоніки залежності; $T_д = \dfrac{T_0}{N_i}$ – період дискретизації, $N_i$ – кількість відліків, $i = \overline{0, N-1}$.

Періодичну залежність можна подати рядом Фур'є:

$$y(t) = \frac{a_0}{2} + \sum_{k=1}^{n} a_k \cos k\omega t + \sum_{k=1}^{n} b_k \sin k\omega t , \qquad (4.69)$$

де $n$ – вища гармоніка апроксимуючого ряду.

*Необхідно, щоб кількість статистично значущих коефіцієнтів була меншою за кількість відліків, тобто* $2n+1 < N$.

Для знаходження оцінок коефіцієнтів використовується метод найменших квадратів, тобто

$$Q = \sum_{i=0}^{N-1} \left[ \hat{y}(t_i) - \tilde{y}(t_i) \right]^2 \to \min,$$

або

$$Q = \sum_{i=0}^{N-1}\left[\frac{\hat{a}_0}{2} + \sum_{k=1}^{n}\hat{a}_k \cos k\omega_0 t_i + \sum_{k=1}^{n}\hat{b}_k \sin k\omega_0 t_i - \tilde{y}(t_i)\right]^2 \to \min.$$

Перепишемо останнє співвідношення у вигляді

$$Q = \sum_{i=0}^{N-1}\left[\frac{\hat{a}_0}{2} + \hat{a}_1 \cos 1\cdot\omega_0 t_1 + \ldots + \hat{a}_n \cos n\omega_0 t_1 + \hat{b}_1 \sin 1\cdot\omega_0 t_1 + \ldots\right.$$
$$\ldots + \hat{b}_n \sin n\omega_0 t_1 + \hat{a}_1 \cos 1\omega_0 t_2 + \ldots + \hat{a}_n \cos n\omega_0 t_2 + \ldots + \hat{b}_1 \sin 1\omega_0 t_2 + \ldots +$$
$$\left. + \hat{b}_n \sin n\omega_0 t_2 + \ldots - \tilde{y}(t_i)\right]^2 \to \min.$$

Для того щоб знайти оцінки коефіцієнтів ряду Фур'є, необхідно взяти частинні похідні від оцінок цих коефіцієнтів і прирівняти їх до нуля:

$$\frac{\partial Q}{\partial \hat{a}_0} = \sum_{i=0}^{N-1} 2\left[\frac{\hat{a}_0}{2} + \hat{a}_1 \cos 1\cdot\omega_0 t_i + \ldots + \hat{a}_n \cos n\omega_0 t_i + \hat{b}_1 \sin 1\cdot\omega_0 t_i + \ldots\right.$$
$$\left. \ldots + \hat{b}_n \sin n\omega_0 t_i + \ldots - \tilde{y}(t_i)\right]\frac{1}{2}\cos 0\cdot\omega_0 t_i = 0.$$

Позначимо доданок, що міститься у квадратних дужках, через $[A]$, тоді

$$\begin{cases} \dfrac{\partial Q}{\partial \hat{a}_1} = \sum\limits_{i=0}^{N-1} 2[\hat{A}]\cdot\cos 1\cdot\omega_0 t_i = 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \dfrac{\partial Q}{\partial b_n} = \sum\limits_{i=0}^{N-1} 2[\hat{A}]\cdot\sin n\omega_0 t_i = 0, \end{cases},$$

де cos 0 = 1 – використовується для формалізації запису.

Запишемо систему $2n + 1$ нормальних рівнянь за відомим правилом (п.4.5):

$$\frac{\hat{a}_0}{2}\sum_{i=0}^{N-1}\cos^2 0\cdot\omega_0 t_i + \hat{a}_1\sum_{i=1}^{N-1}\cos 1\cdot\omega_0 t_i \cos 0\cdot\omega_0 t_i + \ldots +$$
$$+\hat{a}_n\sum_{i=0}^{N-1}\cos n\omega_0 t_i \cos 0\cdot\omega_0 t_i + \hat{b}_1\sum_{i=0}^{N-1}\sin 1\omega_0 t_i \cos 0\cdot\omega_0 t_i + \ldots +$$
$$+\hat{b}_n\sum_{i=0}^{N-1}\sin n\omega_0 t_i \cos 0\cdot w_0 t_i = \sum_{i=0}^{N-1}\tilde{y}_i \cos 0\cdot\omega_0 t_i;$$

$$\frac{\hat{a}_0}{2}\sum_{i=0}^{N-1}\cos 0\cdot\omega_0 t_i\cos 1\cdot\omega_0 t_i+\hat{a}_1\sum_{i=0}^{N-1}\cos^2 1\cdot\omega_0 t_i+\ldots+$$

$$+\hat{a}_n\sum_{i=0}^{N-1}\cos n\omega_0 t_i\cos 1\cdot\omega_0 t_i+\hat{b}_1\sum_{i=0}^{N-1}\sin 1\omega_0 t_i\cos 1\cdot\omega_0 t_i+\ldots+$$

$$+\hat{b}_n\sum_{i=0}^{N-1}\sin n\cdot\omega_0 t_i\cos 1\cdot\omega_0 t_i=\sum_{i=0}^{N-1}\tilde{y}_i\cos 1\omega_0 t_i;$$

$$\frac{\hat{a}_0}{2}\sum_{i=0}^{N-1}\cos 0\cdot\omega_0 t_i\cos n\omega_0 t_i+\ldots+\hat{a}_n\sum_{i=0}^{N-1}\cos^2 n\omega_0 t_i+$$

$$+\hat{b}_1\sum_{i=0}^{N-1}\sin 1\cdot\omega_0 t_i\cos n\omega_0 t_i+\ldots+\hat{b}_n\sum_{i=0}^{N-1}\sin n\omega_0 t_i\cos n\omega_0 t_i=\sum_{i=0}^{N-1}\tilde{y}_i\cos n\omega_0 t_i;$$

$$\frac{\hat{a}_0}{2}\sum_{i=0}^{N-1}\cos 0\cdot\omega_0 t_i\sin 1\cdot\omega_0 t_i+\ldots+\hat{a}_n\sum_{i=0}^{N-1}\cos n\omega_0 t_i\sin 1\cdot\omega_0 t_i+$$

$$+\hat{b}_1\sum_{i=0}^{N-1}\sin^2 1\cdot\omega_0 t_i+\ldots+\hat{b}_n\sum_{i=0}^{N-1}\sin n\omega_0 t_i\sin 1\cdot\omega_0 t_i=\sum_{i=0}^{N-1}\tilde{y}_i\sin 1\cdot\omega_0 t_1;$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\frac{\hat{a}_0}{2}\sum_{i=0}^{N-1}\cos 0\cdot\omega_0 t_i\sin n\omega_0 t_i+\ldots+\hat{a}_n\sum_{i=0}^{N-1}\cos n\omega_0 t_i\sin n\omega_0 t_i+\ldots+$$

$$+\hat{b}_1\sum_{i=0}^{N-1}\sin 1\cdot\omega_0 t_i\sin n\omega_0 t_i\ldots+\hat{b}_n\sum_{i=0}^{N-1}\sin^2 n\omega_0 t_i=\sum_{i=0}^{N-1}\tilde{y}_i\sin n\omega_0 t_i.$$

Дістали систему рівнянь з $2n + 1$ невідомими коефіцієнтами. На головній діагоналі цієї системи під знаком суми маємо елементи, що містять $\sin^2$ або $\cos^2$.

Розглянемо інші елементи цієї системи:

– Відомо, що сума значень періодичної функції (синуса, косинуса), узятих на інтервалі часу, що дорівнює або кратний періоду, дорівнює *нулю*. Тому елементи, які містять суму добутків синусів і косинусів, також будуть дорівнювати нулю:

$$\sum_{i=0}^{N-1}\cos k\omega_0 t_i\sin j\omega_0 t_i=\frac{1}{2}\sum_{i=0}^{N-1}\left[\cos(k\pm j)\omega_0 t_i\pm\sin(k\pm j)\omega_0 t_i\right]=0;$$

$$j,k=\overline{1,n}\ ;\ j\neq k,$$

а також

$$\sum_{i=0}^{N-1}\cos k\omega_0 t_i\cos 0\cdot\omega_0 t_i=\sum_{i=0}^{N-1}\cos k\omega_0 t_i=0.$$

– Аналогічно маємо:

$$\sum_{i=0}^{N-1} \sin k\omega_0 t_i \cdot \sin j\omega_0 t_i = \frac{1}{2}\sum_{i=0}^{N-1}\left[\cos(k-j)\omega_0 t_i - \cos(k+j)\omega_0 t_i\right] = 0,$$

$$j,k = \overline{1,n}\ ;\ j \neq k.$$

Таким чином, усі елементи в лівій частині системи нормальних рівнянь, крім розташованих на головній діагоналі, дорівнюватимуть нулю, тобто розглянута система має *властивість ортогональності*, а отже, дає змогу *незалежно* визначати оцінки коефіцієнтів $\hat{a}_0$, $\hat{a}_k$, $\hat{b}_k$, де $k = \overline{1,n}$.

Для визначення оцінок коефіцієнтів можна скористатися виразом (4.63). Для тригонометричних поліномів цей вираз перепишеться в такий спосіб:

$$\frac{\hat{a}_0}{2} = \frac{\sum_{i=0}^{N-1} \tilde{y}(t_i)\cos 0 \cdot \omega_0 t_i}{\sum_{i=0}^{N-1} \cos^2 0 \cdot \omega_0 t_i}; \tag{4.70}$$

$$\hat{a}_k = \frac{\sum_{i=0}^{N-1} \tilde{y}(t_i)\cos k\omega_0 t_i}{\sum_{i=0}^{N-1} \cos^2 k\omega_0 t_i}; \tag{4.71}$$

$$\hat{b}_k = \frac{\sum_{i=0}^{N-1} \tilde{y}(t_i)\sin k\omega_0 t_i}{\sum_{i=0}^{N-1} \sin^2 k\omega_0 t_i}. \tag{4.72}$$

Розглянемо вираз (4.70). У знаменнику цього виразу маємо $\sum_{i=1}^{N-1} \cos^2 0 \cdot \omega_0 t_i = N,$ оскільки $\cos^2 0 \cdot \omega_0 t_i = 1,$ а в чисельнику $\cos 0 \cdot \omega_0 t_i = 1,$ звідки випливає, що вираз (4.70) набирає вигляду:

$$\frac{\hat{a}_0}{2} = \frac{\sum_{i=0}^{N-1} \tilde{y}(t_i)}{N},$$

або

$$\hat{a}_0 = \frac{2}{N}\sum_{i=0}^{N-1}\tilde{y}(t_i), \tag{4.73}$$

де $\hat{a}_0$ – середнє значення тригонометричного ряду.

Розглянемо вираз (4.71). Знаменник цього виразу:

$$\sum_{i=0}^{N-1}\cos^2 k\omega_0 t_i = \frac{1}{2}\sum_{i=0}^{N-1}\left(1+\cos 2k\omega_0 t_i\right) = \frac{N}{2} + \frac{1}{2}\sum_{i=0}^{N-1}\cos 2k\omega_0 t_i = \frac{N}{2},$$

оскільки другий доданок дорівнює нулю (сума симетричних дискретних відліків за період).

Тоді вираз (4.71) набирає вигляду:

$$\hat{a}_k = \frac{2}{N}\sum_{i=0}^{N-1}\tilde{y}(t_i)\cos k\omega_0 t_i. \tag{4.74}$$

За аналогією з виразом (4.71), враховуючи, що $\sin^2\alpha = 1-\cos^2\alpha,$ знаходимо знаменник виразу (4.72):

$$\sum_{i=0}^{N-1}\sin^2 k\omega_0 t_i = \frac{N}{2}.$$

Отже, вираз (4.72) набирає вигляду:

$$\hat{b}_k = \frac{2}{N}\sum_{i=0}^{N-1}\tilde{y}(t_i)\sin k\omega_0 t_i. \tag{4.75}$$

Знайдені оцінки коефіцієнтів відрізняються від значень їх математичних сподівань. При цьому похибки знайдених коефіцієнтів такі:

$$\alpha_0 = \hat{a}_0 - \grave{a}_0 = \frac{2}{N}\sum_{i=0}^{N-1}\delta(t_i);$$

$$\alpha_k = \hat{a}_k - a_k = \frac{2}{N}\sum_{i=0}^{N-1}\delta(t_i)\cos\frac{2\pi k_i}{N};$$

$$\beta_k = \hat{b}_k - b_k = \frac{2}{N}\sum_{i=0}^{N-1}\delta(t_i)\sin\frac{2\pi k_i}{N}.$$

Бачимо, що кожна з них є лінійною функцією від похибки вимірювання δ, яка за означенням має нормальний закон розподілу з $M\,\delta = 0$ та дисперсією $\sigma^2$, а отже, похибки коефіцієнтів мають також нормальний розподіл. Це необхідно враховувати при подальшій статистичній обробці.

**Перевірка знайдених оцінок коефіцієнтів на статистичну значущість**. У цьому випадку *статистичну значущість* коефіцієнтів визначають за критерієм Стьюдента.

Висувається гіпотеза $H_0$: $a_k = 0$, тоді розрахункове значення коефіцієнта Стьюдента, наприклад для коефіцієнта $a_k$, маємо

$$t_{k\,\mathrm{p}} = \frac{|\hat{a}_k|}{S\{\hat{a}_k\}} \tag{4.76}$$

Аналогічно визначається $t_{\mathrm{p}k}$ і для інших коефіцієнтів.

Для визначення середньоквадратичного відхилення $S\{\hat{a}_k\}$ та інших коефіцієнтів необхідно знайти дисперсії оцінок коефіцієнтів $\hat{a}_0$, $\hat{a}_k$, $\hat{b}_k$, скориставшись виразами (4.70) …(4.75).

Оскільки

$$\tilde{y}(t_i)_0 = y(t_i) + \delta(t_i),$$

то дисперсія оцінки $\hat{a}_0$:

$$D\{\hat{a}_0\} = D\left\{\frac{2}{N}\sum_{i=0}^{N-1}\tilde{y}(t_i)\right\},$$

або

$$D\{\hat{a}_0\} = \frac{4}{N^2}\cdot N\sigma^2 = \frac{4}{N}\sigma^2, \tag{4.77}$$

де $\sigma^2 = D\{\tilde{y}(t_i)\}$.

Дисперсія оцінок $\hat{a}_k$:

$$D\{\hat{a}_k\} = D\left\{\frac{2}{N}\sum_{i=0}^{N-1}\tilde{y}(t_i)cosk\omega_0 t_i\right\} = \frac{4}{N^2}\sigma^2\sum_{i=0}^{N-1}cos^2 k\omega_0 t_i$$

де $\sum_{i=0}^{N-1}\cos^2 k\omega_0 t_i = \frac{N}{2}$.

Тоді

$$D\{\hat{a}_k\} = \frac{2}{N}\sigma^2. \tag{4.78}$$

Узявши до уваги, що

$$\sum_{i=0}^{N-1} \sin^2 k\omega_0 t_i = \frac{N}{2},$$

знайдемо дисперсію оцінок $\hat{b}_k$ за формулою:

$$D\left\{\hat{b}_k\right\} = \frac{2}{N}\sigma^2. \tag{4.79}$$

На підставі наявних експериментальних даних можна знайти незсунену оцінку $\sigma^2$. Оскільки досліди в кожній точці проводять тільки по одному разу, то традиційний шлях (обчислювати середнє для кожної точки $t_i$, а потім знаходити усереднену дисперсію за всіма $N$ точками) неприйнятний. Незсунену оцінку дисперсії можна визначити за формулою:

$$S^2 = \frac{1}{N-(2n+1)} \sum_{i=0}^{N-1} \left[\tilde{y}(t_i) - \hat{y}(t_i)\right]^2. \tag{4.80}$$

Кількість степенів вільності у виразі (4.80) дорівнює $N - (2n + 1)$, бо для обчислення $S^2$ необхідно визначити $(2n + 1)$ оцінок коефіцієнтів рівняння тригонометричної регресії.

На підставі (4.80) вирази (4.77), (4.78) і (4.79) можна записати у вигляді:

$$S^2\left\{\hat{a}_0\right\} = \frac{4}{N}S^2; \tag{4.81}$$

$$S^2\left\{\hat{a}_k\right\} = \frac{2}{N}S^2,\ k = \overline{1,n}; \tag{4.82}$$

$$S^2\left\{\hat{b}_k\right\} = \frac{2}{N}S^2,\ k = \overline{1,n}. \tag{4.83}$$

Дисперсії оцінок коефіцієнтів, обчислені за формулами (4.81), (4.82) і (4.83), застосовують для розрахунку коефіцієнтів Стьюдента за формулою (4.76).

Обчислені значення коефіцієнтів Стьюдента порівнюють із критичним значенням $t_{кр}$, яке буде однаковим для всіх

розрахункових значень, бо кількість степенів вільності $N - (2n + 1)$ одна й та сама.

Якщо $t_{pk} < t_{кр}$, приймають нуль-гіпотезу, тобто вважають оцінку коефіцієнта статистично незначущою.

**Перевірка моделі на відповідність експериментальним даним.**

У випадку, коли *досліди в кожній точці проведено по одному разу*, перевірку отриманої моделі на адекватність об'єкта не можна здійснити. За наявності одного результату в кожній дослідній точці можна оцінити *середню точність отриманого з експериментальних даних наближення* тригонометричного полінома до полінома моделі.

Розглянемо величину:

$$(2n+1)\Sigma^2 = \sum_{i=0}^{N} [\hat{y}(t_0) - y(t_i)]^2,$$

яка несе інформацію про відхилення регресійної тригонометричної моделі, коефіцієнти якої обчислено за експериментальними даними з $(2n + 1)$ рівнянь, від моделі апроксимації, що припускається.

Згідно з цим можна записати вираз для оцінки дисперсії:

$$\Sigma^2 = \frac{\sum_{i=0}^{N} [\hat{y}(t_0) - y(t_i)]^2}{(2n+1)},$$

яка характеризує середньоквадратичне відхилення і так само, як і $S^2$, залежить від впливу похибки вимірювання.

Таким чином, для $\sigma^2$ є дві незалежні оцінки $S^2$ з кількістю степенів вільності $f_2 = (N - 2n - 1)$ та $\Sigma^2$ із кількістю степенів вільності $f_1 = (2n + 1)$. При цьому друга з них більша, бо залежить від похибок вимірювання та неадекватності моделі.

Відношення $$F = \frac{\Sigma^2}{S^2}$$

має розподіл Фішера з $f_1$ та $f_2$ степенями вільності.

Добуток $F_{кр}S$ дає шукану оцінку $\Sigma^2$ середньої точності наближення регресійної тригонометричної моделі до моделі апроксимації, що припускається. Значення $F_{кр}$ знаходять за таблицями $F$-розподілу згідно з вибраною надійною ймовірністю для $f_1$ та $f_2$.

Якщо є можливість проводити дослідження за *m періодів*, то можна перевірити *адекватність* моделі.

Для перевірки адекватності моделі обчислюють коефіцієнт Фішера за формулою (4.47), в якій для цього випадку:

$$S_{ад}^{\ 2} = \frac{m}{N-l}\sum_{i=0}^{N-1}(\bar{y}_i - \hat{y}_i)^2, \qquad (4.84)$$

де $l$ – кількість статистично значущих коефіцієнтів.

Дисперсію відтворюваності, яка характеризує усереднену точність одного вимірювання, можна обчислити за формулою:

$$S_{в}^{\ 2} = \frac{1}{N}\sum_{i=0}^{N-1} S^2\{\tilde{y}_i\},$$

де $S^2\{\tilde{y}_i\}$ – оцінка дисперсії для $i$-ої точки, яку можна визначити як

$$S^2\{\tilde{y}_i\} = \frac{1}{m}\sum_{u=1}^{m}(\tilde{y}_{iu} - \bar{y}_i)^2 .$$

У цьому випадку будуть відмінності і при перевірці коефіцієнтів на статистичну значущість. Для розрахунку $t_{pk}$ замість $S$ необхідно використовувати $S_в$, тоді згідно з (4.76) і (4.82):

$$t_{k\,р} = \frac{|\hat{a}_k|}{S_в}\sqrt{\frac{N}{2}}.$$

Аналогічно буде й для інших коефіцієнтів.

Розрахункове значення коефіцієнта Фішера порівнюють із критичним і приймають відповідне рішення (п. 4.5, 4.8).

Можна обмежитися дослідженням залежності *на одному періоді*, однак при цьому необхідно знати оцінку $S_в^2$, яку було знайдено на підставі додаткового експерименту.

У цьому випадку $S_{\text{в}}^2$ буде оцінкою СКВ приладу:

$$S_{\text{в}}^2 = \frac{1}{m-1}\sum_{p=1}^{m}\left(\tilde{y}_p - \bar{\tilde{y}}\right)^2, p = \overline{1,m},$$

де $\tilde{y}_p$ – результат $p$-го додаткового вимірювання при дослідженні властивостей приладу, а

$$\bar{\tilde{y}} = \frac{1}{m}\sum_{p=1}^{m}\tilde{y}_p .$$

При обчисленні $S_{\text{ад}}^2$ у виразі (4.84) $m$ = 1, оскільки в кожній точці при дослідженні залежності проводять тільки по одному вимірюванню, а замість середнього значення $\bar{y}_i$ беруть $\tilde{y}_i$.

При обчисленні $S_{\text{в}}^2$ кількість степенів вільності дорівнюватиме $m$-1.

## Запитання для самоперевірки

1. В якому випадку доцільно застосовувати тригонометричну регресію?
2. Яким методом визначають коефіцієнти ряду Фур'є?
3. Запишіть вирази для отримання оцінок коефіцієнтів тригонометричної регресії.
4. Як перевірити оцінки коефіцієнтів тригонометричної регресії на статистичну значущість при відомому СКВ?
5. Як перевірити оцінки коефіцієнтів тригонометричної регресії на статистичну значущість при невідомому СКВ?
6. За якої умови можна перевірити отриману тригонометричну модель на адекватність?
7. Чи можна при побудові тригонометричної регресії проводити вимірювання для одного періоду? Як перевіряється відповідність моделі в цьому випадку?

# РОЗДІЛ 5

# СТАТИСТИЧНІ ХАРАКТЕРИСТИКИ ВИПАДКОВИХ ПРОЦЕСІВ

Випадковий процес установлює взаємозв'язок між випадковою величиною $X$ та невипадковим параметром, яким може бути час або простір.

Випадкові процеси характеризують зміну в часі випадкових величин, які в кожний момент $t$ мають певні характеристики.

При вивченні випадкових процесів розглядають окремі його *реалізації* (рис. 5.1).

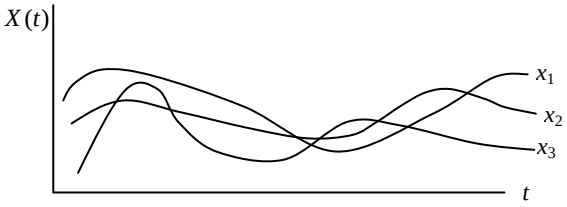



Рис. 5.1. Реалізації випадкових процесів

У загальному випадку, для знаходження ймовірнісних (статистичних) характеристик випадкового процесу необхідно розглядати сукупність реалізацій, яка називається *ансамблем реалізацій.*

**Основні характеристики випадкових процесів.**

***Математичним сподіванням випадкового процесу*** називається невипадкова функція

$$M[X(t)] = m_x(t),$$

значення якої відповідає математичному сподіванню $M[X(t_i)]$ випадкової величини $X(t_i)$, що відповідає значенню параметра в будь-який момент часу $t_i$.

На рис. 5.2 наведено три реалізації $x_1$, $x_2$, $x_3$ випадкового процесу $X(t)$. Математичне сподівання випадкового процесу $m_x(t)$ є *геометричним місцем* центрів розподілів випадкової величини $x(t_i)$ для моменту часу $t_i$.

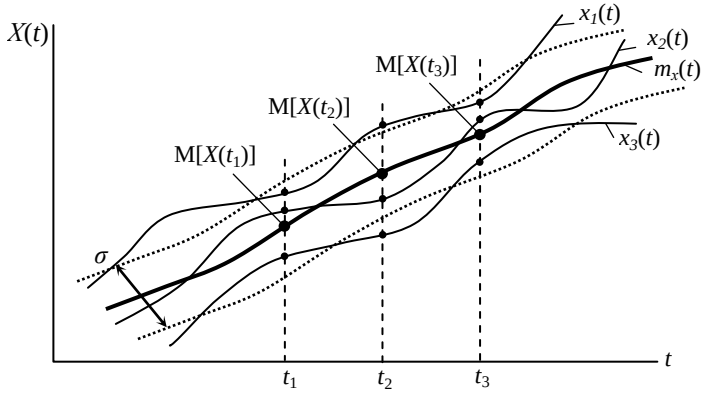



Рис. 5.2. Реалізації випадкового процесу $X(t)$

Математичне сподівання випадкового процесу має властивості, аналогічні властивостям математичного сподівання випадкової величини.

***Дисперсія випадкового процесу***

$$D[X(t)] = D_X(t)$$

є невипадковою функцією часу, значення якої являють собою дисперсії $D[X(t_i)]$ випадкової величини $X(t_i)$, що відповідає цьому значенню параметра для моменту часу $t_i$.

Додатний корінь із дисперсії є *середньоквадратичним відхиленням* (*СКВ*) *випадкового процесу*.

Дисперсія та СКВ характеризують розсіювання можливих реалізацій щодо середнього випадкового процесу. На рис. 5.2 штриховою лінією обмежено зону, в якій найбільш імовірно міститимуться графіки реалізацій відповідних процесів.

***Функція кореляції***. Розглянемо реалізації двох випадкових процесів $X_1(t)$ і $X_2(t)$, зображених на рис. 5.3.

Як бачимо, за характером зміни реалізацій дані процеси різняться, хоча може виявитися, що математичні сподівання й дисперсії цих процесів будуть однакові:

$$m_{x1}(t) = m_{x2}(t),\ D_{x1}(t) = D_{x2}(t).$$

Отже, на відміну від випадкової величини, двох моментів $m_x(t)$ і $D_x(t)$ недостатньо, для того, щоб дістати уявлення про випадковий процес. Для повної характеристики процесів необхідно ввести показник, що дає змогу оцінити *зв'язок* між значеннями випадкового процесу залежно від параметра $t$.

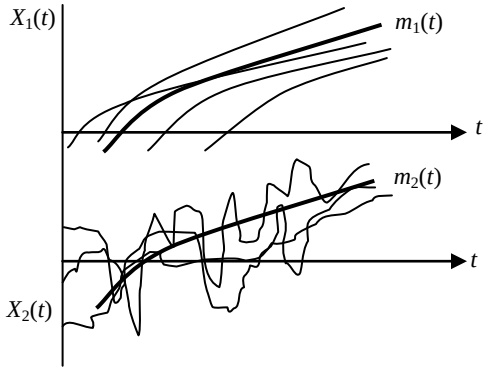



Рис. 5.3. Випадкові процеси, що мають однакові математичні сподівання й дисперсії

Для *характеристики зв'язку* між двома значеннями випадкового процесу, отриманими в моменти часу $t_1$ і $t_2$, використовується момент другого порядку, що називається ***функцією кореляції***

$$K_X(t_1,t_2) = cov\{[x(t_1) - m_x(t)][x(t_2) - m_x(t)]\},$$

Для *характеристики зв'язку* між двома значеннями випадкового процесу, отриманими в моменти часу $t_1$ і $t_2$, використовується момент другого порядку, що називається ***функцією кореляції***

$$K_X(t_1,t_2)=\int_{-\infty-\infty}^{+\infty+\infty} x_1^{\circ}(t_1)x_2^{\circ}(t_2)W(x_1,x_2,t_1,t_2)dx_1dx_2, \qquad (5.1)$$

де $x^{\circ}=x(t)-m_x(t)$ – центрована величина; $W(x_1,x_2,t_1,t_2)$ – щільність імовірностей значень випадкового процесу.

Залежно від зміни в часі випадкові процеси поділяються на *стаціонарні* та *нестаціонарні*.

Параметри ***стаціонарних*** процесів не залежать від часового інтервалу, в якому вони розглядаються. Середнє значення та дисперсія стаціонарних сигналів *постійні* й не залежать від часу, тобто

$$W(x_1, t_1, x_2, t_2) = W(x_3, t_3, x_4, t_4).$$

Якщо випадковий процес стаціонарний, то значення кореляційної функції залежатиме тільки від $\tau = t_2 - t_1$, тобто часового інтервалу між відповідними значеннями випадкового процесу й не залежатиме від абсолютних значень $t_1$ і $t_2$. Таким чином, щільність стохастичного зв'язку між значеннями випадкового процесу в моменти $t_1$ і $t_2$ та в моменти $t_3$ і $t_4$ буде однаковою. У цьому випадку функцію (5.1) називають ***автокореляційною функцією*** і позначають $K_X(\tau)$.

Більшість стаціонарних випадкових процесів мають властивість ***ергодичності***, коли за однією реалізацією випадкового процесу можна визначити його ймовірнісні характеристики. При цьому передбачається, що час дослідження практично великий, а теоретично нескінченний. Випадковий процес задовольняє умови ергодичності, якщо його статистичні характеристики, отримані усередненням за *ансамблем реалізацій*, можуть бути отримані часовим усередненням однієї реалізації за великий період часу.

*Ергодичні процеси є найбільш зручними для дослідження, бо дають змогу безпосередньо застосовувати апарат статистичної обробки до випадкових процесів.*

Для стаціонарного ергодичного процесу вираз (5.1) можна подати як

$$K_X(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{+T} x^{\circ}(t) x^{\circ}(t+\tau) W(x,t)\, dx. \qquad (5.2)$$

Автокореляційна функція дає змогу встановити взаємозв'язок між значеннями процесу в даний момент часу та значеннями в моменти часу, зсуненими на τ.

Оскільки для знаходження автокореляційної функції суттєвим є лише зсув на час τ між вибірковими значеннями $x^{\circ}(t)$ *і* $x^{\circ}(t-\tau)$, то автокореляційна функція є *симетричною*, тобто

$$K_X(-\tau) = K_X(\tau).$$

Автокореляційна функція дає можливість ідентифікувати процес, тобто встановити, яким є процес: гармонійним, випадковим чи їхньою сумою. Наприклад, автокореляційна функція гармонійного процесу є також гармонійна функція.

***Взаємно кореляційна функція*** оцінює зв'язок між значеннями двох випадкових процесів (рис. 5.4).

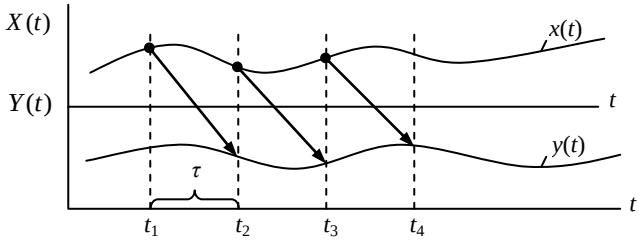



Рис. 5.4. Взаємно кореляційна функція

Для нестаціонарних процесів *взаємно кореляційна функція*

$$K_{XY}(t_1, t_2),$$

для стаціонарних процесів

$$K_{XY}(\tau).$$

Вирази (5.1) і (5.2) справджуються і для взаємно кореляційної функції із заміною $X(t)$ на $Y(t)$ або $X(t+\tau)$ на $Y(t+\tau)$, тобто при обчисленні взаємно кореляційної функції перемножують миттєві значення $x^{\circ}(t)$ та $y^{\circ}(t+\tau)$.

*Взаємно кореляційна функція застосовується для визначення взаємних часових характеристик і параметрів досліджуваних залежностей, наприклад для часу затримки, для виявлення сигналу на фоні шуму тощо.*

На практиці часто використовують ***нормовану кореляційну функцію***. Зважаючи на те, що

$$\rho_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{cov(X,Y)}{\sqrt{\sigma_X^2} \cdot \sqrt{\sigma_Y^2}},$$

для *нестаціонарного процесу* нормована кореляційна функція визначатиметься за формулою:

$$\rho_X(t_1,t_2) = \frac{K_X(t_1,t_2)}{\sqrt{K_X(t_1,t_1) \cdot K_X(t_2,t_2)}},$$

де $$K_X(t_i,t_i) = x^{\circ}(t_i)x^{\circ}(t_i) =$$

$$= [x(t_i) - m_x(t_i)][x(t_i) - m_x(t_i)] = [x(t_i) - m_x(t_i)]^2 = \sigma^2(t_i),$$

тобто відповідає дисперсії випадкового процесу, $i = 1, 2$.

Таким чином,

$$\rho_X(t_1,t_2) = \frac{K_X(t_1,t_2)}{\sigma_X(t_1) \cdot \sigma_X(t_2)}.$$

Аналогічно можна записати вираз і для нормованої взаємно кореляційної функції нестаціонарного процесу:

$$\rho_{XY}(t_1,t_2) = \frac{K_{XY}(t_1,t_2)}{\sigma_X(t_1) \cdot \sigma_Y(t_2)}.$$

Для *стаціонарного процесу*:

– автокореляційна функція:

$$\rho_X(\tau) = \frac{K_X(\tau)}{\sqrt{K_X(0)\cdot K_X(0)}} = \frac{K_X(\tau)}{\sigma_X^2}, \qquad (5.3)$$

де $K_X(0) = K_X(t_1 - t_1) =$

$$= x^{\circ}(t)x^{\circ}(t) = [x(t) - m_X(t)][x(t) - m_X(t)] = [x(t) - m_X(t)]^2 = D_X(t)\ .$$

*Для стаціонарного процесу математичне сподівання та дисперсія – сталі величини і не залежать від моменту відліку і.*

– взаємно кореляційна функція:

$$\rho_{XY}(\tau) = \frac{K_{XY}(\tau)}{\sqrt{K_X(0)\cdot K_Y(0)}} = \frac{K_{XY}(\tau)}{\sigma_X \sigma_Y}. \qquad (5.4)$$

**Знаходження оцінки кореляційної функції.**

***Послідовний метод***. При обробці даних мають скінченну множину значень реалізації ергодичного стаціонарного випадкового процесу. Для визначення автокореляційної функції $K_X$ криву випадкового процесу розбивають на $N$ рівних інтервалів (рис. 5.5). На рисунку $T_0 = t_i - t_{i-1}$ – період дискретизації; $i$ – номер вибірки; $\tau = kT_0$ – абсциса кореляційної функції; $k$ – відносна відстань між вибірками, $k = 0,..., n$; $n$ – кількість ординат кореляційної функції.

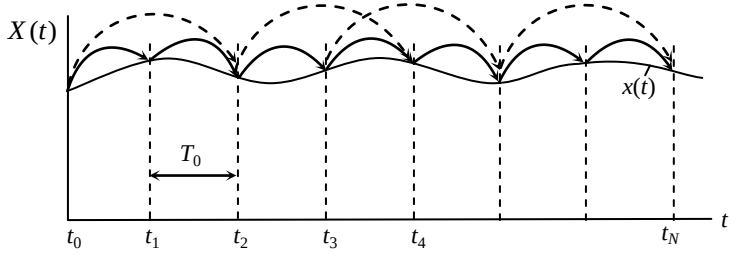



Рис. 5.5. Послідовний метод

Згідно з цим методом спочатку обчислюють ординату кореляційної функції $K_X(0)$, що відповідає нульовому зсуву (k = 0), при цьому кожне значення реалізації $x^\circ(iT_0)$, $i=\overline{1,N}$, множать самого на себе, потім беруть k = 1 і перемножують значення реалізації $\left[x^\circ(iT_0)\right]$ та $\left[x^\circ(iT_0+T_0)\right]$ і т. д. Операцію послідовного множення ілюструють на рис. 5.5 стрілки, що сполучають перемножуванні значення реалізацій $\left[x^\circ(iT_0)\right]$ та $\left[x^\circ(iT_0+kT_0)\right]$.

*При k = 0 обчислюється дисперсія досліджуваного сигналу, оскільки значення реалізації (ордината процесу) $x^\circ(iT_0)$ множиться само на себе.*

Теоретичний вираз для обчислення автокореляційної функції:

$$K_X(kT_0)=\lim_{N\to\infty}\frac{1}{N}\sum_{i=0}^{N}x^\circ(iT_0)x^\circ(iT_0+kT_0). \qquad (5.5)$$

За обмеженої кількості вибіркових значень межа підсумовування становить $N-k$ й оцінка автокореляційної функції подається у вигляді:

$$\hat{K}_X(kT_0)=\frac{1}{N-k}\sum_{i=0}^{N-k}x^\circ(iT_0)x^\circ(iT_0+kT_0). \qquad (5.6)$$

Аналогічно для взаємно кореляційної функції

$$\hat{K}_{XY}(kT_0)=\frac{1}{N-k}\sum_{i=0}^{N-k}x^\circ(iT_0)y^\circ(iT_0+kT_0). \qquad (5.7)$$

Розглянуті вирази (5.6) і (5.7) дають змогу обчислювати кореляційну функцію для *центрованих випадкових процесів*. Якщо центрування здійснюється шляхом формування різниці $x(t)-\hat{m}_X$, де $\hat{m}_X$ – оцінка математичного сподівання процесу $X(t)$, то

$$\hat{K}_X(kT_0)=\frac{1}{N-k}\sum_{i=0}^{N-k}\left[x(iT_0)-\hat{m}_X\right]\left[x(iT_0+kT_0)-\hat{m}_X\right].$$

Перетворимо останній вираз:

$$\hat{K}_X(kT_0)=\frac{1}{N-k}\times$$

$$\times\left[\sum_{i=0}^{N-k}x(iT_0)x(iT_0+kT_0)-\sum_{i=0}^{N-k}x(iT_0)\hat{m}_x-\sum_{i=0}^{N-k}x(iT_0+kT_0)\hat{m}_x+\sum_{i=0}^{N-k}\hat{m}_x^2\right].$$

Узявши до уваги, що

$$\hat{m}_x=\frac{1}{N-k}\sum_{i=0}^{N-k}x(iT_0)=\frac{1}{N-k}\sum_{i=0}^{N-k}x(iT_0+kT_0)\ ,$$

дістанемо:

$$\hat{K}_X(kT_0)=\frac{1}{N-k}\sum_{i=0}^{N-k}x(iT_0)x(iT_0+kT_0)-\hat{m}_x^2. \qquad (5.8)$$

Аналогічно можна дістати вираз для взаємно кореляційної функції:

$$\hat{K}_{XY}(kT_0)=\frac{1}{N-k}=\sum_{i=0}^{N-k}x(iT_0)y(iT_0+kT_0)-\hat{m}_x\hat{m}_y\ . \qquad (5.9)$$

Вирази (28.8) та (28.9) дозволяють обчислити *k*-ту ординату кореляційної функції.

***Інтервал кореляції.*** Побудувавши залежність, що характеризує зміну $K_X(kT_0)$ від *k*, можна помітити, що при збільшенні *k* значення $K_X(kT_0)$ зменшуються (рис 5.6), тобто щільність стохастичного лінійного зв'язку між значеннями випадкового процесу зменшується зі збільшенням часового інтервалу між цими значеннями.

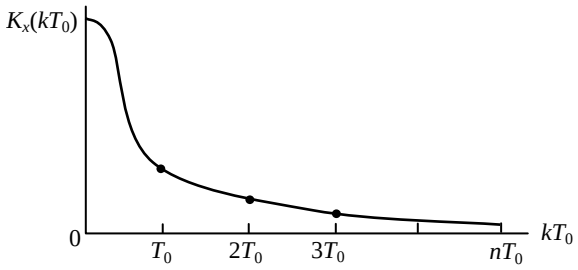



Рис. 5.6. Інтервал кореляції

Для дослідження змін зазначеної залежності зазвичай розглядають нормовану функцію кореляції $\rho_X(\tau)$.

*Інтервалом кореляції* $\tau_k$ випадкової функції $X(t)$ називатимемо інтервал часу, що визначається функціоналом:

$$\tau_k = \int_0^\infty \rho_X(\tau) d\tau \ , \qquad (5.10)$$

Таким чином, інтервалу кореляції відповідає основа прямокутника, площа якого дорівнює площі під нормованою функцією кореляції (рис. 5.7).

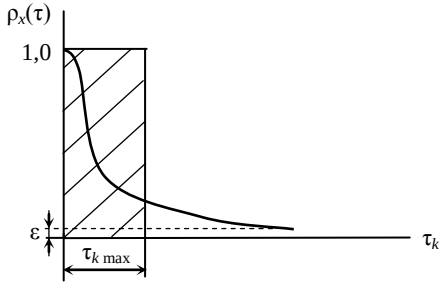



Рис. 5.7. Представлення максимального інтервалу кореляції

На практиці використовують поняття *максимального інтервалу кореляції* $\tau_{k\ \max}$ – часового інтервалу між значеннями випадкового процесу, в якому нормована функція кореляції, спадаючи, досягає деякого малого значення ε і надалі залишається меншою за нього при будь-якому $\tau > \tau_{k\,\max}$.

*За межами $\tau_{k\,\max}$ кореляція настільки мала, що нею можна знехтувати.*

*Вибірки, розподілені $\tau_{k\,\max}$ вважаються практично некорельованими.*

Якщо нормована функція кореляції *знакозмінна* (рис. 5.8), існує *абсолютний інтервал кореляції*, який визначається за виразом:

$$\tau_{ka} = \int_0^{\infty} |\rho_X(\tau_\kappa)| d\tau \qquad (5.11)$$

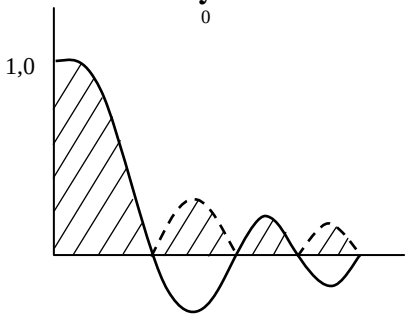



Рис. 5.8. Абсолютний інтервал кореляції

*Величини $\tau_k$ і $\tau_{ka}$ дають лише орієнтовне уявлення про те, на яких інтервалах часу в середньому існує кореляція між значеннями випадкового процесу.*

***Послідовно-паралельний метод обчислення кореляційної функції.*** Розглянемо реалізацію випадкового ергодичного процесу, зображеного на рис. 5.9. Розіб'ємо її на $N$ циклів – часових відрізків тривалістю $T_{\text{ц}} > \tau_{k\,\max}$. Кожний цикл містить $n$ вибіркових значень, віддалених одне від одного на час $T_0$. Останнє вибіркове значення попереднього циклу є першим для наступного. При цьому вибіркові значення з однаковими порядковими номерами у сусідніх циклах будуть практично некорельованими.

Множення вибірок циклу здійснюють у такий спосіб. Значення випадкового процесу для часу $t_0$, тобто $x(0)$, перемножують зі значенням $x(t_1) = x(T_0)$, потім $x(0)$ перемножують зі значенням $x(t_2) = x(2T_0)$ і т. д.

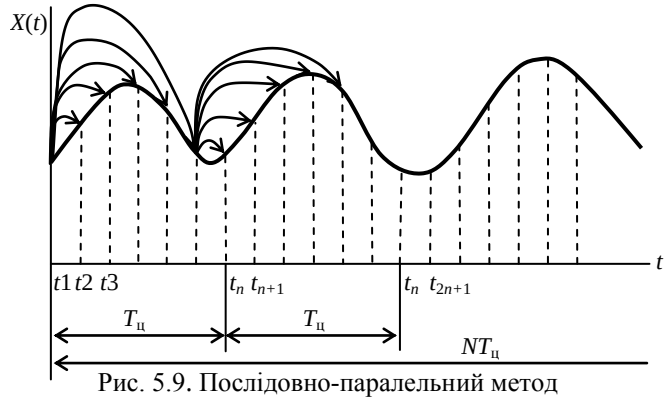


Рис. 5.9. Послідовно-паралельний метод

Так діють доти, доки не буде знайдено добуток значень $x(0)$ і $x(nT_0)$.

Аналогічну процедуру виконують при другому циклі. Ордината $n$-а першого циклу є початковою ординатою другого циклу, вона послідовно перемножується зі значеннями, віддаленими від неї на час $T_0$, $2T_0$,..., $nT_0$.

З огляду на те, що вибіркові значення досліджуваного випадкового процесу, віддалені одне від одного на часовий інтервал $T_ц$, будуть практично некорельованими, у кожному циклі обчислюватимуться поточні значення кожної з $n$ ординат кореляційної функції, які додаються до знайдених на попередніх циклах значень. Таким чином, у кожному циклі паралельно уточнюються точкові значення функції автокореляції $K_X(kT_0)$.

Такий алгоритм дає змогу під час дослідження випадкового процесу відразу обчислювати ординати кореляційної функції на відміну від послідовного алгоритму, коли ці ординати можна

обчислити тільки після завершення експериментального дослідження й одержання всього масиву даних.

**Функція знакової кореляції (кореляційна функція виду «*знак-знак*»).** Знаковою називається функція, для якої справджується таке співвідношення:

$$sgn = X^{\circ}(t) = \begin{cases} +1 \text{ при } X^{\circ}(t) > 0 \\ -1 \text{ при } X^{\circ}(t) < 0 \end{cases}.$$

Оскільки для формування знакової функції достатньо мати найпростіший пристрій – компаратор, то алгоритми обчислення щільності стохастичного зв'язку на основі знакових функцій набувають поширення завдяки простоті апаратурної реалізації.

*Знаковою кореляційною функцією* називається математичне сподівання добутку знакових функцій випадкового центрованого процесу (процесів) для різних моментів часу $t_1$, $t_2$.

Згідно з цим означенням для нестаціонарного процесу:

– знакова *автокореляційна* функція:

$$R_X(t_1, t_2) = M\left\{\operatorname{sgn}\left[X^{\circ}(t_1)\right]\operatorname{sgn}\left[X^{\circ}(t_2)\right]\right\};$$

– знакова взаємно кореляційна функція:

$$R_{XY}(t_1, t_2) = M\left\{\operatorname{sgn}\left[X^{\circ}(t_1)\right]\operatorname{sgn}\left[Y^{\circ}(t_2)\right]\right\}.$$

Функцію знакової кореляції як і кореляційну функцію можна представити, використовуючи двовимірну щільність розподілу

$$R_X(t_1, t_2) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \operatorname{sgn} x^{\circ}(t_1) \cdot \operatorname{sgn} x^{\circ}(t_2) w(x_1, x_2, t_1, t_2) dx_1 dx_2, \quad (5.12)$$

тобто, підсумовується за всіма можливими значеннями.

Розіб'ємо інтеграл (5.12) на чотири складові, враховуючи, що добуток складових з однаковими знаками дає «плюс» 1, а з протилежними – «мінус» 1.

$$R_X(t_1,t_2)=\int_0^{+\infty}\int_0^{+\infty}W(x_1^{\circ},x_2^{\circ},t_1,t_2)dx_1dx_2+\int_{-\infty}^{0}\int_{-\infty}^{0}W(x_1^{\circ},x_2^{\circ},t_1,t_2)dx_1dx_2-$$

$$-\int_0^{+\infty}\int_{-\infty}^{0}W(\overset{\circ}{x}_1,\overset{\circ}{x}_2,t_1,t_2)dx_1dx_2-\int_{-\infty}^{0}\int_0^{+\infty}W(x_1^{\circ},x_2^{\circ},t_1,t_2)dx_1dx_2\,. \quad (5.13)$$

Вираз (5.13) можна переписати у спрощеному виді

$$R_X(t_1,t_2)=p^{++}(t_1,t_2)+p^{--}(t_1,t_2)-q^{+-}(t_1,t_2)-q^{-+}(t_1,t_2)=$$
$$=p(t_1,t_2)-q(t_1,t_2), \quad (5.14)$$

де $p(t_1,t_2)$ – ймовірність того, що знаки в моменти часу $t_1$ й $t_2$ однакові, а $q(t_1,t_2)$ – протилежні.

У виразі (5.14) маємо повну групу подій, тому $p + q = 1$, тоді вираз (5.14) з урахуванням того, що $p - q = p - (1 - p) = 2p - 1$, можна записати

$$R_X(t_1,t_2)=2p(t_1,t_2)-1. \quad (5.15)$$

*На практиці підраховують кількість збігів знаків n, обчислюють оцінку ймовірності збігу знаків $\hat{p}$ як відношення числа n до загальної кількості значень N і отримують оцінку знакової функції кореляції, використовуючи у виразі (5.15) замість p його оцінку $\hat{p}$.*

Якщо випадковий процес характеризується нормальною функцією розподілу з нульовим математичним сподіванням, то з огляду на симетрію такого закону:

$$p^{++}(t_1,t_2)=+p^{--}(t_1,t_2).$$

Тому для такого процесу можна підраховувати тільки збіг «+» або «–», тобто

$$R_X(t_1,t_2)=4p^{++}(t_1,t_2)-1=4p^{--}(t_1,t_2)-1.$$

Для стаціонарного ергодичного випадкового процесу маємо:

– знакова автокореляційна функція

$$R_X(kT_0) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \operatorname{sgn}\left[x^{\circ}(iT_0)\right] \operatorname{sgn}\left[x^{\circ}(iT_0 + kT_0)\right]; \quad (5.16)$$

– знакова взаємно кореляційна функція

$$R_{XY}(kT_0) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \operatorname{sgn}\left[x^{\circ}(iT_0)\right] \operatorname{sgn}\left[y^{\circ}(iT_0 + kT_0)\right]. \quad (5.17)$$

Вирази (5.16) і (5.17) дають уявлення про теоретичне визначення знакової кореляційної функції. На практиці мають обмежений обсяг даних $N$, за якими можна одержати *оцінку знакової кореляційної функції*. Для цього випадку в наведених формулах для ергодичного випадкового процесу верхня межа підсумовування буде $N - k$.

Недоліком знакової функції кореляції є нечутливість до конкретного вигляду закону, якщо він симетричний. Для усунення зазначеного недоліку необхідно конкретизувати закон розподілу.

**Приклад**. Синусоїдний сигнал, сигнал типу «меандр» і «трикутний» сигнал із тими самими періодами, що й синусоїдний, мають однакові знакові кореляційні функції, хоча нормовані кореляційні функції в них різні.

Функцію знакової кореляції можна безпосередньо використати для оцінки лінійного стохастичного зв'язку, однак для техніки вимірювань кореляційної функції необхідно знати співвідношення між функцією знакової кореляції та нормованою кореляційною функцією.

Наприклад, якщо підставимо в рівняння (5.13) вираз двовимірної щільності ймовірностей стаціонарного ергодичного процесу з нормальним розподілом імовірностей і нульовим математичним сподіванням

$$W(x_1, x_2, \tau) = \frac{1}{\sqrt{2\pi} \cdot \sqrt{1 - \rho_X^2(\tau)}} e^{-\frac{x_1^2 - 2x_1 x_2 \rho_X(\tau) + x_2^2}{2\left[1 - \rho_x^2(\tau)\right]}},$$

то дістанемо співвідношення, що встановлює *зв'язок між знаковою кореляційною функцією та нормованою функцією кореляції*

$$R_X(\tau) = \frac{2}{\pi} \arcsin \rho_X(\tau),$$

звідки нормована кореляційна функція буде дорівнювати

$$\rho_X(\tau) = \sin\left[\frac{\pi}{2} R_X(\tau)\right].$$

Таким чином, знаючи $R_X(\tau)$, можна визначити $\rho_X(\tau)$, особливо зважаючи на те, що знакова функція кореляції $R_X(\tau)$ визначається простіше, ніж нормована кореляційна функція.

**Сумо-різницевий метод обчислення кореляційної функції**. Раніше було розглянуто, що для визначення кореляційної функції необхідно виконувати операцію множення миттєвих значень випадкового процесу, що не завжди зручно для апаратурної реалізації. У разі стаціонарного ергодичного процесу за умови, що процес центрований, можна застосувати сумо-різницевий метод обчислення кореляційної функції, який реалізується значно простіше, ніж безпосереднє множення.

Розглянемо вираз для математичного сподівання квадрата суми центрованих процесів:

$$A(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{+T} \left[x^\circ(t) + x^\circ(t+\tau)\right]^2 dt =$$

$$= \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{+T} x^{\circ 2}(t)\,dt + \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{+T} x^{\circ 2}(t+\tau)dt + \lim_{T\to\infty} \frac{2}{2T} \int_{-T}^{+T} x^\circ(t) x^\circ(t+\tau)\,dt.$$

З огляду на те, що

$$x^{\circ 2}(t) = \left[x(t) - m_x\right]^2 = D_X$$

і

$$x^{\circ 2}(t+\tau) = \left[x(t+\tau) - m_x\right]^2 = D_X,$$

тоді

$$A(\tau) = 2D_X + 2K_X(\tau), \qquad (5.18)$$

звідки

$$K_X(\tau) = \frac{A(\tau)}{2} - D_X .$$

Недолік такого алгоритму – необхідно знати дисперсію випадкового процесу.

Розглянемо вираз для математичного сподівання квадрата різниці цих процесів:

$$B(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{+T} \left[x^\circ(t) - x^\circ(t+\tau)\right]^2 dt =$$

$$= \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{+T} x^{\circ 2}(t)dt + \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{+T} x^{\circ 2}(t+\tau)dt -$$

$$- \lim_{T\to\infty} \frac{2}{2T} \int_{-T}^{+T} x^\circ(t)x^\circ(t+\tau)dt ,$$

звідки

$$B(\tau) = 2D_X - 2K_X(\tau). \qquad (5.19)$$

У результаті дістанемо кореляційну функцію з використанням квадрата різниці:

$$K_X(\tau) = D_X - \frac{B(\tau)}{2}.$$

*Застосувавши тільки обчислення квадрата суми або квадрата різниці, можна визначити кореляційну функцію в разі, коли відома дисперсія випадкового процесу, для оцінювання якої потрібно обчислювати кореляційну функцію при* $\tau = 0$, *що є незручним.*

Віднявши (5.19) від (5.18), дістанемо:

$$K_X(\tau) = \frac{A(\tau) - B(\tau)}{4}.$$

Звідси випливає, що для обчислення кореляційної функції не потрібно знати дисперсію випадкового процесу.

Розглянемо алгоритм, що дає змогу обчислювати взаємно кореляційну функцію:

- реалізації випадкового процесу, віддалені одна від одної на час τ, підсумовуються та віднімаються;
- відповідно знайдені суми та різниці підносяться до квадрата і усереднюються;

*Якщо розглядаються аналогові сигнали, то підсумовування та віднімання відбуваються за час T.*
*Якщо розглядаються дискретні значення випадкового процесу, то знаходять суму N миттєвих значень, підносять до другої степені, а потім ділять на кількість значень n.*

- від середнього значення для квадрата суми віднімається середнє значення для квадрата різниці і дістають значення, пропорційне до кореляційної функції.

Варто пам'ятати, що через скінченний час усереднення в разі аналогового сигналу та скінченого обсягу вибіркових значень для дискретного сигналу одержують не кореляційну функцію, а її оцінку.

**Визначення нормованої кореляційної функції з використанням умовного середнього**. Відомо, що рівняння регресії дає наближене уявлення про коефіцієнт кореляції. Існує зв'язок між нормованим коефіцієнтом кореляції і умовним середнім:

$$M[Y \mid x] = MY + \rho_{XY} \frac{\sigma_Y}{\sigma_X}(X - MX).$$

Для випадкового процесу $Y(t)$ математичне сподівання за умови, що $X(t) = d$, подається у вигляді:

$$M[Y(t) \mid X(t)] = m_Y + \rho_{XY} \frac{\sigma_Y}{\sigma_X}(d - m_X). \qquad (5.20)$$

За припущення, що $Y(t) \to X^{\circ}(t+\tau)$, а $X(t) \to X^{\circ}(t)$, доходимо висновку, що $m_X = m_Y = 0$.

У разі стаціонарності випадкового процесу, для якого $\sigma_X = \sigma_Y$, вираз (5.20) можна подати як

$$M\left[X^{\circ}(t+\tau) \mid X^{\circ}(t) = d\right] = \rho_X(\tau) d,$$

звідки нормований коефіцієнт кореляції:

$$\rho_X(\tau) = M\left[X^{\circ}(t+\tau) \mid X^{\circ}(t) = d\right] \frac{1}{d}.$$

Таким чином, нормована кореляційна функція буде пропорційною до математичного сподівання значень центрованого випадкового процесу, віддалених на відстані $\tau$ відносно часу, коли цей процес досяг фіксованого значення *d* (рис. 5.10).

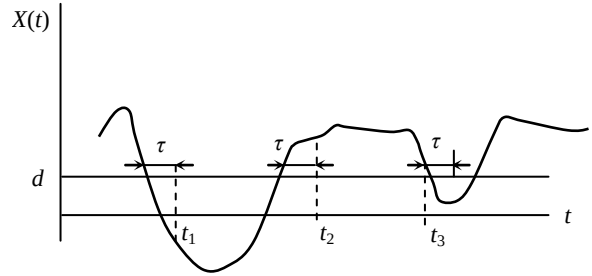



Рис. 5.10. Визначення нормованої кореляційної функції з використанням умовного середнього

На практиці процедуру обчислення математичного сподівання замінюють операцією підсумовування значень із подальшим діленням цієї суми на кількість реалізацій *N*, тобто усереднюють і таким чином дістають оцінку:

$$\hat{\rho}_X(\tau) = \frac{1}{d} \frac{1}{N} \sum_{i=1}^{N} x^{\circ}(t+\tau) \mid x^{\circ}(t) = d.$$

Аналогічно можна визначити і взаємно кореляційну функцію, але для цього необхідно розглядати два випадкові процеси.

**Спектральна щільність випадкового процесу.** Відомо, що майже будь-який сигнал, яким фактично є випадковий процес, можна подати за допомогою ряду Фур'є або інтеграла Фур'є. Згідно з цим можна завжди підібрати таке поєднання гармонічних складових за амплітудою та фазою, які будуть із заданою степеню точності апроксимувати складний сигнал.

Якщо існує нескінченна множина гармонік, то при збільшенні часу спостережень за сигналом (процесом), що відповідає відсутності обмежень на його частотний діапазон, дістанемо неперервну залежність зміни дисперсії з частотою (рис. 5.11).

Якщо сигнал детермінований, амплітуди складових сигналів залишаються незмінними. Для випадкових сигналів (процесів) характерною рисою є те, що амплітуди гармонічних складових можуть змінюватися в часі (майже до нуля, тобто гармоніки може й не бути), причому в загальному випадку дисперсія для кожної гармонійної складової відмінна від дисперсії іншої складової.

У разі подання складного сигналу у вигляді фіксованого (скінченного) ряду гармонік можна говорити *про лінійний спектр дисперсії*, який зображено пунктирними лініями на рис. 5.11. Для лінійного спектра дисперсія всього сигналу буде дорівнювати середньозваженій сумі дисперсій гармонік. Усереднена дисперсія відповідатиме інтегралу, а отже, площі під неперервною кривою.

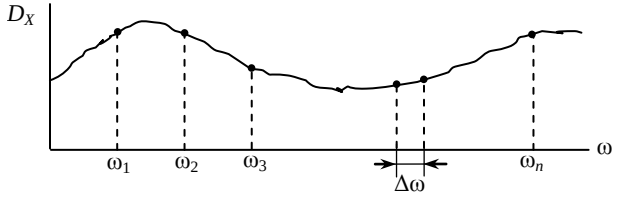



Рис. 5.11. Визначення спектральної щільності

Для суцільного спектра вводиться поняття *спектральної щільності* $S_X(\omega)$, яка відноситься до інтервалу $\Delta\omega$. Аналогічно щільність неперервної випадкової величини визначається так:

$$S_X(\omega) = \lim_{\Delta\omega \to 0} \frac{D_X(\omega, \omega + \Delta\omega)}{\Delta\omega}. \qquad (5.21)$$

*Якщо припустити, що випадковий процес є напругою або струмом, то чисельник виразу* (5.21) *можна розглядати як потужність, що виділяється на опорі в 1 Ом. Це пояснюється тим, що дисперсія – величина, пропорційна до* $I^2$ *або* $U^2$*, а потужність* $P = U^2/R = I^2R$.

*Тому* $S(\omega)$ *називається ще спектральною щільністю потужності. Якщо взяти до уваги, що в чисельнику* (5.21) *одиниця фізичної величини – Вт, а в знаменнику – Гц, то* $S(\omega)$ *має одиницю Вт · с, що відповідає енергії.*

Для апаратурного визначення $S(\omega)$ необхідно мати набір фільтрів із малою смугою пропущення, що є складною задачею.

Існує співвідношення Вінера-Хінчіна, яке пов'язує спектральну щільність та автокореляційну функцію:

$$S(\omega) = \int_{-\infty}^{+\infty} K_X(\tau) \ell^{-j\omega\tau} d\tau \ , \qquad (5.22)$$

згідно з яким $S(\omega)$ можна обчислити, визначивши апаратурним шляхом $K_X$.

За допомогою перетворення Ейлера можна здійснити перехід від показникової форми подання комплексного числа до алгебраїчної:

$$e^{-j\omega\tau} = \cos\omega\tau - j\sin\omega\tau .$$

Скориставшись цим співвідношенням і врахувавши, що спектральна щільність є парною функцією, співвідношення (5.22) можна записати в такий спосіб:

$$S(\omega)=\int_{-\infty}^{+\infty}K_X(\tau)[\cos\omega\tau-j\sin\omega t]d\tau=2\int_{0}^{+\infty}K_X(\tau)cos\omega\tau d\tau\,. \quad (5.23).$$

Існує вираз для *оберненого перетворення*:

$$K_X(\tau)=\frac{1}{2\pi}\int_{-\infty}^{+\infty}S(\omega)e^{j\omega\tau}d\omega\,.$$

У практиці аналізу випадкових процесів використовують *нормовану спектральну щільність*

$$S(\omega)=\frac{S(\omega)}{\int_{-\infty}^{+\infty}S(\omega)d\omega}=\frac{S(\omega)}{K_X(0)}, \quad (5.24)$$

де $\int_{-\infty}^{+\infty}S(\omega)d\omega=K_X(0)$ – постійна складова спектральної щільності, або спектральна щільність, що відповідає нульовій частоті.

Підставивши у вираз (5.24) співвідношення (5.23), дістанемо:

$$S(\omega)=\frac{2\int_{0}^{+\infty}K_X(\tau)cos\omega\tau d\tau}{K_X(0)}=2\int_{0}^{+\infty}\rho_X(\tau)cos\omega\tau d\tau\,. \quad (5.25)$$

У формулі (5.25) $\frac{K_X(\tau)}{K_X(0)}=\rho_X(\tau)$ – *нормований коефіцієнт кореляції*.

*Спектральна щільність дає змогу:*
*– оцінити спектральний склад випадкового сигналу;*
*– виділити діапазон частот, в якому зосереджено основну енергію сигналу;*
*– для нестаціонарних процесів простежити, як змінюється енергетичний спектр у часі.*

**Властивість кореляційної функції сигналу з періодичною складовою**. Знаючи закон зміни автокореляційної функції періодичного процесу, можемо виділити детерміновану складову з випадкового сигналу та визначити її параметри.

Для періодичного сигналу автокореляційну функцію $K_X(\tau)_{\text{пер}}$ можна визначити, скориставшись виразом:

$$K_X(\tau)_{\text{пер}} = \lim_{T_0\to\infty} \frac{1}{T_0} \int_{-T_0/2}^{+T_0/2} x^{\circ}(t)x^{\circ}(t+\tau)dt\,.$$

Для періодичного сигналу результати дослідження на інтервалі, що дорівнює його періоду, такі самі, як і на нескінченно великому проміжку часу, а отже, виконується рівність:

$$K_X(\tau)_{\text{пер}} = \frac{1}{T}\int_{-\frac{T}{2}}^{+\frac{T}{2}} x^{\circ}(t)x^{\circ}(t+\tau)\,dt. \qquad (5.26)$$

З огляду на це можна знайти автокореляційну функцію для періодичного процесу. Припустимо, що

$$x^{\circ}(t) = A\sin(\omega t - \varphi),$$

тоді згідно з (5.26) дістанемо:

$$K_X(\tau)_{\text{пер}} = \frac{A^2}{T}\int_{-T/2}^{+T/2} \sin(\omega t - \varphi)\sin[\omega(t+\tau) - \varphi]dt\,.$$

Візьмемо інтеграл, узявши до уваги, що

$$\sin\alpha\cdot\sin\beta = \frac{1}{2}[\cos(\alpha-\beta) - \cos(\alpha+\beta)],$$

тобто

$$\sin(\omega t - \varphi)\cdot\sin[\omega(t+\tau) - \varphi] = \frac{1}{2}\cos(\omega t) - \cos[\omega(2t+\tau) - 2\varphi].$$

Далі, урахувавши, що інтеграл від періодичної функції, обчислений за період, дорівнює нулю, дістанемо:

$$K_X(\tau)_{\text{пер}} = \frac{2A^2}{T}\cos\omega\tau\,.$$

Таким чином, установлено, що автокореляційна функція періодичного процесу буде змінюватися за законом косинуса з частотою цього періодичного процесу.

Цю властивість широко використовують при статистичній обробці для відокремлення детермінованої періодичної складової сигналу від випадкової його складової. Відомо, що для випадкового сигналу автокореляційна функція досягає максимуму при $\tau = 0$, при збільшенні $\tau$ значення її швидко зменшується. Тому при великих значеннях $\tau$ автокореляційна функція періодичного сигналу, на який накладено шум, буде визначатися тільки детермінованою складовою. Проаналізувавши кореляційну функцію при великих затримках $\tau$, можна визначити частоту «зашумленого» періодичного сигналу.

**Приклад**. При реєстрації кардіограми необхідно виділити корисний сигнал на фоні шумів. Якщо застосувати апарат кореляційного аналізу, можна визначити періодичну детерміновану складову.

## Запитання для самоперевірки

1. Назвіть основні характеристики випадкових процесів. Що характеризує функція кореляції? Які процеси називаються стаціонарними, нестаціонарними, ергодичними?

2. Що характеризують автокореляційна та взаємно кореляційна функції? Що являє собою нормована кореляційна функція?

3. У чому полягає суть послідовного методу знаходження оцінки кореляційної функції?

4. Як здійснюється обчислення оцінки кореляційної функції нецентрованих величин випадкового процесу?

5. Що характеризує інтервал кореляції та абсолютний інтервал кореляції? Що характеризує максимальний інтервал кореляції?

6. У чому полягає послідовно-паралельний метод обчислення кореляційної функції?

7. Що являє собою функція знакової кореляції? Як визначаються знаки функції знакової кореляції?

8. Який зв'язок між знаковою кореляційною функцією та нормованою функцією кореляції і для чого він встановлюється?

9. У чому полягає сумо-різницевий метод обчислення кореляційної функції?

10. Як визначити нормовану кореляційну функцію з використанням умовного середнього?

11. Що характеризує спектральна щільність випадкового процесу?

12. Наведіть співвідношення Вінера-Хінчина. Що характеризує це співвідношення?

13. Які процеси характеризуються лінійним спектром дисперсії?

# РОЗДІЛ 6

# ОСНОВИ ПЛАНУВАННЯ ЕКСПЕРИМЕНТУ

## 6.1. ПОВНИЙ ФАКТОРНИЙ ЕКСПЕРИМЕНТ

Активний експеримент, на відміну від пасивного, полягає не у простому фіксуванні вхідних і вихідних величин, а в активному впливі на об'єкт за заздалегідь обраним планом.

Для знаходження за результатами експерименту зв'язку вихідної величини з керованими факторами, які впливають на цю величину, використовують моделі регресійного аналізу.

Вихідною величиною може бути точність приладу, надійність, ефективність технологічного процесу, показники якості тощо. Факторами можуть бути змінні, які суттєво впливають на вихідну величину, наприклад параметри живлення обладнання, рівень завад, умови перебігу технологічного процесу, зовнішні впливи, механічне навантаження тощо.

*Для проведення експерименту необхідна модель об'єкта, а для уточнення параметрів моделі необхідний експеримент. З одного боку, експеримент дає змогу перевірити і при необхідності уточнити модель, з другого боку, модель диктує, який саме експеримент слід провести.*

Для *реалізації* активного експерименту необхідно виконати такі умови:

– результати спостережень вихідної фізичної величини $\tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_N$ мають являти собою незалежні, нормально розподілені випадкові величини;

– незалежні змінні $x_1, x_2, ..., x_n$ мають вимірюватися з настільки малою похибкою (порівняно з похибкою вимірювання вихідної величини *y*), що нею можна знехтувати;

– дисперсії спостережень вихідної величини $S^2\{\tilde{y}_i\}$ мають бути однорідими, тобто якщо виконувати багаторазові повторні вимірювання величини $y_i$ при деякому певному наборі вхідних значень $x_{i1}, x_{i2}, ..., x_{in}$, то дисперсія $\sigma^2\{y_i\}$ не повинна відрізнятися від дисперсії $\sigma^2\{y_k\}$, отриманої при повторних спостереженнях для будь-якого іншого набору значень незалежних змінних $x_{k1}, x_{k2}, ..., x_{kn}$.

Дані умови припускають, що між вхідними та вихідними величинами існує функціональний зв'язок і метою експерименту є визначення оцінок параметрів, які будуть відрізнятися від їхнього математичного сподівання, передусім через наявність похибок вимірювання вихідної величини або через вплив неврахованих некерованих факторів, а також через обмежений обсяг експериментальних даних.

**Факторний експеримент**. При побудові моделі проводять так звані *факторні* експерименти, в яких, на відміну від *класичних*, відбувається одночасна зміна всіх незалежних вхідних величин.

*При класичному експерименті оператор вивчає вплив кожного фактора при їх почерговому варіюванні. Дослідження взаємодії факторів можливе лише при одночасній зміні їхніх рівнів.*

Експеримент, у результаті якого всі незалежні змінні варіюються на всіх обраних рівнях (набувають усіх можливих значень), називається ***повним факторним експериментом*** (**ПФЕ**).

***План експерименту*** передбачає кількість і умови проведення дослідів. Кількість дослідів *N* при ПФЕ визначається як

$$N = k^n,$$

де $k$ – кількість можливих рівнів, яких може набувати фактор; $n$ – кількість факторів.

На практиці набули поширення експерименти, де незалежні змінні набувають тільки двох значень (як правило, мінімального та максимального), причому в експерименті здійснюються всі можливі комбінації з $n$ факторів. Такі експерименти називаються ***дворівневими***: для них

$$N = 2^n.$$

Оскільки незалежні величини $X_j$ можуть бути різними за фізичною природою й змінюватися в різних динамічних діапазонах, прийнято здійснювати *кодування* незалежних змінних (розд. 4.8). Це дає змогу формалізувати процедуру планування та обробки результатів і з однаковою чутливістю врахувати вплив усіх передбачуваних величин.

Кодування полягає в перенесенні центра координат у точку $X_{j\text{ср}}$, ($j$ – значення фактора, $j$ = 1, ..., $n$), яка надалі називається ***центром плану експерименту***.

Для кодування використовують співвідношення

$$x_j = \frac{X_j - X_{j\text{ср}}}{X_{j\max} - X_{j\text{ср}}} = \frac{X_j - X_{j\text{ср}}}{X_{j\text{ср}} - X_{j\min}}, \qquad (6.1)$$

де $X_{j\text{ср}} = \left(X_{j\max} + X_{j\min}\right)/2$; $X_{j\min}$ і $X_{j\max}$ – граничні значення, яких набувають незалежні змінні – задані або такі, що експериментатор вибирає їх самостійно на підставі апріорної інформації про об’єкт (рис. 6.1).

У кодованій системі на підставі виразу (6.1) виконуватимуться відповідності:

$$X_{j\max} \to x_j = +1;$$
$$X_{j\min} \to x_j = -1;$$
$$X_{j\text{ср}} \to x_j = 0.$$

Надалі будуть використовуватися кодовані змінні.

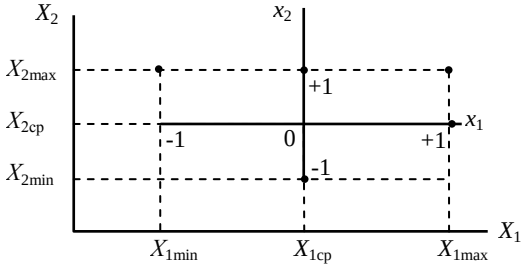


Рис. 6.1. Кодування факторів

*У випадку парної залежності, коли впливає тільки один фактор x, для визначення лінії регресії достатньо виконати два досліди при його граничних значеннях.*
*План експерименту в цьому разі буде* $X=(-1\ +1)^T$, *тобто являтиме собою вектор-стовпець.*

Якщо вхідних величин дві: $x_1$ і $x_2$, тобто реалізується *двофакторний* експеримент, для побудови матриці плану ПФЕ необхідно користуватися таким правилом. Наявний вектор-стовпець плану слід відтворити два рази: один раз – при нижньому рівні фактора (– 1), а другий раз – при його верхньому рівні (+1) і отримати *матрицю плану*.

Розглянувши матрицю плану двофакторного експерименту, побудовану за викладеним правилом, побачимо, що в ній присутні всі $N = 2^n = 4$ комбінацій факторів $x_1$ і $x_2$: –1 –1; –1 +1; +1 –1; +1 +1.

Геометрично план такого експерименту інтерпретується точками, розміщеними у вершинах квадрата (рис. 6.2).

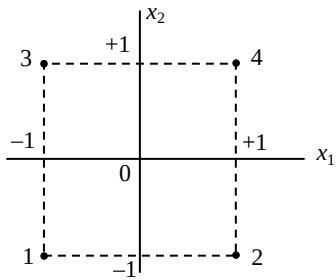


Рис. 6.2. Геометрична інтерпретація плану двох факторного експерименту

Чергування знаків факторів (рівнів їхньої зміни) у відповідному стовпці матриці можна подати формулою $2^{j-1}$, де $j$ – номер фактора.

**Приклад**. Якщо $j = 1$, то $2^0 = 1$, тобто знаки у першому стовпці для $x_1$ чергуються через один; якщо $j = 2$, то $2^1 = 2$, тобто знаки в другому стовпці для $x_2$ чергуються через два, і т. д.

Виходячи з розглянутого правила, можна скласти матрицю плану для трифакторного експерименту. Матрицю плану трифакторного експерименту з доповненнями наведено в табл. 6.1.

Таблиця 6.1. Матриця плану трьохфакторного експерименту з доповненнями

| Номер досліду $i$ | Матриця плану | | | Доповнення для обчислень $a_0$ і коефіцієнтів при взаємодіях факторів | | Значення вихідної величини |
|---|---|---|---|---|---|---|
| | $x1$ | $x_2$ | $x_3$ | $x_0$ | $x_1x_2$ | $\tilde{y}_i$ |
| 1 | –1 | –1 | –1 | +1 | +1 | $\tilde{y}_1$ |
| 2 | +1 | –1 | –1 | +1 | –1 | $\tilde{y}_2$ |
| 3 | –1 | +1 | –1 | +1 | –1 | $\tilde{y}_3$ |
| 4 | +1 | +1 | –1 | +1 | +1 | $\tilde{y}_4$ |
| 5 | –1 | –1 | +1 | +1 | +1 | $\tilde{y}_5$ |
| 6 | +1 | –1 | +1 | +1 | –1 | $\tilde{y}_6$ |
| 7 | –1 | +1 | +1 | +1 | –1 | $\tilde{y}_7$ |
| 8 | +1 | +1 | +1 | +1 | +1 | $\tilde{y}_8$ |

У цьому випадку точки міститимуться у вершинах куба, оскільки $N = 2^3 = 8$.

При побудові матриці плану виходимо з виконання умови його *ортогональності,* тобто забезпечення незалежності знаходження оцінок моделей. Побудована в такий спосіб матриця має три важливі властивості.

- ***Ортогональність***:

$$\sum_{i=1}^{N} x_{ij} x_{ik} = 0 \ , \ j \neq k \ ,$$

де $j,k = \overline{1,n}$ – номери векторів-стовпців відповідних факторів; $i$ – точка факторного простору, в якій виконується експеримент.

Ця властивість забезпечує незалежність знаходження оцінок коефіцієнтів:

$$\hat{a}_j = \frac{\sum_{i=1}^{N} \tilde{y}_i x_{ij}}{\sum_{i=1}^{N} x_{ij}^2}. \qquad (6.2)$$

- ***Симетрія:***

$$\sum_{i=1}^{N} x_{ij} = 0;\ j = \overline{1,n}.$$

Ця властивість забезпечує незалежність знаходження оцінки вільного члена $a_0$.

- ***Нормованість:***

$$\sum_{i=1}^{N} x_{ij}^2 = N;\ j = \overline{1,n}. \qquad (6.3)$$

Ця властивість забезпечує однакову дисперсію оцінки коефіцієнтів $\hat{a}_j$, оскільки знаменник виразу (6.2) буде однаковим для всіх оцінок $\hat{a}_j$, тобто всі коефіцієнти будуть оцінені з однаковою точністю. Таким чином, вираз (6. 3) можна записати як

$$\hat{a}_j = \frac{1}{N} \sum_{i=1}^{N} \tilde{y}_i x_{ij}. \qquad (6.4)$$

**Приклад**. Відповідно до матриці плану двофакторного експерименту знайдено значення вихідної величини $\mathbf{Y} = |\tilde{y}_1, \tilde{y}_2, \tilde{y}_3, \tilde{y}_4|^{\mathbf{T}}$. Оцінки коефіцієнтів при факторах можна знайти згідно з такими виразами:

$$\hat{a}_1 = \frac{1}{4}\left[(-1)\tilde{y}_1 + (+1)\tilde{y}_2 + (-1)\tilde{y}_3 + (+1)\tilde{y}_4\right];$$

$$\hat{a}_2 = \frac{1}{4}\left[(-1)\tilde{y}_1 + (-1)\tilde{y}_2 + (+1)\tilde{y}_3 + (+1)\tilde{y}_4\right].$$

Отже, при обчисленні оцінки коефіцієнта $\hat{a}_j$ необхідно значення $\tilde{y}_i$, знайдене в результаті проведення досліду в *i*-й точці факторного простору, урахувати зі знаком, що відповідає знаку рівня варіювання *j*-го фактора для цього *i*-го рядка матриці плану.

Модель у загальному випадку містить вільний член $a_0$, для обчислення оцінки якого необхідно ввести вектор-стовпець для фіктивної змінної $x_0$, усі елементи якого дорівнюють +1 (див. табл. 6.1).

Зважаючи на те, що критерієм оптимальності плану обрано незалежність визначення оцінок коефіцієнтів, вектор-стовпець $x_0$ також має бути *ортогональним* стосовно інших векторів-стовпців $x_1$, $x_2$, $x_3$, тобто

$$\sum_{i=1}^{N} x_{i0} x_{ij} = 0 \, , \, j = 1,..., n.$$

Оскільки $x_0 = 1$, то властивість ортогональності буде реалізовано, коли

$$\sum_{i=1}^{N} x_{ij} = 0 \ , \qquad (6.5)$$

що відповідає властивості симетрії матриці плану.

Оцінка вільного члена буде визначатися незалежно від оцінок $\hat{a}_j$ згідно з виразом (6.4):

$$\hat{a}_0 = \sum_{i=1}^{N} \tilde{y}_i / N \, .$$

Розглянемо рівняння регресії, в якому припускається наявність взаємодії факторів:

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_{12} x_1 x_2. \qquad (6.6)$$

Якщо модель містить взаємодію факторів, то для оцінки коефіцієнта $\hat{a}_{12}$, що показує вплив цієї взаємодії, матрицю плану доповнюють відповідним вектором-стовпцем. Чергування знаків у

цьому векторі-стовпці відбувається шляхом перемножування знаків вихідних векторів-стовпців $x_1$ і $x_2$ (див. табл. 6.1).

*Жодна з незалежних змінних $x_j$ не є комбінацією (похідною) від інших незалежних змінних. Наявність взаємодії може зумовлюватись тільки наявністю нелінійності функції перетворення досліджуваного об'єкта.*

Побудований у такий спосіб вектор-стовпець для взаємодії $x_1x_2$ має всі три перелічені раніше властивості, тому оцінки коефіцієнтів для взаємодій $\hat{a}_{jk}$, $j \neq k$, вектори-стовпці для факторів яких побудовано за викладеним правилом, можна обчислити згідно з виразом (6.4), тобто

$$\hat{a}_k = \frac{1}{N}\sum_{i=1}^{N} x_{ik}\,\tilde{y}_i\,, \qquad (6.7)$$

де $k$ – стовпець будь-якої доповненої матриці.

Якщо в кожній точці факторного простору проводиться $m$ вимірювань, то у формулі (6.7) замість $\tilde{y}_i$ варто враховувати обчислене значення середнього арифметичного $\overline{y}_i$.

*Знайдені в такий спосіб оцінки коефіцієнтів моделі показують вплив факторів та їхніх взаємодій на вихідну величину. Якщо перед коефіцієнтом стоїть знак «плюс», то зі збільшенням цього фактора вихідна величина збільшується, а якщо стоїть знак «мінус», – навпаки.*

**Рандомізація.** Регресійний аналіз виходить із передумови випадковості похибок, які впливають на формування вхідних і вихідних вимірюваних величин. Однак якщо проводити досліди в тому порядку, як записано рядки матриці плану, побудованої відповідно до правила (тим більш для випадку проведення серії

паралельних дослідів у кожній точці факторного простору), то чим більший порядковий номер фактора, тим при більшій кількості дослідів його рівень залишається незмінним. Для «найстаршого» фактора $n$ його рівень залишається незмінним протягом $N/2$ дослідів. Навіть за відсутності систематичної похибки відтворення $x_j$ фіксоване значення випадкової похибки перетворюється на систематичну похибку, що призводить до зміщення результатів $\tilde{y}_i$. Чим більший номер стовпця матриці, тим сильніший цей вплив.

Відомо, що випадковість величини виявляється у множині її реалізацій. Тому щоб залишити тільки випадкову похибку в заданні рівнів факторів з нульовим математичним сподіванням, *усі досліди здійснюють у випадковому порядку*, тобто виконують ***рандомізацію*** експерименту [7].

Слід зазначити, що бувають випадки, при яких рандомізацію здійснити неможливо, наприклад, коли послідовність умов проведення дослідів є певним параметром.

**Приклад**. Випробування котушки індуктивності із залізним осердям, у якого цикл петлі гістерезису залежить від попередньої робочої точки.

## Запитання для самоперевірки

1. Який експеримент називають повним факторним експериментом?

2. Що розуміють під планом експерименту? Що таке центр плану експерименту? Як визначити кількість дослідів при ПФЕ?

3. Як побудувати матрицю плану ПФЕ? Які властивості вона має?

4. В якому разі вводяться в модель парні та потрійні взаємодії?

5. Яким чином визначається чергування знаків у векторі-стовпці взаємодій факторів для того, щоб оцінка коефіцієнтів при цих взаємодіях визначалася незалежно?

6. Як обчислити вільний член моделі?

7. Що характеризують коефіцієнти при факторах?

8. Що таке рандомізація і для чого її виконують?

## 6.2. ДРОБОВИЙ ФАКТОРНИЙ ЕКСПЕРИМЕНТ

Експеримент, який реалізує певну частину повного факторного експерименту, називається дробовим факторним експериментом.

За значної кількості факторів плани ПФЕ мають надмірність.

**Приклад**. Нехай для моделі

$$y = a_0 + \sum_{j=1}^{6} a_j x_j,$$

потрібно визначити 7 коефіцієнтів. Якщо розглядати повний факторний експеримент, то необхідно здійснити 64 ($N_{\text{ПФЕ}} = 2^6 = 64$) досліди для знаходження семи коефіцієнтів цієї моделі, що є недоцільним.

Зменшити кількість дослідів для знаходження оцінок невідомих коефіцієнтів у більшості практичних випадків можна за допомогою *дробових факторних експериментів* (ДФЕ).

**Побудова матриці дробового факторного експерименту без взаємодії.**

**Приклад**. Нехай необхідно побудувати дробову матрицю за припущення, що ефекти взаємодій відсутні, тобто модель має вигляд:

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3.$$

За такого вигляду залежності невідомими є чотири коефіцієнти $a_j$, $j = 0, 1, 2, 3$, для визначення яких достатньо чотирьох дослідів. Якщо застосовувати ПФЕ, кількість дослідів має бути $N = 2^3$, тобто необхідно зробити 8 дослідів.

Спочатку розглянемо побудову матриці ПФЕ (див. табл. 6.1). Якщо виберемо перші чотири рядки, то фактор $x_3$ має тільки значення нижнього рівня, а отже, не можна одержати повну інформацію про вплив фактора на вихідну величину. Те саме можна сказати й про нижні чотири рядки, де фактор $x_3$ набуває значень тільки верхнього рівня. Можна спробувати вибрати тільки парні чи тільки непарні рядки повної матриці плану, але результат буде також незадовільним.

Виберемо, наприклад, рядки 5-й, 2-й, 3-й, 8-й матриці ПФЕ та побудуємо матрицю плану в такий спосіб:

$$\mathbf{X} = \begin{vmatrix} x_1 & x_2 & x_3 \\ -1 & -1 & +1 \\ +1 & -1 & -1 \\ -1 & +1 & -1 \\ +1 & +1 & +1 \end{vmatrix}.$$

У цій матриці перші два стовпці являють собою матрицю плану двофакторного експерименту виду $2^2$, а третій стовпець відповідає добутку рівнів факторів стовпців $x_1$ і $x_2$.

Побудована в такий спосіб матриця має три властивості: *ортогональності, симетрії та нормованості*. Якщо ця частина матриці плану має зазначені властивості, то її взято не довільно, а за якимось правилом.

Таким чином, експеримент, що реалізує певну частину ПФЕ, називається ***дробовим факторним експериментом***. Матрицю, що утворюється при ДФЕ, називають ***дробовою матрицею планування***. Кількість рядків (дослідів) такої матриці у загальному випадку визначається співвідношенням

$$N = 2^n \cdot 2^{-p} = 2^{n-p},$$

де $n$ – загальна кількість лінійних факторів; $p$ – кількість ***додаткових факторів***; $(n - p)$ – кількість ***основних факторів***.

Для $(n - p)$ основних факторів будується матриця ПФЕ, а для $p$ додаткових факторів рівні варіювання в дослідах вибираються на підставі так званого *генеруючого співвідношення.*

*Для розглянутого прикладу матрицю дробового факторного експерименту можна подати як матрицю розміру $2^{n} \times 2^{-1}$, тобто матриця ДФЕ у цьому випадку є половиною матриці повного факторного експерименту.*

***Генеруюче співвідношення*** (ГС) – це формальна рівність, що показує, знаки яких незалежних змінних необхідно перемножити для отримання знаку додаткового фактора, щоб утворений вектор-стовпець додаткового фактора мав властивості симетрії, нормованості та ортогональності. У розглянутому прикладі вектор-стовпець додаткового фактора побудовано на підставі обраного ГС: $x_3 = x_1 x_2$.

Таким чином, при побудові матриці дробового експерименту необхідно реалізувати частину матриці повного факторного експерименту, кількість дослідів у якій має буде кратною 2 і більше за кількість шуканих коефіцієнтів $l$, тобто $= 2^{n-p} > l$.

*Вибір як основних, так і додаткових факторів є умовним і довільним. При будь-якому поєднанні основних і додаткових факторів матриця не втрачає своїх головних властивостей.*

Побудована в такий спосіб матриця плану ДФЕ дає змогу *незалежно* знаходити оцінки коефіцієнтів лінійної моделі, *якщо в ній немає взаємодій.*

**Побудова матриці дробового факторного експерименту із взаємодіями.**

**Приклад**. Розглянемо модель виду

$$y = a_0 + \sum_{j=1}^{4} a_j x_j + a_{14} x_1 x_4 . \qquad (6.8)$$

За правилом побудови дробової матриці умовно візьмемо як основні фактори $x_1$, $x_2$, $x_3$, а як додатковий – $x_4$. Для основних факторів будується матриця як для повного факторного експерименту. Для визначення додаткового фактора взято ГС: $x_4 = x_1x_2$, тобто знаки векторів-стовпців для додаткового фактора $x_4$ будуть визначатися як добуток $x_1x_2$. Оскільки модель містить значущу взаємодію, то для визначення оцінки коефіцієнта при взаємодії доповнимо матрицю плану вектором-стовпцем цієї взаємодії. Побудовану в такий спосіб матрицю наведено в табл. 6.2.

Таблиця 6.2. Матриця дробового факторного експерименту

| *Номер* досліду, *i* | $x_1$ | $x_2$ | $x_3$ | $x_4 = x_1x_2$ | $x_1x_4$ |
|---|---|---|---|---|---|
| 1 | –1 | –1 | –1 | +1 | –1 |
| 2 | +1 | –1 | –1 | –1 | –1 |
| 3 | –1 | +1 | –1 | –1 | +1 |
| 4 | +1 | +1 | –1 | +1 | +1 |
| 5 | –1 | –1 | +1 | +1 | –1 |
| 6 | +1 | –1 | +1 | –1 | –1 |
| 7 | –1 | +1 | +1 | –1 | +1 |
| 8 | +1 | +1 | +1 | +1 | +1 |

При розгляді матриці можна помітити, що чергування знаків у векторі-стовпці $x_1x_4$ повністю збігається з чергуванням знаків у векторі-стовпці $x_2$.

Зважаючи на те, що оцінки коефіцієнтів обчислюються за тією самою формулою, а кореляція знаків у векторах-стовпцях $x_2$ і $x_1x_4$ дорівнює одиниці, не можна окремо оцінити вплив фактора $x_2$ і взаємодії $x_1x_4$. Можна оцінити тільки їхній *сумарний* вплив, що подається у вигляді або

$$\hat{a}_2^* = (\hat{a}_2 + \hat{a}_{14}),$$

або

$$\hat{a}_{14}^* = (\hat{a}_{14} + \hat{a}_2).$$

Як випливає з цих формул, оцінка сумарного впливу буде *однаковою*. Урахувавши останні рівності вираз 6.8) можна записати як

$$\hat{y} = \hat{a}_0 + \hat{a}_1 x_1 + \hat{a}_2^* x_2 + \hat{a}_3 x_3 + \hat{a}_4 x_4$$

або

$$\hat{y} = \hat{a}_0 + \hat{a}_1 x_1 + \hat{a}_3 x_3 + \hat{a}_4 x_4 + \hat{a}_{14}^* x_1 x_4 .$$

Таким чином, маємо *змішування оцінок*.

*За наявності взаємодії необхідно до проведення експерименту перевіряти матрицю на відсутність змішування коефіцієнтів при факторах першої степені та передбачуваних суттєвих взаємодій.*

Для встановлення факту змішування оцінок коефіцієнтів моделі немає необхідності на основі неправильно вибраного ГС будувати матрицю плану, доповнювати її векторами-стовпцями взаємодій, а потім переконуватися в тому, що обране ГС не дає змоги окремо оцінити вплив можливих факторів і їхніх взаємодій на вихідну величину – оцінки коефіцієнтів будуть мішаними.

Для того, щоб правильно вибрати додатковий фактор і ГС для нього, необхідно:

1) до складання матриці плану визначитися, які взаємодії факторів можуть бути суттєвими, і включити їх у модель;

2) перевірити, чи будуть мішаними при обраному ГС взаємодії з факторами, використовуючи *визначальний контраст.*

***Визначальний контраст*** (ВК) – це формальна рівність, що показує, знаки яких незалежних змінних необхідно перемножити, щоб у всіх рядках матриці ДФЕ дістати +1.

Визначальний контраст будується на підставі генеруючого співвідношення шляхом множення обох його частин на додатковий фактор, тобто на ліву частину ГС.

Для розглянутого прикладу:

$$\text{ГС: } x_4 = x_1x_2;$$

$$\text{ВК: } x_4^2 = x_1x_2x_4.$$

З огляду на те, що $x_j^2 = 1$, ВК: $1 = x_1x_2x_4$.

Для того щоб перевірити, чи змішуватиметься взаємодія з якимось із факторів, необхідно помножити цю взаємодію на визначальний контраст. Для розглянутого прикладу: помножимо обидві частини ВК на взаємодію $x_1x_4$:

$$x_1x_4 = x^2{}_1x_2x_4^2 \quad \text{або} \quad x_1x_4 = x_2.$$

Останнє співвідношення означає, що в розглянутому випадку чергування знаків у стовпці $x_1x_4$ повністю збігається з чергуванням знаків у векторі-стовпці $x_2$, що й спостерігалося при побудові матриці ДФЕ, тобто в цьому випадку ГС обрано неправильно.

Як ГС можна було б узяти й потрійну взаємодію, тобто $x_4 = x_1x_2x_3$. При цьому, як правило, відбуватиметься змішування не тільки з подвійною, а й з потрійною взаємодією. Проте чим більший порядок взаємодії (більше елементів у цій взаємодії), тим менш імовірно, що вона суттєва.

*Отже, якщо модель містить лінійні взаємодії факторів, потрібно правильно вибирати ГС, використовуючи ВК для перевірки незалежності знаходження оцінок коефіцієнтів. Правильний вибір ГС є визначальним при подальшому тлумаченні результатів експерименту.*

**Метод «перевалу» для знаходження незалежних оцінок.** Якщо умови проведення експерименту не можуть підтримуватися постійними (зміна температури, вологості довкілля), це може призвести до зсуву оцінок коефіцієнтів, тому свідомо йдуть на розбиття повної матриці на дробові. При цьому уникають впливу дрейфу факторів, але це може призвести до змішування оцінок.

Для того щоб знаходити оцінки взаємодій *незалежно*, одну частину матриці вибирають при додатному ГС, а другу – при від'ємному.

Припустимо, що модель має вигляд:

$$y = a_0 + \sum_{j=1}^{4} a_j x_j + a_{14} x_1 x_4 + a_{24} x_2 x_4 .$$

Розбиваємо матрицю повного факторного експерименту на дві матриці дробового факторного експерименту за зазначеним правилом.

Виберемо додатне ГС1: $x_4 = x_1 x_2$. Для ГС1 визначальний контраст для перевірки змішування оцінок буде ВК1: $1 = x_1 x_2 x_4$.

Таким чином, при перевірці на змішування оцінок коефіцієнтів дістаємо $x_1 x_4 = x_2$, тобто $x_1 x_4$ змішано з $x_2$. Перевіряємо другу взаємодію: $x_2 x_4 = x_1$, тобто $x_2 x_4$ змішано з $x_1$.

Тоді сумарні оцінки впливу будуть:

$$\hat{a}_2^{\,*} = (\hat{a}_2 + \hat{a}_{14})$$

або

$$\hat{a}_{14}^{\,*} = (\hat{a}_{14} + \hat{a}_2).$$

А також

$$\hat{a}_1^{\,*} = (\hat{a}_1 + \hat{a}_{24})$$

або

$$\hat{a}_{24}^{\,*} = (\hat{a}_{24} + \hat{a}_1).$$

Проводимо другу частину експерименту. Вибираємо від'ємне ГС2: $x_4 = -x_1 x_2$. Для ГС2 визначальний контраст ВК2 $1 = -x_1 x_2 x_4$.

Тоді, виходячи з наявного змішування оцінок, маємо:

$$\hat{a}_2^{\,**} = (\hat{a}_2 - \hat{a}_{14}); \quad \hat{a}_{14}^{\,**} = (\hat{a}_{14} - \hat{a}_2).$$

Аналогічно знаходимо:

$$\hat{a}_1^{\,**} = (\hat{a}_1 - \hat{a}_{24}); \; \hat{a}_{24}^{\,**} = (\hat{a}_{24} - \hat{a}_1).$$

Для знаходження незалежно оцінок коефіцієнтів при факторах і взаємодіях необхідно обчислити півсуму та піврізницю знайдених мішаних оцінок:

$$\frac{\hat{a}_2^* + \hat{a}_2^{**}}{2} = \frac{\hat{a}_2 + \hat{a}_{14} + \hat{a}_2 - \hat{a}_{14}}{2} = \frac{2\hat{a}_2}{2} = \hat{a}_2;$$

$$\frac{\hat{a}_2^* - \hat{a}_2^{**}}{2} = \hat{a}_{14}.$$

Аналогічно знаходимо:

$$\frac{\hat{a}_1^* + \hat{a}_1^{**}}{2} = \hat{a}_1; \qquad \frac{\hat{a}_1^* - \hat{a}_1^{**}}{2} = \hat{a}_{24}.$$

У результаті дістаємо незміщені оцінки коефіцієнтів.

**Побудова матриці ДФЕ із двома додатковими факторами.**

У разі значної кількості факторів і передбачуваного лінійного вигляду моделі надмірність ПФЕ різко зростає.

**Приклад**. Модель має вигляд

$$y = a_0 + \sum_{j=1}^{5} a_j x_j .$$

При кількості факторів $n = 5$ цієї моделі необхідно за результатами експерименту оцінити лише шість коефіцієнтів. З цією метою достатньо провести $2^{n-p} > 6$ дослідів. Виходячи з мінімальної надмірності ДФЕ, вибираємо $N = 8$, тобто $(n - p) = 3$. З огляду на те, що $n = 5$, кількість додаткових факторів $p = 2$, а ДФЕ становить чверть від ПФЕ. Оскільки додаткових факторів при цьому буде два, то необхідно вибирати два генеруючі співвідношення.

Система змішування оцінок у цьому разі буде складнішою, ніж з одним додатковим фактором. Для знаходження системи змішування також використовуються визначальні контрасти.

Нехай для побудови матриці $2^{5-2}$ використовують генеруючі співвідношення:

ГС1: $x_4 = x_1x_2$ ; ГС2: $x_5 = x_1x_2x_3$.

Перш ніж будувати матрицю дробового експерименту й проводити власне експеримент, необхідно перевірити правильність вибору ГС1 і ГС2.

Для цього необхідно, виходячи із ГС, скласти вирази для визначальних контрастів:

ВК1:1= $x_1x_2x_4$; ВК2:1= $x_1x_2x_3x_5$ .

Якщо перемножити між собою ці ВК, то згідно з означенням дістанемо ще один ВК:

$$1= x_3x_4x_5.$$

Для знаходження системи змішування оцінок коефіцієнтів використовується *узагальнений визначальний контраст* (УВК), який будується на основі вихідних ВК.

***Узагальнений визначальний контраст*** (УВК) – це рівність, що містить усі можливі комбінації визначальних контрастів, тобто показує, які комбінації знаків незалежних змінних у всіх рядках матриці дробового експерименту дадуть одиницю.

У загальному випадку УВК містить у собі вихідні визначальні контрасти, отримані із ГС, та їхні добутки по два, по три і т. д.

Для розглянутого прикладу УВК запишеться в такий спосіб:

$$1 = x_1x_2x_4 = x_3x_4x_5 = x_1x_2x_3x_5.$$

*Таким чином, отримане співвідношення містить у собі всі можливі комбінації факторів, добуток яких дає* +1. *Тим самим можна дістати всі можливі варіанти змішування оцінок при факторах та їхніх взаємодіях для обраних ГС.*

Множачи УВК послідовно на $x_j$; $j$ = 1,...,5, дістаємо для наведеного прикладу такі співвідношення:

$$x_1 = x_2x_4 = x_1x_3\ x_4x_5 = x_2x_3\ x_5;$$
$$x_2 = x_1x_4 = x_2x_3\ x_4x_5 = x_1x_3\ x_5;$$
$$x_3 = x_1x_2\ x_3x_4 = x_4x_5 = x_1x_2\ x_5;$$
$$x_4 = x_1x_2 = x_3x_5 = x_1x_2x_3\ x_4\ x_5;$$
$$x_5 = x_1x_2x_4\ x_5 = x_3\ x_4 = x_1x_2\ x_3,$$

за допомогою яких можна оцінити змішування незалежних змінних та їхніх взаємодій.

**Приклад.** Для оцінки $\hat{a}_1^*$ з рівняння $x_1 = x_2x_4 = x_1x_3\ x_4x_5 = x_2x_3\ x_5$; маємо змішування $\hat{a}_1^* = (\ \hat{a}_1 + \hat{a}_{24} + \hat{a}_{1345} + \hat{a}_{235})$.

*Якщо оцінки коефіцієнтів при факторах $\hat{a}_j$ змішані з оцінками коефіцієнтів при значущих парних взаємодіях, то ГС обрано неправильно, або ж даний ДФЕ не дає змоги окремо оцінити коефіцієнти.*

*У цьому випадку необхідно спочатку добудувати отриману чверть матриці до півматриці ПФЕ, а якщо це не допоможе, то добудувати до повної матриці.*

Статистична обробка результатів ДФЕ аналогічна обробці при ПФЕ.

### Запитання для самоперевірки

1. В яких випадках застосовують дробовий факторний експеримент? Як визначається кількість дослідів при ДФЕ?
2. Які властивості повинна мати матриця ДФЕ?
3. Яке правило побудови матриці ДФЕ?
4. Що таке генеруюче співвідношення та визначальний контраст? Як і для чого вони визначаються?
5. В якому випадку може спостерігатися змішування оцінок коефіцієнтів? Як можна уникнути змішування оцінок коефіцієнтів?
6. У чому полягає суть методу «перевалу»?
7. Що таке узагальнюючий визначальний контраст? В яких випадках він застосовується?
8. Як будуються матриці ДФЕ для моделі з двома взаємодіями?

# 6.3. ОБРОБКА ДАНИХ АКТИВНОГО ЕКСПЕРИМЕНТУ

Обробка експериментальних даних має бути спрямована на визначення регресійної моделі, яка б за вибраним критерієм, найкраще відповідала наявним даним

Реєстрація результатів вимірювання вихідної величини *Y* повинна відповідати реально забезпечуваній в досліді точності вимірювання.

Установка рівнів факторів $X_i$ здійснюється згідно плану і має відбуватися у відповідності до теоретичних передумов регресійного аналізу і бути найбільш точною.

Якщо немає впевненості, що умови проведення дослідів сталі або відсутні дані про характеристики похибок застосованих засобів вимірювань, то досліди у кожній точці факторного простору дублюються (проводиться серія дослідів).

Нехай у кожній точці факторного простору, якій відповідає один з рядків матриці плану, проводиться *m* спостережень. У результаті одержують $\hat{y}_{iu}$, де *i* – рядок матриці (точка факторного простору), *u* – поточний номер спостереження в *i*-ій точці.

Для будь-якої *i*-ої точки обчислюється середнє арифметичне значення вихідної величини

$$\overline{y_i} = \sum_{u=1}^{m} \overline{y_{iu}} / m$$

і оцінку дисперсії вихідної величини по рядках (точніше її оцінку):

$$S^2\{y_i\} = \sum_{u=1}^{m} (y_{iu} - \overline{y}_i)^2 / (m-1).$$

Знайдені таким чином оцінки рядкових дисперсій використовуються для перевірки відтворюваності експериментів, що полягає в перевірці однорідності рядкових дисперсій - одного з основних передумов багатофакторного регресійного аналізу. Надалі

ми розглядатимемо етапи обробки результатів експерименту на прикладі двох факторного експерименту, результати якого представлені у наступній таблиці (табл. 6.3).

Таблиця 6.3. Обчислення двохфакторної моделі

| № п/п | $x_0$ | $x_1$ | $x_2$ | $x_1x_2$ | $y_{1i}$ | $y_{2i}$ | $y_{3i}$ | $\overline{y_1}$ | $S^2\{y_i\}$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | +1 | -1 | -1 | +1 | 43 | 35 | 48 | 42 | 43 |
| 2 | +1 | +1 | -1 | -1 | 90 | 86 | 94 | 90 | 16 |
| 3 | +1 | -1 | +1 | -1 | 10 | 16 | 16 | 14 | 12 |
| 4 | +1 | +1 | +1 | +1 | 56 | 54 | 58 | 56 | 4 |

Визначимо середнє значення вихідної величини $\bar{y}_i$ кожній точці (для кожного рядка $m = 3$):

$$\tilde{y}_1 = (43+35+48)/3 = 42;$$

$$\bar{y}_2 = (90+86+96)/3 = 90.$$

$$\overline{y_3} = (10+16+16)/3 = 14;$$

$$\overline{y_4} = (56+54+58)/3 = 56,$$

а також дисперсію вихідної величини для кожного рядка окремо:

$$S^2\{y_1\} = \left[(43-42)^2 + (35-42)^2 + (48-42)^2\right]/2 = 43;$$

$$S^2\{y_2\} = \left[(90-90)^2 + (86-90)^2 + (94-90)^2\right]/2 = 16;$$

$$S^2\{y_3\} = \left[(10-14)^2 + (16-14)^2 + (16-14)^2\right]/2 = 12;$$

$$S^2\{y_4\} = \left[(56-56)^2 + (54-56)^2 + (58-56)^2\right]/2 = 4.$$

Отримані результати занесені до табл. 6.3.

Серед усієї множини обчислених рядкових дисперсій обирається максимальна $S^2\{y_i\}_{\max}$, після чого визначається співвідношення цієї дисперсії до суми всіх порядкових дисперсій $S^2\{y_i\}$. Це значення використовується для обчислення коефіцієнта Кохрена:

$$G_p = \frac{S^2\{y_i\}_{\max}}{\sum_{i=1}^{N} S^2\{y_i\}},$$

показник показує, яку частину від загальної суми рядкових дисперсій становить максимальна серед них – ця частка використовується як міра відмінності між дисперсіями. У випадку ідеальної однорідності рядкових дисперсій коефіцієнт $G_p$ мав би тенденцію до значення $1/N$. Обчислене значення коефіцієнта Кохрена порівнюється з табличним (критичним) значенням *G*-критерію, яке вибирається з таблиць для прийнятого рівня значущості α та відповідно чисел степенів вільності чисельника $f_1$ та знаменника $f_2$:

Звертаючись до табл. *Д* 6 Додатку, вибираємо стовпець $f_1 = (m-1)$ та рядок $f_2 = N$ знаходимо табличне значення коефіцієнта Кохрена, який відповідає $G_{кр}$.

Якщо виконується умова

$$G_p < G_T \tag{6.9}$$

то, з обраним рівнем статистичної значимості α (з достовірністю $(1-\alpha)$, можна стверджувати, що всі рядкові дисперсії вважаються однорідними, тобто умови проведення дослідів були однорідні. В іншому випадку слід відхилити гіпотезу про однорідність рядкових дисперсій, що є порушенням одного з основних припущень регресійного аналізу - подальша статистична обробка результатів експерименту не має сенсу. У випадку створення такої ситуації необхідно збільшити кількість паралельних дослідів або провести експеримент знову, звернувши особливу увагу на правильність та точність установки рівнів вхідних факторів, а також застосувати більш точні прилади чи методи вимірювання.

За даними табл. 6.3 максимальна рядкова дисперсія була отримана в першому експерименті. Визначимо розрахункове значення коефіцієнта.

$$G_p = 43/(43+16+12+4) = 0,57.$$

Відповідно до таблиці, наведеної в додатку (П.1) для $\alpha = 0,05$; $f_1 = 3-1 = 2$ і $f_2 = 4$, знаходимо $G_T = 0,77$; $G_T > G_p$, тобто умова (6.9) виконується.

Після переконання в однорідності, переходять до визначення оцінок коефіцієнтів за формулою

$$\hat{a}_k = \sum_{i=1}^{N} \hat{y}_{ik} x_{ik} / N,$$

де *k* – номер вектора-стовпця.

Для цього використаємо дані, наведені в табл. 6.3. Отримаємо:

$$\hat{a}_o = (42+90+14+56)/4 = 50,5;$$

$$\hat{a}_1 = (-42+90-14+56)/4 = 22,5;$$

$$\hat{a}_2 = (-42-90+14+56)/4 = -15,5;$$

$$\hat{a}_{12} = (42-90-14+58)/4 = 1,5.$$

Знайдені таким чином коефіцієнти рівняння регресії потрібно оцінити з точки зору статистичної значущості. Оцінка проводиться за допомогою *t*-критерію Стьюдента.

Для кожного коефіцієнта $\hat{a}_k$ обчислюється коефіцієнт Стьюдента:

$$tk = |\hat{a}_k - a_k| / S\{\hat{a}_k\},$$

де $S\{\hat{a}_k\}$ – оцінка середнього квадратичного відхилення визначених коефіцієнтів, $a_k$ – математичне сподівання коефіцієнтів, яке, у даному випадку допускається, що дорівнює нулю. Невідомим у

цьому виразі є оцінка дисперсії знайденої за експериментальним даними оцінки коефіцієнта регресії.

Відповідно можна записати:

$$S\{\hat{a}_k\} = S^2\left\{\frac{1}{N}\sum_{i=1}^{N} x_{ik}\hat{y}_i\right\} = \frac{1}{N^2}S^2\left\{\sum_{i=1}^{N} x_{ik}\hat{y}_i\right\}.$$

Приймаючи до уваги, що $N = const$, а $x_{ik} = \pm 1$, останній вираз можна записати:

$$S^2\{\hat{a}_k\} = \frac{1}{N^2}S^2\left\{\sum_{i=1}^{N}\overline{y}_i\right\} = \frac{1}{N^2}\sum_{i=1}^{N}S^2\{\overline{y}_i\}.$$

Відомо, що дисперсія середнього в $m$ разів (кратність проведення дослідів) менша за дисперсію одного вимірювання, тому

$$S^2\left\{\overline{y_i}\right\} = \frac{S^2\{y_i\}}{m} \tag{6.10}$$

або, з врахуванням, маємо:

$$S^2\left\{\overline{y_i}\right\} = \frac{1}{m}\frac{1}{(m-1)}\sum_{u=1}^{m}\left(\tilde{y}_{iu} - \overline{y_i}\right)^2.$$

На підставі вищенаведеного можна записати

$$s^2\{\hat{a}_k\} = \frac{1}{N^2 m}\frac{1}{(m-1)}\sum_{i=1}^{N}\sum_{u=1}^{m}\left(\tilde{y}_{iu} - \overline{y}_i\right)^2.$$

Розглянемо оцінку дисперсії відтворюваності $S_в^2$, яка характеризує точність одного (усередненого) вимірювання і є середньою із усіх вибіркових дисперсій в точках факторного простору

$$s_в^2 = \frac{s^2\{y_i\}}{N} = \frac{1}{N}\frac{1}{(m-1)}\sum_{u=1}^{N}\left(y_{iu} - \overline{y}_i\right)^2.$$

Дисперсія відтворюваності $S_в^2$ ураховує вплив випадкових величин, що призводить до похибки вимірювання вихідної величини.

*У деяких випадках, коли є впевненість, що дисперсії однорідні (мають властивість відтворюваності), оцінкою дисперсії відтворюваності може слугувати одна з рядкових дисперсій або оцінка дисперсії для будь-якої точки факторного простору (зазвичай це відбувається в центрі плану).*

Таким чином, оцінку дисперсії коефіцієнта можна записати у вигляді:

$$S^2\{\hat{a}_k\} = \frac{S_B^2}{Nm}.$$

Обчисливши у такий спосіб $S^2\{\hat{a}_k\}$, можна визначити $t_{Рк}$, які порівнюють із критичним значенням $t_{кр}$. З огляду на те, що плани першого порядку мають властивість нормування, оцінка дисперсії для всіх коефіцієнтів буде однаковою, тому й $t_{кр}$ буде одне й те саме.

Якщо $t_{Рк} < t_{кр}$, гіпотеза приймається, тобто з прийнятим рівнем статистичної значущості α *k*-й коефіцієнт вважається статистично незначущим, і його вилучають з рівняння регресії. Інакше кажучи, не можна виділити вплив $x_k$ фактора або взаємодій на фоні впливу випадкових величин, тому немає сенсу включати цей фактор у модель об'єкта. З огляду на те, що ПФЕ має властивості ортогональності, вилучення будь-якого коефіцієнта з рівняння регресії не вплине на інші коефіцієнти.

Суть перевірки статистичної значущості знайдених оцінок коефіцієнтів полягає в наступному. Зміна вихідної величини залежить від впливу *k*-го члена апроксимуючого полінома та некерованих і неконтрольованих факторів. Вплив *k*-го фактора, відхилення оцінки *k*-го коефіцієнта від нуля, враховується коефіцієнтом:

$$t_k = |a_k| / S\{\hat{a}_k\}.$$

Вплив некерованих або неконтрольованих факторів, а також похибки вимірювання вихідної величини може бути врахований за допомогою дисперсії відтворюваності $s_в^2$, яка має $N(m-1)$ степенів вільності ($N$ степенів вільності «втрачені» при обчисленні середніх по рядках). За вибраним рівнем статистичної значимості $\alpha$, з таблиць розподілу Стьюдента при числі степенів вільності $f = N(m-1)$ знаходять табличне значення коефіцієнта. Знайдене табличне значення порівнюється з розрахунковим значенням коефіцієнта. Якщо виконується нерівність

$$t_{табл} > t_k, \tag{6.11}$$

то приймається нуль-гіпотеза, тобто з прийнятим рівнем статистичної значимості $\alpha$ (статистичною достовірністю $1-\alpha$) і кількістю степенів вільності $f = N(m-1)$ вважається, що знайдений коефіцієнт $\hat{a}_k$ є статистично незначущим і його слід виключити з рівняння регресії.

Отже, при виконанні умови (6.11) не можна визначити (у $100-\alpha$ випадках), чому змінюється вихідна величина: чи впливом $k$-го члена рівняння регресії, чи впливом неврахованих факторів та наявністю випадкової похибки вимірювання вихідної величини.

Для розглянутого прикладу оцінка дисперсії відтворюваності, як оцінка усереднених рядкових дисперсій відповідно до табл. 6.3, буде:

$$S_в^2 = \sum_{i=1}^{N} S^2\{y_i\} \Big/ N = (43+16+12+4)/4 = 18,75.$$

Як уже було зазначено, завдяки властивості нормування оцінки коефіцієнтів будуть знайдені з однаковою дисперсією, тобто

$$S^2\{\hat{a}_k\} = S_в^2 / N_m = 18,75/4 \cdot 3 = 1,56.$$

Тоді

$$S\{\hat{a}_k\} = 1,25.$$

Визначимо розрахункове значення статистики Стьюдента $t_R$ для знайдених оцінок коефіцієнтів $\hat{a}_k$:

$$t_0 = |a_0| / S\{\hat{a}_k\} = 50,5/1,25 = 40,4.$$

Аналогічно отримаємо $t_0 = 22,5/1,25 = 18;$ $t_2 = 15,5/1,25 = 12,4;$ $t_{12} = 1,5/1,25 = 1,2.$

З таблиці в додатку Д.2, при рівні статистичної значущості $\alpha$ – 5 % і кількості степенів вільності $S_в = N(m-1) = 4(3-1) = 8,$ визначимо табличне значення коефіцієнта. Воно дорівнює $t_T = 2,3$. Порівняємо розрахункові значення $t_k$ з табличним $t_T$. Нерівність (6.11) виконується для $t_{12}$. Отже, можна припустити, що коефіцієнт $\hat{a}_{12}$ є статистично незначущим і його можна виключити з рівняння регресії – в розглянутому випадку (для цього об'єкта) вплив парної взаємодії відсутній або незначний.

Однак, перш ніж прийняти гіпотезу $\hat{a}_k = 0,$ необхідно переконатися в правильності постановки експерименту.

Може виявитися, що вибір діапазону зміни незалежної змінної $(X_{k\max} - X_{k\min})$ невеликий, а загальний випадковий шум, накладений на вихідну величину об'єкта, великий. Це також може призвести до статистичної незначущості коефіцієнта. Переконавшись, що з цього погляду експеримент проведений правильно (взяти більш точний вимірювальний прилад, збільшити кількість паралельних експериментів), можна виключити коефіцієнт $\hat{a}_k$ з рівняння регресії. Оскільки повний факторний експеримент має властивість ортогональності, то виключення цього коефіцієнта з рівняння регресії не вплине на знайдені оцінки інших коефіцієнтів.

Для кожного коефіцієнта $\hat{a}_k$ можна визначити *довірчий інтервал*, що накриває істинне значення коефіцієнта $a_k$ із прийнятим рівнем статистичної значущості, для чого застосовують формулу

$$\hat{a}_k - t_m S\{\hat{a}_k\} \le a_k < \hat{a}_k + t_m S\{\hat{a}_k\}.$$

Отриману модель необхідно перевірити на *адекватність* досліджуваному об'єкту, тобто встановити наскільки добре вона відповідає отриманим експериментальним даним.

Висувають гіпотезу $H_0$: модель адекватна. Як статистичну характеристику використовують *F*-статистику, розрахункове значення якої заходять за виразом

$$F_p = \frac{S^2_{ад}}{S^2_в}.$$

Відмінність дисперсії адекватності від нуля може бути обумовлено двома причинами:

1) передбачувана модель дійсно не адекватна фізичному об'єкту (неправильно обраний апроксимуючий поліном);
2) через вплив випадкових величин й обмежений обсяг вибірки.

Для цього потрібно оцінити, наскільки відрізняються середні значення $\hat{y}_i$ – вихідної величини, отриманої в точках факторного простору внаслідок проведення експериментів, від значень $\hat{y}_i$, отриманих з рівняння регресії в тих же точках факторного простору.

Дисперсію адекватності обчислюють за формулою

$$S^2_{ад} = \frac{m}{N-i}\sum_{i=1}^{N}\left(\overline{y}_k - \hat{y}_i\right)^2, \qquad (6.12)$$

де *m* – число паралельних дослідів в *i*-ій точці факторного простору; *l* – число статистично значущих коефіцієнтів моделі; $\overline{y}_i$ – середні значення вихідної величини; $\hat{y}_i$ – значення, отримані шляхом підстановки відповідного знаку $x_{ik}$ (рівня зміни *j*-того фактора або суттєвих взаємодій).

Таким чином можна оцінити відміну середніх значень $\overline{y}_i$ вихідної величини від значень $\hat{y}_i$, отриманих з рівняння регресії в тих же точках факторного простору.

Відмінність $S_{ад}^2$ від нуля пояснюється, в загальному випадку, двома причинами: фактичною неадекватністю рівняння регресії фізичному об'єкту (неправильно обраний апроксимуючий поліном) та наявністю випадкової помилки сприйняття, що характеризується $S_в^2$. Якщо модель адекватна, то оцінка дисперсії адекватності, так само як і оцінка дисперсії відтворюваності, залежать лише від помилки сприйняття вихідної величини, обумовленої сумарними шумами, і у межах будуть однаковими. Тому адекватність отриманої моделі перевіряють шляхом порівняння оцінок двох дисперсій $S_{ад}^2$ і $S_в^2$ та за допомогою критерію Фішера *F:*

$$F_p = S_{ад}^2 / S_в^2 .$$

Розраховане значення $F_p$, порівнюють з табличним значенням $F_T$, яке визначається на рівні статистичної значимості $\alpha$ та кількості степенів вільності $f_{ад} = N - l$ та $f_{ад} = N(m-l)$, обраними в горизонтальному та вертикальному заголовках таблиці відповідно.

Якщо

$$F_p < F_T , \tag{6.13}$$

отримана математична модель з обраним рівнем статистичної значимості $\alpha$ є адекватною експериментальним даним і може бути використана для подальших досліджень. Якщо модель адекватна, то оцінка дисперсії адекватності, так само як і оцінка дисперсії відтворюваності, залежать лише від помилки сприйняття вихідної величини, обумовленої сумарними шумами, і у межах будуть однаковими. Таким чином, $F_p$ дозволяє з'ясувати причину відмінності від нуля $S_{ад}^2$

Знайдене розрахункове значення $F_p$ порівнюється з критичним значенням $F_{кр}$, що визначається за таблицями

*F*-розподілу при рівні статистичної значущості α і числі степенів вільності

$$f_{ad} = N - lu = N(m-1).$$

Якщо $F_p < F_{кр}$, гіпотеза $H_0$ приймається, тобто на підставі наявних дослідних даних можна затверджувати, що модель адекватна об'єкту. Якщо ні, треба змінювати гіпотезу (вводити додаткові фактори, взаємодії тощо).

Переконавшись в адекватності моделі, необхідно представити її в *натуральній системі координат* – використаємо формулу кодування.

Коли дисперсія відтворюваності визначається тільки за проведенням додаткових $m_0$ дослідів у центрі плану, для її обчислення використовують вираз:

$$S^2_{0=} s^2_B = \frac{1}{(m-1)} \sum_{i=1}^{m_0} (\tilde{y}_{0u} - \overline{y}_0)^2 .$$

Це дещо змінює розрахункові формули при статистичній обробці. Так, враховуючи, що $m = 1$ і критичне значення коефіцієнта Стьюдента береться з таблиці для числа степенів вільності $f_в = m_0 - 1$:

$$S^2\{\hat{a}_k\} = S^2_в / N_m .$$

Повернемося до розглянутого прикладу. Було отримано уточнене рівняння регресії $\hat{y} = 50,5 + 22,5x_1 - 15,5x_2$. Визначимо оцінку дисперсії адекватності для отриманої моделі. Спочатку розрахуємо значення $\hat{y}_i$, що відповідає рядкам матриці плану.

Початок форми:

$$\hat{y}_1 = 50,5 + 22,5 \cdot (-1) - 15,5 \cdot (-1) = 43,5;$$

$$\hat{y}_2 = 50,5 + 22,5 \cdot (+1) - 15,5 \cdot (-1) = 88,5;$$

$$\hat{y}_3 = 50,5 + 22,5 \cdot (-1) - 15,5 \cdot (+1) = 12,5;$$

$$\hat{y}_4 = 50,5 + 22,5 \cdot (+1) - 15,5 \cdot (+1) = 57,5.$$

Розрахуємо згідно з формулою (6.12) оцінку дисперсії адекватності:

$$S_{ад}^2 = 3\Big[(42-43,5)^2 + (90-88,5)^2 + (14-12,5)^2 + \\ + (56-57,5)^2\Big] \Big/ (4-3) = 27.$$

Отримане значення оцінки дисперсії адекватності $S_{ад}^2 = 27$ розділимо на оцінку дисперсії відтворюваності $S_в^2 = 18,75$ і отримаємо розрахункове значення коефіцієнта Фішера $S_{ад}^2 = 27/18,75 = 1,44.$

Табличне значення коефіцієнта Фішера (див. п. 3) при рівні статистичної значимості $\alpha = 0,05$ і числі степенів вільності $f_{ад} = 4-3 = 1$ та $f_в = N(m-1) = 4(3-1) = 8$ буде $F_T = 5,32$. Отже, при обраному рівні статистичної значимості $\alpha = 0,05$ отримане в результаті експерименту $\hat{y} = 50,5 + 22,5x_1 - 15,5x_2$ адекватно відображає досліджуваний об'єкт.

Слід зауважити, що ця модель представлена у кодованій системі координат. Щоб отримати її у природній системі, необхідно використовувати формули переходу (6.9).

***Розглянемо приклад*** побудови математичної моделі за результатами експерименту. Припустимо, що на об'єкт впливають три фактори, а саме:

$$X_{1\min} = 4; \qquad X_{1\max} = 8;$$
$$X_{2\min} = 10; \qquad X_{2\max} = 12;$$
$$X_{3\min} = 12; \qquad X_{3\max} = 28$$

пов'язані, у загальному випадку, з вихідною величиною наступною залежністю, яка включає як фактори у першій степені, так і їх взаємодії:

$$Y = A_0 + A_1X_1 + A_2X_2 + A_3X_3 + A_{12}X_1X_2 + A_{13}X_1X_3 + \\ + A_{23}X_2X_3 + A_{123}X_1X_2X_3.$$

Перейдемо до кодованих величин. Для цього скористаємось співвідношенням (6.1)

$$x_j = \frac{X_j - X_{j\text{сер}}}{X_{j\max} - X_{j\text{сер}}} = \frac{X_j - X_{j\text{сер}}}{X_{j\text{сер}} - X_{j\min}}.$$

Виходячи з наявних даних, обчислимо середнє значення $X_{j\,сер} = \left(X_{j\ \max} + X_{J\ \min}\right)/2$ та інтервал варіювання незалежних змінних $\Delta_j = \left(X_{j\ \max} - X_{j\,сер}\right)/2$:

$$X_{1\,сер} = 6; \qquad \Delta_1 = 2;$$

$$X_{2\,сер} = 11; \qquad \Delta_2 = 1;$$

$$X_{3\,сер} = 20; \qquad \Delta_3 = 8.$$

Після підстановки значення $X_{j\,сер}$ та $\Delta_j$ у вираз переходу від натуральної системи координат до кодованої, отримаємо рівняння кодованої моделі:

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_{12}x_1x_2 + a_{13}x_1x_3 + a_{23}x_2x_3 + a_{123}x_2x_3.$$

Оцінки коефіцієнтів цієї моделі будемо знаходити з експериментальних даних, отриманих в результаті проведення ПФЕ розмірності $2^n$, де $n = 3$. Відповідно до відомого правила, побудуємо матрицю повного трьохфакторного експерименту, яка має властивості *ортогональності, симетричності та нормованості* (табл. 6.4).

Таблиця 6.4. Обчислення трьохфакторної моделі

| № | $x_1$ | $x_2$ | $x_3$ | $\tilde{y}_i$ | $x_0$ | $x_1x_2$ | $x_1x_3$ | $x_2x_3$ | $x_1x_2x_3$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | -1 | -1 | -1 | 2 | +1 | +1 | +1 | +1 | -1 |
| 2 | +1 | -1 | -1 | 6 | +1 | -1 | -1 | +1 | +1 |
| 3 | -1 | +1 | -1 | 4 | +1 | -1 | +1 | -1 | +1 |
| 4 | +1 | +1 | -1 | 8 | +1 | +1 | -1 | -1 | -1 |
| 5 | -1 | -1 | +1 | 10 | +1 | +1 | -1 | -1 | +1 |
| 6 | +1 | -1 | +1 | 18 | +1 | -1 | +1 | -1 | -1 |
| 7 | -1 | +1 | +1 | 8 | +1 | -1 | -1 | +1 | -1 |
| 8 | +1 | +1 | +1 | 12 | +1 | +1 | +1 | +1 | +1 |

Вважається, що досліди однорідні. Тому в кожній точці факторного простору можливо виконати лише по одному досліду (серія паралельних дослідів у точках факторного простору не проводиться). Отримані значення вихідної величини $\tilde{y}_i$ наведені у відповідній графі табл. 6.4.

Для визначення оцінок коефіцієнтів рівняння регресії доповнимо матрицю плану (стовпці позначені блакитним фоном) вектор-стовпцями фіктивної змінної $x_0$ та лінійними взаємодіями факторів.

За результатами експерименту, визначимо за виразом (6.7):

$$\hat{a}_j = \frac{1}{N}\sum_{i=1}^{N} \tilde{y}_i x_{ij}$$

оцінки коефіцієнтів передбачуваної моделі:

$$a_0 = \left(2+6+4+8+10+18+8+12\right)/8 = 8{,}5;$$

$$a_1 = \left(-2+6-4+8-10+18-8+12\right)/8 = 2{,}5;$$

$$a_2 = \left(-2-6+4+8-10-18+8+12\right)/8 = -0{,}5;$$

$$a_3 = \left(-2-6-4-8+10+18+8+12\right)/8 = 3{,}5;$$

$$a_{12} = \left(2-6-4+8+10-18-8+12\right)/8 = -0{,}5;$$

$$a_{13} = \left(2-6+4-8-10+18-8+12\right)/8 = 0{,}5;$$

$$a_{23} = \left(2+6-4-8-10-18+8+12\right)/8 = -1{,}5;$$

$$a_{123} = \left(-2+6+4-8+10-18-8+12\right)/8 = -0{,}5.$$

Щоб визначити оцінку дисперсії відтворюваності, а також задля більш достовірної перевірки адекватності отриманої моделі, у центрі плану було додано додаткову серію з $p=3$ дослідів та отримані наступні значення:

$$\tilde{y}_{10} = 8{,}0; \quad \tilde{y}_{20} = 9{,}0; \quad \tilde{y}_{30} = 8{,}8.$$

Середнє значення вихідної величини у центрі плану $\left(x=0\right)$:

$$\tilde{y}_0 = (8+9+8{,}8)/3 = 8{,}6,$$

а дисперсія у центрі плану, яка відповідає оцінці дисперсії відтворюваності, визначається як

$$S^2\{y_0\} = \left|(8{,}6-8)^2 + (9-8{,}6)^2 + (8{,}8-8{,}6)^2\right| / (3-1) = 0{,}28.$$

Оскільки виконується умова нормованості, оцінки коефіцієнтів отриманої моделі будуть знайдені з однаковою дисперсією, тобто

$$S^2\{\hat{a}_k\} = S_в^2 / N \cdot 1 = 0{,}28/8 = 0{,}035,$$

кратність дослідів в кожній *i*-тій точці $\left(i = \overline{1, N}\right)$ дорівнює одиниці $(m=1)$, звідси

$$S\{\hat{a}_k\} \approx 0{,}2.$$

Перевіримо статистичну значущість обчислених коефіцієнтів $\hat{a}_k$. Для цього розрахуємо значення коефіцієнта Стьюдента за виразом $t_k = |\hat{a}_k| / S\{\hat{a}_k\}$. Знайдені значення наведено нижче:

$$t_o = 42{,}5; \quad t_1 = 12{,}5; \quad t_2 = 2{,}5; \quad t_3 = 17{,}5;$$

$$t_{12} = 2{,}5; \quad t_{13} = 2{,}5; \quad t_{23} = 7{,}5; \quad t_{123} = 2{,}5.$$

Табличне (критичне) значення коефіцієнта *t*-Стьюдента будемо знаходити для $\alpha = 0{,}05$ та числі степенів вільності $(p-1) = (3-1) = 2.$ Нагадаємо, що оцінку дисперсії відтворюваності було проведено на підставі серії з $p = 3$ дослідів в одній точці (центрі плану). Це значення дорівнює $t_T = 4{,}3.$

Порівнявши табличне $t_T$ і розрахункове $t_к$ значення коефіцієнтів, встановимо, що незначущими (оскільки $t_к < t_T$) є обчислені оцінки коефіцієнтів $\hat{a}_2$, $\hat{a}_{12}$, $\hat{a}_{13}$ та $\hat{a}_{123}$.

Рівняння регресії, яке містить статистично значущі коефіцієнти, запишеться як:

$$\hat{y} = 8{,}5 + 2{,}5x_1 + 3{,}5x_3 - 1{,}5x_1x_3.$$

Отриману таким чином математичну модель необхідно перевірити на адекватність експериментальним даним. Для цього з початку знайдемо оцінку дисперсії адекватності. Кратність дослідів дорівнює одиниці, тобто $m = 1$, тому

$$S_a^2 = \sum_{i=1}^{N} \left(\tilde{y}_i - \hat{y}_i\right)^2 \Big/ \left(N - l\right),$$

де $N$ = 8 – число дослідів (рядків в матриці плану), $l$ = 4 – число статистично значимих коефіцієнтів моделі.

З початку, визначимо значення вихідної величини, виходячи з рівняння регресії, у точках $i$ від 1 до $N$ факторного простору. Для першої точки, підставляючи кодовані значення $x_1$, $x_3$ та $x_1x_3$ першого рядка матриці плану обчислимо $y_1$:

$$y_1 = 8{,}5 + 2{,}5(-1) + 3{,}5(-1) - 1{,}5(+1) = 1.$$

Аналогічно, отримаємо значення для інших точок плану, які зведемо в табл. 6.5.

Таблиця 6.5. Перевірка адекватності моделі

| | $\tilde{y}_i$ | $\hat{y}_i$ | $\tilde{y}_i - \hat{y}_i$ | $\left(\tilde{y}_i - \hat{y}_i\right)^2$ |
|---|---|---|---|---|
| 1 | 2 | 1 | 1 | 1 |
| 2 | 6 | 6 | 0 | 0 |
| 3 | 4 | 4 | 0 | 0 |
| 4 | 8 | 9 | 1 | 1 |
| 5 | 10 | 11 | 1 | 1 |
| 6 | 18 | 16 | 2 | 4 |
| 7 | 8 | 8 | 0 | 0 |
| 8 | 12 | 13 | 1 | 1 |

Таким чином, при $\sum_{i=1}^{N} \left(\tilde{y}_i - \hat{y}_i\right)^2 = 8$, а $\left(N - l\right) = \left(8 - 4\right) = 4$, знайдемо оцінку дисперсії адекватності:

$$S_{a\partial}^2 = 8/4 = 2.$$

Для перевірки відповідності моделі наявним експериментальним даним скористаємся критерієм Фішера – розрахуємо коефіцієнт Фішера:

$$F_p = S_{ад}^2 / S_в^2 = 2/0,28 = 7,14.$$

Для визначення табличного критичного значення $F_T$ будемо виходити з числа степенів вільності чисельника та знаменника $t_{ад} = (N - l) = 4$ і $t_в = (p - 1) = 2$ та $\alpha = 0,05$. Табличне значення $F_T = 19,3$ суттєво менше від $F_T = 7,14$.

Приходимо до висновку, на підставі на явних даних достовірністю $(1 - \alpha) = 95\%$ отримане рівняння регресії (математична модель) адекватно експериментальним даним.

Так як рівняння регресії представлено у кодованій системі координат, то необхідно перейти у вихідну (натуральну) систему координат, використовуючи формулу переходу для обчислених вище значень $x_{j\,сер}$ та $\Delta_j$. В результаті маємо:

$$\hat{Y} = 8,5 + 2,5\frac{X_1 - 6}{2} + 3,5\frac{X_3 - 20}{8} - 1,5\frac{X_2 - 11}{1} \cdot \frac{X_3 - 20}{8},$$

або, після проміжних математичних перетворень

$$\hat{Y} = 8,5 + 1,25X_1 - 7,5 + 0,44X_3 - 8,75 - 0,19 \cdot X_2X_3 + \\ +3,75X_2 - 2,06X_3 - 41,25.$$

Отримаємо остаточне рівняння регресії:

$$\hat{Y} = -49,0 + 1,25X_1 + 3,75X_2 - 1,62X_3 - 0,19X_2X_3,$$

яке адекватно описує експериментальні дані.

**Оцінка вихідної величини на підставі регресійної моделі.** Установлено, що плани першого порядку через властивість ортогональності дозволяють незалежно оцінювати вплив факторів й їхніх лінійних взаємодій. Розглянемо, з якою точністю оцінюється вихідна величина досліджуваного об’єкта на підставі отриманого рівняння регресії і як по отриманій моделі можна прогнозувати значення вихідної величини.

Припустимо, що модель лінійна й отримане рівняння регресії, що адекватно описує експериментальні дані, має вид:

$$\hat{y} = \hat{a}_0 + \sum_{j=1}^{n} \hat{a}_j x_j \ , \ j = \overline{1,n} \ .$$

Оцінка дисперсії вихідної величини буде

$$S^2\{\hat{y}\} = S^2\left\{ a_0 + \sum_{j=1}^{n} \hat{a}_j x_j \right\}.$$

Внаслідок властивості *ортогональності* (через відсутність кореляції між знайденими оцінками коефіцієнтів) можна записати (дисперсія суми незалежних величин дорівнює сумі їх дисперсій):

$$S^2\{\hat{y}\} = S^2\{\hat{a}_0\} + S^2\left\{ \sum_{j=1}^{n} \hat{a}_j x_j \right\}.$$

Згідно з властивістю *нормованості* випливає, що дисперсії оцінок коефіцієнтів однакові, тобто

$$S^2\{\hat{a}_0\} = S^2\{\hat{a}_j\} = S^2\{\hat{a}\},$$

і тоді

$$S^2\{\hat{y}\} = S^2\{\hat{a}\}\left( 1 + \sum_{j=1}^{n} x_j^2 \right),$$

де $S^2\{\hat{a}\} = \dfrac{S_в^2}{N \times 1}$.

Беручи до уваги, що $x_j$ – невипадкова величина, тобто $\sum_{j=1}^{n} x_j^2 = \text{const}$, можна записати:

$$S^2\{\hat{y}\} = S^2\{\hat{a}\}\left(1 + \rho^2\right), \qquad (6.14)$$

де $\left( \sum_{j=1}^{1} x_j^2 \right) = \rho^2$ – радіус сфери в *n-вимірному* просторі (відповідно до центра плану).

З виразу (6.14) випливає, що дисперсія вихідної величини збільшується пропорційно до квадрата радіуса сфери $\rho^2$ і однакова для всіх *еквідистантних* точок.

Таким чином, чим далі розташована точка від центра плану, тим меншу надійність мають передбачувані на підставі рівняння регресії значення вихідної величини.

Дисперсія вихідної величини, обчислена на підставі здобутої математичної моделі, буде однакова для всіх точок факторного простору, рівновіддалених від центра плану, тобто для точок, що лежать на *гіперсфері*. Вона залежить від радіуса сфери ρ і не залежить від положення точок на ній. Ця властивість називається властивістю *рототабельності*. Плани, які мають таку властивість, називаються *рототабельними* [7]. Для таких планів критерієм оптимальності є однакова точність прогнозування функції відгуку за рівнянням регресії в будь-якому напрямі досліджуваної області факторного простору, що є необхідною умовою при знаходженні екстремальної області.

*Таким чином, дворівневі ортогональні плани першого порядку дають змогу не тільки незалежно оцінювати коефіцієнти математичної моделі, а й з однаковою точністю передбачати* (*у середньому*) *значення вихідної величини для рівновіддалених від центра плану точок незалежно від напряму просування у факторному просторі.*

**Запитання для самоперевірки**

1. Що таке дисперсія відтворюваності? Що вона характеризує і як визначається?

2. У чому полягає сутність перевірки коефіцієнтів на статистичну значущість?

3. Як перевірити адекватність моделі? Які можливі причини неадекватності моделі?

4. Як одержати рівняння регресії в природній системі координат?

5. Які плани називаються рототабельними?

6. Що забезпечує властивість рототабельності планів?

7. Який алгоритм визначення дисперсії відтворюваності, коли відомо, що умови проведення дослідів однорідні?

## 6.4. ЦЕНТРАЛЬНІ КОМПОЗИЦІЙНІ ПЛАНИ ДРУГОГО ПОРЯДКУ

Повний або дробовий факторний експеримент дає змогу одержати незалежні оцінки коефіцієнтів при факторах у першому степені, а також їхніх лінійних взаємодій.

Якщо лінійна модель неадекватна і ми дійдемо висновку, що необхідно в неї ввести фактори у другому степені $x_j^2$, то виявиться, що вектори-стовпці при факторах у другому степені в усіх стовпцях будуть мати +1. У такому разі не можна буде оцінити окремо вплив не тільки $x_j^2$, а й вільного члена, для якого вектор-стовпець також містить тільки +1.

Із теорії апроксимації відомо, що для відтворення нелінійної залежності необхідно, щоб кількість точок принаймні на одну перевищувала порядок залежності.

Таким чином, для одержання регресійної моделі, що включає в себе й фактори в другому степені, необхідно скласти план активного експерименту, який передбачає щонайменше три рівні варіювання кожного фактора. Але кількість дослідів ПФЕ при $k = 3$ різко зростає, тобто ПФЕ буде мати *надмірність*. При цьому трирівневі експерименти не мають властивості ортогональності, тому

необхідно перераховувати параметри, вже отримані при реалізації матриць повного або дробового факторного експерименту.

З огляду на сказане стає актуальною задача оптимального розміщення дослідних точок, за яких кількість дослідів була б мінімальною. Досягти компромісу дають змогу *центральні композиційні плани другого порядку*.

Плани такого виду називаються *композиційними*, оскільки складаються з кількох частин. Такими частинами є:

- «***ядро***» ***плану***, яке містить матрицю ПФЕ або ДФЕ, згідно з якими вже було проведено досліди, але результати розрахунку дали неадекватну модель;
- «***зіркові***» ***точки***, розміщенні на всіх $n$ осях-факторах по обидва боки від центра плану на відстані $\alpha$, що називається ***«зірковим» плечем*** (див. рисунок);
- ***центральні точки*** (центр плану) із кодованим значенням фактора $x_j = 0$, в яких проводиться серія з $N_0$ дослідів (тому плани називаються *центральними*).

*Точки називаються зірковими, оскільки вони зазвичай позначаються хрестиками, як показано на рисунку.*
*При проведенні дослідів у «зіркових» точках одні фактори приймають значення* $-\alpha$ *або* $+\alpha$, *інші в цей час приймають значення* 0.

Отже, кількість точок факторного простору, в яких реалізуються досліди, буде:

$$N_{\text{ЦКП}} = N_{\text{яд}} + N_{\alpha} + N_0;$$

де $N_{\text{яд}} = 2^{n-p}$ (при ПФЕ $p = 0$) – кількість дослідів факторного плану на двох рівнях; $N_{\alpha} = 2n$ – кількість «зіркових» дослідів; $N_0$ – кількість дослідів у центрі плану; $n$ – кількість факторів, які введені у передбачувану модель.

Таким чином

$$N_{\text{ЦКП}} = 2^{n-p} + 2n + 1.$$

У ЦКП другого порядку здійснюється варіювання незалежних змінних на *п'ятьох* рівнях (рис. 6.3):

$$-\alpha;\ -1;\ 0;\ +1;\ +\alpha.$$

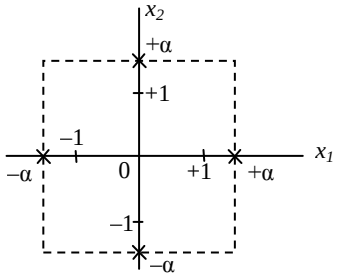



Рис. 6.3. Геометрична інтерпретація ортогонального центрального плану другого порядку

Значення «зіркового» плеча α і кількість дослідів у центральній точці $N_0$ вибирають з умови виконання *критерію оптимальності плану*. Залежно від критерію оптимальності на відміну від планів першого порядку центральні композиційні плани бувають двох видів:

1) *ортогональні*, для яких критерієм оптимальності є ортогональність усіх векторів-стовпців доповненої матриці планування, що забезпечує незалежність оцінок коефіцієнтів рівняння регресії, у тому числі й при факторах у другому степені;

2) *рототабельні*, для яких критерієм оптимальності є однакова точність прогнозування функції відгуку за рівнянням регресії в будь-якому напрямі досліджуваної області факторного простору на одній і тій самій відстані від центра плану.

**Ортогональні центральні композиційні плани**. Це такі плани, критерієм оптимальності яких буде *незалежність знаходження оцінок* коефіцієнтів моделі.

Розглянемо побудову плану двофакторого експерименту. Нехай модель має вигляд:

$$y = a_o + a_1 x_1 + a_2 x_2 + a_{12} x_1 x_2 + a_{11} x_1^2 + a_{22} x_2^2 .$$

Ядро плану (перші чотири рядки) становить матриця дворівневого ПФЕ або ДФЕ. Ядро матриці доповнюється дослідами в «зіркових» точках (досліди з 5-го по 8-й). Кількість таких дослідів 2*n*. В ОЦКП другого порядку достатньо проводити тільки один дослід у центрі плану, тобто $N_0 = 1$ (дослід 9-й).

Побудовану в такий спосіб матрицю наведено в табл. 6.6.

Таблиця 6.6. Матриця ОЦКП другого порядку

| *N* | $x_1$ | $x_2$ | $x_1x_2$ | $x_0$ | $x_1^2$ | $x_1^2$ | |
|---|---|---|---|---|---|---|---|
| 1 | –1 | –1 | +1 | +1 | +1 | 1 | ПФЕ $2^2$ |
| 2 | +1 | –1 | –1 | +1 | +1 | 1 | |
| 3 | –1 | +1 | –1 | +1 | +1 | 1 | |
| 4 | +1 | +1 | +1 | +1 | +1 | 1 | |
| 5 | –α | 0 | 0 | +1 | $\alpha^2$ | 0 | «Зіркові» точки 2*n* |
| 6 | + α | 0 | 0 | +1 | $\alpha^2$ | 0 | |
| 7 | 0 | –α | 0 | +1 | 0 | $\alpha^2$ | |
| 8 | 0 | + α | 0 | +1 | 0 | $\alpha^2$ | |
| 9 | 0 | 0 | 0 | +1 | 0 | 0 | Центр |

Як уже зазначалося, в ортогональних ЦКП другого порядку мають бути ортогональними між собою всі вектори-стовпці (включаючи добудовані вектори-стовпці квадратичних членів і вектор-стовпець фіктивної змінної), необхідні для обчислень відповідних коефіцієнтів моделі другого порядку.

Як бачимо з табл. 6.6, не всі вектори-стовпці є попарно ортогональними. Справді, наприклад:

$$\sum_{i=1}^{N} x_{i0} x_{i1}^2 \neq 0,\ j = \overline{1,k};$$

$$\sum_{i=1}^{N} x_{ij}^2 x_{iu}^2 \neq 0,\ j,u = \overline{1,k};\ j \neq u ,$$

де *j* та *u* – номери векторів-стовпців факторів у другому степені.

Для того щоб досягти ортогональності вектора-стовпця $x_0$, необхідно забезпечити властивість *симетрії*. Це здійснюється шляхом введення постійного *зсуву* β векторів-стовпців при факторах у другому степені з метою виконання умови:

$$\sum_{i=1}^{N} x_{i0}\left(x_{ij}^2 - \beta\right) = 0,\ j = \overline{1,n}\,.$$

Узявши до уваги, що $x_0$ тотожно дорівнює одиниці, дістанемо:

$$\sum_{i=1}^{N}\left(x_{ij}^2 - \beta\right) = 0\,.$$

Звідки $$N\beta = \sum_{i=1}^{N} x_{ij}^2\,,$$

або $$\beta = \frac{\sum\limits_{i=1}^{N} x_{ij}^2}{N} = \frac{\sum\limits_{i=1}^{N} x_{ij}^2}{2^{n-p} + 2n + 1} = \frac{2^{n-p} + 2\alpha^2}{2^{n-p} + 2n + 1}\,.$$

Скориставшись для знаходження $\sum\limits_{i=1}^{N} x_{ij}^2$ даними табл. 6.1, остаточно дістанемо:

$$\beta = \frac{2^{n-p} + 2\alpha^2}{2^{n-p} + 2n + 1}\,. \qquad (6.15)$$

Але навіть при введенні зсуву β перетворені вектори-стовпці факторів у другому степені не будуть між собою ортогональними. Для забезпечення їхньої ортогональності відповідним чином вибирається зіркове плече α.

Наочність вибору α легко простежити на простому прикладі.

**Приклад**. Модель має вигляд:

$$y = a_0 + \sum_{j=1}^{3} a_j x_j + a_{12} x_1 x_2 + \sum_{j=1}^{3} a_{jj} x_j^2\,.$$

Завдання полягає у виборі такого α, щоб

$$\sum_{i=1}^{N}\left(x_{ij}^2-\beta\right)\left(x_{ik}^2-\beta\right)=0\,,\ j\neq k;\ j,k=1,\ldots,N.$$

Матрицю плану ОЦКП другого порядку для цього випадку наведено в табл. 6.7. Кількість дослідів у матриці буде дорівнювати:

$$N = 2^3 + 2 \cdot 3 + 1 = 15.$$

Таблиця 6.7. Приклад побудова ортогональної матриці

| $N$ | $x_1$ | $x_2$ | $x_3$ | $x_1x_2$ | $x_0$ | $x_1^2-\beta$ | $x_2^2-\beta$ | $x_3^2-\beta$ | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | –1 | –1 | –1 | +1 | +1 | 1– β | 1– β | 1– β | "Ядро" плану $2^{n-p}_{p=0}$ |
| 2 | +1 | –1 | –1 | –1 | +1 | 1– β | 1– β | 1– β | |
| 3 | –1 | +1 | –1 | –1 | +1 | 1– β | 1– β | 1– β | |
| 4 | +1 | +1 | –1 | +1 | +1 | 1– β | 1– β | 1– β | |
| 5 | –1 | –1 | +1 | +1 | +1 | 1– β | 1– β | 1– β | |
| 6 | +1 | –1 | +1 | –1 | +1 | 1– β | 1– β | 1– β | |
| 7 | –1 | +1 | +1 | –1 | +1 | 1– β | 1– β | 1– β | |
| 8 | +1 | +1 | +1 | +1 | +1 | 1– β | 1– β | 1– β | |
| **9** | –α | 0 | 0 | 0 | +1 | $\alpha^2-\beta$ | –β | –β | "Зіркові точки" $2n$ |
| 10 | +α | 0 | 0 | 0 | +1 | $\alpha^2-\beta$ | –β | –β | |
| 11 | 0 | –α | 0 | 0 | +1 | –β | $\alpha^2-\beta$ | –β | |
| 12 | 0 | +α | 0 | 0 | +1 | –β | $\alpha^2-\beta$ | –β | |
| 13 | 0 | 0 | –α | 0 | +1 | –β | –β | $\alpha^2-\beta$ | |
| 14 | 0 | 0 | +α | 0 | +1 | –β | –β | $\alpha^2-\beta$ | |
| **15** | 0 | 0 | 0 | 0 | +1 | –β | –β | – β | центр плану |

Умова ортогональності двох перетворених стовпців у другому степені буде:

$$\sum_{i=1}^{N}\left(x_{ij}^2-\beta\right)\left(x_{ik}^2-\beta\right)=0\,.$$

Враховуючи, що

добутків виду $\left(x_j^2-\beta\right)\left(x_k^2-\beta\right) = (1-\beta)^2$ буде $2^{n-p}$;

добутків виду $(-\beta)(\alpha^2-\beta)$ буде 4;

добутків виду $(-\beta)(-\beta)$ буде $(2n - 4)$;

добутків виду (–β)( – β) в центрі плану буде 1, прийдемо до виразу:

$$2^{n-p}(1-\beta)^2 - 4\,\beta(\alpha^2-\beta) + (2n\text{-}4)\beta^2 + \beta^2 = 0, \qquad (6.16)$$

з якого випливає, що умова ортогональності перетворених векторів-стовпців факторів у другому степені виконується при певних значеннях α, β та *n*.

Підставивши в умову ортогональності (6.16) значення β з виразу (6.15), дістанемо залежність α від *n*. Узявши *n* за параметр і послідовно присвоївши йому значення 2, 3, 4, можна визначити відповідне значення α, що забезпечує ортогональність зміщених векторів-стовпців при факторах у другому степені (табл. 6.8).

Таблиця 6.8. Залежність «зіркового» плеча від числа факторів

| *n* | 2 | 3 | 4 | … |
|---|---|---|---|---|
| α | 1,0 | 1,215 | 1,414 | … |

Таким чином, увівши зсув β і вибравши відповідним чином розмір «зіркового» плеча α, можна забезпечити незалежність знаходження оцінок коефіцієнтів моделі, що включає фактори у другому степені.

Раніше зазначалося, що умова нормування забезпечує однакову точність оцінки коефіцієнтів рівняння регресії. Для ортогональних ЦКП другого порядку ця умова не виконується. Рівність, на підставі якої можна незалежно визначати оцінки коефіцієнтів, набирає вигляду:

$$\hat{a}_k = \frac{\sum_{i=1}^{N} x_{ik}\tilde{y}_i}{\sum_{i=1}^{N} x_{ik}^2} \qquad (6.17)$$

причому для різних груп коефіцієнтів знаменники у (6.17) будуть різними.

Розглянемо знаменники виразу (6.17) для кожної групи коефіцієнтів:

При визначенні вільного члена $\hat{a}_0$:

$$\sum_{i=1}^{N} x_{i0}^2 = 2^{n-p} + 2n + 1 = N;$$

– при визначенні оцінок групи коефіцієнтів при лінійних факторах $\hat{a}_j$, ( $j = \overline{1,n}$):

$$\sum_{i=1}^{N} x_{ij}^2 = 2^{n-p} + 2\alpha^2;$$

– при визначенні оцінок групи коефіцієнтів при факторах у другому степені $\hat{a}_{jj}$:

$$\sum_{i=1}^{N} \left(x^2{}_{ij} - \beta\right)^2 = 2^{n-p}(1-\beta)^2 + 2\left(\alpha^2 - \beta\right)^2 + (2n-1)\beta^2;$$

– при визначенні оцінок коефіцієнтів при лінійних взаємодіях $\hat{a}_{ju}$:

$$\sum_{i=1}^{N} \left(x_{ij} x_{iu}\right)^2 = 2^{n-p}.$$

Виходячи з виразу (6.16) і матриці ОЦКП, одержимо модель у кодованій системі координат виду

$$\hat{y} = \hat{a}_0^* + \sum_{j=1}^{n} \hat{a}_j x_j + \sum_{j,u=1, j\neq u}^{n} \hat{a}_{ju} x_j x_u + \ldots + \sum_{j=1}^{n} \hat{a}_{jj} \left(x_j^2 - \beta\right),$$

яка включає в себе зсув β.

Перетворимо цей вираз і подамо його у вигляді:

$$\hat{y} = \hat{a}_0 + \sum_{j=1}^{n} \hat{a}_j x_j + \sum_{j,u=1, j\neq u}^{n} \hat{a}_{ju} x_j x_u + \sum_{j=1}^{n} \hat{a}_{jj} x_j^2,$$

де $$\hat{a}_0 = \hat{a}_0^* - \beta \sum_{j=1}^{n} \hat{a}_{jj}.$$

З огляду на те, що властивість нормування не виконується, при обчисленні дисперсії *k*-ї групи знайдених оцінок потрібно використовувати вираз:

$$S^2\{\hat{a}_k\} = \frac{S_в^2}{m\sum_{i=1}^{N} x_{ik}^2},$$

де $m$ – кількість паралельних дослідів.

Оцінки дисперсій коефіцієнтів для кожної з чотирьох груп обчислюють за такими формулами:

$$S^2\{\hat{a}_j\} = \frac{S_в^2}{m}\left[\frac{1}{2^{n-p}+2\alpha^2}\right]; \tag{6.18}$$

$$S^2\{\hat{a}_{ju}\} = \frac{S_в^2}{m}\left[\frac{1}{2^{n-p}}\right]; \tag{6.19}$$

$$S^2\{\hat{a}_{jj}\} = \frac{S_в^2}{m}\left[\frac{1}{2^{n-p}(1-\beta)^2+2\left(\alpha^2-\beta\right)^2+(2n-1)\beta^2}\right]; \tag{6.20}$$

$$S^2\{a_0\} = S^2\{\hat{a}_0^*\} + \beta S^2\left\{\sum_{j=1}^{n}\hat{a}_{jj}\right\} = \\ = \frac{S_в^2}{m}\left[\frac{1}{2^{n-p}+2n+1} + \frac{\beta n}{2^{n-p}(1-\beta)^2+2\left(\alpha^2-\beta\right)^2+(2n-1)\beta^2}\right]. \tag{6.21}$$

Щоб оцінити статистичну значущість коефіцієнтів моделі, розраховують коефіцієнти Стьюдента для кожної із зазначених груп коефіцієнтів за формулою:

$$t_{\text{Р}\,k} = \frac{|\hat{a}_k|}{S\{\hat{a}_k\}}, \tag{6.22}$$

де $k$ дорівнює відповідно 0; $j$; $ju$; $jj$.

Визначивши значущі коефіцієнти моделі, необхідно уточнити вираз для моделі з урахуванням значущих коефіцієнтів, перевірити модель на адекватність і для адекватної моделі записати вираз її у природній системі координат.

Якщо отримане на підставі ОЦКП рівняння регресії виявилося адекватним із прийнятим рівнем значущості, його можна

подалі досліджувати аналітичними методами й уточнювати область екстремуму. Для цього необхідно взяти частинні похідні за всіма $n$ факторами, прирівняти їх до нуля й розв'язати здобуту систему $n$ рівнянь. Отриманий опис математичної області екстремуму можна також використати для прогнозування функції відгуку.

### Запитання для самоперевірки

1. В яких випадках застосовують ЦКП?
2. Чому плани називаються центральними, композиційними?
3. Що називається центром плану? Що таке «ядро» плану, «зіркові» точки?
4. Як розрахувати кількість дослідів для матриці ЦКП?
5. На скількох рівнях варіюються фактори при ЦКП?
6. Як побудувати матрицю ЦКП експерименту другого порядку?
7. Які властивості має матриця ЦКП?
8. Які характерні ознаки матриці ОЦКП?
9. Як домагаються незалежності знаходження оцінки коефіцієнта $\hat{a}_0$?
10. Від чого залежить розмір «зіркового» плеча α і з яких міркувань вибирають розміри цього плеча?
11. Як визначити розмір «зіркового» плеча α у природній системі координат?
12. Які властивості не виконуються для матриць ОЦКП і до чого це призводить?
13. Яка з груп однорідних коефіцієнтів визначається з найменшою точністю?

# РОЗДІЛ 7

# ОСНОВИ ДИСПЕРСІЙНОГО АНАЛІЗУ

У багатьох практичних завданнях необхідно встановити, наскільки суттєвим є вплив того чи іншого фактора на значення вихідної величини об'єкта.

У багатьох випадках дослідження об'єктів вплив деяких вхідних величин на вихідну неможливо оцінити кількісно. При цьому оператора може цікавити питання, наскільки суттєвим є вплив того чи іншого фактора на розкид результатів спостережень вихідної величини. Факторами, які можуть зацікавити дослідника, можуть бути верстати, що працюють паралельно, вимірювальні прилади, оператори, зміни, якість вихідної сировини, місця розташування об'єктів тощо. Для вивчення впливу таких факторів на вихідну величину, їх загального оцінювання, ранжирування та визначення серед них суттєвих недоцільно застосовувати методи регресійного аналізу, оскільки вони передбачають вимірювання рівнів досліджуваних факторів.

*Суть дисперсійного аналізу* полягає в поданні загальної дисперсії результату у вигляді суми дисперсій, які відображають можливий вплив факторів. Дисперсія тут використовується як найпростіша міра розсіювання, яка дає змогу порівнювати фактор, що вивчається, і фактор випадковості.

Припустимо, що розсіювання зумовлене спільною дією випадкових причин і зміною рівнів факторів. Тоді, отримавши оцінку загальної дисперсії відгуку й оцінки дисперсій факторів, можна знайти оцінку залишкової дисперсії (дисперсії відтворюваності), а далі, використовуючи статистичні критерії порівняння дисперсій, ранжувати фактори за степеню їхнього впливу на розсіювання відгуку.

**Загальна постановка задачі**. Нехай:

– вихідна величина (ознака, відгук) згідно з фізичними властивостями залежить від $n$ факторів, які не мають кількісного опису, та від їхніх парних взаємодій;

– кожний фактор може варіюватися на кількох рівнях (експеримент проводять кілька операторів, застосовують різні методи вимірювань тощо);

– кожну дослідну ситуацію можна спостерігати кілька разів (реалізується серія паралельних спостережень).

Потрібно визначити, якою мірою на вихідну величину впливає той чи інший якісний фактор. При цьому припускають, що ознака в загальному випадку є випадковою величиною, розподіленою за нормальним законом, дисперсія її у всіх дослідах однорідна, тобто висувається вимога виконання *умови відтворюваності* дослідів.

Залежно від кількості факторів, які враховуються при дисперсійному аналізі, статистичні моделі поділяються на *однофакторні*, *двофакторні* та *багатофакторні*.

**Однофакторний дисперсійний аналіз**. Вважається, що результат експерименту може залежати тільки від одного фактора та випадкових величин із математичним сподіванням, що дорівнює нулю.

При застосуванні однофакторного дисперсійного аналізу перевіряється гіпотеза $H_0$ про відсутність впливу, що його вносить зміна рівня досліджуваного фактора.

Альтернативною гіпотезою в цьому разі є гіпотеза $H_1$: вплив рівнів досліджуваного фактора суттєвий. Якщо приймається гіпотеза $H_0$, то це означає, що розсіювання результатів зумовлене лише похибкою експерименту.

**Приклад**. Деяка продукція постачається від різних $N$ постачальників. Необхідно з'ясувати, чи однакова якість цієї продукції, що визначається за якимось *узагальненим* показником,

тобто необхідно перевірити гіпотезу $H_0$: узагальнені показники якості однакові.

Для того щоб оцінити продукцію за результатами випробувань, для кожного її різновиду накопичуються масиви результатів. Бажано, щоб по кожному виду продукції розміри $m$ масивів і паралельність спостережень в них були однакові.

Масив даних можна подати у вигляді табл. 7.1, де $i=\overline{1,N}$ – рівень зміни досліджуваного фактора (номер постачальника); $j$ – номер поточного паралельного досліду оцінки показника (продукції), $j=\overline{1,m}$.

Таблиця 7.1. Масив даних при однофакторному дисперсійному аналізі

| $j$ / $i$ | 1 | 2 | … | $m$ | $\overline{y}_i$ |
|---|---|---|---|---|---|
| 1 | $\tilde{y}_{11}$ | $\tilde{y}_{12}$ | … | $\tilde{y}_{1m}$ | $\overline{y}_1$ |
| 2 | $\tilde{y}_{21}$ | $\tilde{y}_{22}$ | … | $\tilde{y}_{2m}$ | $\overline{y}_2$ |
| … | … | … | … | … | … |
| $N$ | $\tilde{y}_{N1}$ | $\tilde{y}_{N2}$ | … | $\tilde{y}_{Nm}$ | $\overline{y}_N$ |
| | | | | | $\overline{y}=\frac{1}{N}\sum_{i=1}^{N}\overline{y}_i$ |

Неодмінним припущенням при виконанні дисперсійного аналізу є забезпечення *однорідності умов* проведення випробувань та їхньої *відтворюваності.*

*Однорідність умов проведення випробувань характеризується однаковістю впливу факторів в усіх дослідах. Відтворюваність умов проведення випробувань характеризується для всіх різновидів продукції однаковістю умов, в яких перевіряється вплив фактора.*

*Спільна обробка результатів можлива лише в разі виконання умов однорідності та відтворюваності.*

Для реалізації умов однорідності й відтворюваності необхідно, щоб випробування виконувалися одним оператором, на одному обладнанні, протягом малого відрізка часу. Не мають значення характеристики устаткування, кваліфікація оператора, а важливо тільки те, щоб вони були ті самі.

*Перед проведенням дисперсійного аналізу необхідно провести перевірку однорідності дисперсій, для чого застосовується G-критерій.*

Розсіювання результатів випробувань, наведених у табл. 7.1, може зумовлюватися впливом випадкових величин та суттєвою розбіжністю між узагальненими показниками продукції.

Розсіювання значень для продукції, виробленої *i*-м постачальником (*i*-й рядок таблиці), визначається впливом випадкових величин. Для врахування цього впливу необхідно обчислити середнє значення ознаки:

$$\bar{y}_i = \frac{1}{m}\sum_{j=1}^{m}\tilde{y}_{ij} .$$

Обчислені в такий спосіб значення $\bar{y}_i$ можуть відрізнятися між собою через вплив випадкових величин та обмежений обсяг спостережень, а також через наявність розбіжностей між узагальненими показниками продукції від різних постачальників.

Розсіювання ознак – вихідних величин $\tilde{y}_{ij}$ – можна оцінити відносно генерального середнього:

$$\bar{y} = \frac{1}{Nm}\sum_{i=1}^{N}\sum_{j=1}^{m}\tilde{y}_{ij} = \frac{1}{N}\sum_{i=1}^{N}\bar{y}_i .$$

Розсіювання результатів випробування відносно генерального середнього буде найбільш надійною оцінкою

розсіювання, оскільки визначається за всією сукупністю спостережень:

$$Q_0^2 = \sum_{i=1}^{N}\sum_{j=1}^{m}(\tilde{y}_{ij} - \bar{y})^2 \ . \qquad (7.1)$$

Внесемо в дужки виразу (7.1) складову $\pm\bar{y}_i$, що характеризує середнє значення ознаки для продукції, яка постачається *i*-м виробником (вироблена на *i*-й установці):

$$\bar{y}_i = \frac{1}{m}\sum_{j=1}^{m}\tilde{y}_{ij} \ .$$

Розбіжність між ознаками $\tilde{y}_{ij}$ продукції від одного виробника може залежати від випадкової похибки виробництва, відтворення технологічного процесу тощо.

Після групування вираз (7.1) набере вигляду:

$$Q_0^2 = \sum_{i=1}^{N}\sum_{j=1}^{m}\left[\left(\tilde{y}_{ij} - \bar{y}_i\right) + \left(\bar{y}_i - \bar{y}\right)\right]^2$$

або після перетворення:

$$Q_0^2 = \sum_{i=1}^{N}\sum_{j=1}^{m}(\bar{y}_{ij} - \bar{y}_i)^2 + 2\sum_{i=1}^{N}\sum_{j=1}^{m}(\tilde{y}_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y}) + \sum_{i=1}^{N}\sum_{j=1}^{m}(\bar{y}_i - \bar{y})^2 \ . \qquad (7.2)$$

Другий доданок у правій частині виразу (7.2):

$$\sum_{i=1}^{N}(\bar{y}_i - \bar{y})\sum_{j=1}^{m}(\tilde{y}_{ij} - \bar{y}_i)$$

включає суму відхилень значень випадкової величини $\tilde{y}_{ij}$ відносно середнього $\bar{y}_i$, обчисленого для цих даних, і дорівнює нулю.

Тоді повна сума матиме вигляд:

$$Q_0^2 = \underbrace{\sum_{i=1}^{N}\sum_{j=1}^{m}(\tilde{y}_{ij} - \bar{y}_i)^2}_{Q_\varepsilon^2} + \underbrace{\sum_{i=1}^{N}\sum_{j=1}^{m}(\bar{y}_i - \bar{y})^2}_{Q_x^2} \ . \qquad (7.3)$$

Перший доданок $Q_\varepsilon^2$ виразу (7.3) визначається впливом випадкових факторів, що виявляється в розкиді результатів ознак відносно середнього значення цієї величини.

Другий доданок $Q_x^2$ виразу (7.3) в основному зумовлений розбіжностями між ознаками (якістю продукції). Але доданок $Q_x^2$ залежить також від впливу випадкових величин $\varepsilon$, оскільки обмежений обсяг спостережень не дає змоги повністю виключити цей вплив.

З огляду на сказане можна записати:

$$Q_0^2 = Q_\varepsilon^2 + Q_x^2.$$

Для перевірки гіпотези $H_0$ застосовуємо критерій Фішера, згідно з яким знаходиться відношення двох дисперсій:

$$F_\text{Р} = \frac{S_x^2}{S_\varepsilon^2}.$$

Тут $S_x^2$ – оцінка дисперсії, яка може залежати від впливу випадкових величин і суттєвої розбіжності якості продукції. Її можна обчислити за формулою:

$$S_x^2 = \frac{Q_x^2}{N-1},$$

де $Q_x^2 = m\sum_{i=1}^{N}(\bar{y}_i - \bar{y})^2$.

Оцінку дисперсії впливу випадкових величин в умовах відтворюваності однорідності випробувань знаходимо за формулою:

$$S_\varepsilon^2 = \frac{Q_\varepsilon^2}{N(m-1)}.$$

Обчислене значення $F_\text{р}$ порівнюється з $F_\text{кр}$, знайденим для степенів вільності $f_x = N-1$, $f_\varepsilon = N(m-1)$ та рівнем статистичної значущості $\alpha$. Якщо $F_\text{р} < F_\text{кр}$, то з імовірністю $1-\alpha$ можна для

наявних даних стверджувати, що відмінність між дисперсіями несуттєва і викликана впливом випадкових величин (приймається гіпотеза $H_0$). Якщо $F_р > F_{кр}$, можна стверджувати, що отримані результати суттєво розбігаються, тобто наявні дані дають підстави стверджувати, що існує істотна розбіжність в якості продукції, і потрібно шукати причину цього.

У тому разі, коли необхідно з'ясувати *причину* суттєвої розбіжності даних, наприклад оператора та обладнання, на результати випробувань, потрібно аналізувати вплив *двох або кількох факторів*.

**Двофакторний дисперсійний аналіз**. При однофакторному дисперсійному аналізі дані тільки групуються за різними рівнями одного фактора. Для випадку оцінки впливу двох факторів необхідно враховувати і спосіб їх взаємодії, тобто *вид моделі.*

Існують два види взаємодії факторів $X_1$ і $X_2$ – *ієрархічна та перехресна.* Відповідно розглядають ієрархічну та перехресну класифікацію.

При ***ієрархічній*** класифікації розрізняють фактори основної групи і фактори підгруп, причому кожний рівень одного основного фактора може бути пов'язаний з множиною рівнів другого фактора – фактора підгрупи.

При ***перехресній*** класифікації кожний рівень одного фактора може поєднуватися з усіма рівнями іншого фактора, і упорядкування в цьому разі, на відміну від ієрархічної класифікації, неможливе.

**Приклад**. Припустимо, що продукція надходить від різних постачальників. Завдання полягає в оцінюванні якості різновидів продукції, на яку в загальному випадку можуть впливати два фактори: обладнання, на якому виробляється продукція, й оператори, які працюють на цьому обладнанні. Якщо є інформація або припущення, що оператори мають однакову кваліфікацію, в цьому випадку застосовується ієрархічна схема випробувань, згідно з якою кожний оператор закріплений за відповідною установкою без

додаткових обмежень. Якщо оператори мають різну кваліфікацію, то для забезпечення умов відтворюваності випробувань необхідно, щоб на кожній установці працювали оператори з різною кваліфікацією, тобто застосовується перехресна схема випробувань.

***Ієрархічна схема проведення випробувань.*** Припустимо, що є чотири однотипові установки (фактор $X_1$), на яких працюють у три зміни 12 операторів (фактор $X_2$). Таким чином, реалізується двофакторний експеримент, де фактор $X_1$ має чотири рівні, а фактор $X_2$ – дванадцять рівнів (рис. 7.1). За такої постановки задачі добір операторів для однієї установки ніяк не пов'язаний з добором операторів для іншої установки. Вважається, що оператори проходили однакову підготовку і мають однакову кваліфікацію. Висувається гіпотеза $H_0$: оператори мають однакову кваліфікацію.

У загальному випадку кількість операторів, які працюють на $N$ установках буде $\sum_{i=1}^{N} n_i$, де $i = \overline{1, N}$ – кількість установок, $n_i$ – кількість операторів, що працюють на $i$ - ій установці (у нашому випадку $N = 4$, $n_i = 3$). Кожний з операторів виробив продукцію, кількість якої – $m_{ik}$, де $k = \overline{1, n_i}$ – поточний номер оператора, що працює на $i$-ій установці.

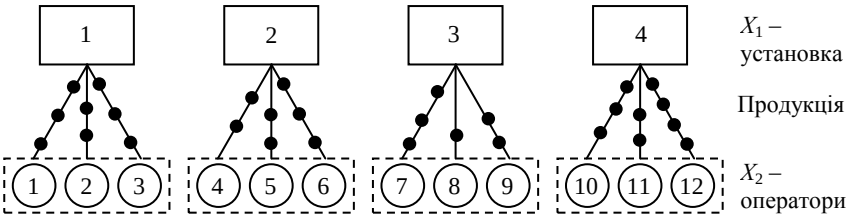



Рис. 7.1. Ієрархічна схема

Для аналізу впливу факторів на показники продукції розглянемо повну суму квадратів відхилень результатів від генерального середнього:

$$Q_0^2 = \sum_{i=1}^{N}\sum_{k=1}^{n_i}\sum_{j=1}^{m_{ik}}(\tilde{y}_{ikj} - \bar{y})^2 \ , \qquad (7.4)$$

де $\tilde{y}_{ikj}$ – значення досліджуваної ознаки *j*-ї одиниці продукції, виготовленої *k*-м оператором, закріпленим за *i*-ю установкою; $\bar{y}$ – середнє значення ознаки для всієї продукції.

Для проведення дисперсійного аналізу згідно з поставленим завданням необхідно відокремити дві складові повної дисперсії:

– *міжгрупову дисперсію*, яка визначається між середніми груповими значеннями $\bar{y}_i$, тобто характеризує складову, зумовлену розсіюванням між ознаками продукції, отриманої на різних установках;

– *усереднену дисперсію* ознак продукції $\bar{y}_{ik}$, виготовленої $n_i$ операторами, які працюють на *i*-й установці.

Для цього, введемо у вираз (7.4) $\pm\bar{y}_i$ та $\pm\bar{y}_{ik}$, отримаємо:

$$Q_0^2 = \sum_{i=1}^{N}\sum_{k=1}^{n_i}\sum_{j=1}^{m_{ik}}\left[(\bar{y}_i - \bar{y}) + (\bar{y}_{ik} - \bar{y}_i) + (\tilde{y}_{ikj} - \bar{y}_{ik})\right]^2.$$

Оскільки між доданками, що містяться в дужках, немає кореляції, подвоєна сума добутків цих доданків буде дорівнювати нулю (як і для однофакторного аналізу). У результаті дістанемо:

$$Q_0^2 = \underbrace{\sum_{i=1}^{N}\sum_{k=1}^{n_i}\sum_{j=1}^{m_{ik}}\left[(\bar{y}_i - \bar{y})^2\right.}_{Q_{x1}^2} + \underbrace{\sum_{i=1}^{N}\sum_{k=1}^{n_i}\sum_{j=1}^{m_{ik}}(\bar{y}_{ik} - \bar{y}_i)^2}_{Q_{x2}^2} + \qquad (7.5)$$

$$+ \underbrace{\left.\sum_{i=1}^{N}\sum_{k=1}^{n_i}\sum_{j=1}^{m_{ikl}}(\tilde{y}_{ikj} - \bar{y}_{ik})\right]^2}_{Q_\varepsilon^2}$$

де $\bar{y}_{ik} = \frac{1}{m_{ik}} \sum_{j=1}^{m_{ik}} \tilde{y}_{ikj}$ – середнє значення ознаки – вихідної величини, отримане *k*-м оператором на *i*-й установці; $\bar{y}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} \bar{y}_{ik}$ – середнє значення ознаки, отриманої трьома операторами, які працюють на *i*-й установці.

У формулі (7.5) перший доданок зумовлений впливом випадкових величин і можливим впливом установок; другий доданок – впливом випадкових величин і можливим впливом операторів; третій доданок – тільки впливом випадкових величин.

Для того щоб оцінити суттєвість впливу установки на продукцію, необхідно обчислити $S_{x_1}^2$ і $S_\varepsilon^2$, а потім перевірити їх відношення за критерієм Фішера.

Позначимо кількість продукції, виготовленої на $i$ -й установці трьома операторами:

$$R_i = \sum_{k=1}^{n_i} m_{ik} .$$

Тоді загальна кількість продукції на всіх установках буде:

$$R = \sum_{i=1}^{N} R_i .$$

Оцінку дисперсії, зумовленої впливом випадкових величин, знаходять як

$$\frac{Q_\varepsilon^2}{f_\varepsilon} .$$

Оскільки для обчислення оцінки цієї дисперсії необхідно було для кожного оператора знайти середнє значення $\bar{y}_{ik}$, то кількість степенів вільності $f_\varepsilon$ при обчисленні $S_\varepsilon^2$ визначається як

$f_\varepsilon = R - \sum_{i=1}^{N} n_i$ З огляду на те, що для обчислення $S_{x_1}^2 = \frac{Q_{x_1}^2}{f_{x_1}}$ потрібно було знайти $\bar{y}$, кількість степенів вільності буде $f_{x_1} = N - 1$.

Розрахункове значення коефіцієнта Фішера:

$$F_{\text{P}(x_1)} = \frac{S_{x_1}^2}{S_\varepsilon^2}.$$

Далі за відомим алгоритмом визначають $F_{\text{кр}}$ для $f_{x_1}, f_\varepsilon$ і обраного $\alpha$, порівнюють з $F_{\text{p}(x_1)}$ і приймають відповідне рішення.

Доданок $Q_{x_2}^2$ несе інформацію про можливе відхилення рівня кваліфікації операторів, а також про вплив випадкових факторів. Узявши відношення $\frac{S_{x_2}^2}{S_\varepsilon^2}$, можна оцінити вплив наявної розбіжності у кваліфікації операторів, тобто суттєвий чи ні цей вплив.

Розглянемо вираз:

$$Q_{x_2}^2 = \sum_{i=1}^{N} \sum_{k=1}^{n_i} m_{ik} (\bar{y}_{ik} - \bar{y}_i)^2,$$

з якого випливає, що для обчислення $Q_{x_2}^2$ необхідно визначити середнє значення ознаки продукції, отриманої на *N* установках, тому кількість степенів вільності буде

$$f_{x_2} = \sum_{i=1}^{N} n_i - N.$$

Аналогічно розглянутому для фактора $X_1$ перевіряємо суттєвість розбіжності між дисперсіями $S_{x_2}^2$ й $S_\varepsilon^2$, а отже, про можливий вплив операторів на якість продукції (розсіювання ознаки).

***Перехресна схема проведення випробувань.*** Припустимо, що оператори мають *різну кваліфікацію* і з трьох операторів, які працюють на кожній з установок, один має 1-й розряд,

другий – 2-й, а третій – 3-й. Для забезпечення достовірності результатів випробувань необхідно, щоб на кожній з установок було задіяно операторів усіх рівнів кваліфікації, тобто потрібно задіяти по чотири оператори з 1-м, 2-м та 3-м розрядами. Тоді за умови попереднього прикладу реалізується перехресна класифікація з чотирма рівнями фактора $X_1$ «установка» і трьома рівнями фактора $X_2$ «розряд». Схематичну модель з перехресними зв'язками наведено на рис. 7.2. На відміну від ієрархічної структури схема взаємодій факторів буде більш складною.

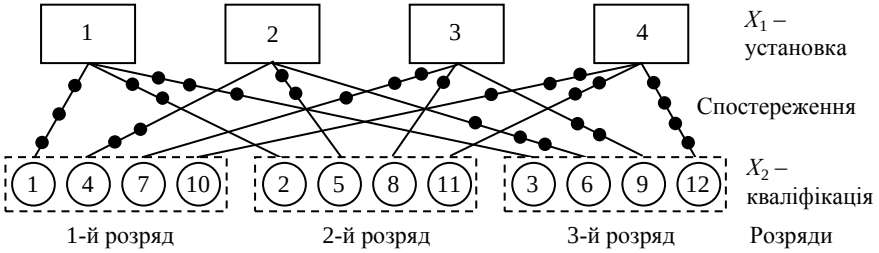



Рис. 7.2 Перехресна схема

Вихідними є експериментальні дані, які заносять у табл. 7.2.

Дані позначають трьома індексами $\tilde{y}_{ikj}$, де $i$ – рівень першого фактора (номер установки, $i = \overline{1, N_1}$); $k$ – рівень другого фактора (розряд оператора, $k = \overline{1, N_2}$); $j$ – порядковий номер одного з $m_{ik}$ спостережень для сполучення $i$-го рівня першого фактора з $k$-м рівнем другого. Вважається, що кожний оператор виготовляє однакову кількість продукції $m$.

Для кожної комірки табл. 7.2, тобто для фіксованих значень $i$, $k$ можна знайти середнє значення:

$$\bar{y}_{ik} = \frac{1}{m} \sum_{j=1}^{m} \tilde{y}_{ikj},$$

яке може бути використане для оцінювання впливу випадкових факторів.

Таблиця 7.2. Масив даних для перехресної схеми

| $x_2$ | | | | | | | |
|---|---|---|---|---|---|---|---|
| $x_1$ | 1 | 2 | … | $k$ | … | $N_2$ | |
| 1 | $\tilde{y}_{111}$ …..... $\tilde{y}_{11m}$ | $\tilde{y}_{121}$ …..... $\tilde{y}_{12m}$ | … | $\tilde{y}_{1k1}$ …..... $\tilde{y}_{1km}$ | … | $\tilde{y}_{1N_21}$ …..... $\tilde{y}_{1N_2m}$ | $\bar{y}_1$ |
| 2 | $\tilde{y}_{211}$ …..... $\tilde{y}_{21m}$ | $\tilde{y}_{221}$ …..... $\tilde{y}_{22m}$ | … | $\tilde{y}_{2k1}$ …..... $\tilde{y}_{2km}$ | … | $\tilde{y}_{2N_21}$ …..... $\tilde{y}_{2N_2m}$ | $\bar{y}_2$ |
| … | … | … | … | … | … | … | … |
| $i$ | $\tilde{y}_{i11}$ … $\tilde{y}_{i1m}$ | $\tilde{y}_{i21}$ …..... $\tilde{y}_{i2m}$ | … | $\tilde{y}_{ik1}$ …..... $\tilde{y}_{ikm}$ | … | $\tilde{y}_{iN_21}$ …..... $\tilde{y}_{iN_2m}$ | $\bar{y}_i$ |
| … | … | … | … | … | … | … | … |
| $N_1$ | $\tilde{y}_{N_111}$ … $\tilde{y}_{N_11m}$ | $\tilde{y}_{N_121}$ …..... $\tilde{y}_{N_12m}$ | … | $\tilde{y}_{N_1k1}$ …..... $\tilde{y}_{N_1km}$ | … | $\tilde{y}_{N_1N_21}$ …..... $\tilde{y}_{N_1N_2m}$ | $\bar{y}_{N_1}$ |
| | $\bar{y}_1'$ | $\bar{y}_2'$ | … | $\bar{y}_k'$ | … | $\bar{y}_{N_2}'$ | $\bar{y}$ |

Для кожного рядка таблиці ($X_1$ = const) можна знайти середнє за всіма значеннями фактора $X_2$ ,тобто середнє значення ознаки продукції, виготовленої усіма операторами на всіх установках:

$$\bar{y}_i = \frac{1}{N_2}\sum_{k=1}^{N_2}\bar{y}_{ik} = \frac{1}{N_2 m}\sum_{k=1}^{N_2}\sum_{j=1}^{m}\tilde{y}_{ikj}\,.$$

Для кожного стовпця таблиці ($X_2$ = const) можна знайти $\bar{y}_k'$, яке характеризує середнє значення показника продукції, отриманої всіма операторами $k$-ї кваліфікації :

$$\bar{y}_k' = \frac{1}{N_1}\sum_{i=1}^{N_1}\bar{y}_{ik} = \frac{1}{N_1 m}\sum_{i=1}^{N_1}\sum_{j=1}^{m}\tilde{y}_{ikj}\,.$$

Генеральне середнє визначають як

$$\bar{y} = \frac{1}{N_1 N_2 m}\sum_{i=1}^{N_1}\sum_{k=1}^{N_2}\sum_{j=1}^{m}\tilde{y}_{ikj} = \frac{1}{N_1 N_2}\sum_{i=1}^{N_1}\sum_{k=1}^{N_2}\bar{y}_{ik} = \frac{1}{N_1}\sum_{i=1}^{N_1}\bar{y}_i = \frac{1}{N_2}\sum_{k=1}^{N_2}\bar{y}_k'\,,$$

і в загальному випадку буде залежати:

– від стану обладнання;

– розряду оператора;

– впливу випадкових величин;

– наявності перехресних зав'язків, тобто взаємодії між $X_1$ і $X_2$.

Розглянемо вираз для повної суми квадрата відхилень від загального середнього, додавши складову $\pm \bar{y}_{ik}$:

$$Q_0^2 = \sum_{i=1}^{N_1}\sum_{k=1}^{N_2}\sum_{j=1}^{m}(\tilde{y}_{ikj} - \bar{y})^2 = \sum_{i=1}^{N_1}\sum_{k=1}^{N_2}\sum_{j=1}^{m}\left[(\tilde{y}_{ikj} - \bar{y}_{ik}) + (\bar{y}_{ik} - \bar{y})\right]^2. \quad (7.6)$$

Як уже було встановлено, добуток двох некорельованих величин дорівнює нулю, тому вираз (7.6) можна переписати у вигляді:

$$Q_0^2 = \underbrace{\sum_{i=1}^{N_1}\sum_{k=1}^{N_2}\sum_{j=1}^{m}(\tilde{y}_{ikj} - \bar{y}_{ik})^2}_{Q_\varepsilon^2} + \underbrace{\sum_{i=1}^{N_1}\sum_{k=1}^{N_2}\sum_{j=1}^{m}(\bar{y}_{ik} - \bar{y})^2}_{Q_1}. \quad (7.7)$$

Перший доданок виразу (7.7) являє собою суму розсіювань результатів у комірці таблиці відповідно до середнього для цієї комірки. Оскільки досліди в кожній комірці проводилися при фіксованих значеннях $X_1$ і $X_2$, то розсіювання значень $\tilde{y}_{ikj}$ відносно середнього $\bar{y}_{ik}$ зумовлене впливом випадкових величин.

Розглянемо окремо другий доданок виразу (7.7). Розіб'ємо його на відповідні суми (необхідно пам'ятати, що попарні добутки дають нуль), увівши в нього складові $\pm\bar{y}_i$, $\pm\bar{y}_k'$, $\pm\bar{y}$:

$$\sum_{i=1}^{N_1}\sum_{k=1}^{N_2}\sum_{j=1}^{m}(\bar{y}_{ik} - \bar{y}) = \underbrace{\sum_{i=1}^{N_1}\sum_{k=1}^{N_2}\sum_{j=1}^{m}(\bar{y}_i - \bar{y})}_{Q_{x1}^2} +$$

$$+ \underbrace{\sum_{i=1}^{N_1}\sum_{k=1}^{N_2}\sum_{j=1}^{m}(\bar{y}_k' - \bar{y})}_{Q_{x2}^2} + \underbrace{\sum_{i=1}^{N_1}\sum_{k=1}^{N_2}\sum_{j=1}^{m}(\bar{y}_{ik} - \bar{y}_i - \bar{y}_k' + \bar{y})}_{Q_{x1x2}^2}.$$

У кінцевому результаті дістанемо три складові суми. Перша з них – $Q_{x_1}^2$ пропорційна до впливу випадкових величин та можливого впливу характеристик установок, друга – пропорційна до впливу випадкових величин і можливої суттєвої розбіжності рівнів кваліфікації операторів, а третя несе інформацію про можливий вплив взаємозв'язку між розрядами операторів і характеристиками установок, а також щодо впливу випадкових величин. Як і при ієрархічній схемі організації випробувань, для виявлення можливого впливу факторів (установки, оператора) та їхніх взаємодій необхідно оцінити відношення пов'язаних з ними дисперсій та дисперсії, зумовленої впливом тільки випадкових величин. Для цього необхідно виконати такі дії:

1. Обчислити $S_{x_1}^2 = \dfrac{Q_{x_1}^2}{f_{x_1}}$,

де $f_{x_1} = N_1 - 1$.

2. Обчислити $S_{x_2}^2 = \dfrac{Q_{x_2}^2}{f_{x_2}}$,

де $f_{x_2} = N_2 - 1$.

3. Обчислити $S_{x_1x_2}^2 = \dfrac{Q_{x_1x_2}^2}{f_{x_1x_2}}$,

де $f_{x_1x_2} = N_1N_2 - N_1 - N_2 + 1 = (N_1 - 1)(N_2 - 1)$.

4. Обчислити $S_\varepsilon^2 = \dfrac{Q_\varepsilon^2}{f_\varepsilon}$;

де $f_\varepsilon = N_1N_2(m-1)$.

Перевірка гіпотез про значущість впливів факторів та їх взаємодій здійснюється за допомогою критерію Фішера. Знаходять відповідне розрахункове значення $F_\text{Р}$, взявши відношення $S_{x_1}^2$, або $S_{x_2}^2$, або $S_{x_1x_2}^2$ до $S_\varepsilon^2$. Знайдені у такий спосіб розрахункові значення коефіцієнтів Фішера порівнюють з критичними для відповідного

числа степенів вільності (для чисельників $f_{x_1}$, або $f_{x_2}$, або $f_{x_1x_2}$) і $f_\varepsilon$ – для знаменника. Якщо $F_\text{Р} > F_T$, то відповідний фактор $X_1$ або $X_2$ або їх взаємодія $X_1X_2$ суттєво вплинули на значення ознаки.

Перевірка гіпотез про значущість впливів факторів та їхніх взаємодій здійснюється за допомогою критерію Фішера. Знаходять відповідне розрахункове значення $F_\text{р}$, взявши відношення $S^2_{x_1}$, або $S^2_{x_2}$, або $S^2_{x_1x_2}$ до $S^2_\varepsilon$. Знайдені в такий спосіб розрахункові значення коефіцієнтів Фішера порівнюють з критичними для відповідної кількості степенів вільності (для чисельників $f_{x_1}$, або $f_{x_2}$, або $f_{x_1x_2}$) і $f_\varepsilon$ – для знаменника. Якщо $F_\text{р} > F_\text{кр}$, то відповідний фактор $X_1$ або $X_2$ або їх взаємодія $X_1X_2$ суттєво вплинули на значення ознаки.

**Латинські квадрати**. При зростанні кількості факторів, які можуть впливати, різко зростає обсяг досліджень і аналіз експериментальних даних ускладняються. Застосування так званих латинських квадратів дає змогу значно зменшити обсяг досліджень і спростити обробку даних [8].

План дослідження виду «класичний латинський квадрат» дає змогу досліджувати вплив трьох факторів, які варіюються на $N$ рівнях, але використовується тільки $N^2$ комбінацій рівнів факторів замість $N^3$ можливих комбінацій, причому існування розбіжностей можна перевірити для кожного з трьох факторів. Розмірність латинського квадрата визначається кількістю рівнів варіювання факторів: якщо $N = 3$, то говорять про латинський квадрат $3 \times 3$.

*Латинськими квадрати називаються тому, що рівні варіювання третього фактора позначаються не арабськими цифрами, а латинськими літерами.*

**Приклад**. Розглянемо трифакторний експеримент, реалізований на чотирьох рівнях. Перший фактор **(**устаткування) змінюється на $i = \overline{1,4}$ рівнях; другий фактор (оператор) – на $k = \overline{1,4}$ рівнях; а для третього фактора (день тижня) рівні позначаються латинськими літерами, тобто *l* = *A*,*B*,*C*,*D*.

План типу латинського квадрата наведено в табл. 7.3. Рівні варіювання двох факторів подано у вигляді рядків (фактор 1) та стовпців (фактор 2), а третій фактор, рівні якого позначено літерами, заносять у комірки квадрата 4 × 4 в такий спосіб.

Таблиця 7.3. Латинський квадрат

| Фактор 1 | Фактор 2 | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| 1 | A | B | C | D |
| 2 | D | A | B | C |
| 3 | C | D | A | B |
| 4 | B | C | D | A |

У перший рядок табл. 7.3 (перший рівень фактора 1) в упорядкованій послідовності заносяться літери позначення рівнів третього фактора. Порядок розташування рівнів третього фактора в наступних рядках визначається шляхом циклічного зсуву елементів попереднього рядка праворуч і переведення останнього елемента на перше місце. Побудований у такий спосіб план має на діагоналях ті самі рівні третього фактора (на головній діагоналі розміщено рівень *A*), а рівні-літери, що відповідають рівню третього фактора розподілені у матриці плану так, що в кожному рядку і кожному стовпці кожна літера трапляється тільки один раз.

*По діагоналі розміщені однакові рівні третього фактора. Жодна комбінація рівнів усіх факторів не повторюється в жодній комірці матриці.*

Відповідно до матриці реалізуються досліди, в результаті яких дістають значення $\tilde{y}_{ikj}$. На підставі наявних результатів можна обчислити:

$$\bar{y}_i = \frac{1}{N}\sum_{k=1}^{N}\tilde{y}_{kj};$$

$$\bar{y}_k = \frac{1}{N}\sum_{i=1}^{N}\tilde{y}_{ij};$$

$$\bar{y}_j = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N}\tilde{y}_{ikj}.$$

Крім того, за результатами випробувань можна обчислити й генеральне середнє:

$$\bar{y} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{k=1}^{N}\tilde{y}_{ikj}\ .$$

Для виявлення впливу трьох розглянутих факторів на досліджувану ознаку виходять з повної суми відхилень результатів від генерального середнього:

$$Q_0^2 = \sum_{i=1}^{N}\sum_{k=1}^{N}(\tilde{y}_{ikj} - \bar{y})^2\ .$$

Крім того, обчислюють вираз:

$$Q_{x_1}^2 = \sum_{i=1}^{N}(\bar{y}_i - \bar{y})^2\ ,$$

що несе інформацію про можливий вплив першого фактора.

Аналогічно обчислюють:

$$Q_{x_2}^2 = \sum_{k=1}^{N}(\bar{y}_k - \bar{y})^2\ ,$$

що несе інформацію про можливий вплив другого фактора, і вираз:

$$Q_{x_3}^2 = \sum_{j=1}^{N}(\bar{y}_j - \bar{y})^2$$

несе інформацію про вплив третього фактора.

На відміну від двофакторних випробувань латинські квадрати не дають змоги виділяти окремо вплив випадкових величин ε та взаємодії факторів, тому за випадкове розсіювання $Q_\varepsilon^2$ береться різниця: $Q_\varepsilon^2 = Q_0^2 - Q_{x_1}^2 - Q_{x_2}^2 - Q_{x_3}^2$ .

У такий спосіб обчислюються всі необхідні відхилення для визначення суттєвості впливу розглянутих факторів. Для визначення суттєвості впливу факторів необхідно на підставі розрахованих відхилень обчислити відповідні оцінки дисперсії, визначивши кількість степенів вільності для розсіювань:

$$f_{x_1} = f_{x_2} = f_{x_3} = N - 1;$$

$$f_{\varepsilon} = (N-1)(N-2).$$

Потім за відомою схемою розраховують відповідні коефіцієнти Фішера, які оцінюють суттєвість впливу фактора шляхом порівняння розрахованого коефіцієнта $F_{\text{р}}$ з критичним значенням $F_{\text{кр}}$.

**Греко-латинські квадрати**. Існують практичні випадки, коли кількість досліджуваних факторів необхідно збільшити до чотирьох. Це здійснюється шляхом накладення на основний латинський квадрат другого латинського квадрата тієї самої розмірності $N \times N$ і ортогонального вихідному, тобто кожна літера одного латинського квадрата один раз з'являється на тій самій позиції, як і кожна літера другого латинського квадрата. Таким чином, кожна літера першого квадрата трапиться у вигляді комбінації тільки один раз з кожною літерою другого квадрата. Щоб розрізнити рівні факторів, які входять до вихідного латинського й ортогонального квадратів, у другому латинському квадраті вживаються грецькі літери (звідси назва *греко-латинський квадрат*).

Методи побудови греко-латинських квадратів різні. Деякі з них відрізняються простотою, але не є загальними і не дозволяють отримати повної множини греко-латинських квадратів, тобто не для всіх квадратів *N*-го порядку можна визначити ортогональні квадрати.

Дисперсійний аналіз для плану типу греко-латинського квадрата аналогічний до розглянутого раніше латинського квадрата. Відмінність полягає в тому, що необхідно обчислювати ще суму

квадратів для четвертого фактора $Q_{x_4}^2$. Крім того, кількість степенів вільності залишкової суми квадратів зменшується на $N - 1$. Тому при $N = 3$ кількість степенів вільності для залишкової суми квадратів дорівнює нулю і греко-латинські квадрати 3×3 для автономних досліджень будувати недоцільно.

### Запитання для самоперевірки

1. Сформулюйте задачу дисперсійного аналізу.
2. У чому полягають умови однорідності та відтворюваності?
3. Яким чином визначається вплив факторів при дисперсійному аналізі?
4. Як знайти оцінку загальної дисперсії та її складових при однофакторному дисперсійному аналізі?
5. Як здійснити перевірку нульової гіпотези при однофакторному дисперсійному аналізі?
6. Як забезпечується відтворюваність умов при проведенні перехресної схеми випробувань?
7. Чим відрізняється ієрархічна схема зав'язків факторів від перехресної?
8. Яку фізичну інтерпретацію мають складові дисперсії при ієрархічній класифікації?
9. Для чого застосовуються латинські квадрати?
10. У чому полягає відмінність латинських та греко-латинських квадратів?
11. Як будуються латинські квадрати?
12. Для чого потрібно перевіряти однорідність дисперсій при аналізі?

# СПИСОК ЛІТЕРАТУРИ


1. Володарський Є. Т. Теорія та практика експериментальних досліджень : навч. посіб. / Є. Т. Володарський, Л. О. Кошева. – Вінниця : ФОП Барановська Т. П., 2023. – 298 с.

2. Дорогань-Писаренко Л. О. Статистика : навч. посіб. / Л. О. Дорогань-Писаренко, О. В. Єгорова, А. І. Рудич. – Полтава : РВВ ПДАУ, 2021. – 300 с.

3. Куц Ю. В. Спеціальні розділи математики. Курс лекцій : навч. посіб / Ю. В. Куц, Ю. Ю. Лисенко. – Київ : КПІ ім. Ігоря Сікорського, 2022. – 180 с

4. Пашинський В. А. Статистичні методи в інженерних дослідженнях : навч. посіб. / В. А. Пашинський. – Кропивницький : ЦНТУ, 2020. – 106 с.

5. Горват А. А. Методи обробки експериментальних даних з використанням MS EXCEL / А. А. Горват, О. О. Молнар, В. В Міньковић. – Ужгород : Видавництво УжНУ «Говерла», 2019. – 182 с.

6. Мішура Ю. С. Випадкові процеси: теорія, статистика, застосування : підруч. / Ю. С. Мішура, Г. М. Ральченко, К. В. Шевченко. – 2-ге вид., випр. і допов. – Київ : ВПЦ «Київський університет», 2021. – 496 с.

7. Павленко П. М. Математичне моделювання систем і процесів: навч. посіб. / П. М. Павленко, С. Ф. Філоненко, О. М. Чередніков, В. В. Трейтяк. – Київ : НАУ, 2017. – 392 с.

8. Johnson, N. L. Statistics & Experimental Design in Engineering & the Physical Sciences (Wiley Series in Probability & Mathematical Statistics) / Johnson N. L., F. C. Leone. – John Wiley & Sons, 1977. – 1082 pp.

9. Cooper. G. R. Probabilistic Methods of Signal and System Analysis / G. R. Cooper, C. D. McGillem. – 2rd Ed. – Bailliere Tindall / W. B. Saunders, 1986. – 408 pp.

10. Kendall M. The Advanced Theory of Statistics. Vol. 3 : Design and Analysis and Time-Series / M. Kendall. – 3rd Ed. – Arnold, 1968. – 567 pp.

11. Statistics (Handbook of Applicable Mathematics, Vol. 6, Part B) / by Walter Ledermann (Editor), Emlyn Lloyd (Editor). – Wiley, 1984. – 522 pp.

12. Таблиця значень функції Лапласа [Електронний ресурс]. – Режим доступу: https://studfile.net/preview/7071470/page:8/. Назва з екрана.

# ДОДАТОК

## Статистичні таблиці

Таблиця Д.1. Нормальний розподіл

| z | 0,00 | 0,01 | 0,02 | 0,03 | 0,04 | 0,05 | 0,06 | 0,07 | 0,08 | 0,09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0,0 | 0,500 | 0,504 | 0,508 | 0,512 | 0,516 | 0,520 | 0,524 | 0,528 | 0,532 | 0,536 |
| 0,1 | 0,540 | 0,544 | 0,548 | 0,552 | 0,556 | 0,560 | 0,564 | 0,568 | 0,571 | 0,575 |
| 0,2 | 0,579 | 0,583 | 0,587 | 0,591 | 0,595 | 0,599 | 0,603 | 0,606 | 0,610 | 0,614 |
| 0,3 | 0,618 | 0,623 | 0,625 | 0,629 | 0,633 | 0,637 | 0,640 | 0,644 | 0,648 | 0,652 |
| 0,4 | 0,655 | 0,659 | 0,663 | 0,666 | 0,670 | 0,674 | 0,677 | 0,681 | 0,684 | 0,688 |
| 0,5 | 0,691 | 0,695 | 0,698 | 0,702 | 0,705 | 0,709 | 0,712 | 0,716 | 0,719 | 0,722 |
| 0,6 | 0,726 | 0,729 | 0,732 | 0,736 | 0,739 | 0,742 | 0,745 | 0,749 | 0,752 | 0,755 |
| 0,7 | 0,758 | 0,761 | 0,764 | 0,767 | 0,770 | 0,773 | 0,776 | 0,779 | 0,782 | 0,785 |
| 0,8 | 0,788 | 0,791 | 0,794 | 0,797 | 0,799 | 0,802 | 0,805 | 0,808 | 0,811 | 0,813 |
| 0,9 | 0,816 | 0,819 | 0,820 | 0,824 | 0,826 | 0,829 | 0,832 | 0,834 | 0,836 | 0,839 |
| 1,0 | 0,841 | 0,844 | 0,846 | 0,848 | 0,851 | 0,853 | 0,855 | 0,858 | 0,860 | 0,862 |
| 1,1 | 0,864 | 0,866 | 0,869 | 0,871 | 0,873 | 0,875 | 0,877 | 0,879 | 0,881 | 0,883 |
| 1,2 | 0,885 | 0,887 | 0,889 | 0,891 | 0,892 | 0,894 | 0,896 | 0,898 | 0,900 | 0,901 |
| 1,3 | 0,903 | 0,905 | 0,906 | 0,908 | 0,910 | 0,911 | 0,913 | 0,915 | 0,916 | 0,918 |
| 1,4 | 0,919 | 0,921 | 0,922 | 0,924 | 0,925 | 0,926 | 0,928 | 0,929 | 0,931 | 0,932 |
| 1,5 | 0,933 | 0,934 | 0,936 | 0,937 | 0,938 | 0,939 | 0,941 | 0,942 | 0,943 | 0,944 |
| 1,6 | 0,945 | 0,946 | 0,947 | 0,948 | 0,949 | 0,950 | 0,951 | 0,952 | 0,953 | 0,954 |

Продовження табл. Д.1

| z | 0,00 | 0,01 | 0,02 | 0,03 | 0,04 | 0,05 | 0,06 | 0,07 | 0,08 | 0,09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1,7 | 0,955 | 0,956 | 0,957 | 0,958 | 0,959 | 0,960 | 0,961 | 0,962 | 0,962 | 0,963 |
| 1,8 | 0,964 | 0,965 | 0,966 | 0,966 | 0,967 | 0,968 | 0,968 | 0,969 | 0,970 | 0,971 |
| 1,9 | 0,971 | 0,972 | 0,972 | 0,973 | 0,974 | 0,974 | 0,975 | 0,975 | 0,976 | 0,977 |
| 2,0 | 0,977 | 0,978 | 0,978 | 0,979 | 0,979 | 0,980 | 0,980 | 0,981 | 0,981 | 0,982 |
| 2,1 | 0,982 | 0,982 | 0,983 | 0,983 | 0,984 | 0,984 | 0,985 | 0,985 | 0,985 | 0,986 |
| 2,2 | 0,986 | 0,986 | 0,987 | 0,987 | 0,987 | 0,988 | 0,988 | 0,988 | 0,989 | 0,989 |
| 2,3 | 0,989 | 0,989 | 0,990 | 0,990 | 0,990 | 0,991 | 0,991 | 0,991 | 0,991 | 0,992 |
| 2,4 | 0,992 | 0,992 | 0,992 | 0,992 | 0,993 | 0,993 | 0,993 | 0,993 | 0,993 | 0,994 |
| 2,5 | 0,994 | 0,994 | 0,994 | 0,994 | 0,994 | 0,995 | 0,995 | 0,995 | 0,995 | 0,995 |
| 2,6 | 0,995 | 0,995 | 0,996 | 0,996 | 0,996 | 0,996 | 0,996 | 0,996 | 0,996 | 0,996 |
| 2,7 | 0,996 | 0,997 | 0,997 | 0,997 | 0,997 | 0,997 | 0,997 | 0,997 | 0,997 | 0,997 |
| 2,8 | 0,997 | 0,997 | 0,997 | 0,997 | 0,998 | 0,998 | 0,998 | 0,998 | 0,998 | 0,998 |
| 2,9 | 0,998 | 0,998 | 0,998 | 0,998 | 0,998 | 0,998 | 0,998 | 0,998 | 0,998 | 0,999 |
| 3,0 | 0,999 | 0,999 | 0,999 | 0,999 | 0,999 | 0,999 | 0,999 | 0,999 | 0,999 | 0,999 |

Таблиця Д.2. Щільність імовірностей нормованого та центрованого нормального розподілу

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0,0 | $3989^{-4}$ | 3989 | 3989 | 3988 | 3986 | 3984 | 3982 | 3980 | 3977 | 3973 |
| 0,1 | $3970^{-4}$ | 3965 | 3961 | 3956 | 3951 | 3945 | 3939 | 3932 | 3925 | 3918 |
| 0,2 | $3910^{-4}$ | 3902 | 3894 | 3885 | 3876 | 3867 | 3857 | 3847 | 3836 | 3825 |
| 0,3 | $3814^{-4}$ | 3802 | 3790 | 3778 | 3765 | 3752 | 3739 | 3725 | 3712 | 3697 |
| 0,4 | $3683^{-4}$ | 3668 | 3653 | 3637 | 3621 | 3605 | 3589 | 3572 | 3555 | 3538 |
| 0,5 | $3521^{-4}$ | 3503 | 3485 | 3467 | 3448 | 3429 | 3410 | 3391 | 3372 | 3352 |
| 0.6 | $3332^{-4}$ | 3312 | 3292 | 3271 | 3251 | 3230 | 3209 | 3187 | 3166 | 3141 |
| 0,7 | $3123^{-4}$ | 3101 | 3079 | 3056 | 3034 | 3011 | 2989 | 2966 | 2943 | 2920 |
| 0,8 | $2897^{-4}$ | 2874 | 2850 | 2827 | 2803 | 2780 | 2756 | 2732 | 2709 | 2685 |
| 0,9 | $2661^{-4}$ | 2637 | 2613 | 2589 | 2565 | 2541 | 2516 | 2492 | 2468 | 2444 |
| 1,0 | $2420^{-4}$ | 2396 | 2371 | 2347 | 2323 | 2299 | 2275 | 2251 | 2227 | 2203 |
| 1,1 | $2179^{-4}$ | 2155 | 2131 | 2107 | 2083 | 2059 | 2035 | 2012 | 1989 | 1965 |
| 1,2 | $1942^{-4}$ | 1919 | 1895 | 1872 | 1849 | 1826 | 1804 | 1781 | 1758 | 1736 |
| 1,3 | $1714^{-4}$ | 1691 | 1669 | 1647 | 1626 | 1604 | 1582 | 1561 | 1539 | 1518 |
| 1,4 | $1497^{-4}$ | 1476 | 1456 | 1435 | 1415 | 1394 | 1374 | 1354 | 1334 | 1315 |
| 1,5 | $1295^{-4}$ | 1276 | 1257 | 1238 | 1219 | 1200 | 1182 | 1163 | 1145 | 1127 |
| 1,6 | $1109^{-4}$ | 1092 | 1074 | 1057 | 1040 | 1023 | 1006 | 9893 | 9728 | 9566 |
| 1,7 | $9405^{-5}$ | 9246 | 9089 | 8933 | 8780 | 8628 | 8478 | 8329 | 8183 | 8038 |
| 1,8 | $7895^{-5}$ | 7754 | 7614 | 7477 | 7341 | 7206 | 7074 | 6943 | 6814 | 6687 |
| 1,9 | $6562^{-5}$ | 6438 | 6316 | 6195 | 6077 | 5960 | 5844 | 5730 | 5618 | 5508 |
| 2,0 | $5399^{-5}$ | 5292 | 5186 | 5082 | 4980 | 4879 | 4780 | 4682 | 4586 | 4491 |
| 2,1 | $4398^{-5}$ | 4307 | 4217 | 4128 | 4041 | 3955 | 3871 | 3788 | 3706 | 3626 |
| 2,2 | $3547^{-5}$ | 3470 | 3394 | 3319 | 3246 | 3174 | 3103 | 3034 | 2965 | 2898 |
| 2,3 | $2833^{-5}$ | 2768 | 2705 | 2643 | 2582 | 2522 | 2463 | 2406 | 2349 | 2294 |
| 2,4 | $2239^{-5}$ | 2186 | 2134 | 2083 | 2033 | 1984 | 1936 | 1888 | 1842 | 1797 |
| 2,5 | $1753^{-5}$ | 1709 | 1667 | 1625 | 1585 | 1545 | 1506 | 1468 | 1431 | 1394 |
| 2,6 | $1358^{-5}$ | 1323 | 1289 | 1256 | 1223 | 1191 | 1160 | 1130 | 1100 | 1071 |
| 2,7 | $1042^{-5}$ | 1014 | 9871 | 9606 | 9347 | 9094 | 8846 | 8605 | 8370 | 8140 |
| 2,8 | $7915^{-6}$ | 7697 | 7483 | 7274 | 7071 | 6873 | 6679 | 6491 | 6307 | 6127 |
| 2,9 | $5953^{-6}$ | 5782 | 5616 | 5454 | 5296 | 5143 | 4993 | 4847 | 4705 | 4567 |
| 3,0 | $4432^{-6}$ | 4301 | 4173 | 4049 | 3928 | 3810 | 3695 | 3584 | 3475 | 3370 |

Таблиця Д.3. *t*-розподіл Стьюдента

| ν | α = 0,50 | α = 0,25 | α = 0,1 | α = 0,05 | α = 0,02 | α = 0,01 |
|---|---|---|---|---|---|---|
| 1 | 1,000 | 2,41 | 6,31 | 12,7 | 31,82 | 63,7 |
| 2 | 0,816 | 1,60 | 2,92 | 4,30 | 6,97 | 9,92 |
| 3 | 0,765 | 1,42 | 2,35 | 3,18 | 4,54 | 5,84 |
| 4 | 0,741 | 1,34 | 2,13 | 2,78 | 3,75 | 4,60 |
| 5 | 0,727 | 1,30 | 2,01 | 2,57 | 3,37 | 4,03 |
| 6 | 0,718 | 1,27 | 1,94 | 2,45 | 3,14 | 3,71 |
| 7 | 0,711 | 1,25 | 1,89 | 2,36 | 3,00 | 3,50 |
| 8 | 0,706 | 1,24 | 1,86 | 2,31 | 2,90 | 3,36 |
| 9 | 0,703 | 1,23 | 1,83 | 2,26 | 2,82 | 3,25 |
| 10 | 0,700 | 1,22 | 1,81 | 2,23 | 2,76 | 3,17 |
| 11 | 0,697 | 1,21 | 1,80 | 2,2 | 2,72 | 3,11 |
| 12 | 0,695 | 1,21 | 1,78 | 2,18 | 2,68 | 3,05 |
| 13 | 0,694 | 1,20 | 1,77 | 2,16 | 2,65 | 3,01 |
| 14 | 0,692 | 1,20 | 1,76 | 2,14 | 2,62 | 2,98 |
| 15 | 0,691 | 1,20 | 1,75 | 2,13 | 2,60 | 2,95 |
| 16 | 0,690 | 1,19 | 1,75 | 2,12 | 2,58 | 2,92 |
| 17 | 0,689 | 1,19 | 1,74 | 2,11 | 2,57 | 2,90 |
| 18 | 0,688 | 1,19 | 1,73 | 2,10 | 2,55 | 2,88 |
| 19 | 0,688 | 1,19 | 1,73 | 2,09 | 2,54 | 2,86 |
| 20 | 0,687 | 1,18 | 1,73 | 2,09 | 2,53 | 2,85 |
| 21 | 0,686 | 1,18 | 1,72 | 2,08 | 2,52 | 2,83 |
| 22 | 0,686 | 1,18 | 1,72 | 2,07 | 2,51 | 2,82 |
| 23 | 0,685 | 1,18 | 1,71 | 2,07 | 2,50 | 2,81 |
| 24 | 0,685 | 1,18 | 1,71 | 2,06 | 2,49 | 2,80 |
| 25 | 0,684 | 1,18 | 1,71 | 2,06 | 2,49 | 2,79 |
| 26 | 0,684 | 1,18 | 1,71 | 2,06 | 2,48 | 2,78 |
| 27 | 0,684 | 1,18 | 1,71 | 2,05 | 2,47 | 2,77 |
| 28 | 0,683 | 1,17 | 1,70 | 2,05 | 2,47 | 2,76 |
| 30 | 0,683 | 1,17 | 1,70 | 2,04 | 2,46 | 2,75 |
| 120 | 0,677 | 1,16 | 1,66 | 1,98 | 2,36 | 2,62 |
| ∞ | 0,674 | 1,15 | 1,64 | 1,96 | 2,33 | 2,58 |
| *f* | α/2 = 0,25 | α/2 = 0,125 | α/2 = 0,05 | α/2 = 0,025 | α/2 = 0,01 | α/2 = 0,005 |

Таблиця Д.4. $\chi^2$-розподіл Пірсона

| ν | α=0,99 | 0,95 | 0,9 | 0,5 | 0,1 | 0,05 | 0,01 |
|---|---|---|---|---|---|---|---|
| 1 | 0,0001 | 0,0039 | 0,0158 | 0,455 | 2,71 | 3,84 | 6,64 |
| 2 | 0,0201 | 0,103 | 0,211 | 1,39 | 4,61 | 5,99 | 9,21 |
| 3 | 0,115 | 0,352 | 0,584 | 2,37 | 6,25 | 7,81 | 11,3 |
| 4 | 0,297 | 0,711 | 1,06 | 3,36 | 7,78 | 9,49 | 13,3 |
| 5 | 0,554 | 1,15 | 1,61 | 4,35 | 9,24 | 11,1 | 15,1 |
| 6 | 0,872 | 1,64 | 2,20 | 5,35 | 10,6 | 12,6 | 16,8 |
| 7 | 1,24 | 2,17 | 2,83 | 6,35 | 12,0 | 14,1 | 18,5 |
| 8 | 1,65 | 2,73 | 3,49 | 7,34 | 13,4 | 15,5 | 20,1 |
| 9 | 2,09 | 3,33 | 4,17 | 8,34 | 14,7 | 16,9 | 21,7 |
| 10 | 2,56 | 3,94 | 4,87 | 9,34 | 16,0 | 18,3 | 23,2 |
| 11 | 3,05 | 4,57 | 5,58 | 10,3 | 17,3 | 19,7 | 24,7 |
| 12 | 3,57 | 5,23 | 6,3 | 11,3 | 18,5 | 21,0 | 26,2 |
| 13 | 4,11 | 5,89 | 7,04 | 12,3 | 19,8 | 22,4 | 27,7 |
| 14 | 4,66 | 6,57 | 7,79 | 13,3 | 21,1 | 23,7 | 29,1 |
| 15 | 5,23 | 7,26 | 8,55 | 14,3 | 22,3 | 25 | 30,6 |
| 16 | 5,81 | 7,96 | 9,31 | 15,3 | 23,5 | 26,3 | 32,0 |
| 17 | 6,41 | 8,67 | 10,1 | 16,3 | 24,8 | 27,6 | 33,4 |
| 18 | 7,01 | 9,39 | 10,9 | 17,3 | 26,0 | 28,9 | 34,8 |
| 19 | 7,63 | 10,1 | 11,7 | 18,3 | 27,2 | 30,1 | 36,2 |
| 20 | 8,26 | 10,9 | 12,4 | 19,3 | 28,4 | 31,4 | 37,6 |
| 21 | 8,90 | 11,6 | 13,2 | 20,3 | 29,6 | 32,7 | 38,9 |
| 22 | 9,54 | 12,3 | 14,0 | 21,3 | 30,8 | 33,9 | 40,3 |
| 23 | 10,2 | 13,1 | 14,8 | 22,3 | 32,0 | 35,2 | 41,6 |
| 24 | 10,9 | 13,8 | 15,7 | 23,3 | 33,2 | 36,4 | 43,0 |
| 25 | 11,5 | 14,6 | 16,5 | 24,3 | 34,4 | 37,7 | 44,3 |
| 26 | 12,2 | 15,4 | 17,3 | 25,3 | 35,6 | 38,9 | 45,6 |
| 27 | 12,9 | 16,2 | 18,1 | 26,3 | 36,7 | 40,1 | 47,0 |
| 28 | 13,6 | 16,9 | 18,9 | 27,3 | 37,9 | 41,3 | 48,3 |
| 29 | 14,3 | 17,7 | 19,8 | 28,3 | 39,1 | 42,6 | 49,6 |
| 30 | 15,0 | 18,5 | 20,6 | 29,3 | 40,3 | 43,8 | 50,6 |
| 40 | 22,2 | 26,5 | 29,1 | 39,3 | 51,8 | 55,6 | 63,7 |
| 60 | 37,5 | 43,2 | 46,5 | 59,3 | 74,4 | 79,1 | 88,4 |

Таблиця Д.5. Квантиль $F$-розподілу ($\alpha = 0,05$)

| ν1 | ν2 | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 15 | 20 | 24 | 30 | 40 | ∞ |
| 1 | 161,4 | 199,5 | 215,7 | 224,6 | 230,2 | 234 | 236,8 | 238,9 | 240,5 | 241,9 | 243,9 | 245,9 | 248 | 249,1 | 250,1 | 251,1 | 254,3 |
| 2 | 18,51 | 19 | 19,16 | 9,25 | 19,3 | 19,33 | 19,35 | 19,37 | 19,38 | 19,4 | 19,41 | 19,43 | 19,45 | 19,45 | 19,46 | 19,47 | 19,5 |
| 3 | 10,13 | 9,55 | 9,28 | 9,12 | 9,01 | 8,94 | 8,89 | 8,85 | 8,81 | 8,79 | 8,78 | 8,7 | 8,66 | 8,64 | 8,62 | 8,59 | 8,53 |
| 4 | 7,71 | 6,94 | 6,59 | 6,39 | 6,26 | 6,16 | 6,09 | 6,04 | 6 | 5,96 | 5,91 | 5,86 | 5,8 | 5,77 | 5,75 | 5,72 | 5,63 |
| 5 | 6,61 | 5,79 | 5,41 | 5,19 | 5,05 | 4,95 | 4,88 | 4,82 | 4,77 | 4,74 | 4,68 | 4,62 | 4,56 | 4,53 | 4,5 | 4,46 | 4,36 |
| 6 | 5,99 | 5,14 | 4,76 | 4,53 | 4,39 | 4,28 | 4,21 | 4,15 | 4,10 | 4,06 | 4,00 | 3,94 | 3,87 | 3,84 | 3,81 | 3,77 | 3,67 |
| 7 | 5,59 | 4,74 | 4,35 | 4,12 | 3,97 | 3,87 | 3,79 | 3,73 | 3,68 | 3,64 | 3,57 | 3,51 | 3,44 | 3,41 | 3,38 | 3,34 | 3,23 |
| 8 | 5,32 | 4,46 | 4,07 | 3,84 | 3,69 | 3,58 | 3,5 | 3,44 | 3,39 | 3,35 | 3,28 | 3,22 | 3,15 | 3,12 | 3,08 | 3,04 | 2,93 |
| 9 | 5,12 | 4,26 | 3,86 | 3,63 | 3,48 | 3,37 | 3,29 | 3,23 | 3,18 | 3,14 | 3,07 | 3,01 | 2,94 | 2,9 | 2,86 | 2,83 | 2,71 |
| 10 | 4,96 | 4,1 | 3,71 | 3,48 | 3,33 | 3,22 | 3,14 | 3,07 | 3,02 | 2,98 | 2,91 | 2,85 | 2,77 | 2,74 | 2,7 | 2,66 | 2,54 |
| 11 | 4,84 | 3,98 | 3,59 | 3,36 | 3,2 | 3,09 | 3,01 | 2,95 | 2,9 | 2,85 | 2,79 | 2,72 | 2,65 | 2,61 | 2,57 | 2,53 | 2,4 |
| 12 | 4,75 | 3,89 | 3,49 | 3,26 | 3,11 | 3,00 | 2,91 | 2,85 | 2,8 | 2,75 | 2,69 | 2,62 | 2,54 | 2,51 | 2,47 | 2,43 | 2,3 |
| 13 | 4,67 | 3,81 | 3,41 | 3,18 | 3,03 | 2,92 | 2,83 | 2,77 | 2,71 | 2,67 | 2,6 | 2,53 | 2,46 | 2,42 | 2,38 | 2,34 | 2,21 |
| 14 | 4,6 | 3,74 | 3,34 | 3,11 | 2,96 | 2,85 | 2,76 | 2,7 | 2,65 | 2,6 | 2,53 | 2,46 | 2,39 | 2,35 | 2,31 | 2,27 | 2,13 |
| 15 | 4,54 | 3,68 | 3,29 | 3,06 | 2,9 | 2,79 | 2,71 | 2,64 | 2,59 | 2,54 | 2,48 | 2,4 | 2,33 | 2,29 | 2,25 | 2,2 | 2,07 |
| 16 | 4,49 | 3,63 | 3,24 | 3,01 | 2,85 | 2,74 | 2,66 | 2,59 | 2,54 | 2,49 | 2,42 | 2,35 | 2,28 | 2,24 | 2,19 | 2,15 | 2,01 |
| 17 | 4,45 | 3,59 | 3,2 | 2,96 | 2,81 | 2,7 | 2,61 | 2,55 | 2,49 | 2,45 | 2,38 | 2,31 | 2,23 | 2,19 | 2,15 | 2,1 | 1,96 |
| 18 | 4,41 | 3,55 | 3,16 | 2,93 | 2,77 | 2,66 | 2,58 | 2,51 | 2,46 | 2,41 | 2,34 | 2,27 | 2,19 | 2,15 | 2,11 | 2,06 | 1,92 |
| 19 | 4,38 | 3,52 | 3,13 | 2,9 | 2,74 | 2,63 | 2,54 | 2,48 | 2,42 | 2,38 | 2,31 | 2,23 | 2,16 | 2,11 | 2,07 | 2,03 | 1,88 |

| v1 | v2 | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 15 | 20 | 24 | 30 | 40 | ∞ |
| 20 | 4,35 | 3,49 | 3,1 | 2,87 | 2,71 | 2,6 | 2,51 | 2,45 | 2,39 | 2,35 | 2,28 | 2,2 | 2,12 | 2,08 | 2,04 | 1,99 | 1,84 |
| 21 | 4,32 | 3,47 | 3,07 | 2,84 | 2,68 | 2,57 | 2,49 | 2,42 | 2,37 | 2,32 | 2,25 | 2,18 | 2,1 | 2,05 | 2,01 | 1,96 | 1,81 |
| 22 | 4,3 | 3,44 | 3,05 | 2,82 | 2,66 | 2,55 | 2,46 | 2,4 | 2,34 | 2,3 | 2,23 | 2,15 | 2,07 | 2,03 | 1,98 | 1,94 | 1,78 |
| 23 | 4,28 | 3,42 | 3,03 | 2,8 | 2,64 | 2,53 | 2,44 | 2,37 | 2,32 | 2,27 | 2,2 | 2,13 | 2,05 | 2,01 | 1,96 | 1,91 | 1,76 |
| 24 | 4,26 | 3,4 | 3,01 | 2,78 | 2,62 | 2,51 | 2,42 | 2,36 | 2,3 | 2,25 | 2,18 | 2,11 | 2,03 | 1,98 | 1,94 | 1,89 | 1,73 |
| 25 | 4,24 | 3,39 | 2,99 | 2,76 | 2,6 | 2,49 | 2,4 | 2,34 | 2,28 | 2,24 | 2,16 | 2,09 | 2,01 | 1,96 | 1,92 | 1,87 | 1,71 |
| 26 | 4,23 | 3,37 | 2,98 | 2,74 | 2,59 | 2,47 | 2,39 | 2,32 | 2,27 | 2,22 | 2,15 | 2,07 | 1,99 | 1,95 | 1,9 | 1,85 | 1,69 |
| 27 | 4,21 | 3,35 | 2,96 | 2,73 | 2,57 | 2,46 | 2,37 | 2,31 | 2,25 | 2,2 | 2,13 | 2,06 | 1,97 | 1,93 | 1,88 | 1,84 | 1,67 |
| 28 | 4,2 | 3,34 | 2,95 | 2,71 | 2,56 | 2,45 | 2,36 | 2,29 | 2,24 | 2,19 | 2,12 | 2,04 | 1,96 | 1,91 | 1,87 | 1,82 | 1,65 |
| 29 | 4,18 | 3,33 | 2,93 | 2,7 | 2,55 | 2,43 | 2,35 | 2,28 | 2,22 | 2,18 | 2,1 | 2,03 | 1,94 | 1,9 | 1,85 | 1,81 | 1,64 |
| 30 | 4,17 | 3,32 | 2,92 | 2,69 | 2,53 | 2,42 | 2,33 | 2,27 | 2,21 | 2,16 | 2,09 | 2,01 | 1,93 | 1,89 | 1,84 | 1,79 | 1,62 |
| 40 | 4,08 | 3,23 | 2,84 | 2,61 | 2,45 | 2,34 | 2,25 | 2,18 | 2,12 | 2,08 | 2 | 1,92 | 1,84 | 1,79 | 1,74 | 1,69 | 1,51 |
| 60 | 4,00 | 3,15 | 2,76 | 2,53 | 2,37 | 2,25 | 2,17 | 2,1 | 2,04 | 1,99 | 1,92 | 1,84 | 1,75 | 1,7 | 1,65 | 1,59 | 1,39 |
| 120 | 3,92 | 3,07 | 2,68 | 2,45 | 2,29 | 2,17 | 2,09 | 2,02 | 1,96 | 1,91 | 1,83 | 1,75 | 1,66 | 1,61 | 1,55 | 1,5 | 1,25 |
| | 3,84 | 3 | 2,6 | 2,37 | 2,21 | 2,1 | 2,01 | 1,94 | 1,88 | 1,83 | 1,75 | 1,67 | 1,57 | 1,52 | 1,46 | 1,39 | 1,00 |

Таблиця Д.6. $\alpha$-5 %-і межі для визначення $G$-коефіцієнта

| | $f_1$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $f_2$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 16 | 36 |
| 2 | 0,9985 | 0,9750 | 0,9392 | 0,9057 | 0,8584 | 0,8534 | 0,8332 | 0,8159 | 0,801 | 0,788 | 0,7341 | 0,6602 |
| 3 | 0,9669 | 0,8709 | 0,7977 | 0,7457 | 0,7071 | 0,6771 | 0,6530 | 0,6333 | 0,6167 | 0,6025 | 0,5466 | 0,4748 |
| 4 | 0,9065 | 0,7679 | 0,6841 | 0,6287 | 0,5895 | 0,5598 | 0,5365 | 0,5175 | 0,5017 | 0,4884 | 0,4366 | 0,3720 |
| 5 | 0,8412 | 0,6838 | 0,5981 | 0,5440 | 0,5063 | 0,4783 | 0,4564 | 0,4387 | 0,4241 | 0,4118 | 0,3645 | 0,3066 |
| 6 | 0,7808 | 0,6161 | 0,6321 | 0,4803 | 0,4447 | 0,4148 | 0,3980 | 0,3817 | 0,3682 | 0,3568 | 0,3135 | 0,2612 |
| 7 | 0,7271 | 0,5612 | 0,4800 | 0,4307 | 0,3907 | 0,3726 | 0,3555 | 0,3384 | 0,3254 | 0,3254 | 0,2756 | 0,2273 |
| 8 | 0,6798 | 0,5157 | 0,4377 | 0,3910 | 0,3595 | 0,3362 | 0,3165 | 0,3043 | 0,2926 | 0,2829 | 0,2462 | 0,202 |
| 9 | 0,6385 | 0,4775 | 0,4027 | 0,3584 | 0,3286 | 0,3067 | 0,2901 | 0,2768 | 0,2659 | 0,2568 | 0,2226 | 0,1820 |
| 10 | 0,6020 | 0,4450 | 0,3733 | 0,3311 | 0,3029 | 0,2823 | 0,2666 | 0,2541 | 0,2439 | 0,2353 | 0,2332 | 0,1655 |
| 12 | 0,6410 | 0,3924 | 0,3264 | 0,2880 | 0,2624 | 0,2439 | 0,2299 | 0,2187 | 0,2087 | 0,2020 | 0,1737 | 0,1403 |
| 15 | 0,4709 | 0,3346 | 0,2758 | 0,2419 | 0,2195 | 0,2034 | 0,1911 | 0,1315 | 0,1736 | 0,1671 | 0,1429 | 0,1144 |
| 20 | 0,3894 | 0,2705 | 0,2205 | 0,1921 | 0,1835 | 0,1602 | 0,1601 | 0,1422 | 0,1357 | 0,1303 | 0,1108 | 0,0879 |
| 24 | 0,3434 | 0,2354 | 0,1907 | 0,1656 | 0,1493 | 0,1374 | 0,1286 | 0,1216 | 0,116 | 0,1113 | 0,0942 | 0,0742 |
| 30 | 0,2929 | 0,1980 | 0,1593 | 0,1377 | 0,1237 | 0,1137 | 0,1061 | 0,1002 | 0,0958 | 0,0921 | 0,0771 | 0,0604 |
| 40 | 0,2370 | 0,1576 | 0,1259 | 0,1032 | 0,0968 | 0,0887 | 0,0827 | 0,0780 | 0,0745 | 0,0713 | 0,0595 | 0,0462 |
| 70 | 0,1737 | 0,1131 | 0,0395 | 0,0766 | 0,0623 | 0,0583 | 0,0582 | 0,0520 | 0,0487 | 0,0411 | 0,0316 | 0,0234 |
| 120 | 0,0998 | 0,0632 | 0,0495 | 0,0419 | 0,0371 | 0,0337 | 0,0312 | 0,0292 | 0,0279 | 0,0266 | 0,0218 | 0,0165 |

Таблиця Д.7. Критичні значення для випробувань Уїлкоксона

| $n_2$ | $n_2$ | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| 2 | – | – | – | – | 8,0 | 9,0 | 10,0 | 10,0 | 11,0 | 12,0 | 13,0 |
| 3 | – | 7,5 | 8,0 | 9,5 | 10,0 | 11,5 | 12,0 | 13,5 | 14,0 | 15,5 | 16,0 |
| 4 | 8,0 | 9,0 | 10,0 | 11,0 | 12,0 | 13,0 | 15,0 | 16,0 | 17,0 | 18,0 | 19,0 |
| 5 | 9,0 | 10,5 | 12,0 | 12,5 | 14,0 | 15,5 | 17,0 | 18,5 | 19,0 | 20,5 | 22,0 |
| 6 | | | 13,0 | 15,0 | 16,0 | 17,0 | 19,0 | 20,0 | 22,0 | 23,0 | 25,0 |
| 7 | | | | 16,5 | 18,0 | 19,5 | 21,0 | 22,5 | 24,0 | 25,5 | 27,0 |
| 8 | | | | | 19,0 | 21,0 | 23,0 | 25,0 | 26,0 | 28,0 | 29,0 |
| 9 | | | | | | 22,5 | 25,0 | 26,5 | 28,0 | 30,5 | 32,0 |
| 10 | | | | | | | 27,0 | 29,0 | 30,0 | 32,0 | 34,0 |
| 11 | | | | | | | | 30,5 | 33,0 | 34,5 | 37,0 |
| 12 | | | | | | | | | 35,0 | 37,0 | 39,0 |
| 13 | | | | | | | | | | 38,5 | 41,0 |
| 14 | | | | | | | | | | | 43,0 |

Таблиця Д.8. Межі значущості для критеріїв, основаних на серіях

$U_{0,025}$

| $N_2$ | $N_1$ | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| 2 | | | | | | | | | | | | | | | | | |
| 3 | | | | | | | | | | | | | | | | | |
| 4 | | | | | | | | | | | | | | | | | |
| 5 | | | 2 | 2 | | | | | | | | | | | | | |
| 6 | | 2 | 2 | 3 | 3 | | | | | | | | | | | | |
| 7 | | 2 | 2 | 3 | 3 | 3 | | | | | | | | | | | |
| 8 | | 2 | 3 | 3 | 3 | 4 | 4 | | | | | | | | | | |
| 9 | | 2 | 3 | 3 | 4 | 4 | 5 | 5 | | | | | | | | | |
| 10 | | 2 | 3 | 3 | 4 | 5 | 5 | 5 | 6 | | | | | | | | |
| 11 | | 2 | 3 | 3 | 4 | 5 | 5 | 6 | 6 | 7 | | | | | | | |
| 12 | 2 | 2 | 3 | 4 | 4 | 5 | 6 | 6 | 7 | 7 | 7 | | | | | | |
| 13 | 2 | 2 | 3 | 4 | 5 | 5 | 6 | 6 | 7 | 7 | 8 | 8 | | | | | |
| 14 | 2 | 2 | 3 | 4 | 5 | 5 | 6 | 7 | 7 | 8 | 8 | 9 | 9 | | | | |
| 15 | 2 | 3 | 3 | 4 | 5 | 6 | 6 | 7 | 7 | 8 | 8 | 9 | 9 | 10 | | | |
| 16 | 2 | 3 | 3 | 4 | 5 | 6 | 6 | 7 | 8 | 8 | 9 | 9 | 10 | 10 | 11 | | |
| 17 | 2 | 3 | 4 | 4 | 5 | 6 | 7 | 7 | 8 | 9 | 9 | 10 | 10 | 11 | 11 | 11 | |
| 18 | 2 | 3 | 4 | 5 | 5 | 6 | 7 | 8 | 8 | 9 | 9 | 10 | 10 | 11 | 11 | 12 | 12 |
| 19 | 2 | 3 | 4 | 5 | 6 | 6 | 7 | 8 | 8 | 9 | 10 | 10 | 11 | 11 | 12 | 12 | 13 |
| 20 | 2 | 3 | 4 | 5 | 6 | 6 | 7 | 8 | 9 | 9 | 10 | 10 | 11 | 12 | 12 | 13 | 13 |

$U_{0,975}$

| $N_2$ | $N_1$ | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| 1 | | | | | | | | | | | | | | | | | |
| 2 | 4 | | | | | | | | | | | | | | | | |
| 3 | 5 | 6 | | | | | | | | | | | | | | | |
| 4 | 5 | 7 | 8 | | | | | | | | | | | | | | |
| 5 | 5 | 7 | 8 | 9 | | | | | | | | | | | | | |
| 6 | 5 | 7 | 8 | 9 | 10 | | | | | | | | | | | | |
| 7 | 5 | 7 | 9 | 10 | 11 | 12 | | | | | | | | | | | |
| 8 | 5 | 7 | 9 | 10 | 11 | 12 | 13 | | | | | | | | | | |
| 9 | 5 | 7 | 9 | 11 | 12 | 13 | 13 | 14 | | | | | | | | | |
| 10 | 5 | 7 | 9 | 11 | 12 | 13 | 14 | 15 | 15 | | | | | | | | |
| 11 | 5 | 7 | 9 | 11 | 12 | 13 | 14 | 15 | 16 | 16 | | | | | | | |
| 12 | 5 | 7 | 9 | 11 | 12 | 13 | 15 | 15 | 16 | 17 | 18 | | | | | | |
| 13 | 5 | 7 | 9 | 11 | 13 | 14 | 15 | 16 | 17 | 18 | 18 | 19 | | | | | |
| 14 | 5 | 7 | 9 | 11 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 19 | 20 | | | | |
| 15 | 5 | 7 | 9 | 11 | 13 | 14 | 15 | 17 | 17 | 18 | 20 | 20 | 21 | 21 | | | |
| 16 | 5 | 7 | 9 | 11 | 13 | 15 | 16 | 17 | 18 | 19 | 20 | 20 | 21 | 22 | 23 | | |
| 17 | 5 | 7 | 9 | 11 | 13 | 15 | 16 | 17 | 18 | 19 | 21 | 21 | 22 | 22 | 23 | 24 | |
| 18 | 5 | 7 | 9 | 11 | 13 | 15 | 16 | 17 | 18 | 19 | 21 | 21 | 22 | 23 | 24 | 24 | 25 |
| 19 | 5 | 7 | 9 | 11 | 13 | 15 | 16 | 17 | 19 | 20 | 22 | 22 | 22 | 23 | 24 | 25 | 25 |
| 20 | 5 | 7 | 9 | 11 | 13 | 15 | 16 | 17 | 19 | 20 | 22 | 22 | 23 | 24 | 24 | 25 | 26 |

# ЗМІСТ