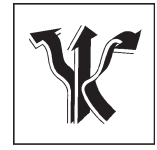
МІЖРЕГІОНАЛЬНА
АКАДЕМІЯ УПРАВЛІННЯ ПЕРСОНАЛОМ



МАУП

П. І. Бідюк, Б. П. Ткач, Т. Харрінгтон

МАТЕМАТИЧНА СТАТИСТИКА

Навчальний посібник

Київ
ДП «Видавничий дім «Персонал»
2018

УДК 519.766.4–004.942

Рецензенти: *Ф. Г. Гаращенко,* д-р техн. наук, проф. (Київ. нац. ун-т імені Тараса Шевченка);
М. В. Андрєєв, д-р фіз.-мат. наук, проф. (Ін-т прикладного системного аналізу НТУУ "КПІ");
І. І. Дахно, д-р екон. наук, проф.

Схвалено Вченою радою Міжрегіональної Академії управління персоналом (протокол № 5 від 27.05.15)

Математична статистика: навч. посіб. / П. І. Бідюк, Б. П. Ткач, Т. Харрінгтон. — К.: ДП «Вид. дім «Персонал», 2018. — 348 с. — Бібліогр.: с. 346.

ISBN 978-617-02-0234-5

Висвітлено сучасні методи статистичного аналізу даних, представлених часовими перерізами або часовими рядами. Розглянуто методику статистичного аналізу даних у вигляді чотирьох основних етапів, методи збору та попередньої обробки даних. Наведено основи побудови регресійних моделей процесів довільної природи. Запропоновано методику застосування розв'язків дискретних рівнянь типу авторегресії та авторегресії з ковзним середнім до аналізу поведінки випадкових процесів з детермінованими складовими та знаходження оцінок коротко- і середньострокових прогнозів динаміки їх розвитку. Розглянуто елементи теорії прийняття статистичних рішень, моделювання ризику та перевірки гіпотез. Наведено методики факторного і дискримінантного аналізу. Окремий розділ присвячено оптимальному статистичному оцінюванню станів динамічних систем за допомогою фільтра Калмана з використанням дискретних моделей у просторі станів, що враховують збурення станів та похибки вимірів. Подано процедуру статистичного моделювання фільтра, яка надає можливість формувати прогнозуючі розподіли, а також приклади застосування методів обробки статистичних даних до розв'язування реальних задач.

Для студентів, аспірантів та викладачів, а також для інженерів, які спеціалізуються у галузі розв'язання задач статистичного аналізу даних, математичного моделювання і прогнозування динаміки розвитку фінансово-економічних процесів та процесів іншої природи, представлених статистичними або експериментальними даними.

ISBN 978-617-02-0234-5

ЗМІСТ

ВСТУП

Статистичний аналіз даних (САД) виконується з метою побудови прогнозуючих моделей та прийняття рішень на основі оцінок прогнозів. Його застосовують у соціально-економічних, фінансових, технічних та інших системах для коротко- і середньострокового прогнозування обсягів виробництва, накопичення продукції на складах, оцінювання альтернативних економічних стратегій, формування бюджетів підприємств і держави, оцінювання, прогнозування та менеджменту ризиків різної природи, а також для розв'язання багатьох інших задач. Загалом, коло задач, що розв'язуються методами математичної статистики, дуже широке. Кількість різних підходів, методів, моделей та методик статистичного аналізу даних також надзвичайно велика, а тому сьогодні доцільно описувати в одній публікації один із напрямів розвитку САД, наприклад, методи попередньої обробки даних — перший напрям; регресійний аналіз — другий напрям; методи оцінювання параметрів математичних і статистичних моделей — ще один напрям і т. ін. Однак, такий підхід до висвітлення САД потребує досить великих матеріальних витрат на публікації. Тому ця публікація розрахована на вступний семестровий курс для студентів, окремі розділи будуть корисні також аспірантам. Побудова моделей, описаних у книзі, орієнтована на використання даних у формі часових рядів.

На сьогодні у спеціальній літературі описано досить багато методів аналізу статистичних та експериментальних даних з метою побудови математичних і статистичних моделей для прогнозування розвитку лінійних стаціонарних та нелінійних нестаціонарних процесів на основі використання даних у вигляді часових рядів. Найбільш поширеними серед них є регресійні методи — авторегресія (АР), авторегресія з ковзним середнім (АРКС), авторегресія з інтегрованим ковзним середнім (АРІКС), лінійна та нелінійна множинна регресія, квантильна регресія, регресійні дерева, метод групового урахування аргументів (МГВА), нейромережі різноманітних структур, байєсівські моделі і мережі, нечіткі множини, нечіткі нейромережі та ін. Відносно "універсальними" методами моделювання і прогнозування є МГВА і нечіткі нейромережі. Однак практика показує, що одного, навіть відносно універсального методу, не завжди достатньо для досягнення повноти аналізу досліджуваного процесу з подаль-

шою метою використання результатів аналізу для прийняття рішень.

Кожний метод аналізу даних має свої недоліки і переваги стосовно обчислювальних витрат, характеристик точності (адекватності) моделей і оцінок прогнозів. Так, висока (практично прийнятна) точність прогнозу за допомогою МГВА або нейромережі іноді досягається за рахунок високих обчислювальних витрат, які можуть виявитися неприйнятними в окремих випадках. Це особливо стосується застосування отриманої моделі у системі керування реального часу, де модель необхідна для оцінювання прогнозу керованих процесів і синтезу керуючого впливу у кожному періоді дискретизації вимірів. Суттєвий виграш стосовно обчислювальних витрат можна досягти у такому випадку, наприклад, за допомогою набагато простішої моделі авторегресії з ковзним середнім (АРКС) або авторегресії з інтегрованим ковзним середнім (АРІКС), перевагами якої є відносна простота структури та можливості її оперативної адаптації (структури і параметрів) до змінних характеристик процесу у реальному часі.

Іншим поширеним типом даних є часові перерізи, які характеризують стан процесу (об'єкта, суб'єкта) на вибраний момент часу. Наприклад, поточний стан макроекономіки країни можна характеризувати множиною змінних (350–400 змінних залежно від постановки задачі), які вимірюють на заданий момент часу, наприклад, на кінець року. Надалі ці дані можна використати для побудови статистичної моделі у формі багатовимірного розподілу або регресії спеціального типу. Якщо змінні, віднесені до часового перерізу, реєструвати через однакові проміжки часу (наприклад, кожні півроку), то далі можна перейти до представлення у формі часового ряду.

Загалом економетричні статистичні дані або експериментально отримані виміри містять множину структурних компонент, математичний опис яких потребує застосування сучасних методів ймовірнісно-статистичного аналізу, наведених у посібнику.

Метою даної роботи є: 1) — представлення основних етапів статистичного аналізу даних; 2) — аналіз основних описових статистик даних (моди, медіани, середнього, варіації); 3) — опис положення окремого спостереження у ряду розподілу (процентилі, *z*-оцінки, *T*-оцінки), групування даних та їх статистичні параметри; 4) — аналіз нормального розподілу, як одного із самих поширених у САД; 5) — аналіз процесу прийняття статистичних рішень і перевірка гіпотез; 6) — представлення методики побудови математичних (регресій-

них) моделей на основі статистичних даних; 7) — аналіз стаціонарності досліджуваних процесів; 8) — представлення методів факторного і дискримінантного аналізу; 9) — докладний аналіз методу оптимального оцінювання і прогнозування стану процесу (об'єкта) з урахуванням статистичних характеристик (коваріацій) випадкових зовнішніх збурень і шумів (похибок) вимірів за допомогою моделей у просторі станів (за допомогою фільтра Калмана).

Запропоновано нове визначення структури математичної моделі, яке складається з п'яти елементів і сприяє поглибленому розумінні студентами цього важливого поняття, яке використовується практично в усіх розділах навчального посібника. У посібнику запропонована оригінальна методика побудови регресійних моделей, коректне застосування якої дає можливість будувати лінійні та нелінійні математичні моделі високого ступеня адекватності процесам. Для встановлення ступеня адекватності побудованої моделі об'єкта, що моделюється, наведено належну множину статистичних критеріїв якості (коефіцієнт детермінації, *t*-статистика Стьюдента, статистика Дарбіна-Уотсона, інформаційний критерій Акайке, статистика Байєса-Шварца та ін.). Ще одна множина критеріїв запропонована для аналізу якості коротко- та середньострокових прогнозів, які оцінюються на основі створеної моделі (середня квадратична похибка, середня абсолютна похибка у процентах, коефіцієнт Тейла та ін.). Також представлена оригінальна методика знаходження оцінок багатокрокових прогнозів за допомогою функцій прогнозування, які можна отримати на основі побудованих математичних моделей АР та АРКС.

У посібнику запропонована методика знаходження розв'язків різницевих рівнянь (АР та АРКС). Розв'язки необхідні для аналізу досліджуваних процесів, використання їх у системах керування та імітаційного моделювання, а також для знаходження оцінок прогнозів, необхідних для прийняття рішень на їх основі.

Наведена у посібнику методика побудови і застосування оптимального фільтра буде корисною для студентів, аспірантів та інженерів, які зацікавлені у створенні багатовимірних (матричних) причинно-наслідкових моделей. Моделі такого типу знаходять широке застосування у моделюванні та прогнозуванні об'єктів керування у дослідженні та менеджменті ризиків, у системах діагностики в техніці та медицині, системах підтримки прийняття рішень різного призначення тощо. Попередній досвід побудови та використання оптималь-

них фільтрів свідчить про те, що це сучасний потужний інструмент статистичної обробки даних в умовах впливу на досліджувані об'єкти випадкових збурень і наявності похибок (шумів) вимірів. Інструментарій такого типу також надає можливість прогнозування розвитку процесів різної природи.

Автори сподіваються, що посібник загалом буде корисним для студентів, аспірантів, інженерів, викладачів та всіх тих, хто займається практичними задачами поглибленого статистичного аналізу даних, математичного моделювання і прогнозування на основі статистичних та експериментальних даних у різних галузях техніки, економіки, фінансів, екології, соціальних дослідженнях і т. ін.

ПЕРЕЛІК СКОРОЧЕНЬ

АКФ – автокореляційна функція
АР – авторегресія
АРКС – авторегресія з ковзним середнім
АРІКС – авторегресія з інтегрованим ковзним середнім
ВВ – випадкова величина
ВВП – валовий внутрішній продукт
ВП – випадковий процес
ЕПП – економіка перехідного періоду
ІІ – індекс інфляції
КС – ковзне середнє
КФ – кореляційна функція
КЧК – коефіцієнт часткової кореляції
МВК – модель випадкового кроку
МКП – модель коригування похибки
ММ – математична модель
ММП – метод максимальної правдоподібності
МНК – метод найменших квадратів
МР – множинна регресія
МС – математичне сподівання
НКФ – нелінійна кореляційна функція
НСП – нестаціонарний процес
ОФ – оптимальний фільтр
ППР – правило прийняття рішень
ПС – простір станів
ПОВС – похибка оцінки вектора стану
РМНК – рекурсивний метод найменших квадратів
РР – різницеве рівняння
САП – середня абсолютна похибка
САД – статистичний аналіз даних

САПВ – середня абсолютна похибка у відсотках
СКП – середня квадратична похибка
СП – стаціонарний процес
ФК – фільтр Калмана
ЦФ – цифровий фільтр
ЧАКФ – часткова автокореляційна функція

Розділ 1

ЧОТИРИ ЕТАПИ СТАТИСТИЧНОГО АНАЛІЗУ

Математична статистика — це мистецтво і наука збору, обробки даних і аналізу отриманих результатів з метою прийняття коректних обґрунтованих рішень технічного, ділового, політичного, персонального або іншого характеру.

Статистичні методи необхідно розглядати як важливу трудомістку частину процесу прийняття рішень, яка дає можливість приймати обґрунтовані тактичні і стратегічні рішення. Ці методи ґрунтуються на знаннях та інтуїції фахівців і глибокому аналізі наявної інформації.

Коректне використання статистичних методів надає значні переваги практично в усіх напрямах людської діяльності: в конкуренції за підвищення якості та продаж нової продукції, дає можливість отримати високоякісні оцінки прогнозів, обґрунтувати мікро- і макроекономічні рішення, рішення стосовно раціонального ведення домашнього господарства, а також розв'язати багато інших задач, у тому числі персонального характеру.

1.1. Задачі, які розв'язують методами математичної статистики

1. Аналіз стану ринку і прийняття рішення стосовно номенклатури та обсягу випуску продукції виробничого підприємства.
2. Аналіз, прогнозування і управління соціально-економічними процесами і системами (регіональний, галузевий та державний рівні).
3. Управління якістю продукції на виробництві.
4. Управління якістю навчання на всіх рівнях підготовки персоналу.
5. Автоматичне керування технологічними процесами і технічними системами — побудова адекватних моделей, застосування методів статистичного керування, прогнозування розвитку процесів і аналіз якості керування за допомогою множини відповідних статистичних параметрів.

6. Прогнозування і планування розвитку процесів різної природи на всіх рівнях ієрархії прийняття управлінських рішень.
7. Статистична підтримка прийняття експертних рішень з використанням високорозвинених комп'ютерних систем підтримки прийняття рішень (СППР), особливо з використанням інтелектуальних СППР.
8. Керування фізичними експериментами, аналіз даних і поглиблене дослідження отриманих результатів.
9. Виконання соціальних досліджень — надзвичайно популярний напрям досліджень внаслідок значного ускладнення і прискорення розвитку соціально-економічних процесів.
10. Підтримка прийняття особистих рішень та рішень стосовно ведення домашнього господарства.
11. Контроль стану навколишнього середовища з використанням двох типів моделей — на основі диференціальних рівнянь з частинними похідними і рівнянь авторегресії (АР), авторегресії з ковзним середнім (АРКС), множинної регресії.
12. Аналіз даних і прийняття рішень в генетиці, біології (існують такі напрями досліджень: біостатистика, біоінформатика), психології (наприклад, у США виходить міжнародний *Journal of Experimental Psychology*, який має велику популярність у всьому світі).

На початковій стадії виконання статистичного аналізу власне даних для аналізу, як правило, немає або ще навіть не прийнято рішення стосовно того, які дані необхідно аналізувати. Це питання вирішують на першому етапі *планування збору даних* таким чином, щоб отримати дійсно корисні, максимально повні та інформативні дані.

На другому етапі *висувають гіпотези* та виконують *попередню обробку* і *дослідження* даних.

Третій етап — *оцінювання необхідних статистичних параметрів та інших невідомих величин*, параметрів математичних моделей і т. ін.

На останньому, четвертому, етапі виконується *перевірка висунутих гіпотез* — дані використовують для прийняття рішення щодо відповідності апріорно висунутого припущення дійсній ситуації. Розглянемо етапи статистичного аналізу докладніше.

Чотири етапи статистичного аналізу даних:

1. Планування експерименту і збір даних. Результат — корисні інформативні та повні дані стосовно функціонування (протікання) процесу.

2. Попередня обробка та дослідження даних, формулювання гіпотез стосовно типів розподілів, значущості оцінок, якості (адекватності) моделей, настання можливих ситуацій і т. ін.
3. Оцінювання параметрів математичних і статистичних моделей. Результат — адекватні процесу математичні і статистичні моделі, що підтверджуються відповідними статистичними параметрами якості.
4. Перевірка раніше сформульованих гіпотез, прийняття рішень стосовно управління процесами.

Розглянемо деякі терміни, що стосуються статистичних даних і моделей, які оцінюються на їх основі.

Математичні моделі, які ми будемо розглядати, це моделі у вигляді рівнянь різних типів: диференціальних, різницевих, алгебраїчних.

Статистичні моделі — це моделі у формі розподілів ймовірностей випадкових величин.

Статистичні дані у формі вимірів, прив'язаних до конкретних моментів часу, тобто *часових рядів,* розглядаються у математичній статистиці як випадкові процеси з детермінованими складовими:

$$y(k) = \text{Константа} + \text{Детермінована складова} + \\ + \text{Випадкова складова}.$$

Детермінована складова процесу надає можливість будувати регресійні моделі та моделі деяких інших типів. Так, наявність автокореляції часового ряду дає можливість будувати авторегресійні моделі, порядок яких визначається за допомогою автокореляційної і часткової автокореляційної функцій.

Статистичні дані у формі вимірів, які стосуються деякого вибраного моменту часу (наприклад, конкретного місяця року), називають *часовими перерізами*.

Поява випадкової складової у вимірах змінних досліджуваних процесів зумовлена такими причинами:

- наявність випадкових збурень, які, як правило, неможливо виміряти;
- некоректне формування структури моделі: у праву частину рівняння можуть бути введені "зайві" змінні, тобто такі, що формально корельовані із залежною змінною, але фактично на неї не впливають або впливають дуже слабо;

- можливо, що у праву частину рівняння помилково не введені незалежні змінні, які фактично мають досить сильний вплив на залежну;
- похибки обчислень, які завжди виникають при оцінюванні параметрів моделей.

1.2. Планування збору даних (перший етап статистичного аналізу)

Планування збору даних (планування експерименту) — перший етап статистичного аналізу даних. Планування збору даних в соціально-економічних, фінансових та маркетингових дослідженнях називають *плануванням вибіркового дослідження*.

У процесі планування вибіркового дослідження розв'язують задачі, які можна розділити на групи.

Група 1. **Визначення цілей дослідження та їх представлення у конкретному вигляді**

До цієї групи відносять такі завдання:

- визначити область (галузь) дослідження (економіка, політика, соціологія і т. ін.);
- визначити головну мету, наприклад, оцінювання та прогнозування стану галузі промисловості або стійкості виробничого підприємства (його близькість до банкрутства);
- аналізувати структуру досліджуваного процесу — з яких елементів він складається (наприклад, університет складається з факультетів, інститутів та кафедр, а підприємство може мати цехи, відділи, філії і т. ін.);
- встановити перелік змінних і параметрів, які необхідно виміряти чи обчислити;
- визначити показники (змінні), які будуть представляти остаточний результат дослідження, тобто, що буде отримано після закінчення роботи (наприклад, оцінки коротко- або середньострокового прогнозу розвитку процесу, рейтинг кандидата у президенти або середня успішність студентів університету для вибраних спеціальностей та курсів).

Група 2. **Створення загального плану дослідження**

Для створення загального плану дослідження розв'язують такі задачі:

- вибрати одиницю, стосовно якої буде виконуватись дослідження (наприклад, сім'я, окремі особи або групи, підприємство, район чи область);
- визначити: чи будуть вибрані об'єкти досліджуватись за вибіркою чи за генеральною сукупністю даних (виконання збору максимального можливого об'єму даних);
- вибрати метод отримання інформації стосовно об'єкта (за допомогою інтерв'ю, Інтернету, поштою або спеціальної апаратури чи шляхом простого спостереження);
- вибрати тип бази даних для накопичення інформації;
- оцінити вартість кожної операції;
- встановити вимоги щодо точності результатів дослідження (як правило, збільшення об'єму вибірки сприяє підвищенню точності результату, але підвищує вартість дослідження).

***Група 3*. Планування характеру вибірки**

У процесі планування характеру вибірки даних необхідно розглянути такі задачі:

- встановити необхідність стратифікації, тобто необхідно досліджувати об'єкт на одному рівні чи більше (наприклад, успішність в університеті може досліджуватись на таких рівнях: молодші курси, старші курси, аспіранти, друга вища освіта та курси підвищення кваліфікації);
- визначити вибіркову одиницю, її характер і величину (наприклад, вибірковою одиницею може бути один студент або група студентів);
- встановити кількість і типи стадій збору даних (наприклад, збір даних по семестрах);
- оцінити кількість і типи фаз виконання кожної стадії (наприклад, послідовність збору даних по групах);
- встановити кількість вибіркових одиниць, які будуть взяті на кожній стадії і фазі;
- оцінити вартість виконання операцій стосовно збору даних;
- встановити метод відбору одиниць спостережень — ймовірнісний або на основі суджень, експертний, тобто якість даних можна оцінювати на основі знаходження ймовірностей появи некоректних значень (наприклад, відомо, що процент недостовірних даних може сягати 10–15 %) або ж якість даних оцінюють експерти;

- встановити способи подолання труднощів, які можуть виникнути в процесі отримання вибіркових даних (дані не можна отримати у випадку небажання їх давати або у випадку захворювання суб'єкта дослідження і т. ін.);
- вибрати (розробити) методи обчислення необхідних статистичних оцінок за вибірковими даними.

***Група 4*. Безпосередній збір даних**

Безпосередній збір даних передбачає виконання таких операцій:

- розробка форми реєстрації даних (анкета, таблиця, виборчий бюлетень, щоденник, цифровий реєстратор і т. ін.);
- вибір методу вибору, навчання та контролю роботи дослідника (інтерв'юера, спостерігача і т. ін.);
- вибір альтернативного (відмінного від вибіркового) методу збору спостережень у випадку виникнення труднощів із отриманням даних стосовно досліджуваного об'єкта;
- безпосередній збір даних прямим або непрямим (опосередкованим) методом.

Дослідження технічних об'єктів і технологічних процесів має свої особливості. Перший етап називають *плануванням* та *виконанням експерименту*. На цьому етапі виконують такі дії:

- визначають мету дослідження;
- визначають об'єкт дослідження та можливості виконання експерименту з цим об'єктом;
- вибирають змінні для вимірювання, їх граничні значення та створюють методику збору даних;
- визначають тип бази даних та інструментальну систему для її створення;
- визначають необхідний для дослідження процесу обсяг даних;
- визначають число учасників і час проведення дослідження;
- вибирають технічне устаткування для виконання дослідження, засоби обчислювальної техніки;
- визначають вартість дослідження.

Вибіркові статистичні дослідження особливо корисні тоді, коли є великі групи людей, підприємств або інших об'єктів чи процесів (*генеральна сукупність*), які необхідно дослідити, але повне дослідження провести неможливо. Для того щоб отримати неідеальне, але практично корисне розуміння ситуації щодо генеральної сукупності, мож-

на відібрати невелику групу (*вибірку*) даних, яка складається із деяких, але не всіх, об'єктів генеральної сукупності.

Процес узагальнення результатів дослідження на всю генеральну сукупність називають статистичним висновком або рішенням

Випадкова вибірка з генеральної сукупності є одним із кращих способів накопичення даних, оскільки генеральна сукупність, як правило, занадто велика, щоб її вивчати повністю.

1.3. Попередня обробка і дослідження даних (другий етап статистичного аналізу даних)

Попередня обробка і дослідження даних — важлива частина вибіркового дослідження, оскільки кваліфікований попередній аналіз дає можливість уникнути багатьох трудомістких обчислювальних процедур і скоротити час виконання дослідження.

Метою попередньої обробки і дослідження даних є:

- приведення даних до формату, який сприяє підвищенню їх якості з точки зору оцінювання параметрів моделей, тобто підвищення ступеня обумовленості матриць вимірів;
- попереднє оцінювання структури моделі.

Попередня обробка передбачає виконання таких операцій над зібраними статистичними даними.

1. Візуальне дослідження даних

За допомогою графіків, діаграм та описових статистик попередньо виявляють наявність можливих *трендів* (інтегрованість та коінтегрованість процесів), *сезонних ефектів*, *екстремальних значень*.

Тренд — поточне середнє значення процесу, що вказує на його довгострокові зміни. Як правило, аналіз тренду виконують при розв'язанні задачі довгострокового прогнозування. Короткострокові зміни визначаються коливаннями, які накладаються на тренд.

Називають нестаціонарний процес із трендом по-різному:

Процес із трендом = Інтегрований процес =
= Процес із одиничними коренями.

Термін "інтегрований" виникає від зовнішньої подібності графіка тренду з виходом електронного пристрою — *інтегратора*. При подачі на вхід інтегратора сигналу у вигляді константи, на його виході фор-

мується сигнал у вигляді прямої з додатним нахилом (лінійний тренд).

Термін "процес з одиничними коренями" означає, що серед коренів характеристичного рівняння, записаного для рівняння АР(p) або АРКС (p, q), є хоча б один одиничний корінь. Кількість одиничних коренів відповідає порядку тренду. Як приклад, розглянемо рівняння авторегресії першого порядку, АР(1)

$$y(k) = a_0 + a_1 y(k-1) + \varepsilon(k).$$

Однорідне рівняння для цієї моделі має вигляд

$$y(k) - a_1 y(k-1) = 0.$$

Підставимо в це однорідне рівняння однорідний розв'язок загального вигляду $y^h(k) = A\lambda^k$ і отримаємо

$$A\lambda^k - a_1 A\lambda^{k-1} = 0.$$

Поділимо обидві частини останнього рівняння на $A\lambda^{k-1}$ і отримаємо характеристичне рівняння першого порядку для однорідного рівняння АР(1)

$$\lambda - a_1 = 0.$$

Звідси корінь характеристичного рівняння: $\lambda = a_1$. Якщо $a_1 = 1$, то таке рівняння АР(1) описує процес з лінійним трендом (тренд першого порядку), який позначають через $I(1)$. Це можна показати, якщо знайти розв'язок рівняння

$$y(k) = a_0 + y(k-1) + \varepsilon(k).$$

Знайдемо розв'язок ітераційним методом при початкових умовах $y(0) = y_0$ і $\varepsilon(0) = \varepsilon_0$:

$$y(1) = a_0 + y_0 + \varepsilon(1);$$
$$y(2) = a_0 + y(1) + \varepsilon(2) = a_0 + a_0 + y_0 + \varepsilon(1) + \varepsilon(2).$$

За індукцією для довільного моменту часу k отримаємо розв'язок:

$$y(k) = y_0 + a_0 k + \sum_{i=1}^{k} \varepsilon(i),$$

який містить лінійний тренд $a_0 k$.

Візуальний аналіз дає можливість (приблизно) встановити наявність *гетероскедастичності* — коли дисперсія процесу змінюється у часі за лінійним або складнішим законом. Різні рівні дисперсії на

різних ділянках ряду даних свідчать про наявність гетероскедастичності і необхідності застосування спеціальних моделей для опису самих даних і дисперсії процесу. Моделі дисперсії необхідно будувати для того, щоб отримати можливість обчислювати оцінки прогнозів дисперсії і стандартного відхилення (стандартне відхилення називають ще волатильністю, тобто ступенем мінливості досліджуваного процесу).

За допомогою візуального аналізу можна приблизно оцінити тип розподілу за гістограмою та описовими статистичними параметрами (ексцес, асиметрія, статистика Жарк-Бера, статистика Колмогорова-Смірнова та ін.), а також методами, які ґрунтуються на використанні функцій-ядер.

Деякі формули для розрахунку описових статистик

Приклади статистичних параметрів, які використовуються з метою встановлення наближення розподілу до нормального.

Критерій узгодженості χ^2. Це критерій загального типу, який ґрунтується на порівнянні емпіричної гістограми розподілу випадкової величини з її теоретичною щільністю. Діапазон зміни експериментальних даних розбивають на m інтервалів і обчислюють статистику:

$$\chi^2 = \sum_{k=1}^{m} \frac{\left(n_k - Np_k\right)^2}{Np_k},$$

де n_k — кількість значень випадкової величини, які попадають в k-й інтервал; m — кількість інтервалів; $N = \sum_{k=1}^{m} n_k$ — об'єм вибірки; $p_k = F(x_{k+1}) - F(x_k)$ — теоретична ймовірність попадання випадкової величини в k-й інтервал; $F(x)$ — гіпотетичний теоретичний закон розподілу ймовірностей випадкової величини.

Дисперсія статистики χ^2 визначається за формулою [4]

$$\mathrm{var}\left(\chi^2\right) = 2\left(m-1\right) + \frac{1}{N}\left(\sum_{k=1}^{m}\frac{1}{p_k} - m^2 - 2m + 2\right).$$

Якщо $\sum_{k=1}^{m}\frac{1}{p_k} << N$ і $m << N$, то $\mathrm{var}(\chi^2)$ співпадає з дисперсією випадкової величини, яка має розподіл χ^2. На цій основі прийнято вважати, що статистика χ^2 має розподіл, близький до розподілу *хі*-квадрат.

Коефіцієнт асиметрії (*skewness*) — характеризує симетричність (хвостів) розподілу і розраховується за формулою

$$S = \frac{1}{N}\sum_{k=1}^{N}\left[\frac{y(k)-\overline{y}}{\sigma}\right]^3,$$

де $\overline{y}$ — вибіркове середнє; σ — стандартне відхилення процесу. Якщо $S > 0$, то правий хвіст розподілу довший, а при $S < 0$ довшим є лівий хвіст розподілу; якщо $S = 0$, то розподіл симетричний.

Ексцес (коефіцієнт гостровершинності або *kurtosis*) — характеризує відмінність форми розподілу від нормального і розраховується за формулою

$$K = \frac{1}{N}\sum_{k=1}^{N}\left[\frac{y(k)-\overline{y}}{\sigma}\right]^4.$$

$K = 3$ для нормального розподілу; якщо $K > 3$, то форма розподілу буде "гострішою" від нормального; при $K < 3$ форма розподілу буде "плоскішою" від нормального.

Статистика Жак-Бера (*Jarque-Bera*) [4] — тестова статистика, яка показує, наскільки близьким є емпіричний ряд розподілу до нормального. Це різниця між значеннями S і K для досліджуваного ряду та нормального розподілу:

$$JB = \frac{N-p}{6}\left[S^2 + \frac{1}{4}(K-3)^2\right],$$

де p — кількість коефіцієнтів, використаних для побудови моделі ряду даних. При нуль-гіпотезі щодо нормальності розподілу статистика Жак-Бера має розподіл χ^2 з двома ступенями вільності. Ймовірність, пов'язана із статистикою Жак-Бера, показує ймовірність справедливості нуль-гіпотези. *Мала ймовірність (близька до нуля) свідчить про те, що нуль-гіпотезу щодо нормальності розподілу необхідно відхилити.*

Шляхом візуального аналізу можна приблизно встановити наявність *екстремальних значень,* тобто значень, які суттєво перевищують всі інші значення вибірки, але не відносяться до похибок вимірів. Необхідно встановити походження наявних екстремальних значень і вибрати метод їх обробки. Як правило, за вибраною методикою зменшують амплітуди екстремальних значень з метою наближення процесу до стаціонарного.

Крім того, виявляють наявність нелінійностей у статистичних даних. *Нелінійні процеси* вимагають застосування спеціальних моделей для опису та оцінювання параметрів.

Візуальний аналіз дає можливість визначити наявність *перехідних та стаціонарних режимів* функціонування процесу. Перехідний — це, як правило, відносно короткостроковий режим функціонування процесу, який може характеризуватись значними коливаннями значень змінних. Перехідні процеси у макроекономіці можуть тривати десятиліття, а перехідні процеси у технічних системах можуть тривати від десятків мікросекунд до десятків хвилин. Стаціонарний режим — це режим, у якому досліджуваний об'єкт може функціонувати тривалий час (кілька годин, днів, років, десятиліть — для макроекономіки).

Встановлюється наявність коливань: *слабких*, що зумовлені похибками вимірів, і *сильних,* що зумовлені різноманітними випадковими впливами або збуреннями.

2. Заповнення пропусків даних, якщо вони є

Для заповнення пропусків даних використовують такі методи: *просте усереднення, екстраполяція, інтерполяція, прогнозування* та ін.

Приклад прогнозування без розв'язку рівнянь

Структура різницевого рівняння така, що воно дає змогу виконувати прогнозування на один крок (один період дискретизації вимірів) без додаткових перетворень. Тобто в праву частину необхідно підставити минулі значення змінних і обчислити оцінку прогнозу головної змінної в лівій частині. Але для того, щоб знайти оцінку прогнозу на більше число кроків, необхідно застосувати деякі попередні перетворення різницевих рівнянь. Розглянемо деякі можливі підходи до формування функцій прогнозування та обчислення оцінок прогнозованих значень.

Як приклад, розглянемо рівняння АР(1):

$$y(k) = a_0 + a_1 y(k-1) + \varepsilon(k),\ E[\varepsilon(k)] = 0. \qquad (1.3.1)$$

Збільшимо незалежну змінну часу на одну одиницю і запишемо рівняння знову

$$y(k+1) = a_0 + a_1 y(k) + \varepsilon(k+1). \qquad (1.3.2)$$

Якщо коефіцієнти a_0, a_1 відомі, то можна знайти умовне математичне сподівання на основі відомої інформації до моменту k включно:

$$E_k[y(k+1)] = \mathrm{E}_k[y(k+1)|\, y(k), y(k-1), ..., \varepsilon(k), \varepsilon(k-1), ...] = \\ = a_0 + a_1 E_k[y(k)] = a_0 + a_1 y(k), \qquad (1.3.3)$$

оскільки $y(k)$ в момент k є відомою константою.

По аналогії запишемо рівняння (1.3.1) для моменту $k + 2$:

$$y(k + 2) = a_0 + a_1 y(k + 1) + \varepsilon (k + 2). \qquad (1.3.4)$$

і знайдемо умовне математичне сподівання:

$$E_k[y(k + 2)] = a_0 + a_1 E_k[y(k + 1)] = a_0 + a_1 E_k[a_0 + a_1 y(k)] = \\ = a_0 + a_0 a_1 + a_1^2 y(k).$$

Для наступного моменту часу маємо:

$$E_k[y(k + 3)] = a_0 + a_0 a_1 + a_0 a_1^2 + a_1^3 y(k).$$

Таким чином, для загального випадку прогнозування на s кроків можна записати:

$$E_s\left[y(k+s)\right] = a_0\left(\sum_{i=0}^{s-1} a_1^i\right) + a_1^s y(k) = a_0 \sum_{i=0}^{s-1} a_1^i + a_1^s y(k). \qquad (1.3.5)$$

Отримане рівняння називають функцією прогнозування на довільне число кроків.

3. Обробка екстремальних значень

Екстремальними називають значення, які суттєво відрізняються від середнього значення вибірки, але не відносяться до помилкових.

Мета виявлення та обробки екстремальних значень — наблизити процес до стаціонарного для того, щоб можна було застосувати теорію аналізу стаціонарних процесів.

Методи виявлення та опису екстремальних значень

- за розподілами середніх для екстремальних значень;
- за допомогою характеристичних екстремальних значень;
- з використанням розподілів вимірів типу експоненціального, логістичного, логнормального;
- шляхом вибору та оцінювання параметрів спеціальних функцій для опису розподілу.

4. Нормування даних

Нормування даних можна виконувати за допомогою різних процедур, але при цьому важливо, щоб перетворення вимірів було лінійним. Розглянемо деякі підходи до нормування даних [1–4].

1. Нормування додатних значень шляхом їх *логарифмування*:

$$y_l(k) = \ln[y(k)].$$

2. Нормування діленням на максимальне значення:

$$y_{\text{норм}}(k)=\frac{y(k)}{|Y_{\max}|}\Rightarrow -1\le y(k)\le +1. \qquad (1.3.7)$$

Нормоване значення може бути визначене також і в іншому діапазоні: $-10 \le y(k) \le 10$.

3. Досить хороші результати нормування при оцінюванні множинної регресії:

$$y(k) = \beta_0 + \beta_1 x_1(k) + \beta_2 x_2(k) + ... + \beta_p x_p(k) + \varepsilon(k) \qquad (1.3.8)$$

можна досягти завдяки одночасному нормуванню і центруванню даних таким чином:

$$x_{i\text{H}}(k)=\frac{x_i(k)-\overline{x}_i}{\sqrt{S_x}}=\frac{x_i(k)-\overline{x}_i}{\left(\dfrac{\sum\limits_{k=1}^{N}\left(x_i(k)-\overline{x}_i\right)^2}{N-1}\right)^{1/2}},\quad k=1,\ldots,N,\quad i=1,\ldots,p;$$

$$y_{\text{H}}(k)=\frac{y(k)-\overline{y}}{\sqrt{S_y}}=\frac{y(k)-\overline{y}}{\left(\dfrac{\sum\limits_{k=1}^{N}\left(y_i(k)-\overline{y}\right)^2}{N-1}\right)^{1/2}}, \qquad (1.3.9)$$

де $x_i(k)$ — значення i-го стовпчика матриці вимірів (виміри незалежних змінних); $y(k)$ — виміри залежної змінної; $x_{i\text{H}}(k)$, $y_{i\text{H}}(k)$ — нормовані значення змінних; $\overline{x}_i, \overline{y}$ — вибіркові середні значення незалежних і залежної змінних, відповідно; N — кількість вимірів; p — кількість незалежних змінних (регресорів) x_i. Якщо ввести позначення для центрованих змінних

$$\tilde{x}_i(k)=x_i(k)-\overline{x}_i;\quad \tilde{y}(k)=y(k)-\overline{y}, \qquad (1.3.10)$$

то регресія для центрованих змінних матиме вигляд:

$$\tilde{y}(k)=\beta_1\tilde{x}_1(k)+\beta_2\tilde{x}_2(k)+\ldots+\beta_p\tilde{x}_p(k)+\varepsilon(k). \qquad (1.3.11)$$

Якщо підставити (1.3.9) в (1.3.11), то рівняння множинної регресії прийме вигляд:

$$y_{\text{H}}(k)S_y^{1/2}=\beta_1 S_1^{1/2}x_{1\text{H}}(k)+\beta_2 S_2^{1/2}x_{2\text{H}}(k)+\ldots+\beta_p S_p^{1/2}x_{p\text{H}}(k)+\varepsilon'(k). \qquad (1.3.12)$$

Тепер поділимо ліву і праву частини на $S_y^{1/2}$:

$$y_{\text{Н}}(k) = \alpha_1 x_{1\text{Н}}(k) + \alpha_2 x_{2\text{Н}}(k) + \ldots + \alpha_p x_{p\text{Н}}(k) + \varepsilon''(k), \quad (1.3.13)$$

де $\alpha_1 = \beta_1 \left(S_1 / S_y\right)^{1/2}, \ldots, \alpha_p = \beta_p \left(S_p / S_y\right)^{1/2}$. Отримане рівняння (1.3.13) — це рівняння для нормованих вимірів.

У результаті центрування і нормування покращується ступінь зумовленості матриці вимірів, яка вимірюється відношенням:

$$\eta = \left| \frac{\lambda_{\max}}{\lambda_{\min}} \right|,$$

де $\lambda_{\max}$, $\lambda_{\min}$ — максимальне і мінімальне власні числа матриці вимірів. Для забезпечення належних умов оцінювання параметрів необхідно задовольнити умову: $\eta < 10$ (емпірично встановлена рекомендація).

5. Фільтрація даних від шумів

За необхідністю виконується фільтрація даних від шумів. Для розв'язання цієї важливої задачі застосовують *цифрові* або *оптимальні* фільтри.

Цифровий фільтр (ЦФ) можна представити, наприклад, рівнянням типу АР(p):

$$y(k) = a_1 y(k-1) + a_2 y(k-2) + \ldots + a_p y(k-p).$$

Фільтр має амплітудно-частотну характеристику (АЧХ), яка визначається значеннями коефіцієнтів рівняння.

Мета застосування ЦФ: пропустити корисну частину спектра і затримати шумову або просто непотрібну для аналізу складову.

Оптимальний фільтр потребує використання моделі процесу, представленої у просторі станів. Основна задача застосування оптимального фільтра полягає в обчисленні оптимальних оцінок стану досліджуваного процесу із врахуванням впливу на його функціонування випадкових збурень та похибок (шумів) вимірів.

Так, нестаціонарна лінійна система описується у дискретному часі рівняннями із змінними у часі коефіцієнтами (тобто коефіцієнти або параметри моделі залежать від часу k) [13; 14]:

$$\mathbf{x}(k) = \mathbf{A}(k, k-1)\,\mathbf{x}(k-1) + \mathbf{B}(k, k-1)\,\mathbf{u}(k-1) + \mathbf{w}(k),$$

де $\mathbf{x}(k)$ — n-вимірний вектор станів системи; $\mathbf{u}(k-1)$ — m-вимірний вектор детермінованих вхідних величин (сигнали керування); $\mathbf{w}(k)$ — n-вимірний вектор випадкових зовнішніх збурень; $\mathbf{A}(k, k-1)$ —

$(n \times m)$ матриця динаміки системи (вона містить коефіцієнти, що характеризують динаміку, тобто швидкість зміни станів у часі); $\mathbf{B}(k, k-1)$ — $(n \times m)$ матриця коефіцієнтів керування. Подвійний часовий аргумент у вигляді $(k, k-1)$ означає, що величина з цим аргументом використовується в момент k, але її значення ґрунтується на попередніх даних, які відомі на момент $k-1$, включно. Далі будемо записувати для простоти матриці $\mathbf{A}$ і $\mathbf{B}$ з одним аргументом, тобто $\mathbf{A}(k)$ та $\mathbf{B}(k)$. Очевидно, що стаціонарна система описується матрицями з постійними коефіцієнтами, які записують просто $\mathbf{A}$ і $\mathbf{B}$. Оскільки матриця $\mathbf{A}$ зв'язує поточний стан із попереднім, то її називають ще перехідною матрицею станів. Нагадаємо, що дискретний час k зв'язаний з неперервним часом t періодом дискретизації вимірів T_s: $t = kT_s$. Використання поняття дискретного часу відповідає характеру зібраних даних і дає можливість спростити обчислення.

У класичній постановці задачі оптимальної фільтрації послідовність зовнішніх збурень $\mathbf{w}(k)$ задовольняє властивостям білого гаусового шуму з нульовим середнім значенням і коваріаційною матрицею $\mathbf{Q}$, тобто статистики шуму мають вигляд [5; 6]:

$$E\left[\mathbf{w}(k)\right] = 0, \ \ \forall k;$$

$$E\left[\mathbf{w}(k)\mathbf{w}^T(j)\right] = \mathbf{Q}(k)\delta_{kj},$$

де δ_{kj} — дельта-функція Кронекера, що визначається: $\delta_{kj} = \begin{cases} 0 \text{ для } k \neq j \\ 1 \text{ для } k = j \end{cases}$;

$\mathbf{Q}(k)$ — додатно визначена коваріаційна матриця зовнішніх збурень стану розмірності $(n \times n)$. Діагональні елементи матриці є дисперсією компонент вектора збурень $\mathbf{w}(k)$.

Початковим станом системи $\mathbf{x}_0$ будемо вважати випадкові змінні з відомими статистиками:

$$E\left[\mathbf{x}_0\right] = \bar{\mathbf{x}}_0; \ \ E\left[\mathbf{x}_0\mathbf{x}_0^T\right] = \mathbf{M}; \ \ E\left[\mathbf{w}(k)\mathbf{x}_0^T\right] = 0, \ \ \forall k.$$

Нехай вектор вимірів $\mathbf{z}(k)$ вихідних змінних доступний у будь-який момент часу t_k, а його компоненти лінійно зв'язані з вектором стану і на них впливає шум вимірів, тобто

$$\mathbf{z}(k) = \mathbf{H}(k)\,\mathbf{x}(k) + \mathbf{v}(k),$$

де $\mathbf{H}(k)$ — матриця спостережень вимірності $(r \times n)$, $\mathbf{v}(k)$ — r-вимірний вектор випадкових величин шуму вимірів з відомими статистиками:

$$E\left[\mathbf{v}(k)\right] = 0, \ \ E\left[\mathbf{v}(k)\mathbf{v}^T(j)\right] = \mathbf{R}(k)\delta_{kj},$$

де $\mathbf{R}(k)$ — додатно визначена коваріаційна матриця шумів вимірів вимірності $(r \times r)$, діагональні елементи якої є дисперсіями адитивного шуму в кожному каналі вимірів. Шум вимірів також задовольняє властивостям білого гаусового шуму. Він вважається некорельованим із зовнішнім збуренням $\mathbf{w}(k)$ і початковим станом системи, тобто

$$E\left[\mathbf{v}(k)\mathbf{w}^T(j)\right]=0 \quad \forall k, j;$$

$$E\left[\mathbf{v}(k)\mathbf{x}_0^T\right]=0 \quad \forall k.$$

Для визначеної вище системи з вектором стану $\mathbf{x}(k)$ необхідно знайти оцінку стану $\hat{\mathbf{x}}(k)$ в момент t_k як лінійну комбінацію оцінки в момент $\hat{\mathbf{x}}(k-1)$ і самого останнього виміру (статистичних даних) t_{k-1}.

Оцінка $\hat{\mathbf{x}}(k)$ повинна обчислюватися як найкраща за мінімумом середнього значення суми квадратів оцінок похибок. Інакше кажучи, оцінка повинна бути такою, щоб

$$E\left[\left(\mathbf{x}(k)-\mathbf{x}(k)\right)^T\left(\mathbf{x}(k)-\mathbf{x}(k)\right)\right]=\min_{\mathrm{K}},$$

де $\mathbf{x}(k)$ — точне значення вектора стану, яке може бути обчислене за допомогою детермінованої складової математичної моделі процесу; $\mathbf{K}$ — оптимальний матричний коефіцієнт фільтра, який необхідно обчислити в результаті розв'язання оптимізаційної задачі.

Таким чином, фільтр необхідно будувати та використовувати для уточнення оцінок стану процесу в умовах впливу випадкових зовнішніх збурень та наявності шумів (похибок) вимірів.

На сьогодні оптимальні фільтри — це невід'ємна складова комп'ютерних систем обробки експериментальних даних.

6. Кореляційний аналіз даних

Кореляційний аналіз даних необхідно виконувати для встановлення наявних зв'язків між значеннями однієї вибірки даних та між значеннями кількох вибірок.

Кореляція характеризує наявність (відсутність) лінійної або нелінійної залежності між змінними.

Коефіцієнт кореляції, а загалом *кореляційна функція*, дають можливість встановити зв'язок між змінними, наприклад, r_{yx} — коефіцієнт кореляції між змінними y та x. Кореляція може бути лінійною або нелінійною залежно від типу взаємозв'язку, який фактично існує між змінними. Досить часто на практиці розглядають тільки лінійну ко-

реляцію, але більш глибокий аналіз потребує використання для дослідження функціонування процесів нелінійних залежностей. Складну нелінійну залежність часто можна спростити (лінеаризувати), але знати про її існування необхідно для того, щоб побудувати адекватну модель процесу.

Вибірковий коефіцієнт кореляції між двома змінними обчислюється за формулою [10]:

$$r_{yx} = \frac{1}{N-1} \frac{\sum_{k=1}^{N} \left\{ \left[y(k) - \overline{y} \right] \left[x(k) - \overline{x} \right] \right\}}{\sigma_x \sigma_y},$$

або

$$r_{yx} = \frac{\sum_{k=1}^{N} \left\{ \left[y(k) - \overline{y} \right] \left[x(k) - \overline{x} \right] \right\}}{\sqrt{\sum_{k=1}^{N} \left[y(k) - \overline{y} \right]^2} \sqrt{\sum_{k=1}^{N} \left[x(k) - \overline{x} \right]^2}}.$$

де $-1 < r_{yx} < +1$; $\sigma_x \sigma_y$ — стандартні відхилення для змінних x і y, відповідно.

Якщо необхідно одночасно обчислити кореляцію між кількома змінними (наприклад, між ендогенною та кількома екзогенними), то формують кореляційну матрицю.

7. Попереднє визначення структури математичних моделей

Однією із складових прикладної статистики є регресійний аналіз даних, тобто побудова математичних моделей на основі статистичних даних і їх використання для прогнозування і керування процесами різної природи.

Важливим завданням процесу моделювання є коректний вибір структури моделі. Поняття структури математичної моделі включає в себе такі елементи:

- розмірність моделі (кількість рівнянь, що описують процес або об'єкт);
- порядок моделі (максимальний порядок різницевого або диференціального рівняння, яке входить у модель);
- нелінійність та її тип (може бути нелінійність стосовно змінних або нелінійність стосовно параметрів);
- час (лаг) запізнення (по входу) та його оцінка;
- збурення стану та його тип (детерміноване або випадкове);
- можливі обмеження на змінні і параметри моделі.

Оцінювання структури моделі процесу називають ще структурною ідентифікацією. Воно виконується на основі всієї наявної інформації про функціонування процесу, а також на основі результатів кореляційного аналізу наявних експериментальних (статистичних) даних.

1.4. Оцінювання параметрів статистичних і математичних моделей (третій етап статистичного аналізу)

На цьому етапі статистичного аналізу оцінюють параметри вибраних типів моделей, а також інших невідомих величин. Наприклад, об'єми продажу деякого продукту, реакцію населення на новий продукт, зміну продуктивності виробничого підприємства, рівень браку у виробничому процесі.

Відповідно до типів моделей (лінійні та нелінійні) вибирають методи оцінювання параметрів. Найпоширеніші з них — метод найменших квадратів (МНК) і метод максимальної правдоподібності (ММП).

Звичайний метод найменших квадратів для оцінювання лінійних та псевдолінійних моделей ґрунтується на використанні квадратичного критерію:

$$\min_{\hat{\theta}} J = \sum_{k=1}^{N} \left[y(k) - \hat{y}(k) \right]^2 = \sum_{k=1}^{N} e^2(k),$$

де $\hat{y}(k)$ — оцінка залежної змінної по побудованій моделі, наприклад, по АР(2): $\hat{y}(k) = \hat{a}_0 + \hat{a}_1 y(k-1) + \hat{a}_2 y(k-2)$. Формула МНК має вигляд:

$$\hat{\theta} = \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{y},$$

де $\hat{\theta}$ — вектор параметрів; $\mathbf{X}$ — матриця вимірів екзогенних змінних у правій частині рівняння; $\mathbf{y} = \mathbf{y}(k)$ — вектор вимірів ендогенної змінної. Наприклад, для регресійної моделі множинної регресії

$$y(k) = a_0 + a_1 x_1(k) + a_2 x_2(k) + \ldots + a_m x_m(k) + \varepsilon(k)$$

матриця вимірів має вигляд:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1(1) & \ldots & x_m(1) \\ 1 & x_1(2) & \ldots & x_m(2) \\ \vdots & & & \vdots \\ 1 & x_1(N) & \ldots & x_m(N) \end{bmatrix}.$$

Статистичні дані регресорів (стовпчики матриці **X**) не повинні бути висококорельованими, оскільки їх висока корельованість веде до виродженості і неможливості знайти обернену матрицю. Для зменшення корельованості регресорів застосовують спеціальні методи ортогоналізації, наприклад, метод головних компонент (МГК).

Метод максимальної правдоподібності (ММП) ґрунтується на максимізації функції правдоподібності [10]:

$$L = \left(2\pi\sigma^2\right)^{-N/2} \prod_{k=1}^{N} \exp\left(-\frac{\left[y(k) - \mathbf{X}(k)\beta\right]^2}{2\sigma^2} \right);$$

або логарифмованої функції правдоподібності:

$$\ell = -\frac{N}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\prod_{k=1}^{N}\left[y(k) - \mathbf{X}(k)\beta\right]^2 =$$

$$= -\frac{N}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{k=1}^{N}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) =$$

$$= -\frac{N}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\mathbf{e}^T\mathbf{e},$$

де $\mathbf{e} = \mathbf{y} - \mathbf{X}\beta$; σ^2 — вибіркова дисперсія; β — параметр моделі.

Існують також рекурсивні версії МНК і ММП.

Формулювання гіпотез стосовно оцінок параметрів

Статистичні оцінки — це тільки *припущення щодо можливих значень*, а тому вони часто бувають неточними. Однак, якщо вони достатньо близькі до істини, то служать поставленій меті. Якщо відома їх точність, то можна вирішити, в якій мірі їх варто приймати до уваги.

Верхнє і нижнє значення оцінки невідомої величини дає *довірчий інтервал*, який дає нам впевненість, що оцінка лежить у визначеному діапазоні значень.

Для отриманих статистичних оцінок параметрів моделі формулюють гіпотези стосовно їх статистичної значущості. Наприклад, для оцінок регресійної моделі формулюють таку гіпотезу:

$H_0 : \hat{a}_i = 0$ — оцінка статистично незначуща;

$H_0 : \hat{a}_i \neq 0$ — оцінка статистично значуща.

Кожна гіпотеза формулюється як твердження, яке може бути вірним або невірним. У результаті перевірки гіпотези (на основі об-

робки даних) ми приймаємо або відкидаємо попередньо висунуту гіпотезу.

Існує стандартна процедура перевірки гіпотез, яку використовують у техніці, психології, освіті, суспільних науках та багатьох інших областях. Ця процедура буде розглянута у даному розділі.

Необхідно підкреслити той важливий факт, що *результат експерименту, який підтверджує справедливість висунутої гіпотези, майже ніколи не може бути основою для прийняття цієї гіпотези*. Водночас *результат, несумісний з висунутою гіпотезою, є цілком достатнім для її відхилення як неправильної*. Очевидно, що наведене твердження потребує обґрунтування.

Приклад 1

Припустимо, що середній коефіцієнт розумового розвитку (КРР) деякої генеральної сукупності людей $\mu = 100$ (це гіпотетичне значення). Результат взятої випадкової вибірки дає результат $\mu_i = 102$, який є сумісним з висунутою гіпотезою. Однак цей результат є також сумісним з припущенням, що $\mu = 101$ або $\mu = 99$ і, звичайно, є сумісним з гіпотезою, що $\mu = 102$. Таким чином, на основі даного результату не можна віддати перевагу гіпотезі, що $\mu = 100$.

Припустимо тепер, що інша випадкова вибірка дала середнє $\mu = 135$. *Якщо об'єм вибірки був достатньо великим, то можна показати таке: якщо початкова гіпотеза правильна, то ми практично ніколи не отримали б подібного результату*. На основі цього висновку отриманий результат цілком обґрунтовано можна використати як підтвердження неправильності висунутої гіпотези і ризик помилки при цьому буде мінімальним.

Все сказане вище ґрунтується на тому факті, що результат експерименту, який є сумісним з висунутою гіпотезою, виявляється також сумісним з іншими гіпотезами. Це веде до того, що подібний результат не може бути використано для обґрунтування вибору деякої гіпотези порівняно з іншими. Однак ми завжди можемо отримати результат, який розбігається з висунутою гіпотезою і може спричинити значні сумніви щодо її правильності або достовірності.

При перевірці гіпотез остаточний висновок можна зробити тільки в тому випадку, якщо ми можемо відхилити висунуту гіпотезу. Таким чином, мета експерименту повинна полягати в тому, щоб відхилити сформульовану гіпотезу.

Мета експерименту повинна полягати у тому, щоб відхилити сформульовану гіпотезу

А це означає, що початкова гіпотеза повинна формулюватись як альтернатива тому, у що ми віримо і що ми хочемо отримати.

Якщо ми зможемо відхилити висунуту гіпотезу (довести її неправильність), то тим самим продемонструємо справедливість того твердження, в яке дійсно віримо.

1.5. Перевірка сформульованих гіпотез (четвертий етап статистичного аналізу)

На даному етапі ми перевіряємо істинність гіпотез, висунутих на третьому етапі. Це можуть бути гіпотези стосовно статистичної значущості отриманих оцінок невідомих величин, тобто оцінок параметрів статистичних і математичних моделей.

Іншими прикладами гіпотез, які можна було б перевірити за допомогою статистичних даних, можуть бути такі:

1. Ви переможете на виборах, які відбудуться через 10 днів.
2. Похибка оцінки параметра є меншою деякої величини.
3. Рівень виробничого браку є меншим, ніж його очікують споживачі.

При перевірці гіпотез велике значення має мінімізація можливості появи ситуацій, коли випадкова збіжність ряду факторів може призвести до відхилення правильної гіпотези. Тому перед тим як відхилити гіпотезу експериментатор вимагає, щоб ймовірність отримання відповідного вибіркового значення була дуже малою.

У деяких областях науки прийнято відхиляти гіпотезу тільки у тих випадках, коли випадкове вибіркове значення може зустрічатись не частіше ніж 5 разів на 100 експериментів. В інших областях гіпотези відхиляються, якщо ймовірність появи відповідного вибіркового значення не перевищує 0,01. Очевидно, що експериментатор повинен прямувати до того, щоб ймовірність появи вибіркового значення, яке вказує на неправильність висунутої гіпотези, була досить малою.

Ймовірність появи вибіркового значення, яке вказує на неправильність сформульованої гіпотези, вибирається експериментатором і називається *рівнем значущості* експерименту

Вибір рівня значущості *повинен відбуватись до збору експериментальних даних*, оскільки результати експерименту не повинні впливати на величину вибраного критерію. Після визначення рівня значущості і обробки даних дослідник припускає, що його гіпотеза є правильною і визначає — більшою чи меншою вибраного рівня значущості буде ймовірність отриманого результату.

Якщо ймовірність отриманого результату перевищить рівень значущості, то експериментатор не зможе відхилити висунуту гіпотезу, справедливо вважаючи, що даний результат є в достатній мірі сумісним з нею. Наприклад, припустимо, що вибрано рівень значущості 0,05, а в результаті експерименту отримано 52 герби при 100 підкиданнях монети. Якщо експериментатор встановить, що відхилення в два герби від очікуваного значення зустрічається більше ніж у 5 % випадків, то він не зможе відхилити висунуту гіпотезу.

Якщо розраховане значення ймовірності деякого результату виявилось меншим рівня значущості, то висунуту гіпотезу можна відхилити. Наприклад, припустимо, що при використанні рівня значущості 0,05, підкидання монети привело до 96 гербів, а розрахунок показав, що відхилення цього значення від очікуваного має ймовірність появи менше 0,05.

Логіка, якою керується експериментатор, полягає в такому: "*Після формулювання даної гіпотези я отримав такий неймовірний результат, що не можу в нього повірити*". Після цього експериментатор може вважати своє початкове припущення (теорію) доведеним і, таким чином, може сподіватись, що завжди буде отримувати результати, ймовірність появи яких при даних обставинах буде меншою рівня значущості.

Вибір рівня значущості означає також, що введено визначене правило прийняття і неприйняття гіпотез. *Рівень значущості показує ймовірність відхилення деякого показника від його очікуваного значення, при якому дослідник може відхилити висунуту гіпотезу*.

Вибраний *рівень значущості вказує на точне значення ймовірності помилки першого роду* у випадку, якщо гіпотеза дійсно правильна. Наприклад, якщо рівень значущості дорівнює 0,01, то це означає, що у випадку правильності гіпотези в одному випадку із 100 буде робитись помилка першого роду на основі отриманого результату.

Інакше кажучи, *рівень значущості означає величину ризику зробити помилку першого роду*. Чим менший рівень значущості, тим меншою є ймовірність припущення помилки першого роду (відхилити

правильну гіпотезу). Однак, чим менший рівень значущості, тим більшою є ймовірність помилки другого роду, якщо гіпотеза виявиться помилковою. Таким чином, *вибір рівня значущості означає вибір правила прийняття рішення при перевірці гіпотези*.

Сформулюємо процедуру перевірки гіпотез у вигляді наведених нижче кроків.

1. Формулювання нульової гіпотези, яку необхідно перевірити. Відхилення сформульованої гіпотези дає можливість вважати початкові припущення (теорію) правильними. (У прикладі з монетою метою може бути встановлення факту неповноцінності монети. Таким чином, необхідно перевірити гіпотезу про те, що монета є повноцінною.)
2. Вибір рівня значущості. (Наприклад, рівень значущості 5 % означає, що ймовірність припуститись помилки першого роду складає 0,05.)
3. Виконати експеримент і обчислити необхідний статистичний параметр.
4. Припускаючи, що сформульована гіпотеза є вірною, необхідно визначити ймовірність відхилення знайденого значення статистичного параметра від його очікуваного значення.
5. Якщо за припущення істинності гіпотези розрахунки показують, що ймовірність відхилення отриманого вибіркового статистичного показника від очікуваного значення перевищує рівень значущості, то *відхилити висунуту гіпотезу неможливо*. Якщо за припущення стосовно істинності гіпотези розрахунки показують, що ймовірність відхилення отриманого вибіркового статистичного показника від очікуваного значення є меншою рівня значущості, то *висунута гіпотеза відхиляється*.

Приклад 2 (задача другого типу)

Чи буде середній коефіцієнт розумового розвитку (КРР) 64-х *десятилітніх хлопчиків, які збираються стати інженерами*, відрізнятись від середнього коефіцієнта генеральної сукупності. Нехай дослідження виконується в місті, в якому середній КРР десятилітніх хлопчиків $\mu = 100$, а стандартне відхилення $\sigma = 20$.

Необхідно показати, що десятилітні хлопчики, які збираються стати інженерами, відрізняються за рівнем свого інтелектуального розвитку від середнього коефіцієнта генеральної сукупності. За нульгіпотезу прийнято:

H_0 : вибрана група — типовий представник генеральної сукупності;
H_1 : вибрана група відрізняється від генеральної сукупності.

Виберемо за рівень значущості 0,05 і будемо сподіватись, що зможемо відхилити висунуту гіпотезу. Нехай із генеральної сукупності вибрано 64 хлопчика, які збираються стати інженерами. Встановлено, що середній КРР для них складає: $\mu_{гр} = 108$.

Ми вважаємо, що нуль-гіпотеза є вірною, тобто, їхні 64 КРР можна розглядати як випадкову вибірку із великої сукупності спостережень, і є однією з можливих вибірок обсягом 64 кожна. З теорем для середнього ряду розподілу, утвореного із середніх значень (будуть розглянуті нижче) випливає, що при таких умовах розподіл середніх значень цих вибірок буде мати середнє $\mu = 100$, а стандартне відхилення

$$\sigma_{гр} = \frac{\sigma}{\sqrt{N}} = \frac{20}{\sqrt{64}} = 2{,}5.$$

Цей розподіл представлено на рис. 1.1, на якому середнє $\mu_{гр} = 108$ позначено хрестиком.

З рис. 1.1 видно, що за припущення істинності нульової гіпотези ми отримали вибіркове середнє, величина якого перевищує точку рівноваги нормального розподілу на 8 пунктів, а стандартне відхилення випадкової вибірки складає 2,5. Інакше кажучи, припускаючи істинність нульової гіпотези, ми повинні зробити висновок, що отримане вибіркове середнє на 3,2 (8/2,5) стандартного відхилення перевищує очікуване значення середнього, тобто 100.

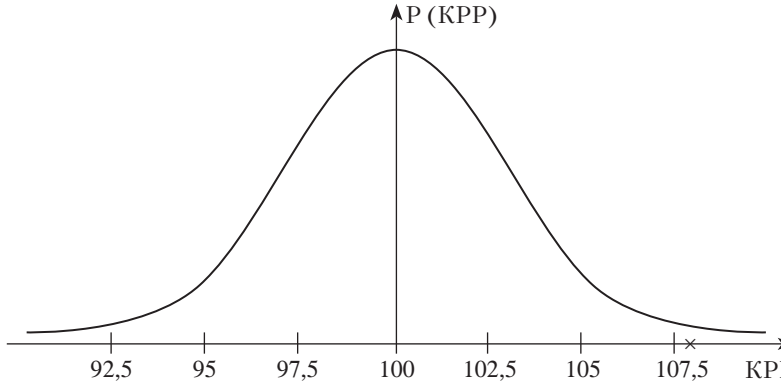


Рис. 1.1. **Форма розподілу для КРР при** $\mu_{гр} = 108$

З таблиці для площ нормального розподілу (див. Додаток) знайдемо, що ймовірність отримання значення, *z*-оцінка якого $\left(z = \frac{x_i - \bar{x}}{\sigma_x} \right)$ відрізняється більше ніж на 3 одиниці (в обидва боки) від точки рівноваги, складає менш як 0,001. Таким чином, припустивши істинність нульової гіпотези, ми отримали вибіркове значення, яке настільки сильно відрізняється від очікуваної величини, що ймовірність його появи є менш як 0,001. Отже, цей результат свідчить про те, що ми повинні відхилити нульову гіпотезу і прийняти альтернативну. Тобто хлопчики, які у майбутньому збираються працювати інженерами, в середньому відрізняються за рівнем свого інтелектуального розвитку від інших хлопчиків їхнього віку.

У даному випадку було б корисно застосувати формулу обчислення *z*-оцінки вибіркового середнього розподілу вибіркових середніх. Ці оцінки дають можливість швидше розв'язувати подібні задачі. Так, *z*-оцінка вибіркового середнього обчислюється за формулою:

$$z = \frac{\text{Оцінка відхилення середнього}}{\text{Стандартне відхилення середніх}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{N}}, \qquad (1.5.1)$$

де $\bar{x}$ — значення вибіркового середнього (у нашому прикладі $\bar{x}$ = 108); μ — середнє розподілу вибіркових середніх (ми припускали, що μ = 100); σ — стандартне відхилення КРР для генеральної сукупності (у нашому прикладі σ = 20).

Підставляючи конкретні значення у формулу (1.5.1), отримаємо:

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{N}} = \frac{108 - 100}{20 / \sqrt{64}} = 3{,}2.$$

Звертаючись тепер до таблиці площ нормального розподілу (при *z* = 3,2), відхиляємо нуль-гіпотезу на рівні значущості 0,05.

Приклад 3 (задача третього типу)

Нехай *необхідно перевірити гіпотезу стосовно того, що студенти п'ятого курсу НТУ КПІ є типовими представниками всіх студентів п'ятого курсу за вмінням виконувати курсові проекти* з проектування комп'ютерних інформаційних систем.

Перевіримо нульову гіпотезу для даного випадку на рівні значущості 0,01 (1 %). За основу взято результати виконання курсових проектів у всіх університетах Києва. Середня оцінка за курсовий про-

ект по місту виявилась рівною $\mu = 72$ (при максимальному значенні 100); стандартне відхилення $\sigma = 12$.

У групі, яку аналізували, налічувалось $N = 36$ студентів при середньому значення оцінки $\mu_{гр} = 74$. Тобто, необхідно визначити, чи суттєво відрізняється середнє для вибраної групи від середнього генеральної сукупності.

Припустимо, що сформульована гіпотеза вірна (студенти цієї групи є типовими представниками студентів 5-го курсу по Києву) і що отримані 36 оцінок за курсові проекти можна розглядати як випадкову вибірку із всієї сукупності оцінок. Це означає, що ми можемо розглядати вибіркове середнє для групи студентів, $\mu_{гр} = 74$, як одне із значень теоретичного розподілу середніх різних вибірок об'ємом $N = 36$ кожна. Відповідно до наведеної нижче теореми розподіл таких вибіркових середніх є нормальним. При цьому середнє розподілу середніх дорівнює 72 (таке саме значення має середнє генеральної сукупності), а стандартне відхилення:

$$\sigma_{рс} = \frac{\sigma}{\sqrt{N}} = \frac{12}{\sqrt{36}} = 2.$$

Таким чином, отримане вибіркове середнє для вибраної групи студентів, $\mu_{гр} = 74$, на $z = 1$ перевищує теоретично очікуване значення, тобто:

$$z_{гр} = \frac{\bar{x} - \mu}{\sigma/\sqrt{N}} = \frac{74 - 72}{12/\sqrt{36}} = 1{,}0.$$

Відповідно до таблиці площ нормального розподілу (див. Додаток), *ймовірність розбіжності між середнім генеральної сукупності і середнім вибраної групи студентів* є більшою вибраного нами рівня значущості, а тому *нуль-гіпотеза приймається*. Тобто середнє генеральної сукупності і середнє вибраної групи відрізняються несуттєво.

Отже, на основі виконаного експерименту *не можна зробити висновок, що студенти п'ятого курсу НТУ КПІ відрізняються за рівнем виконання курсових проектів* з проектування інформаційних систем від студентів інших університетів.

1.6. Суцільний та вибірковий методи збору даних

Будемо називати набір значень деякої величини *рядом розподілу* або розподілом значень цієї величини. Наприклад, множина оцінок, отриманих на екзамені або ряд числових значень, що характеризують

зміну в часі рейтингу кандидата у президенти. В останньому випадку кажуть, що існує часовий ряд даних, що характеризує вибрану змінну на деякому (вибраному) часовому інтервалі. При цьому дані можуть накопичуватись через однакові (постійні) або різні (змінні) часові інтервали, які називають *періодами дискретизації* даних (вимірів). У більшості випадків статистичні (експериментальні) дані накопичують через однаковий часовий інтервал, що спрощує їх подальшу обробку. Якщо ж період дискретизації змінюється у часі, то такі дані потребують застосування спеціальних методів обробки і спеціальних моделей для опису.

Існує два основних підходи до збору експериментальних або статистичних даних:

- метод суцільних спостережень;
- вибірковий метод.

Метод *суцільних спостережень* полягає у визначенні максимально можливого об'єму спостережень та формуванні бази вимірів із включенням в неї усіх елементів даної сукупності. Таку базу даних називають *генеральною сукупністю*. Наприклад, для визначення успішності студентів з прикладної статистики в КПІ формується база даних, яка вміщує оцінки всіх студентів усіх факультетів і груп, які вивчають прикладну статистику.

Теорія вибіркового методу передбачає формування статистичного висновку щодо всієї можливої сукупності даних на основі аналізу деякої обмеженої вибірки, взятої з генеральної сукупності.

Наприклад, статистичний висновок стосовно успішності студентів з прикладної статистики (ПС) в університеті формується на основі вибірки оцінок студентів, взятих з тих факультетів, де вивчають ПС. Можна взяти випадкову вибірку, яка включає по 10 студентів з кожної групи, і знайти середню оцінку для сформованої таким чином вибірки.

Вибірковий метод є основним методом статистичного аналізу, оскільки, як правило, формування генеральної сукупності є фізично неможливим або занадто дорогим. Крім того, результати дослідження вибірок із генеральної сукупності можна коректно узагальнити на генеральну сукупність за допомогою відповідних обчислювальних процедур.

Випадковий і типовий відбори даних

Для того щоб правильно виконувати статистичні дослідження, необхідно розуміти можливості статистичних методів. Якщо необхідно

зробити узагальнюючі висновки на основі деякої вибірки даних, остання повинна бути *представницькою* (репрезентативною), тобто *вона повинна представляти*:

- *усі соціальні групи населення*, якщо, наприклад, виконується визначення рейтингу політичної фігури;
- *представників студентів усіх курсів і всіх спеціальностей*, якщо аналізується успішність студентів університету загалом;
- *повні дані щодо визначеного режиму роботи об'єкта*, який досліджується (наприклад, перехідного режиму).

Синонімом слова представницька є *репрезентативна* (від англ. *representative*). Наприклад, вибірка студентів не буде представницькою, якщо ми включимо тільки половину факультетів, де вивчають ПС. Може виявитись, що друга половина, яка не включена у вибірку, суттєво відрізняється від першої. Якщо виконується аналіз рейтингу кандидатів у президенти, то у вибірку необхідно включити всі наявні соціальні групи населення. Інакше опитування може призвести до так званих зміщених оцінок, тобто оцінок, далеких від реальності.

У випадках, коли необхідно виконати статистичний аналіз роботи *технічного об'єкта*, дані повинні повністю охоплювати спостереженнями ті режими його функціонування, що цікавлять дослідника. Наприклад, якщо це *перехідний режим*, то необхідно так спланувати експеримент, щоб перехідний процес був повністю відображений зібраними числовими даними.

Існує два підходи до формування вибірок даних з генеральної сукупності:

- *власне випадкова вибірка*, коли кожна одиниця сукупності має однакові шанси бути відібраною для аналізу;
- *типовий відбір*.

Випадкову вибірку студентів можна сформувати наступним чином: згенерувати випадкові числа у діапазоні 1–30 (за числом студентів у групі) і взяти перших кілька чисел або написати на однакових аркушах прізвища всіх студентів факультету, які вивчають деякий предмет, і наугад витягати ці аркуші. У такому випадку кожна одиниця сукупності має однаковий шанс щодо включення у вибірку.

Різновидом такої техніки відбору є *типовий відбір*. Техніка цього методу гарантує, що вибірка буде містити такі самі пропорції елементів різних груп, як і в генеральній сукупності, з якої вони взяті. Наприклад, нехай нас цікавить думка студентів університету щодо загального рівня навчання (підготовки) у даному університеті.

Типовий відбір передбачає, що у сформованій вибірці будуть представлені частини (пропорції) студентів з кожного курсу і факультету і що *ці пропорції будуть такими самими, як і в генеральній сукупності*. Тобто, якщо на першому курсі навчається 20 % студентів університету, то першокурсники повинні скласти 20 % вибірки даних. Якщо рівень компетентності студентів різний на різних курсах, то це також необхідно врахувати у вибірці.

1.7. Достовірність статистичних досліджень

Статистика — *надзвичайно могутній інструмент аналізу даних різної природи і її силу можна використати коректно або некоректно*. Є багато прикладів некоректного застосування статистичних методів, що проявилось у формулюванні висновків, протилежних до дійсності. Тому коректність застосування методів статистичного аналізу даних — це ключовий момент, який забезпечує правильність остаточного результату.

Наприклад, у 1936 р. в США журнал "*Literary Digest*" помилково визначив вибірку даних для аналізу. Журнал виконав телефонне *опитування своїх передплатників* і спрогнозував, що *Альфред Лендон* легко переможе свого суперника *Франкліна Рузвельта*. Однак, Рузвельт переміг Лендона у 46 із 48 штатів. Справа в тому, що 1936 р. — це один із років економічної депресії і дозволити собі телефон і передплату журналу могли тільки люди з хорошим матеріальним станом. Тобто вибірка, визначена журналом, не представляла всіх прошарків виборців у США — вона *не була представницькою*. Передплатники журналу у своїй більшості збирались голосувати за Лендона, у той час як основна частина генеральної сукупності була за Рузвельта. Після виборів журнал "*Literary Digest*" швидко втратив свою популярність і досить швидко перестав існувати.

Звичайно, можна навести багато інших прикладів некоректного застосування статистичних методів і не тільки у минулому, а й сьогодні. На жаль, бувають випадки, коли некоректне застосування цих методів зумовлене не тільки поганими знаннями, а й цілеспрямованим навмисним формуванням некоректних висновків на основі зібраних даних.

Очевидно, що майже всі статистичні процедури можна виконати некоректно як у результаті незнання, так і навмисно. Однак, як свідчить практика, навмисне спотворення методів формування статис-

тичного висновку веде тільки до негативних результатів. Якщо ми будемо стверджувати, що зростання ВВП України становить 12 % на рік, а фактичне зростання — 4–5 %, то добробут населення і довіра до влади від цього не підвищаться.

Зазначимо, що невірні статистичні дані і некоректно виконаний статистичний аналіз можуть привести до прийняття принципово неправильних тактичних і стратегічних рішень з катастрофічними наслідками

Випадки неправильного використання звичайних статистичних методів можна знайти в американській книзі "*Як обманювати за допомогою статистики*".

Іноді кажуть, що за допомогою статистики можна довести, що завгодно. Однак, якраз це "що завгодно" пов'язане з випадками некоректного застосування статистичних методів. Деякі компанії виконують дослідження, за допомогою яких доводять, що їх продукт переважає за своїми якостями аналогічні продукти, що випускаються конкуруючими фірмами. Очевидно, що висновки будь-якого дослідження можна підігнати під бажаний результат за рахунок зміни техніки збору даних або неправильної інтерпретації встановлених залежностей. У повсякденному житті кожний з нас зустрічається із статистичними даними та висновками, що знаходяться в протиріччі між собою. Тому немає нічого дивного у тому, що багато людей висловлюють недовіру до статистики. Зазначимо, що те ж саме може відноситись до будь-якої науки.

Однак, у випадках використання статистикою недостовірних даних *винні не цифри і не використані методи, а ті виконавці*, які мають схильність обманювати суспільство за їх допомогою. Чесні наміри і знання статистичних методів завжди дають корисні змістовні результати.

Іноді можна зустріти заперечення проти застосування статистичних методів у зв'язку з тим, що отримувані за ними висновки носять узагальнений характер і не відображають індивідуальних якостей людей. Так, висновок стосовно того, що індивідууми з вищою освітою матимуть деякий рівень середньої зарплати, не сприймається деякими людьми. Вони вважають, що їх це не стосується і що висновок неправильний. Так само як і у випадку голосування за деякого визначеного кандидата у президенти не всі представники однієї соціальної групи голосуватимуть за одного кандидата.

З цього приводу можна сказати, що статистичні висновки є узагальнюючими, але вони *не абсолютні*. Загальне судження ніколи не залишається незмінним. Якщо ми будемо наводити приклади, що будуть у протиріччі з висновком, то тим самим будемо придавати висновку *форму абсолютного судження*. Розв'язок цієї проблеми полягає у тому, що кожний індивідуум повинен сприймати справедливість статистичного судження, але в той самий час повинен розуміти, що ніхто не забирає у нього свободи вибору за допомогою цього судження.

Загалом можна сказати, *що основною метою застосування статистичного аналізу є обґрунтоване прийняття рішень на основі аналізу даних*. Якщо ми хочемо, щоб рішення були об'єктивними і якісними, то необхідно знати і вміти користуватись методами статистичного аналізу та іншими методами прийняття рішень, що доповнюють його. Доповнюючими методами є *методи оптимізації, експертне оцінювання альтернатив, нечіткі множини, нейромережі, байєсівський підхід* та ін. Зазначимо, що сукупне застосування кількох методів аналізу даних і прийняття рішень, а також відповідних критеріїв аналізу їх якості значно підвищує ймовірність отримання кращих рішень з множини можливих альтернатив.

1.8. Контрольні питання і вправи

1. Які задачі можна розв'язувати методами математичної статистики?
2. Назвіть чотири етапи статистичного аналізу даних.
3. Що означає генеральна сукупність даних?
4. Яку вибірку даних називають випадковою?
5. Які моделі називають математичними, а які статистичними?
6. Назвіть два основних види статистичних даних.
7. Які існують причини появи випадкової складової у вимірах (статистичних даних)?
8. Яка складова вимірів дає можливість будувати математичні моделі досліджуваних процесів?
9. Які задачі розв'язують на етапі планування збору даних?
10. Яка мета попередньої обробки і дослідження даних?
11. Дайте означення тренду. Що означають терміни "інтегрований" і "процес з одиничними коренями"?
12. Для чого виконується візуальний аналіз даних?

13. Які процеси називають гетероскедастичними?
14. Поясніть такі статистики: критерій узгодженості χ^2, асиметрію, ексцес.
15. Як розраховується статистика Жак-Бера? Що означають знайдені значення?
16. Дайте визначення прогнозу. Запишіть функцію прогнозування для заповнення пропусків даних з використанням рівняння авторегресії з ковзним середнім АРКС(1, 1):

$$y(k) = a_0 + a_1 y(k - 1) + b_1 \varepsilon(k - 1) + \varepsilon(k).$$

17. Які значення відносять до екстремальних? Як вони впливають на статистичні характеристики вибірки даних?
18. Схарактеризуйте особливості перехідного та стаціонарного режимів функціонування досліджуваних об'єктів.
19. Які методи заповнення пропусків даних ви знаєте?
20. Що таке нормування даних і яким чином воно виконується? Чи можна виконувати нормування даних з використанням нелінійних перетворень?
21. Яка мета кореляційного аналізу даних? Яким чином розраховується вибіркова дискретна функція взаємної кореляції двох змінних?
22. Що таке автокореляція? Яка причина існування ненульової автокореляції? Як розраховується автокореляційна функція?
23. Як розраховується часткова автокореляційна функція? У чому полягає відмінність часткової АКФ від звичайної?
24. Які елементи включає у себе поняття структури математичної моделі? Яким чином можна оцінити порядок моделі авторегресії за умови наявності статистичних даних у вигляді часового ряду?
25. Які методи оцінювання параметрів математичних і статистичних моделей широко вживаються на практиці?
26. Який критерій оптимальності використовують при оцінюванні параметрів за методом найменших квадратів (МНК)?
27. Наведіть вираз для оцінювання параметрів за МНК, поясніть всі змінні, які в нього входять. Як формується матриця вимірів при оцінюванні моделі множинної регресії?
28. Оцінки параметрів (коефіцієнтів) моделі, які отримують за допомогою МНК, — це випадкові чи детерміновані величини? Дайте ґрунтовне пояснення.

29. Який критерій оптимальності використовують при застосуванні методу максимальної правдоподібності? Наведіть приклад функції правдоподібності.
30. Як потрібно формулювати нуль-гіпотезу при застосуванні теорії перевірки гіпотез до аналізу отриманих результатів статистичного дослідження?
31. Опишіть і поясніть послідовність дій при перевірці гіпотез у загальному випадку.
32. Опишіть і поясніть послідовність дій при перевірці гіпотези стосовно значущості оцінок параметрів регресійної моделі у статистичному значенні.
33. Поясніть, який смисл має рівень значущості при перевірці гіпотез.
34. Що означають помилки першого і другого роду при перевірці гіпотез? Поясніть на прикладі.
35. Що означає термін "стандартна похибка" у регресійному аналізі даних? Як вона знаходиться? Чи пов'язана стандартна похибка з точністю оцінювання параметрів регресійних моделей?
36. Що означає "рівень значущості експерименту"? Виходячи з яких міркувань його вибирають?
37. Як пов'язаний рівень значущості експерименту з помилками першого і другого роду?

Розділ 2

МОДА, МЕДІАНА, СЕРЕДНЄ І ВАРІАЦІЯ

2.1. Означення

Нехай досліджується певна сукупність об'єктів. У результаті проведення спостереження отримуємо певні значення ознаки у кожного з цих об'єктів. Ці значення ознаки можуть бути дискретними або належати певним інтервалам.

Досліджувані об'єкти групуються за значеннями ознаки в упорядковані ряди розподілу. Якщо значення ознаки дискретні, то маємо дискретний варіаційний ряд розподілу. Якщо ж значеннями ознаки є певні інтервали, то одержуємо інтервальний варіаційний ряд розподілу.

Кожний варіаційний ряд розподілу має дві характеристики – значення ознаки і частоту.

Середину ряду розподілу або центр ряду розподілу характеризують мода, медіана і середні значення.

Означення 2.1. *Мода* — це значення ознаки, яке найчастіше зустрічається в ряді розподілу, тобто має найбільшу частоту.

Означення 2.2. Якщо в ряді розподілу два значення ознаки мають однакову найбільшу частоту, то ряд розподілу називається *бімодальним*.

У випадку дискретного ряду розподілу легко визначити моду як значення ознаки з найбільшою частотою.

У випадку інтервального ряду розподілу легко визначається модальний інтервал значень ознаки як інтервал, який відповідає найбільшій частоті. Значення моди M_0 визначається за формулою [16]:

$$M_0 = x_0 + h\frac{f_{m_0} - f_{m_0-1}}{\left(f_{m_0} - f_{m_0-1}\right) + \left(f_{m_0} - f_{m_0+1}\right)},$$

де x_0 — лівий кінець модального інтервалу, f_{m_0} — частота модального ряду, f_{m_0-1} — частота інтервалу, що стоїть перед модальним, а f_{m_0+1} — частота інтервалу, що стоїть після модального. Оскільки при побудові інтервального ряду розподілу довжина інтервалів береться однаковою, то h — довжина модального інтервалу.

Означення 2.3. *Медіана* — це значення ознаки, яке припадає на середину впорядкованого ряду розподілу і поділяє його навпіл — на дві рівні за обсягом частини.

Для знаходження медіани використовують накопичені або кумулятивні частоти, які отримують для кожного інтервалу, починаючи з другого, шляхом додавання до частоти інтервалу накопичену частоту попереднього інтервалу в порядку зростання значень ознаки.

У дискретному ряді розподілу медіаною буде значення ознаки, накопичена частота якого буде першою, що перевищує половину обсягу сукупності.

В інтервальному ряді розподілу спочатку визначається медіанний інтервал як інтервал, накопичена частота якого дорівнює половині обсягу сукупності або буде першою, що перевищує половину обсягу сукупності.

Значення медіани M_e визначається за формулою:

$$M_t = x_e + h \cdot \frac{0{,}5\sum_{i=1}^{n} f_i - S_{m_e-1}}{f_{m_e}},$$

де x_e — лівий кінець медіанного інтервалу, h — довжина медіанного інтервалу, f_{m_e} — частота медіанного інтервалу, S_{m_0-1} — накопичена частота інтервалу, що стоїть перед медіанним.

Означення 2.4. *Середнім арифметичним* називається значення ознаки, яке отримують як відношення суми всіх значень ознаки до суми всіх частот.

Якщо всі значення ознаки мають частоту рівну одиниці, то отримують просте середнє арифметичне за формулою:

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

де x_i (i_n = 1, 2, ..., n) — i-те значення ознаки, а n — обсяг всієї сукупності.

У випадку, коли значення ознаки x_i зустрічається n_i раз, то знаходять середнє арифметичне зважене за формулою:

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i},$$

причому $\sum_{i=1}^{k} n_i = n.$

2.2. Дві властивості середнього

1. Середнє значення будь-якої вибірки даних — це її *єдина точка рівноваги*.

Якщо, наприклад, числові значення вибірки даних представляють собою вагу деякого набору гирьок, то середнє буде точкою рівноваги. У деяких випадках в економіці середнє вважають середньостроковим або довгостроковим прогнозом (умова довгострокової рівноваги).

У випадку, коли розглядають *теоретичне середнє* деякого випадкового процесу, то його називають *математичним сподіванням*. Тобто для економічного процесу, що описується змінною $y(k)$, прогнозом буде його *математичне сподівання*:

$$E[y(k)] = \mu_y,$$

де E — оператор математичного сподівання або визначення середнього; k — дискретний час; μ_y — математичне сподівання змінної $y(k)$, яке можна обчислити різними способами. Одним із способів є, наприклад, математичний опис ряду даних рівнянням авторегресії другого порядку:

$$y(k) = a_0 + a_1 y(k - 1) + a_2 y(k - 2) + \varepsilon(k)$$

і знаходження розв'язку цього рівняння, який дасть можливість обчислити математичне сподівання та оцінки прогнозів.

2. *Сума відхилень від середнього дорівнює нулю* (табл. 2.1). Це означає, що ніяке інше число не може бути точкою рівноваги ряду розподілу.

Таблиця 2.1

Знаходження відхилень від середнього

Елемент ряду розподілу	Середнє	Різниця
4	8	–4
5	8	–3
5	8	–3
7	8	–1
8	8	0
10	8	2
17	8	9
	Сума різниць	0

Теорема 2.1. *Для будь-якої множини даних суми різниць між їх значеннями та середнім дорівнює нулю. Навпаки, якщо сума різниць між цими елементами і деяким числом дорівнює нулю, то це число є середнім.*

Доведення:

$$\sum_{k=1}^{N}\left[x(k)-\mu_X\right]=\sum_{k=1}^{N}x(k)-\sum_{k=1}^{N}\mu_X=\sum_{k=1}^{N}x(k)-N\frac{1}{N}\sum_{k=1}^{N}x(k)=0,$$

оскільки $\sum_{k=1}^{N}\mu_X=N\mu_X$.

Можна легко встановити, що сума квадратів відхилень від середнього також буде мати мінімальне значення у порівнянні з іншими можливими відхиленнями. Це можна сформулювати такою теоремою.

Теорема 2.2. *Для будь-якої множини значень ознаки сума квадратів їх відхилень від середнього приймає найменше можливе значення у порівнянні із сумою квадратів відхилень від інших точок. І навпаки, точка, в якій мінімізується сума квадратів відхилень, є середнім значенням.*

Теорему 2.2 не можна застосувати для знаходження середнього. Однак, наведений результат дуже важливий для теорії оцінювання. Його називають властивістю "*найменших квадратів*". Мінімізація суми квадратів відхилень від середнього дає можливість отримати конкретні вирази для обчислення параметрів випадкових процесів. Насамперед, це відомий метод найменших квадратів (МНК), який широко застосовується для оцінювання параметрів (коефіцієнтів) лінійних та псевдолінійних моделей (до псевдолінійних відносять моделі, нелінійні стосовно змінних, наприклад, моделі поліноміального типу).

2.3. Вплив зміни значень ряду розподілу на середнє

Теорема 2.3. *Якщо кожне значення ознаки ряду розподілу збільшити або зменшити на деяку константу, то і середнє відповідно збільшиться або зменшиться на цю ж константу.*

Теорема 2.4. *Якщо кожне значення ознаки ряду розподілу помножити на деяку константу, то і значення середнього помножиться на цю константу* (табл. 2.2). *Якщо кожне значення ознаки ряду розподілу*

поділити на деяку константу, то і значення середньої поділиться на цю ж константу.

Таблиця 2.2

Знаходження відхилень від середнього

Значення ознаки ряду розподілу	Середнє	Різниця
$4 \times 3 = 12$	$8 \times 3 = 24$	–12
$5 \times 3 = 15$	24	–9
$5 \times 3 = 15$	24	–9
$7 \times 3 = 21$	24	–3
$8 \times 3 = 24$	24	0
$10 \times 3 = 30$	24	6
$17 \times 3 = 51$	24	27
	Сума різниць	0

2.4. Деякі приклади застосування середнього, медіани і моди

Внаслідок чутливості середнього до зміни значень ознаки ряду його можна використовувати як вимір, що характеризує всі елементи досліджуваної сукупності.

Середнє можна використати як вимір, що відображає, наскільки великими є у своїй масі елементи даної сукупності. Одним із найпоширеніших методів порівняння елементів двох різних груп полягає у порівнянні їх середніх.

Так, коли ми говоримо, що середнє число попадань у гравців команди **А** в середньому є більшим ніж у команди **Б**, то можна стверджувати, що гравці команди **А** загалом мають вищу майстерність, ніж гравці команди **Б**.

Якщо середнє число дітей в сім'ї в одній країні складає 2,5, а в іншій — 1,5, то це важливе і об'єктивне свідчення того, що в першій країні демографічна ситуація є кращою, ніж у другій. Якщо в першій країні населення поступово збільшується, то в другій воно зменшується.

Тобто *середнє використовують для порівняння між собою груп даних* і визначення того, яка група є переважаючою за даним параметром.

В економетричному аналізі *безумовне математичне сподівання* використовують як довгостроковий прогноз. Його можна знайти, наприклад, як математичне сподівання розв'язку рівняння, що описує динаміку процесу.

Умовне математичне сподівання використовують для визначення короткострокового і середньострокового прогнозів.

Середнє часто використовують для *оцінювання ступеня зміщення вибірки даних від нуля*. Так, математичне сподівання шумової складової вказує на те, чи виконується припущення щодо її центрованості.

Середнє визначають також *для того, щоб видалити його з даних і працювати тільки з відхиленнями*. Такий підхід, як правило, сприяє покращанню якості оцінок математичних моделей завдяки покращанню ступеня обумовленості матриць вимірів.

Медіана ніяк не залежить від крайніх (за значеннями) елементів, що іноді робить її дуже важливим показником. Медіана дає особливо важливу інформацію щодо ряду розподілу у тих випадках, коли *відносно невелике число елементів суттєво відрізняється від загальної маси спостережень*.

Наприклад, нехай у містечку проживає 5 тис. жителів. Припустимо, що всі вони фактично заробляють не більше 8 тис. грн на рік. Однак, власники магазинів та фірм з виробництва м'ясопродуктів і круп'яних виробів, розміщених у цьому містечку, отримують майже стільки, скільки всі інші жителі містечка разом взяті.

Медіана ряду розподілу доходів дає реальнішу картину рівня життя населення містечка, ніж середнє. Те, що на медіану не впливають великі доходи, отримувані окремими особами, перетворює її у *представницький* або *репрезентативний* показник.

Так, якщо річний медіанний рівень доходу становить 5 тис. грн, то це означає, що половина жителів заробляє менше 5 тис. грн на рік, а інша половина — більше 5 тис. грн. При цьому середнє, на величину якого впливають великі доходи власників магазинів та підприємств, може становити 10 тис. грн на рік. Отже, лише невелике число жителів може мати середній дохід.

Значення *моди* дає важливу інформацію для виробників різних товарів, конструкторів, власників магазинів, які поставляють свої товари на конкретний ринок. Так, наприклад, виробник годинників повинен знати, за якою ціною найчастіше купують його продукцію для того, щоб приділити увагу виробництву годинників саме такої вартості.

Власник магазину одягу повинен знати, які розміри костюмів є найбільш розповсюдженими у даному районі для того, щоб відповідно формувати запаси на складі.

Смисл показника моди полягає в тому, що він вказує на максимальну частоту значень деякого показника, яка нерідко свідчить про найбільшу популярність того чи іншого виробу або послуги.

2.5. Позначення та знаходження середнього арифметичного

У подальшому будемо користуватися такими позначеннями:

$X = \{x_1, x_2, ..., x_N\}$ або $\{x(k)\}$, $k = 1, ..., N$ — ряд спостережень довжиною N; $x_1, x_2, ..., x_N$ — елементи ряду;

μ_X — середнє значення ряду елементів X;

$\sum X = \sum_{k=1}^{N} x(k)$ — сума значень елементів ряду;

N_X — число значень ряду розподілу X; якщо розглядається один ряд елементів, то його обсяг позначається просто через N.

Отже, середнє арифметичне будемо розраховувати за формулою:

$$\mu_X = \frac{\sum X}{N} = \frac{\sum_{k=1}^{N} x(k)}{N} = \frac{1}{N}\sum_{k=1}^{N} x(k).$$

2.6. Визначення поточного середнього

Досить часто виникає необхідність знаходження так званого поточного значення середнього. Наприклад, у випадку, коли середнє змінюється у часі (середнє, що є функцією часу, називають *трендом*). Формулу для обчислення поточного середнього можна легко отримати за допомогою формули для арифметичного середнього:

$$\bar{x}(N) = \frac{1}{N}\sum_{k=1}^{N} x(k) = \frac{1}{N}\sum_{k=1}^{N-1} x(k) + \frac{1}{N}x(N) = \frac{1}{N}\frac{N-1}{N-1}\sum_{k=1}^{N-1} x(k) + \frac{1}{N}x(N) =$$

$$= \frac{N-1}{N}\frac{1}{N-1}\sum_{k=1}^{N-1} x(k) + \frac{1}{N}x(N) = \bar{x}(N-1) - \frac{1}{N}\bar{x}(N-1) + \frac{1}{N}x(N) =$$

$$= \bar{x}(N-1) + \frac{1}{N}\left[x(N) - \bar{x}(N-1)\right],$$

де N — поточне значення.

Таким чином, остаточно рекурсивна формула для обчислення поточного середнього має вигляд:

$$\bar{x}(N) = \bar{x}(N-1) + \frac{1}{N}\left[x(N) - \bar{x}(N-1)\right].$$

2.7. Математичне сподівання дискретної випадкової змінної

Математичне сподівання (МС) випадкової змінної приблизно дорівнює її середньому значенню, фактично, *це очікуване значення середнього змінної*. Для розв'язку багатьох задач достатньо знати математичне сподівання.

Означення 2.5. *Математичним сподіванням дискретної випадкової величини називають суму добутків її можливих значень на ймовірності цих значень.*

Якщо випадкова величина X може приймати тільки значення x_1, x_2, ..., x_n з відповідними ймовірностями p_1, p_2, ..., p_n, то математичне сподівання цієї величини $E[X] = E[x(k)]$ визначається за формулою:

$$E[x(k)] = x_1 p_1 + x_2 p_2 + ... + x_n p_n,$$

де E — оператор математичного сподівання.

З визначення випливає, що математичне сподівання дискретної випадкової величини є сталою величиною.

Приклад 2.1. Знайти математичне сподівання числа появ події A в одному випробуванні (досліді), якщо ймовірність події A дорівнює p, тобто $p(A) = p$.

Розв'язок. Випадкова величина X — число появ події A в одному випробуванні — може приймати тільки два значення: $x_1 = 1$ (подія A наступила) з ймовірністю p і $x_2 = 0$ (подія A не наступила) з ймовірністю $q = 1 - p$. Таким чином, математичне сподівання числа появ події A:

$$E[X] = 1 \cdot p + 0 \cdot q = p.$$

Отже, *математичне сподівання числа появ події в одному випробуванні дорівнює ймовірності появи цієї події*. Цим результатом ми скористаємось нижче.

Математичне сподівання є більшим найменшого значення і меншим найбільшого значення випадкової величини. Тобто, її можливі значення на числовій осі знаходяться зліва і справа від математично-

го сподівання. У цьому смислі МС *характеризує розміщення розподілу*, а тому його часто називають *центром розподілу*.

Властивості математичного сподівання

1. *Математичне сподівання постійної величини дорівнює цій величині*

$$E[c] = c.$$

2. *Постійний множник можна виносити за знак математичного сподівання*

$$E[cX] = c\,E[X].$$

3. *Математичне сподівання добутку двох незалежних випадкових величин дорівнює добутку їх математичних сподівань*

$$E[XY] = E[X]\,E[Y].$$

4. *Математичне сподівання суми двох випадкових величин дорівнює сумі МС доданків*

$$E[X + Y] = E[X] + E[Y].$$

Наслідок. *Математичне сподівання суми кількох випадкових величин дорівнює сумі математичних сподівань доданків.*

Приклад 2.2. Виконується три постріли з ймовірностями попадання у ціль, рівними відповідно $p_1 = 0{,}4$; $p_2 = 0{,}3$ і $p_3 = 0{,}6$. Знайти МС загального числа попадань у ціль.

Розв'язок. Числом попадань при першому пострілі є випадкова величина X_1, яка може приймати тільки два значення: 1 (попадання) з ймовірністю $p_1 = 0{,}4$ і 0 (промах) з ймовірністю $q = 1 - 0{,}4 = 0{,}6$.

МС числа попадань при першому пострілі дорівнює ймовірності попадання (попередній приклад), тобто, $E[X_1] = 0{,}4$. Аналогічно знайдемо МС числа попадань при другому і третьому пострілах: $E[X_2] = 0{,}3$ і $E[X_3] = 0{,}6$.

Загальне число попадань також є випадкова величина, яка є сумою попадань у кожному з трьох пострілів:

$$X = X_1 + X_2 + X_3\,.$$

Таким чином, шукане МС знайдемо за властивістю МС суми:

$$E[X] = E[X_1 + X_2 + X_3] = E[X_1] + E[X_2] + E[X_3] = \\ = 0{,}4 + 0{,}3 + 0{,}6 = 1{,}3 \quad \text{(попадань)}.$$

2.8. Інші види середнього

Далеко не в усіх випадках типові ознаки процесів можна характеризувати за допомогою середнього арифметичного. Тому застосовують інші форми середнього, наприклад, *середнє квадратичне* і *середнє кубічне*:

$$\mu_2 = \sqrt{\frac{1}{N}\sum_{k=1}^{N} x^2(k)}, \quad \mu_3 = \sqrt[3]{\frac{1}{N}\sum_{k=1}^{N} x^3(k)},$$

середнє геометричне

$$G = \mu_{\text{геом}} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n},$$

та *середнє гармонійне*

$$H = \frac{N}{\sum_{k=1}^{N}\left[x(k)\right]^{-1}}.$$

При визначенні середніх величин для значень площі і об'єму необхідно узгоджувати середні значення для лінійних розмірів та середні для отриманих за ними значень площі і об'єму. Обчислене за лінійними розмірами середнє арифметичне, як правило, дає менші значення площі та об'єму, ніж ті, які отримують за фактичними лінійними розмірами. У таких випадках обчислюють середнє квадратичне для значень лінійного параметра поверхні і середнє кубічне для значень лінійного параметра об'єму.

Приклад 2.3. Розглянемо приклад *використання середнього квадратичного* при визначенні залежності площі листків *примули* від їх лінійного параметра. Форма участків поверхні листків може бути довільною, але однаковою або близькою у рамках вибірки, що усереднюється. Значення площі листків у рамках цієї вибірки повинні визначатися за однією формулою. Вихідні дані для аналізу наведені у табл. 2.3.

Таблиця 2.3

Розміри листків примули

Довжина листка x, см	Фактична площа листка y, см^2
3	7
5	20
6	30
8	55
$\sum x = 22$; $\mu_x = 5{,}50$; $\mu_{x2} = 5{,}788$	$\sum y = 112$; $\mu_y = 28{,}00$

Покажемо, що користуватись середнім арифметичним при знаходженні площі листка буде некоректно. Площа листків примули з хорошим наближенням може бути обчислена за наступною емпіричною формулою [4]:

$$y = 0{,}6728 \cdot x^{2{,}1229},$$

де y — площа листка; x — довжина листка.

Середнє арифметичне рядів x і y дорівнюють:

$$\mu_x = 5{,}50 \text{ см і } \mu_y = 28{,}00 \text{ см}^2.$$

Середня площа листка з використанням середнього арифметичного довжини листка буде дорівнювати:

$$y = 0{,}6728 \cdot 5{,}50^{2{,}1229} = 25{,}00 \text{ см}^2,$$

тобто, середнє арифметичне довжини листка не дає правильного середнього значення площі листка. Середнє квадратичне довжини листка:

$$\mu_{x2} = \sqrt{\frac{3^2 + 5^2 + 6^2 + 8^2}{4}} = 5{,}7879 \approx 5{,}788.$$

Тепер визначимо середню площу листка через середнє квадратичне:

$$y = 0{,}6728 \cdot 5{,}78790^{2{,}1229} = 28{,}000 \text{ см}^2,$$

тобто, це значення повністю відповідає фактичному значенню площі листків. Таким чином, середнє квадратичне точніше передає залежність площі від лінійного параметра об'єкта дослідження.

Приклад 2.4. Розглянемо приклад *використання середнього геометричного*. Нехай нас цікавить середній за рік щомісячний темп приросту ваги птахів, яких розводять на фермі. При цьому вимірюється відносний приріст ваги за кожний місяць:

- за перший місяць — в 1,23 раза;
- за другий місяць — в 1,65 раза;
- за третій місяць — в 1,08 раза і т. д.

У такому випадку більш адекватно середньомісячний темп приросту буде виражатись не середнім арифметичним, а середнім геометричним. Адекватність описання потрібно розуміти наступним чином. Якщо мова йде про *абсолютний приріст*, то адекватною середньою характеристикою процесу буде середнє арифметичне, тому що воно до-

рівнює загальному приросту за рік, поділеному на число місяців; і навпаки, для знаходження повного абсолютного значення приросту необхідно помножити середній приріст за місяць на число місяців.

Якщо мова йде про *відносний приріст*, то характер обчислень змінюється. Дійсно, нехай на початку року загальна вага птахів на фермі дорівнює 1000 кг. Оскільки за перший місяць вага збільшилась у 1,23 раза, то на кінець першого місяця загальна вага становить 1230 кг. За другий місяць вага збільшилась у 1,65 раза. Тобто, на кінець другого місяця загальна вага птахів становитиме $1230 \cdot 1{,}65 \approx 2030$ кг, а на кінець третього місяця вона складе $2030 \cdot 1{,}08 \approx 2190$ кг.

Визначаючи середній за квартал щомісячний темп приросту ваги як середнє геометричне, отримаємо:

$$\mu_{\text{геом}} = \sqrt[3]{1{,}23 \cdot 1{,}65 \cdot 1{,}08} \approx 1{,}30,$$

а середнє арифметичне буде таким:

$$\mu_X = \frac{1{,}23 + 1{,}65 + 1{,}08}{3} = 1{,}32.$$

Тепер правильний загальний приріст ваги за квартал отримаємо тільки у випадку, якщо скористаємось середнім геометричним:

Вага в кінці кварталу = $1000 \cdot 1{,}30 \cdot 1{,}30 \cdot 1{,}30 = 2190$ кг.

Приклад 2.5. *Середнє гармонійне* застосовують, зокрема, при усередненні величин, які представляють собою зміни швидкості протікання досліджуваних процесів. Наприклад, при дослідженні приросту довжини або діаметру пагінців, при усередненні індексів інфляції, або при дослідженні зміни концентрації речовин у міліграмах на літр.

Розглянемо визначення середнього гармонійного для концентрації деякої шкідливої речовини у воді озера біля промислового підприємства в мг/л. Нехай виміри концентрації речовини, виконані у чотирьох кутах та посередині озера, виявились наступними:

$$X = [5 \quad 7 \quad 10 \quad 13 \quad 15].$$

Середнє арифметичне цих концентрацій становить:

$$\mu_X = \frac{5 + 7 + 10 + 13 + 15}{5} = \frac{50}{5} = 10 \text{ мг/л}.$$

Середнє гармонійне цих самих чисел дорівнює:

$$H = \frac{5}{1/5 + 1/7 + 1/10 + 1/13 + 1/15} = \frac{50}{0{,}587} = 8{,}52 \text{ мг/л}.$$

Прямим підрахунком можна показати, що у даному випадку точніше значення має середнє гармонійне. Так, загальний об'єм п'яти зразків води (по одному зразку з кожної вказаної точки озера) при умові, що в кожному зразку буде міститись 100 мг речовини, дорівнює:

$$\frac{100}{5}+\frac{100}{7}+\frac{100}{10}+\frac{100}{13}+\frac{100}{15}=58{,}7 \text{ мг.}$$

Таким чином, фактична маса речовини у даному об'ємі дорівнює: 500 мг.

Загальна маса речовини, визначена в об'ємі 58,7 мг через середнє гармонійне H, у цьому об'ємі становить:

$$58{,}7 \cdot 8{,}52 = 500{,}1 \text{ мг,}$$

а маса, визначена через середнє арифметичне, дорівнює:

$$58{,}7 \cdot 10 = 587 \text{ мг.}$$

Тобто, середнє гармонійне дає похибку 0,1 мг, а середнє арифметичне веде до похибки 87 мг.

2.9. Варіація

Елементи ряду розподілу можуть суттєво різнитися між собою і можуть бути дуже схожими. Наприклад, якщо ми виміряємо артеріальний тиск кожного з нас, то отримаємо ряд, значення якого досить близькі між собою. Однак, якщо утворити ряд значень, який описує формування цін на біржові акції, то ці значення можуть дуже сильно різнитись між собою. Можна сказати, що *варіація або розсіювання* даних у другому випадку є набагато більшою, ніж у першому.

Варіація є надзвичайно важливим поняттям, яке характеризує взаємозв'язок між всіма елементами ряду розподілу, взятого в цілому. Далі покажемо, що зміна значення одного елемента ряду приводить до зміни варіації ряду в цілому. Величину варіації або ступеня розсіювання елементів ряду можна, в деякій мірі, оцінити візуально, але це буде тільки якісна суб'єктивна оцінка. Необхідно мати об'єктивну числову міру варіації.

2.10. Виміри варіації: середнє відхилень та дисперсія

Виміри варіації повинні вказувати на ступінь розсіювання елементів ряду. Вони мають задовольняти певним вимогам. Сформулюємо ці вимоги або властивості показника.

Перша властивість значення показника розсіювання має бути *не дуже великим*, якщо значення елементів ряду, на основі яких він розраховується, *не дуже сильно відрізняються* одне від одного. І навпаки, його значення повинне бути великим, якщо значення елементів ряду сильно розсіянні.

Друга властивість показника розсіювання полягає в тому, що його значення *не повинне залежати від числа елементів ряду*. Точніше кажучи, нам не потрібен показник розсіювання, значення якого зросталo б тільки внаслідок *збільшення* числа елементів ряду. Він повинен відображати лише подібність або розбіжність між самими числами і не залежати від їх кількості.

Третя властивість даного показника полягає у наступному: оскільки він характеризує тільки ступінь розсіювання даних, то він *не повинен залежати від значення середнього*. Показник середнього ніяк не пов'язаний із варіацією елементів і його величина не повинна впливати на величину розсіювання даних.

Перший вимір варіації — середнє значення відхилень від середнього

Якщо знайти суму всіх абсолютних відхилень від середнього і поділити її на число елементів ряду, то знайдемо вимір варіації, яку називають *середнім відхиленням*.

Однак середнє відхилення використовується досить рідко внаслідок того, що цей вимір є недостатньо інформативним. Існує інший підхід до визначення виміру варіації, який ґрунтується на використанні квадратів відхилень.

Другий вимір варіації — дисперсія

Для того щоб сума квадратів відхилень не була чутливою до числа елементів ряду, суму квадратів відхилень необхідно розділити на число елементів. Отриманий показник називають *дисперсією*.

Означення 2.6. *Дисперсією називають середнє суми квадратів різниць між елементами ряду розподілу та їх середнім.*

Таким чином, дисперсію можна знайти за формулою:

$$\mathrm{var}\left[y(k)\right]=\sigma_y^2=\frac{1}{N-1}\sum_{k=1}^{N}\left[y(k)-\mu_y\right]^2,$$

де var — позначення дисперсії; μ_y — середнє значення ряду розподілу $\{y(k)\}$. Ділення суми квадратів відхилень від середнього на $N-1$ за-

безпечує незміщеність оцінки дисперсії. Якщо, скажімо, ряд A має більшу дисперсію ніж ряд $Б$, то варіація ряду A є вищою. Скориставшись математичним сподіванням, отримаємо наступний вираз для дисперсії:

$$\text{var}\,[y(k)] = E\,[y(k) - E(y(k))]^2 = E\,[Y - E(Y)]^2.$$

Розрахункову формулу для дисперсії можна представити також в іншому вигляду:

$$\sigma_y^2 = \frac{\sum_{k=1}^{N}\left[y(k)-\mu_y\right]^2}{N-1} = \frac{\sum y^2(k) - 2\mu_y\sum y(k) + \sum \mu_y^2}{N-1} =$$

$$= \frac{\sum y^2(k)}{N-1} - 2\mu_y^2 + \left(\frac{\sum y(k)}{N-1}\right)^2 = \frac{\sum y^2(k)}{N-1} - 2\mu_y^2 + \mu_y^2 =$$

$$= \frac{\sum y^2(k)}{N-1} - \mu_y^2$$

або

$$\sigma_y^2 = \frac{\sum y^2(k)}{N-1} - \frac{\left(\sum y(k)\right)^2}{(N-1)^2}.$$

Використовуючи оператор математичного сподівання, останню формулу для дисперсії можна записати у вигляді:

$$\text{var}\left[y(k)\right] = \sigma_y^2 = E\left[y^2(k)\right] - \left\{E\left[y(k)\right]\right\}^2,$$

тобто, *дисперсія дорівнює різниці між математичним сподіванням квадрату випадкової величини $y(k)$ та квадратом її математичного сподівання.*

Приклад 2.6. Знайти дисперсію випадкової величини X, яка задана таким розподілом:

$$X = [2 \quad 3 \quad 5]; \quad p = [0{,}1 \quad 0{,}6 \quad 0{,}3].$$

Розв'язок. Математичне сподівання $E\,[X] = E\,[x(k)]$:

$$E\,[X] = 2 \cdot 0{,}1 + 3 \cdot 0{,}6 + 5 \cdot 0{,}3 = 3{,}5.$$

Знайдемо квадрати значень змінної $X^2 = [4 \quad 9 \quad 25]$ і математичне сподівання $E\,[X^2]$:

$$E\,[X^2] = 4 \cdot 0{,}1 + 9 \cdot 0{,}6 + 25 \cdot 0{,}3 = 13{,}3.$$

Шукана дисперсія:

$$\text{var}\,[X] = E[X^2] - \{E[X]\}^2 = 13{,}3 - (3{,}5)^2 = 1{,}05.$$

2.11. Знаходження незміщеної оцінки дисперсії

Коректне знаходження значень статистичних параметрів вибірок даних розглядається в теорії оцінювання, яка є самостійною дисципліною для вивчення. Однак деякі положення цієї теорії розглядають у курсі математичної статистики з метою ознайомлення з методами коректного оцінювання статистичних параметрів (статистик). *Оцінити* — означає знайти значення статистики шляхом застосування процедур оцінювання до експериментальних даних.

Наприклад, можна взяти вибірку коефіцієнтів розумового розвитку (КРР) для вибраної соціальної групи при $N = 50$ і знайти середнє $\bar{x} = 115$. Очевидно, що знайдена оцінка є добрим наближенням лише для даної вибірки. Якщо вибірку зменшити або розширити, то середнє арифметичне отримує інше значення. Можна було б, також, знайти середнє для мінімального і максимального обсягів вибірки. Воно може відрізнятися від 115. Існують різні критерії вибору кращих оцінок, але найбільш важливим серед них є критерій *незміщеності.*

Оператор оцінювання параметра має властивість незміщеності, якщо середнє вибіркових оцінок, отриманих з незалежних випадкових вибірок, наближається до істинного значення параметра при необмеженому зростанні кількості вибірок.

Нехай є 5 різних вибірок значень КРР із генеральної сукупності. Мета полягає в тому, щоб оцінити невідоме середнє значення розподілу. Для наявних 5-ти вибірок вибіркові середні мають наступні значення: $\bar{x}_1 = 108$, $\bar{x}_2 = 107$, $\bar{x}_3 = 113$, $\bar{x}_4 = 115$ і $\bar{x}_5 = 105$. Середнє значення цих оцінок дорівнює 109,6.

Якщо збільшувати число вибірок, то середнє оцінок середнього буде наближатися до істинного значення середнього розподілу. Так, середнє для 5000 вибіркових оцінок середніх буде відрізнятись від істинного середнього не більше, ніж на 1/2 стандартного відхилення розподілу. Для того щоб переконатись у цьому, розглянемо середнє 5000 значень середніх як середнє вибірки обсягом $N = 25000$. Якщо навіть стандартне відхилення значень КРР перевищує 20, то існує невеликий шанс, що середнє ряду, що складається з 25000 елементів, буде відрізнятись від істинного середнього значення розподілу біль-

ше, ніж на 1/2 стандартного відхилення, що доводиться наступною теоремою.

Теорема 2.5. *Нехай з однієї нескінченної сукупності формується нескінченне число випадкових вибірок однакового обсягу N. Розподіл середніх значень цих вибірок має стандартне відхилення (стандартне відхилення середніх), що дорівнює стандартному відхиленню вихідної сукупності, діленому на* $\sqrt{N}$, тобто

$$\sigma_M = \frac{\sigma}{\sqrt{N}},$$

де σ — стандартне відхилення генеральної сукупності, з якої сформовано вибірки для обчислення середніх; σ_M — стандартне відхилення середніх.

На основі сказаного можна зробити висновок, що чим більше ми маємо оцінок середніх для різних вибірок, тим ближчим буде середнє цих оцінок до істинного значення середнього розподілу. Тому кажуть, що арифметичне середнє дає оцінку середнього розподілу, яка має властивість незміщеності.

Незміщена оцінка дисперсії

Розглянемо приклад з визначенням статистичних параметрів для значень КРР, наведених у табл. 2.4. Нехай істинне середнє розподілу відоме: $\mu = 110$ (наприклад, його можна знайти за допомогою моделі процесу).

Таблиця 2.4

КРР і варіації, обчислені за допомогою відомого середнього

№ пор.	КРР	$x - \mu$	$(x - \mu)^2$	Сума 5-ти квадратів різниць
1	2	3	4	5
1	90	–20	400	
2	100	–10	100	
3	105	– 5	25	
4	145	35	1225	
5	100	–10	100	1850
6	120	10	100	
7	110	0	0	
8	100	–10	100	
9	115	5	25	

Закінчення табл. 2.4

1	2	3	4	5
10	90	–20	400	625
11	105	–5	25	
12	120	10	100	
13	95	–15	225	
14	130	20	400	
15	115	5	25	775

Квадрати різниць, необхідні для обчислення дисперсії, наведено в четвертому стовпчику. Якщо для оцінювання дисперсії взяти першу вибірку з 5-ти значень, то оцінка дисперсії приймає значення: 1850/5 = 370.

Однак реальна ситуація відрізняється від наведеної тим, що істинне середнє значення розподілу невідоме. Таким чином, знову візьмемо першу вибірку КРР із 5-ти значень і знайдемо для неї вибіркове арифметичне середнє: $\mu_1 = 108$. Тепер дисперсія першої вибірки з 5-ти значень:

$$\sigma_1^2 = \frac{1}{5}\sum_{i=1}^{5}[x_i - 108]^2 = \frac{1830}{5} = 366,$$

тобто отримане значення оцінки дисперсії є меншим, ніж при використанні істинного значення середнього. Цей результат не випадковий.

Теорема 2.6. *Для кожної вибірки сума квадратів відхилень від середнього менша суми квадратів відхилень від будь-якої іншої точки.*

Для наступних п'яти членів другого стовпчика табл. 2.4 вибіркове середнє: $\mu_2 = 107$, а сума квадратів відхилень від цього вибіркового середнього дорівнює 580, тобто є меншою 625.

Таким чином, ми переконались, що неможливо отримати незміщену оцінку дисперсії розподілу за допомогою звичайної середньої суми квадратів відхилень від вибіркової оцінки середнього. Така оцінка буде завжди зміщеною. Для подальшого розгляду необхідно ввести поняття *числа ступенів вільності*, яке відіграє важливу роль у теорії оцінювання.

Поняття кількості ступенів вільності

Розглянемо різницю між числом незалежних фактів і загальним числом фактів, що містяться в експериментальних даних. Наприклад,

нехай абітурієнт отримав 85 балів із 100 можливих, тобто втратив 15 балів. У даному випадку у нас є три числових факти: 85, 100 і 15. Однак з них тільки два є незалежними, оскільки третій можна обчислити з двох інших.

Факт називають незалежним від іншого або від групи інших фактів, якщо він несе нову інформацію, яку неможливо отримати з групи фактів, що порівнюється з ним.

У подальшому будемо використовувати термін *ступені вільності* для *позначення незалежної інформації*, тобто інформації, яку неможливо отримати (вивести) ні з якої іншої групи даних, що аналізуються. Отже, інформація щодо абітурієнта містить два ступеня вільності.

Розглянемо тепер коректну процедуру оцінювання вибіркової дисперсії. Нехай є вибірка з двох даних: 90 і 100. Для них вибіркове середнє становить 95, а відхилення від середнього: –5 і +5. Нагадаємо, що сума відхилень повинна дорівнювати нулю. Таким чином, відхилення є залежними величинами, оскільки $(x_1 - \bar{x}) + (x_2 - \bar{x}) = 0$. Звідси можна зробити висновок, що вибірка з двох чисел дає тільки один факт щодо значення дисперсії. Хоча вибірка містить два елементи, оцінка дисперсії базується тільки на одному ступені вільності.

Якщо вибірка складається з N елементів і *відоме середнє розподілу* генеральної вибірки, то для кожного з них також можна обчислити значення відхилення від середнього (наприклад, середнє можна обчислити за моделлю процесу). Таким чином, ми будемо мати N незалежних значень відхилень для обчислень дисперсії, тобто N ступенів вільності.

Якщо середнє розподілу невідоме (реальна ситуація), то знаходячи вибіркове середнє (центральну точку), ми можемо визначити тільки $N - 1$ незалежних відхилень від цієї точки. Це зумовлено тим, що сума всіх відхилень повинна дорівнювати нулю незалежно від довжини вибірки. Наприклад, якщо сума будь-яких $N - 1$ відхилень дорівнює –5, то виключене з розгляду відхилення буде дорівнювати +5. Таким чином, у загальному випадку, N різниць (відхилень) містять $N - 1$ ступінь вільності й визначення оцінки дисперсії також ґрунтується на $N - 1$ ступені вільності. Отримана таким чином оцінка дисперсії буде незміщеною. Сказане може бути сформульоване у вигляді теореми.

Теорема 2.7. *Незміщена оцінка дисперсії випадкової вибірки з N елементів визначається як сума квадратів відхилень всіх членів вибірки від вибіркового середнього, поділена на $N - 1$*:

$$s_x^2 = \frac{1}{N-1}\sum_{k=1}^{N}\left[x(k)-\bar{x}\right]^2,$$

де s_x^2— вибіркова дисперсія випадкової величини x, тобто це незміщена оцінка дисперсії генеральної сукупності σ_x^2.

Зазначимо ще раз, що дільник $N - 1$ необхідно застосовувати коректно. Якщо експериментальні дані є вибіркою з деякого розподілу, дисперсію якого необхідно оцінити, то для знаходження оцінки дисперсії необхідно застосовувати дільник $N - 1$. У такому випадку знайдене значення s^2 є вибірковою дисперсією, а s — вибіркове стандартне відхилення.

Якщо ж набір даних представляє всю вибірку повністю, то необхідно застосовувати дільник N, а обчисленою величиною буде σ^2, а не s^2.

2.12. Властивості дисперсії

Оскільки константа не має розсіювання, то її дисперсія повинна дорівнювати нулю.

Властивість 1. *Дисперсія постійної величини дорівнює нулю*:

$$\text{var}\,[c] = 0.$$

Доведення. За визначенням дисперсії

$$\text{var}\,[c] = E\,\{[c - E(c)]^2\}.$$

Оскільки математичне сподівання константи (перша властивість МС) дорівнює самій константі, то

$$\text{var}\,[c] = E\,[(c - c)^2] = E\,[0] = 0.$$

Властивість 2. *Постійний множник, піднесений до квадрату, виноситься за знак дисперсії*:

$$\text{var}\,[cX] = c^2\ \text{var}\,[X].$$

Доведення. За визначенням дисперсії маємо:

$$\text{var}\,[cX] = E\,\{[cX - E(cX)]^2\}.$$

Скористаємось другою властивістю математичного сподівання — постійний множник можна виносити за символ оператора. У результаті отримаємо:

$$\text{var}\,[cX] = E\,\{[cX - cE(X)]^2\} = E\,\{c^2[X - E(X)]^2\} = \\ = c^2\,E\,\{[X - E(X)]^2\} = c^2\ \text{var}\,[X],$$

тобто, $\text{var}\,[cX] = c^2\,\text{var}\,[X]$.

Властивість 3. *Дисперсія суми двох незалежних випадкових величин дорівнює сумі дисперсій цих величин*:

$$\text{var}\,[X + Y] = \text{var}\,[X] + \text{var}\,[Y].$$

Доведення. За формулою для знаходження дисперсії маємо:

$$\text{var}\,[X + Y] = E\,[X + Y]^2 - [E\,(X + Y)]^2.$$

Розкриємо дужки і скористаємось властивостями математичного сподівання щодо суми кількох величин та добутку двох незалежних випадкових величин:

$$\text{var}\,[X + Y] = E\,[X^2 + 2XY + Y]^2 - [E\,(X) + E(Y)]^2 = \\ = E\,[X^2] + 2\,E\,[X]\,E\,[Y] + E\,[Y^2] - E^2\,[X] - 2E\,[X]\,E\,[Y] - E^2\,[Y] = \\ = \{E(X^2) - [E\,(X)]^2\} + \{E(Y^2) - [E\,(Y)]^2\} = \\ = \text{var}\,[X] + \text{var}\,[Y].$$

Наслідок 1. *Дисперсія суми кількох взаємно незалежних випадкових величин дорівнює сумі дисперсій цих величин.*

Наслідок 2. *Дисперсія суми постійної величини і випадкової дорівнює дисперсії випадкової величини.*

Властивість 4. *Дисперсія різниці двох незалежних випадкових величин дорівнює сумі їх дисперсій*:

$$\text{var}\,[X - Y] = \text{var}\,[X] + \text{var}\,[Y].$$

Доведення. За третьою властивістю

$$\text{var}\,[X - Y] = \text{var}\,[X] + \text{var}\,[-Y].$$

Згідно з другою властивістю маємо:

$$\text{var}\,[X - Y] = \text{var}\,[X] + (-1)^2\,\text{var}\,[Y] = \text{var}\,[X] + \text{var}\,[Y].$$

2.13. Стандартне або середнє квадратичне відхилення

Ще одним показником варіації ряду є *стандартне* або *середнє квадратичне відхилення*, яке обчислюється за допомогою дисперсії.

Означення 2.7. *Стандартним відхиленням називають показник, що дорівнює квадратному кореню з дисперсії, взятому із знаком плюс.* У наших позначеннях це буде σ_y.

Стандартне відхилення має ті самі властивості, що і дисперсія. Його застосування буде розглянуте нижче, але можна сказати, що його, як і дисперсію, часто використовують, наприклад, як міру ризику при аналізі фінансових та економічних процесів. Воно часто використовується в системах статистичного аналізу і контролю якості продукції, а також у багатьох інших випадках. Розраховується стандартне відхилення за формулою:

$$\sigma_y = \sqrt{\frac{\sum_{k=1}^{N}\left[y(k)-\mu_y\right]^2}{N-1}}.$$

Середнє квадратичне відхилення суми незалежних випадкових величин

У процесі виконання статистичного аналізу досить часто необхідно розглядати кілька випадкових величин. Нехай відомі стандартні відхилення для кількох взаємно незалежних випадкових величин. Необхідно знайти стандартне відхилення суми значень цих величин. Відповідь на це запитання дає наступна теорема.

Теорема 2.8. *Стандартне відхилення суми скінченого числа взаємно незалежних випадкових величин дорівнює квадратному кореню із суми квадратів стандартних відхилень цих величин*:

$$\sigma_{X_1+X_2+\ldots+X_n} = \sqrt{\sigma^2_{X_1} + \sigma^2_{X_2} + \ldots + \sigma^2_{X_n}}.$$

Доведення. Позначимо через X суму випадкових величин:

$$X = X_1 + X_2 + \ldots + X_n$$

і знайдемо дисперсію цієї суми:

$$\text{var}\,[X] = \text{var}\,[X_1] + \text{var}\,[X_2] + \ldots + \text{var}\,[X_n],$$

а звідси стандартне відхилення:

$$\sigma_X = \sqrt{\text{var}[X_1] + \text{var}[X_2] + \ldots + \text{var}[X_n]},$$

або остаточно

$$\sigma_X = \sqrt{\sigma^2_{X_1} + \sigma^2_{X_2} + \ldots + \sigma^2_{X_n}}.$$

2.14. Вплив зміни значень елементів ряду на дисперсію

Нагадаємо, що додавання константи до всіх значень ряду приводить до зміни середнього на таку саму величину. Однак додавання константи до кожного елемента ряду не впливає на значення дисперсії ряду розподілу, оскільки при обчисленні варіації середнє вираховується з кожного елемента. Має місце теорема.

Теорема 2.9. *Збільшення або зменшення кожного елемента ряду розподілу на деяку константу не впливає на значення варіації ряду розподілу і, відповідно, не впливає на значення дисперсії та стандартного відхилення.*

З іншого боку, множення кожного елемента ряду на константу впливає на варіацію цього ряду. Приклад наведено в табл. 2.5. У результаті множення кожного елемента ряду на 5 дисперсія збільшилась у 25 разів, а стандартне відхилення — у 5 разів.

Таблиця 2.5

Вплив множення на константу елементів ряду на дисперсію

Початковий ряд розподілу			Характеристики ряду, отриманого множенням кожного елемента на 5		
Елемент ряду	Середнє	Відхилення від середнього	Елемент ряду	Середнє	Відхилення від середнього
1	2	3	1	2	3
4	8	–4	20	40	–20
5	8	–3	25	40	–15
5	8	–3	25	40	–15
7	8	–1	35	40	–5
8	8	0	40	40	0
10	8	2	50	40	10
17	8	9	85	40	45
Дисперсія		17,143	Дисперсія		428,57
Стандартне відхилення		4,14	Стандартне відхилення		20,70

Має місце така властивість стандартного відхилення.

Теорема 2.10. *При множенні кожного елемента ряду розподілу на деяку константу стандартне відхилення множиться на абсолютну*

величину цієї константи, а дисперсія — на квадрат цієї константи. Аналогічні наслідки мають місце при діленні значень ряду на константу.

2.15. Застосування дисперсії

Дисперсія і стандартне відхилення знаходять дуже широке застосування у прикладних і теоретичних задачах. Наведемо кілька прикладів їх практичного використання.

1. Дисперсія є важливою *мірою неоднорідності* даних. Так, дисперсія оцінок за тест, який щоденно виконується студентом протягом 10 днів, свідчить про різну ступінь його зосередженості в процесі відповіді на запитання. Якщо аналізуються доходи жителів міста, то висока дисперсія буде свідчити про велику різницю у рівнях доходів.

2. Дисперсія часто використовується як *міра ризику* у фінансах. Наприклад, високе значення дисперсії доходів від цінних паперів свідчить про те, що існує високий ризик не отримати ці доходи і навіть втратити кошти, витрачені на придбання цінних паперів.

3. Дисперсію використовують також як *міру інформативності* сигналів та рядів даних. Так, нульове значення дисперсії свідчить про відсутність інформації, тобто статистичні дані є сталі або нулі. Збільшення дисперсії говорить про те, що дані містять інформацію про зміни, що відбуваються у процесі чи об'єкті.

4. Дисперсію використовують також для опису поведінки нестаціонарних процесів. Існує клас процесів, нестаціонарних стосовно дисперсії (значення дисперсії є функцією часу), які отримали назву *гетероскедастичні* процеси, тобто процеси, для яких $\text{var}\,[y(k)] \neq \text{const}$. Використовуючи *дисперсію як змінну, що характеризує поведінку процесу*, можна побудувати математичну модель, яка описує динаміку дисперсії. Така модель дає можливість прогнозувати значення дисперсії на задане число кроків і приймати рішення на основі цього прогнозу. Наприклад, це може бути рішення щодо купівлі цінних паперів. Для прогнозування значення дисперсії при виконанні аналізу часових рядів користуються *умовною дисперсією*:

$$\sigma^2(k) = \frac{1}{k-1}\sum_{i=1}^{k}\left[y(i) - \mu_y\right]^2, \quad k = 2, 3, \ldots .$$

5. Дисперсію також використовують у *системах статистичного аналізу якості* продукції. Системи аналізу і контролю якості — це

невід'ємна складова сучасного виробництва, оскільки вона забезпечує неперервний контроль та підвищення якості продукції. При цьому однією з основних мір якості є стандартне відхилення розмірів деталей від нормативів.

6. Дисперсія може служити за *вимір подібності процесів*, які описуються вибірками даних однакового змісту. Наприклад, якщо необхідно порівняти коефіцієнти розумового розвитку для різних груп студентів чи населення, то одним із вимірів може бути дисперсія (стандартне відхилення).

2.16. Однаково розподілені взаємно незалежні випадкові величини

Відомо, що за допомогою закону розподілу можна знайти числові характеристики випадкової величини. Розглянемо випадок, коли аналізуються кілька випадкових величин, що мають однакові розподіли з однаковими числовими характеристиками.

Нехай $X_1, X_2, ..., X_n$ — взаємно незалежні випадкові величини з однаковими законами розподілу і однаковими характеристиками (математичне сподівання, дисперсія та інші можливі параметри). З практичної точки зору, як правило, є цікавим дослідження середнього арифметичного цих величин.

Позначимо середнє арифметичне вказаної множини величин через $\bar{X}$:

$$\bar{X} = \left(X_1 + X_2 + \ldots + X_n\right)/n.$$

Три наступні положення встановлюють зв'язок між числовими характеристиками середнього арифметичного $\bar{X}$ та відповідними характеристиками кожної з випадкових величин.

1. *Математичне сподівання середнього арифметичного однаково розподілених (тобто з однаковими статистичними характеристиками) взаємно незалежних випадкових величин дорівнює математичному сподіванню кожної з величин, тобто* $\bar{x}_1 = \bar{x}_2 = \ldots = \bar{x}_n = \bar{x}_0$, де

$$\bar{x}_0 = E\left[\bar{X}\right].$$

Доведення. Винесемо постійний множник за знак математичного сподівання:

$$E\left[\bar{X}\right] = E\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right] = \frac{E\left[X_1\right] + E\left[X_2\right] + \ldots + E\left[X_n\right]}{n}.$$

Оскільки МС всіх величин однакові і дорівнюють $\bar{x}_0$, то

$$E\left[\bar{X}\right] = \frac{n\bar{x}_0}{n} = \bar{x}_0.$$

2. *Дисперсія середнього арифметичного n однаково розподілених взаємно незалежних випадкових величин є в n разів меншою дисперсії кожної з величин, тобто* $\sigma^2_{X_1} = \sigma^2_{X_2} = \ldots = \sigma^2_{X_n} = \sigma^2_0$:

$$\mathrm{var}\left[\bar{X}\right] = \sigma^2_{\bar{X}} = \frac{\sigma^2_0}{n}.$$

Доведення. Винесемо постійний множник у квадраті за знак дисперсії:

$$\mathrm{var}\left[\bar{X}\right] = \mathrm{var}\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right] = \frac{\mathrm{var}\left[X_1\right] + \mathrm{var}\left[X_2\right] + \ldots + \mathrm{var}\left[X_n\right]}{n^2}$$

або

$$\mathrm{var}\left[\bar{X}\right] = \frac{n\sigma^2_0}{n^2} = \frac{\sigma^2_0}{n}.$$

3. *Стандартне відхилення середнього арифметичного n однаково розподілених взаємно незалежних випадкових величин є в* $\sqrt{n}$ *разів меншим стандартного відхилення* σ_0 *кожної з величин*:

$$\sigma\left[\bar{X}\right] = \frac{\sigma_0}{\sqrt{n}}.$$

Доведення. Оскільки $\mathrm{var}\left[\bar{X}\right] = \sigma^2_0/n$, то середнє квадратичне відхилення для $\bar{X}$ дорівнює:

$$\sigma\left[\bar{X}\right] = \sqrt{\mathrm{var}\left[\bar{X}\right]} = \sqrt{\sigma^2_0/n} = \sigma_0/n.$$

Оскільки дисперсія і середнє квадратичне є виміром розсіювання випадкової величин, то із наведених вище результатів можна зробити висновок, що середнє арифметичне суми кількох взаємно незалежних випадкових величин має значно менше розсіювання, ніж кожна величина окремо.

Приклад 2.7. При дослідженні фізичних та хімічних процесів для визначення оцінок змінних роблять кілька вимірів, а потім знаходять середнє арифметичне отриманих чисел, яке приймають за наближене значення (оцінку) вимірюваної величини. За припущення, що виміри виконуються в одних і тих самих умовах, необхідно довести наступне:

А) середнє арифметичне дає надійніший результат ніж окремі виміри;

Б) збільшення числа вимірів приводить до підвищення надійності результату.

Розв'язок. А) Відомо, що окремо взяті виміри однієї і тієї самої змінної дають неоднакові значення. Результат кожного виміру залежить від багатьох випадкових впливів (зміна температури, вплив випадкових шумових сигналів, вплив сторонніх магнітних полів та ін.). Усі випадкові впливи неможливо врахувати наперед.

Тому ми розглядаємо результати n окремих вимірів як випадкові величини x_1, x_2, ..., x_n. Оскільки виміри виконуються за допомогою одного і того приладу і за однією методикою, то вважаємо, що вони мають однаковий розподіл ймовірностей. Крім того, вони взаємно незалежні, тобто результат кожного окремого виміру не залежить від інших вимірів.

Раніше було показано, що середнє арифметичне випадкових величин має менше розсіювання, ніж кожна величина, взята окремо. Тобто середнє арифметичне виявляється ближчим до істинного значення сигналу, ніж результат окремого виміру. З цього випливає, що середнє арифметичне кількох вимірів є надійнішим результатом, ніж окремо взятий вимір.

Б) Також було встановлено, що із зростанням числа окремих випадкових величин, розсіювання середнього арифметичного зменшується. Це означає, що із збільшенням числа вимірів їх середнє арифметичне буде все менше відрізнятись від істинного значення, що підвищує надійність результату.

Наприклад, якщо стандартне відхилення одного виміру становить 6 одиниць, а всього виконано 36 вимірів, то стандартне відхилення середнього арифметичного цих вимірів дорівнює тільки 1, дійсно:

$$\sigma\left(\bar{X}\right) = \sigma_0 / \sqrt{n} = \frac{6}{\sqrt{36}} = 1.$$

Таким чином, середнє арифметичне кількох вимірів є набагато ближчим до істинного значення, ніж результат окремо взятого виміру.

2.17. Початкові і центральні моменти

Розглянемо дискретну випадкову величину X, задану таким табличним законом розподілу:

X	1	2	5	100
P	0,6	0,2	0,19	0,01

Знайдемо для неї математичне сподівання:

$$E[X] = 1 \cdot 0{,}6 + 2 \cdot 0{,}2 + 5 \cdot 0{,}19 + 100 \cdot 0{,}01 = 2{,}95.$$

Запишемо тепер закон розподілу для X^2:

X^2	1	4	25	10000
P	0,6	0,2	0,19	0,01

і знайдемо математичне сподівання для X^2:

$$E[X^2] = 1 \cdot 0{,}6 + 4 \cdot 0{,}2 + 25 \cdot 0{,}19 + 10000 \cdot 0{,}01 = 106{,}15.$$

Видно, що математичне сподівання $E[X^2] >> E[X]$. Це пояснюється тим, що після піднесення до квадрату значення 100 привело до появи значення 10000, хоча ймовірність цього значення досить незначна (0,01).

Таким чином, перехід від $E[X]$ до $E[X^2]$ дозволив краще врахувати вплив на математичне сподівання великого можливого значення, яке має незначну ймовірність. Отже, якби змінна X мала декілька великих малоймовірних значень, то перехід до X^2, а тим більше до X^3, X^4 і так далі, дав змогу ще більше "підсилити" роль великих, але малоймовірних значень. Саме з цієї точки зору корисно обчислювати математичне сподівання цілої додатної степені (дискретної і неперервної) випадкової величини.

Початковим моментом порядку p випадкової величини X називають математичне сподівання величини X^p:

$$m_p = E[X^p].$$

Зокрема,

$$m_1 = E[X],\ m_2 = E[X^2].$$

Користуючись визначенням моментів, запишемо вираз для обчислення дисперсії у вигляді:

$$\text{var}\,[X] = m_2 - m_1^2.$$

Крім моментів випадкової змінної X доцільно розраховувати моменти відхилення $X - E[X]$, тобто розглядати статистичні характеристики відхилень від середнього.

Центральним моментом порядку p випадкової змінної X називають математичне сподівання величини $(X - E[X])^p$:

$$\mu_p = E[(X - E[X])^p].$$

Зокрема,

$$\mu_1 = E[(X - E[X])] = 0;$$
$$\mu_2 = E[(X - E[X])^2] = \text{var}[X].$$

Можна легко отримати співвідношення, які зв'язують початкові і центральні моменти. Наприклад, порівнюючи два останні вирази, отримаємо:

$$\mu_2 = m_2 - m_1^2.$$

Якщо скористатись визначенням центрального моменту і властивостями математичного сподівання, то можна отримати наступні формули:

$$\mu_3 = m_3 - 3m_2 m_1 - 2m_1^3;$$
$$\mu_4 = m_4 - 4m_3 m_1 + 6m_2 m_1^2 - 3m_1^4.$$

Зазначимо, що моменти вищих порядків у теорії і практиці статистичного аналізу застосовують досить рідко.

Примітка. Моменти, розглянуті у даному параграфі, називають теоретичними. Моменти, які обчислюють на основі експериментальних спостережень, називають вибірковими або емпіричними.

2.18. Поняття групових і загальних статистичних характеристик

Групова і загальна середні

Нехай всі значення деякої кількісної змінної X (немає значення — генеральної сукупності чи вибіркової) розділені на кілька груп. (Саме такі ситуації часто зустрічаються на практиці.) Розглядаючи кожну групу як самостійну сукупність, можна знайти її середнє арифметичне.

Груповою середньою називають середнє арифметичне значень змінної, що належать одній групі.

Тепер введемо спеціальний термін для середнього всієї сукупності даних.

Загальним середнім $\bar{x}$ називають середнє арифметичне значень змінної, які належать всій сукупності.

Якщо відомі групові середні і об'єми груп, то можна знайти загальне середнє: *загальне середнє дорівнює середньому арифметичному групових середніх, зваженому за об'ємами груп.*

Приклад 2.8. Знайти загальне середнє сукупності, яка складається з двох наступних груп:

Група	перша		друга	
Значення змінної	1	6	1	5
Частота	10	15	20	30
Об'єм	10 + 15 = 25		20 + 30 = 50	

Розв'язок. Обчислимо групові середні:

$$\bar{x}_1 = (10 \cdot 1 + 15 \cdot 6) / 25 = 4;$$

$$\bar{x}_2 = (20 \cdot 1 + 30 \cdot 5) / 50 = 3,4.$$

Загальне середнє на основі групових середніх:

$$\bar{x} = (25 \cdot 4 + 50 \cdot 3,4) / (25 + 50) = 3,6.$$

Примітка. Для спрощення розрахунку загального середнього сукупності великого об'єму доцільно розбити її на кілька груп, знайти групові середні і за ними — загальне середнє.

Відхилення від загального середнього та його властивості

Розглянемо сукупність значень деякої кількісної змінної X (немає значення — генеральної сукупності чи вибіркової) об'єму N.

Значення змінної	x_1	x_2	...	x_l
Частота	n_1	n_2	...	n_l

При цьому $\sum_{i=1}^{l} n_i = N$. Знайдемо загальне середнє:

$$\bar{x} = \frac{\sum_{i=1}^{l} n_i x_i}{N},$$

а звідси маємо:

$$\sum_{i=1}^{l} n_i x_i = N \bar{x}. \qquad (*)$$

Зазначимо, що оскільки $\bar{x}$ — постійна величина, то

$$\sum_{i=1}^{l} n_i \bar{x} = \bar{x} \sum_{i=1}^{l} n_i = \bar{x} N. \qquad (**)$$

Відхиленням називають різницю $x_i - \overline{x}$ між значенням змінної та загальним середнім.

Теорема 2.11. *Сума добутків відхилень на відповідні частоти дорівнює нулю*:

$$\sum_{i=1}^{l} n_i \left(x_i - \overline{x} \right) = 0.$$

Доведення. Враховуючи (*) і (**), отримаємо:

$$\sum_{i=1}^{l} n_i \left(x_i - \overline{x} \right) = \sum_{i=1}^{l} n_i x_i - \sum_{i=1}^{l} n_i \overline{x} = \overline{x}N - \overline{x}N = 0.$$

Наслідок. *Середнє значення відхилення дорівнює нулю*. Дійсно,

$$\frac{\sum_{i=1}^{l} n_i \left(x_i - \overline{x} \right)}{\sum_{i=1}^{l} n_i} = \frac{0}{N} = 0.$$

2.19. Контрольні питання і вправи

1. Дайте означення моди і медіани.
2. Які дві властивості має середнє арифметичне?
3. Як впливають зміни (збільшення, зменшення, множення і ділення) значень деякого ряду розподілу на середнє?
4. Запишіть формулу для знаходження вибіркового середнього. Наведіть приклади використання середнього.
5. Виведіть рекурсивну формулу для знаходження поточного середнього.
6. Чому дорівнює математичне сподівання добутку двох незалежних випадкових змінних? Доведіть результат.
7. Чому дорівнює математичне сподівання суми двох випадкових змінних? Доведіть результат.
8. Запишіть формули для знаходження середнього квадратичного і середнього кубічного, в яких випадках їх доцільно застосовувати?
9. Наведіть вираз для знаходження середнього геометричного і приклад його практичного використання.
10. В якому випадку необхідно використовувати середнє гармонійне?
11. Які є виміри варіації елементів вибірки?

12. Запишіть формулу для обчислення вибіркової дисперсії.
13. Яким чином забезпечується незміщеність оцінок параметрів, що знаходяться за статистичними даними?
14. Запишіть і поясніть вираз для знаходження стандартного відхилення середніх значень випадкових вибірок однакової потужності N.
15. Який факт називають незалежним? Поясніть на прикладі.
16. Поясніть термін "кількість ступенів вільності"?
17. Які властивості має дисперсія? Доведіть, що
$$\mathrm{var}[X + Y] = \mathrm{var}[X] + \mathrm{var}[Y].$$
18. Дайте означення стандартного відхилення.
19. Як впливають зміни значень елементів вибірки на дисперсію?
20. Які процеси називають гетероскедастичними? Наведіть приклад гетероскедастичного процесу.
21. Чи можна використовувати дисперсію як формальний вимір інформативності? Поясніть на прикладі.
22. Яке застосування має дисперсія у фінансах?
23. Яким чином пов'язана дисперсія середнього арифметичного n однаково розподілених взаємно незалежних випадкових змінних з дисперсією кожної змінної?
24. Дайте означення початкового моменту порядку p випадкової змінної X.
25. Дайте означення центрального моменту порядку p випадкової змінної X.

Розділ 3

ОПИС ПОЛОЖЕННЯ ОКРЕМОГО СПОСТЕРЕЖЕННЯ В РЯДУ РОЗПОДІЛУ, ГРУПУВАННЯ ДАНИХ

3.1. Процентильні ранги і процентилі

Необхідність описати положення окремого спостереження в ряду розподілу зустрічається досить часто. Наприклад, необхідно встановити, яке місце за успішністю займає один із студентів у своїй групі. У даному випадку елементами розподілу є числові значення оцінок студентів конкретної групи. Можна сказати, наприклад, що студент *A* отримав другу за величиною оцінку в групі. Однак така інформація не дає можливості визначити наскільки успішно навчається студент *A*, тому що в групі могло бути 5 студентів і 35. Для того щоб встановити, яке місце займає в групі окремий студент, необхідно знати число студентів у цій групі.

Місцеположення деякого спостереження у розподілі визначається шляхом присвоєння йому значення, яке називають *процентильним рангом*.

Наприклад, нехай у групі всього 20 студентів і студент *A* отримав вищу оцінку, ніж п'ятнадцятеро його товаришів по групі, а інші четверо студентів отримали оцінки вищі, ніж *A*. Таким чином, 75 % оцінок (15 з 20) нижчі, ніж в *A*. Сама оцінка студента *A* складає 5 % від загального числа оцінок, а 20 % оцінок є вищими, ніж у *A*. Ці процентні співвідношення показані на рис. 3.1.

75 %	5 %	20 %

Рис. 3.1. **Процентні співвідношення оцінок у групі**

Місце студента *A* в групі вкажемо наступним чином: додамо до 75 % (оцінки, які є меншими, ніж у студента *A*) ще 2,5 %, тобто половину від процентного значення оцінки *A*. Для цього можна умовно провести вертикальну лінію через центр області, що відноситься до

оцінки *А*. У результаті знайдемо, що процентильний ранг студента *А* дорівнює 77,5 %:

$$\text{Пр} = 75 + 2{,}5 = 77{,}5\ \%.$$

Означення 3.1. *Процентильним рангом деякого спостереження ряду розподілу називається сума відсотків, що відносяться до спостережень, які стоять у розподілі перед ним, і половини відсотків, які відносяться безпосередньо до нього.*

Нехай в іншій групі, яка налічує 50 студентів, студент *Б* отримав оцінку, яка є вищою, ніж у 17 його товаришів, тобто 34 % оцінок у розподілі є нижчими від оцінки *Б*. Оскільки на одного студента припадає 2 %, то процентильний ранг *Б* становить:

$$\text{Пр} = 34 + 1 = 35\ \%.$$

Можна сказати, що студент *А* навчається у своїй групі краще, ніж студент *Б* у своїй, оскільки його процентильний ранг 77,5 % набагато перевищує 35 %.

Процентильні ранги дають можливість порівнювати успішність студентів, незважаючи на те, що строгість викладачів та їх методи оцінювання можуть бути різними. Таким чином, *процентильні ранги дають можливість співставляти між собою елементи різних розподілів.*

Показник студента *А* є вищим, тому що він перегнав більшу, ніж студент *Б,* частину студентів у своїй групі. При цьому припускається, що рівень конкуренції в обох групах однаковий, хоча на практиці це припущення часто не виконується. Зазначимо, що при порівнянні студентів *А* і *Б* ми не користувалися самими числовими значеннями членів рядів розподілу, тобто дійсними значеннями оцінок, отриманих студентами.

Процентильний ранг елемента розподілу часто виявляється змістовнішим, ніж власне значення цього елемента.

Наприклад, нехай випускник КПІ претендує на місце в аналітичному відділі банку. В результаті проходження спеціального тестування для виявлення здібностей до роботи в аналітичному відділі встановлено, що його коефіцієнт розумового розвитку становить 115 балів (КРР = *IQ* = *Intelligence Quotient*). Це велике значення, але воно ще не є підставою для позитивного рішення щодо прийняття випускника на роботу; для цього потрібна додаткова інформація. Якщо буде надана додаткова інформація, що випускник має 94-й процентиль по успішності серед студентів, які вивчали банківську справу, то, судячи

з результату тестування, він має перспективу влаштуватись на роботу в банк.

Слово *процентиль* відноситься безпосередньо до елемента розподілу, тобто до значення, яке знаходиться між двома елементами. Елемент розподілу з процентильним рангом 35 називають тридцять п'ятим процентилем, а елемент з процентильним рангом 77 — сімдесят сьомим процентилем.

Означення 3.2. *Елемент розподілу з певним процентильним рангом називається відповідним процентилем.*

3.2. Нормовані відхилення (*z*-оцінки)

Процентильні ранги не дають відповіді на деякі питання. Наприклад, припустимо, що студенти *А* і *Б* є кращими в своїх групах і мають однакові процентильні ранги. При цьому один з них може значно переважати своїх конкурентів, а перевага іншого в своїй групі є незначною. Оскільки процентильні ранги обох студентів однакові, то виявити між ними різницю неможливо. Для того щоб її виявити, необхідно ввести новий спосіб визначення положення елемента в ряду розподілу, який буде показувати, наскільки далеко знаходиться даний елемент від точки рівноваги (середнього) і по який бік від нього.

Розглянемо розподіл з шести елементів (оцінки студентів): 4, 6, 7, 10, 15, 18. Середнє дорівнює:

$$\mu_x = (4 + 6 + 7 + 10 + 15 + 18)/6 = 10.$$

Знайдемо відхилення для елементів ряду і стандартне відхилення (табл. 3.1).

Таблиця 3.1

Відхилення оцінок від середнього для 1-ї групи

№ пор.	Оцінка	Середнє	Відхилення
1	4	10	–6
2	6	10	–4
3	7	10	–3
4	10	10	0
5	15	10	5
6	18	10	8
Стандартне відхилення = 5			

Відхилення оцінки кожного студента вказує на місце його оцінки в розподілі для його групи, але відхиленням не можна скористатись для порівняння студентів двох різних груп. Причина полягає в тому, що в іншій групі оцінки можуть ставитись по іншому принципу. Таким чином, відхилення від середнього на кілька балів може мати різний смисл у різних групах. Розглянемо для прикладу оцінки, отримані в іншій групі (табл. 3.2).

У другій групі оцінки розсіяні більше, ніж в першій, про що свідчать відхилення оцінок.

Таблиця 3.2

Відхилення оцінок від середнього для 2-ї групи

№ пор.	Оцінка	Середнє	Відхилення
1	20	50	–30
2	30	50	–20
3	35	50	–15
4	50	50	0
5	75	50	25
6	90	50	40
Стандартне відхилення = 25			

Порівнюючи табл. 3.1 і 3.2, можна зробити висновок, що оцінки у другій групі є набагато вищими. Так, краща оцінка у першій групі, яка дорівнює 18, у другій групі буде меншою найнижчої. Відповідні відхилення оцінок від середнього також є набагато більшими у другій групі.

Для порівняння успішності студентів цих груп *не можна використовувати відхилення, тому що вони мають різну "вагу" для різних груп*. Однак, для цієї мети можна скористатись двома параметрами — *відхиленням* і *стандартним відхиленням*.

Означення 3.3. *Відношення відхилення від середнього до стандартного відхилення називають z-оцінкою.*

Вона розраховується за формулою:

$$z_k = \frac{x_k - \mu_k}{\sigma_x}, \quad k = 1, \ldots, N.$$

Так, *z*-оцінка 6-го студента першої групи: $z_{16} = 8/5 = 1{,}6$, а *z*-оцінка 5-го студента другої групи: $z_{25} = 25/25 = 1$. Оскільки *z*-оцінка 6-го студента першої групи є вищою, ніж 5-го студента другої групи, то можна сказати, що перший з них займає вище місце у своїй групі, ніж другий у своїй.

Таким чином, z-оцінка показує, на скільки одиниць стандартних відхилень конкретний елемент є більшим або меншим середнього його ряду розподілу.

Наприклад, якщо *z*-оцінка елемента дорівнює 1,5, то він перевищує середнє на 1,5 одиниці стандартного відхилення. Елемент, який має *z*-оцінку — 0,5, є на 5/10 одиниці стандартного відхилення меншим середнього його розподілу. Таким чином, *z*-оцінки можна використовувати для порівняння елементів різних розподілів. На відміну від значень відхилень *z*-оцінки не залежать від "цінності" балів, що виставляються студентам. Причина цього полягає у тому, що *z*-оцінка характеризує відносне місцеположення елемента із врахуванням ступеня коливання всього ряду розподілу.

3.3. Середні і *z*-оцінки

У випадках, коли в різних розподілах статистичних даних використовують різні одиниці виміру, для знаходження їх середніх є доцільними *z*-оцінки. Нехай група студентів отримує два завдання: *А* і *Б*. Максимальне число балів за виконання завдання *А* становить 100, а за завдання *Б* — 10. Нехай студент *С*1 отримує за завдання *А* 76 балів, а за завдання *Б* — 8; студент *С*2 отримує 83 бали за завдання *А* і 1 бал за завдання *Б*, тобто не виконує його. *С*1 отримує вищу оцінку, тому що він виконав два завдання, а *С*2 — тільки одне.

Формально можна підійти до проблеми оцінювання виконаних завдань студентами наступним чином. Середня оцінка *С*2 за два завдання становить: (83 + 1) /2 = 42, а середня оцінка *С*1 становить: (76 + 8)/2 = 42, тобто формальні середні арифметичні однакові. Однак, чи буде це правильна оцінка?

Інтуїтивне відчуття того, що *С*1 заслуговує вищої оцінки не підтверджується середнім арифметичним отриманих балів. Фактично, усереднювати в даному випадку не можна, тому що "вага" або "ціна" одного бала в одному тесті була вищою, ніж в іншому. *С*2 випередив *С*1 у тесті *А* на сім балів, але *С*1 набрав на сім балів більше у тесті *Б*, в якому кожний бал ціниться набагато вище (за рахунок того, що максимальна оцінка становить 10).

Якщо ж перетворити оцінки кожного студента в *z*-оцінки, то можна отримати рівноцінні одиниці виміру, незважаючи на те, що початкові одиниці виміру оцінок можуть бути різними. Такий підхід можливий завдяки тому, що *z*-оцінки мають однаковий смисл для будь-

якого ряду розподілу; фактично, це величини, які не мають одиниць виміру.

У загальному випадку, коли елементи різних розподілів оцінюються за допомогою різних одиниць виміру, застосування звичайних середніх втрачає сенс.

3.4. Порівняння *z*-оцінок і процентилів

Процентильні ранги і *z*-оцінки безпосередньо не зв'язані між собою і не можна отримати значення іншого, якщо відомий один з них. Теоретично, елемент ряду розподілу, який має визначений процентильний ранг, може мати будь-яке значення *z*-оцінки. Розглянемо приклад, наведений у табл. 3.3. Розподіл складається з елементів: 2, 3, 6, 8, 11 із середнім 6.

Таблиця 3.3

Порівняння *z*-оцінок з процентильними рангами

Елемент	*z*-оцінка	Процентильний ранг
2	–1,21	10
3	–0,91	30
6	0,00	50
8	0,61	70
11	1,52	90

Два елементи цього ряду (2 і 3) розташовані зліва від середнього, а тому вони мають від'ємні *z*-оцінки. Два інших (8 і 11) розташовані справа від середнього, а тому мають додатні *z*-оцінки. Один елемент (6) дорівнює середньому, а тому його *z*-оцінка дорівнює нулю.

Елемент 8 має процентильний ранг 70, а *z*-оцінка цього елемента дорівнює 0,61. Додатна *z*-оцінка даного елемента свідчить про те, що його значення перевищує середнє. Однак, в іншому випадку може мати місце ситуація, коли елемент з процентильним рангом 70 буде знаходитись нижче середнього і мати від'ємну *z*-оцінку (табл. 3.4, середнє дорівнює 4).

Елемент 3 має процентильний ранг 70, оскільки він перевищує 70 % елементів розподілу, але його *z*-оцінка від'ємна, тому що він є меншим середнього. Існують ще більше "незбалансовані" ряди, в яких елемент може мати процентильний ранг 90, але все-таки бути меншим середнього.

Таблиця 3.4

Порівняння *z*-оцінок з процентильними рангами

Елемент	*z*-оцінка	Процентильний ранг
0	–0,78	10
1	–0,59	30
2	–0,39	50
3	–0,20	70
14	1,96	90

Висновок. *Не існує чітких правил для визначення залежності між z-оцінкою і процентильним рангом. Кожний з цих показників є окремою характеристикою ряду. Процентильний ранг вказує на те, скільки елементів ряду розташовано нижче даного спостереження, але не дає інформації щодо його зв'язку із середнім. У свою чергу, z-оцінка вказує на те, як розташований даний елемент по відношенню до середнього значення ряду, але не містить інформації щодо проценту елементів ряду, розташованих нижче або вище нього.*

Примітка. У більшості рядів розподілу, які мають місце на практиці, елемент з процентильним рангом 70–80 буде мати додатну *z*-оцінку. Аналогічно, якщо для елемента x_i *z*-оцінка $z_i \geq 1$, то він, як правило, буде більшим щонайменше половини інших елементів.

3.5. Середнє і стандартне відхилення ряду, утвореного із *z*-оцінок

Теорема 3.1. *Середнє ряду розподілу z-оцінок завжди дорівнює нулю, а стандартне відхилення — одиниці, тобто* $\mu_z = 0$, $\sigma_z = 1$.

Доведення:

$$\mu_z = E\left[\frac{x_k - \mu_x}{\sigma_x}\right] = \frac{E[x_k] - \mu_x}{\sigma_x} = \frac{\mu_x - \mu_x}{\sigma_x} = 0;$$

$$\operatorname{var}[z_k] = E\left[\left(\frac{x_k - \mu_x}{\sigma_x}\right)^2\right] = \frac{E\left[(x_k - \mu_x)^2\right]}{\sigma_x^2} = \frac{\sigma_x^2}{\sigma_x^2} = 1.$$

Приклад. Розглянемо ряд розподілу X = [4 7 8 10 11 20] із середнім μ_x = 10 і стандартним відхиленням σ_x = 5. Розрахункові значення *z*-оцінок та їх характеристики наведені у табл. 3.5.

Таблиця 3.5

z-оцінки та їх характеристики

Елемент	Відхилення	z-оцінка
4	–6,0	–1,2
7	–3,0	–0,6
8	–2,0	–0,4
10	0	0,0
11	1,0	0,2
20	10,0	2,0
$\mu_x = 10, \sigma_x = 5$		$\mu_z = 0, \sigma_z = 1$

Зазначимо, що *отриманий результат не залежить від випадкових характеристик ряду розподілу*. Так, середнє відхилень завжди дорівнює нулю, а їх стандартне відхилення — стандартному відхиленню вихідного ряду розподілу. Очевидно, що середнє ряду, отриманого в результаті ділення відхилень на стандартне відхилення також буде нульовим. Оскільки всі відхилення від середнього ряду розподілу діляться на їх власне стандартне відхилення, то і стандартне відхилення середнього ряду буде поділеним само на себе. У результаті стандартне відхилення z-оцінок буде дорівнювати 1.

3.6. Стандартні або *T*-оцінки

Оскільки z-оцінки містять десяткові знаки і можуть бути додатними або від'ємними, то це не зовсім зручно при аналізі цих оцінок. Часто буває зручніше замінити z-оцінки додатними числами. Такі оцінки називають стандартними або T-оцінками, а розраховують їх за формулою:

$$T_k = 50 + 10z_k = 50 + 10 \cdot \frac{x_k - \mu_x}{\sigma_x}.$$

Тобто, z-оцінки множать на 10 і округляють до цілого, а додавання числа 50 перетворює всі значення z-оцінок у додатні числа. Теоретично T-оцінка може бути від'ємною, але таких z-оцінок практично немає. Так, $T_k = 0$ при $z_k = -5{,}0$, а від'ємними стандартні оцінки будуть при $z_k < -5{,}0$.

T-оцінки можна застосовувати для розв'язування таких самих задач, що і z-оцінки. Наприклад, за їх допомогою можна порівняти

успішність студентів різних груп. Очевидно, що середнє T-оцінок $\mu_T = 50$, а стандартне відхилення $\sigma_T = 10$.

3.7. Застосування процентильних рангів і стандартних оцінок

1. *Процентилі і процентильні ранги використовують як характеристики рівня освіти, для визначення різниці між індивідуумами в психології* та інших областях, де застосовують бали та оцінки. Наприклад, приймальним комісіям університетів було б доцільно розглядати спочатку процентильні ранги абітурієнтів, а вже потім числові значення оцінок. Процентильний ранг вказує на місце, яке займає учень у своєму класі, а оцінка вказує на рівень успішності та на достовірність визначення оцінок у школі.

2. Економісти розраховують процентильні ранги рівнів *доходу на жителя країни* з метою порівняння з іншими країнами.

Можна порівнювати також процентильні ранги рівня *промислового виробництва* різних міст у масштабах однієї країни.

3. Зручніше z-оцінки трансформувати в стандартні T-оцінки, оскільки вони містять суттєву інформацію. Так, якщо студент отримав стандартну оцінку за тест $T = 60$, то (на основі виразу для її знаходження) можна зробити висновок, що його оцінка є на одне стандартне відхилення вищою середньої оцінки групи.

3.8. Часові ряди і часові перерізи даних

Усі статистичні дані можна розділити на два великих класи: часові ряди і часові перерізи. Обидва типи даних зустрічаються дуже часто на практиці.

Означення 3.4. *Часовий ряд — це послідовність значень випадкової величини, яка отримана за деяким визначеним часовим інтервалом* (періодом дискретизації T_s).

Період дискретизації найчастіше постійний, але в окремих випадках він може бути змінним, тобто, $T_s \neq \text{const}$. Він може бути змінним у випадках, коли неможливо виконати збір даних з постійним періодом, або у даних є пропуски. Крім того, період збору даних може бути взагалі *нерегулярним*, тобто значення T_s носить не прогнозований характер. У таких випадках кажуть, що *дані носять нерегуляр-*

ний характер, і для того, щоб робити статистичний висновок на основі таких даних, необхідно застосовувати спеціальні методи.

Прикладами часових рядів є статистичні дані, які характеризують зміну в часі економічних процесів, виміри ознак, що характеризують поведінку технічної системи (швидкість, прискорення, амплітуди коливань) або технологічного процесу (температура протікання реакції, концентрація розчину, рівень рідини в ємкості і т. ін.).

Означення 3.5. *Часові перерізи — це дані, що характеризують вибраний процес у деякий конкретний момент або відрізок часу* (цим відрізком може бути день, тиждень, місяць чи навіть рік).

Наприклад, може бути цікавим дослідження успішності студентів на кінець семестру, тобто на кінець січня або червня. У такому випадку дані щодо успішності збирають протягом кількох днів у кінці визначеного місяця і вони характеризують вибрану групу студентів саме на кінець конкретного періоду часу. Якщо визначається рейтинг того чи іншого кандидата в президенти, то дані також збирають протягом кількох днів. Результати опитування визначають як опитування на конкретну дату. Перепис населення виконують, наприклад, протягом місяця і отриманий результат перепису вважається дійсним на кінець конкретного місяця і року.

3.9. Дискретні і неперервні величини

У загальному випадку ознаки можуть приймати різні за своєю природою та величиною значення. Ознака "вік" за своєю природою є неперервною, оскільки час є неперервною категорією, але ми користуємось дискретними значеннями цієї величини. Коли досліджують зміни функціонування довгострокової чи короткострокової пам'яті людини, то вибирають індивідуумів конкретного віку, наприклад, 10, 20, 30, 40, 50, 60 і 70 років. Можна вважати, що в даному випадку період дискретизації експериментальних даних буде становити 10 років (T_s = 10 років).

У технічних системах період дискретизації може складати дуже короткі проміжки часу, наприклад, десятки мікросекунд (мкс) або мілісекунд (мс). Так, у системах автоматичного керування двигунами внутрішнього згоряння він складає близько 20 мс. У соціально-економічних системах процеси мають, в основному, неперервний характер, але дані завжди збирають дискретно. Інтервал між будь-

якими двома сусідніми числами включає всі значення, які знаходяться між ними. Однак є змінні, які приймають тільки дискретні значення. Наприклад, число деталей, які виробляє робітник за одиницю часу.

Означення 3.6. *Якщо всередині деякого інтервалу ознака теоретично може приймати тільки деякі, визначені тим чи іншим способом, значення, то така ознака на даному інтервалі змінюється дискретно.*

Наприклад, число голосів, поданих за того чи іншого кандидата в президенти. Сімейний стан може визначатись чотирма якісними змінними: *одинокий, жонатий, розведений, вдівець*. Змінна, яка відображає відношення до військової служби: *військовозобов'язаний* і *не військовозобов'язаний*.

Означення 3.7. *Якщо всередині деякого інтервалу ознака теоретично може приймати будь-яке значення, то ця величина змінюється в даному інтервалі неперервно.*

Ознака "вік" теоретично є неперервною, оскільки може приймати будь-яке значення в інтервалі від нуля до нескінченності. Для людини ця нескінченність обмежена значенням приблизно 150 років, а для деяких дерев це може бути кілька тисяч років. Вік Сонячної системи обчислюється мільярдами років.

На практиці ми користуємось тільки деякими обмеженими значеннями з інтервалу визначення неперервної ознаки і округлюємо фактичні значення до зручних для користування. Наприклад, *вік людини* округлюють до *років*. Здібності людини також є неперервною змінною, але коефіцієнт розумового розвитку завжди округлюють до цілого числа.

У технічних та багатьох інших системах ми можемо покращати точність вимірювань завдяки використанню нових точних приладів.

3.10. Табулювання і графічне зображення дискретних величин

Ряд значень дискретної змінної зручно представляти у вигляді *таблиці частот*. Що означає термін "частота"?

Означення 3.8. *Частотою будь-якого значення, що може приймати ознака в ряду розподілу, називається число, яке показує, скільки разів дане значення зустрічається у вибірці.*

У табл. 3.6 наведено приклад значень частот для ряду розподілу сімейного стану жінок на невеликій фабриці.

Таблиця 3.6

Ряд розподілу сімейного стану жінок

Сімейний стан	Частота
Одинокі	15
Заміжні	25
Розведені	20
Вдови	5

Зображення ряду розподілу можна спростити, якщо скористатися *стовпчиковою діаграмою*.

Стовпчикова діаграма — один з найбільш зручних та популярних видів графіків, які використовують для графічного представлення ряду розподілу. Зазначимо, що між окремими стовпчиками є проміжки. На рис. 3.2 наведено стовпчикову діаграму для ряду з табл. 3.6.

3.11. Округлювання значень

Оскільки вимірювання значень ознак неможливо здійснити з абсолютною точністю, то їх округлюють. Це дає можливість суттєво спростити необхідні розрахунки.

Округлення означає, що ознаці приписується значення найближчої, вибраної нами точки.

Так, всю область визначення змінної можна розбити на інтервали і в подальшому аналізі, значенню, яке попадає в конкретний інтервал, приписувати значення середини цього інтервалу. Таким чином,

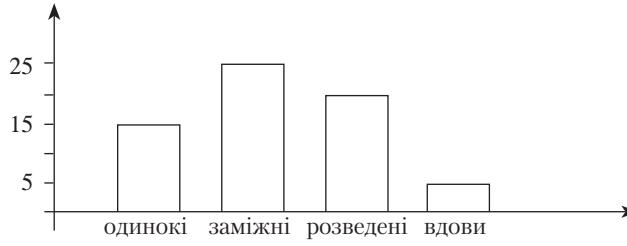


Рис. 3.2. **Стовпчикова діаграма для ряду розподілу сімейного стану жінок**

при округлюванні змінні приймають значення з обмеженого (скінченого числа значень). Розглянемо, наприклад, ряд значень зросту людини (рис. 3.3).

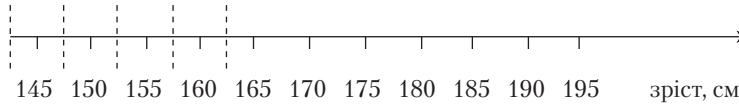


Рис. 3.3. **Інтервали та їх середні для значень ознаки "зріст"**

Так, всім величинам, які попадають в інтервал 152,5÷157,5 (верхня границя не включається в інтервал), приписується значення 155. Тим, які попадають в інтервал 157,5÷162,5, приписують 160 і т. ін.

Таким чином, округлювання передбачає групування значень в інтервали і присвоєння згрупованим даним значень середини інтервалу.

Техніку роботи з округленими значеннями змінної називають *методом групування*.

3.12. Групування даних і побудова групової таблиці частот

Групування і табулювання сукупності спостережень складається з чотирьох наступних процедур.

1. Розрахувати *варіаційний розмах*, який дорівнює різниці між найбільшим і найменшим значеннями ряду спостережень. Наприклад, якщо в групі студентів максимальна оцінка дорівнює 100, а мінімальна 3, то варіаційний розмах дорівнює: 100 – 3 = 97.

2. Визначити кількість *інтервалів* та *їх довжину*. Зручніше довжину інтервалу вибирати рівною непарному числу для того, щоб середина кожного інтервалу дорівнювала цілому значенню.

Так, у прикладі з оцінками групи студентів зручно взяти інтервали довжиною 9 балів. Всього буде 11 інтервалів, які охоплюють 99 балів. Важливо, щоб сумарна довжина всіх інтервалів дорівнювала, щонайменше, варіаційному розмаху. Таким чином, в інтервалі 3÷100 знаходиться 98 цілих чисел, тобто на одне число більше від варіаційного розмаху.

3. Визначити *граничне значення самого верхнього інтервалу*. В нашому прикладі верхній інтервал охоплює значення 92÷100. Визначаючи крайні точки інтервалу, ми включаємо їх у даний інтервал. Тоб-

то ширина інтервалу 92÷100 складає 9 одиниць. Наступний інтервал буде охоплювати значення 83÷91. Продовжуючи процедуру визначення інтервалів, встановимо, що самий нижній інтервал включає значення 2÷10 (табл. 3.7).

Таблиця 3.7

Приклад інтервального варіаційного ряду розподілу

Інтервал	Центр
92–100	96
83–91	87
74–82	78
65–73	69
56–64	60
47–55	51
38–46	42
29–37	33
20–28	24
11–19	15
2–10	6

4. *Табулювання частот інтервалів,* тобто визначення числа значень, які відносяться до кожного інтервалу. Наприклад, таблиця частот для ряду розподілу оцінок, розглянутих вище, може мати вигляд, наведений у табл. 3.8.

Таблиця 3.8

Групова таблиця частот ряду розподілу

Інтервал	Центр	Частота
1	2	3
92–100	96	60
83–91	87	140
74–82	78	160
65–73	69	120
56–64	60	140
47–55	51	80
38–46	42	119

Закінчення табл. 3.8

1	2	3
29–37	33	81
20–28	24	50
11–19	15	32
2–10	6	18
	Всього = 1000	

З табл. 3.8 видно, що 60 студентів мають оцінки в інтервалі 92÷100 (включаючи оцінки 92 і 100), а оцінку в інтервалі 83÷91 (включаючи 83) мають 140 студентів. Тепер ми можемо оперувати з таким рядом, що має 60 елементів із значенням 96, 140 елементів із значенням 87 і т. д. Фактично, ми округлили значення змінної до найближчого, наперед визначеного числа.

3.13. Гістограма і полігон ряду розподілу

Далі розглянемо метод графічного зображення неперервного ряду розподілу на основі інформації, яка міститься в груповій таблиці частот. Повернемося до табл. 3.8. Оцінку за тест можна розглядати як неперервну змінну, значення якої можуть відрізнятися. Теоретично, оцінка може приймати значення 91,6 або 92,4 і т. ін.

Фактично, оцінки, які перевищують 91,5, відображають як 92 і вище. Отже, нижньою границею верхнього інтервалу буде 91,5. Аналогічно 91,5 є і верхньою границею інтервалу 83÷91, а нижньою границею цього інтервалу є значення 82,5. Звідси випливає, що показники оцінок, які були вміщені в інтервал 83÷91, при більш точному вимірюванні можуть потрапити в інтервал 82,5÷91,5.

За аналогією можна визначити границі всіх інтервалів, представлених у табл. 3.8. Значення цих границь наведено в табл. 3.9.

Ряд розподілу оцінок за тест можна представити тепер у графічній формі (див. рис. 3.4).

Гістограма схожа на стовпчикову діаграму, але на стовпчиковій діаграмі прямокутники розташовані окремо один від одного, що вказує на дискретний характер розподілу. Прямокутники на гістограмі стоять впритул, що свідчить про те, що ряд розподілу є неперервним.

Ще одним способом графічного представлення неперервного ряду є *частотний полігон*. Частоти інтервалів позначають точками, розташованими над центрами інтервалів, які з'єднують між собою прямими лініями (рис. 3.5).

Таблиця 3.9

Групова таблиця частот

Інтервал	Границя	Центр	Частота
92–100	91,5–100,5	96	60
83–91	82,5–91,5	87	140
74–82	73,5–82,5	78	160
65–73	64,5–73,5	69	120
56–64	55,5–64,5	60	140
47–55	46,5–55,5	51	80
38–46	37,5–46,5	42	119
29–37	28,5–37,5	33	81
20–28	19,5–28,5	24	50
11–19	10,5–19,5	15	32
2–10	1,5–10,5	6	18
		Всього = 1000	

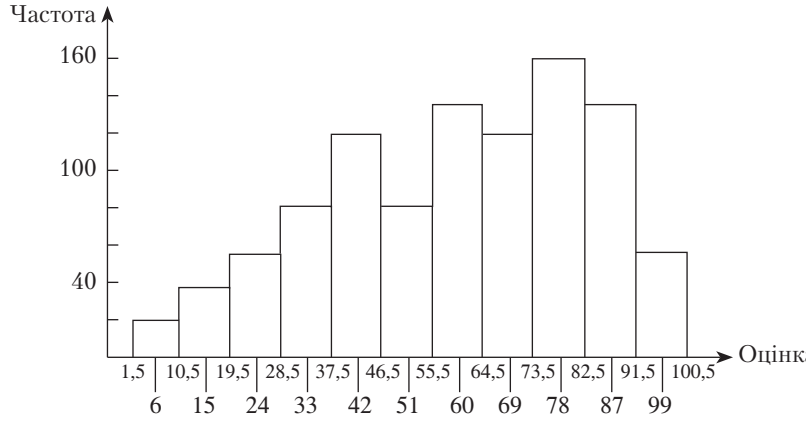


Рис. 3.4. **Гістограма ряду розподілу оцінок за тест**

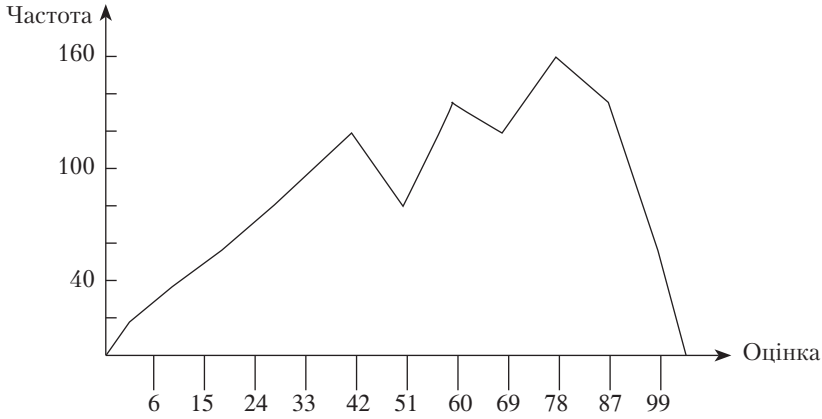


Рис. 3.5. **Частотний полігон ряду розподілу оцінок за тест**

Якщо при побудові гістограми зменшувати довжину інтервалів, то число прямокутників збільшуватиметься, а зміна висоти прямокутників, що стоять рядом, буде більш рівномірною. Гістограма нагадуватиме криву лінію. Подальше зменшення довжини інтервалу веде до того, що різниця між площею гістограми і площею фігури, до якої прямує графік, стане довільно малою. Таким чином, *можна перейти до неперервного ряду розподілу*. Однак, для реалізації такої процедури необхідно мати досить велике число спостережень.

3.14. Контрольні питання і вправи

1. Дайте означення процентильного рангу, наведіть приклад. Поясніть термін “процентиль”.
2. Які значення називають z-оцінками? Порівняйте середнє і z-оцінку деякого ряду розподілу.
3. Які статистичні характеристики (середнє і дисперсія) має ряд, утворений з z-оцінок? Доведіть результат.
4. Яким чином знаходяться T-оцінки? Чим вони зручніші у використанні?
5. Наведіть приклади застосування процентильних рангів.
6. Наведіть приклади застосування T-оцінок.
7. Дайте означення часового ряду і часового перерізу статистичних даних.

8. Які ознаки відносять до дискретних, наведіть приклади. Чи можна віднести до дискретних ознак квартальні значення валового внутрішнього продукту?
9. Які ознаки називають неперервними? Наведіть приклади. Ознака *вік людини* — це дискретна, чи неперервна ознака?
10. Як розраховується варіаційний розмах ряду спостережень. Наведіть приклад.
11. З яких кроків складається процедура групування і табулювання сукупності спостережень?
12. Поясніть на прикладі побудову гістограми.
13. Яким чином будується частотний полігон ряду розподілу?
14. Як будується стовпчикова діаграма? Наведіть приклади застосування стовпчикових діаграм.
15. Наведіть приклади застосування гістограм.
16. Побудуйте стовпчикову діаграму для успішності студентів у вашій групі.

Розділ 4

СТАТИСТИЧНІ ПАРАМЕТРИ ДЛЯ ЗГРУПОВАНИХ ДАНИХ

4.1. Визначення медіани і процентилів з гістограми

Як ми зазначали раніше, медіана — це значення ознаки, яке припадає на середину впорядкованого ряду розподілу і поділяє його на дві рівні за обсягом частини.

Медіана є 50-м процентилем. Медіана не обов'язково повинна співпадати з конкретним значенням ряду розподілу. Про це необхідно пам'ятати при визначенні медіани для ряду розподілу упорядкованих даних.

Розглянемо наведений приклад з оцінками за тест (табл. 4.1 та відповідна гістограма на рис. 4.1). Для того щоб медіану знайти, необхідно вказати на горизонтальній осі рис. 4.1 значення ознаки, яке перевищує кількість оцінок 500.

Таблиця 4.1

Групова таблиця частот

Інтервал	Границя	Центр	Частота
92–100	91,5–100,5	96	60
83–91	82,5–91,5	87	140
74–82	73,5–82,5	78	160
65–73	64,5–73,5	69	120
56–64	55,5–64,5	60	140
47–55	46,5–55,5	51	80
38–46	37,5–46,5	42	119
29–37	28,5–37,5	33	81
20–28	19,5–28,5	24	50
11–19	10,5–19,5	15	32
2–10	1,5–0,5	6	18
		Всього = 1000	

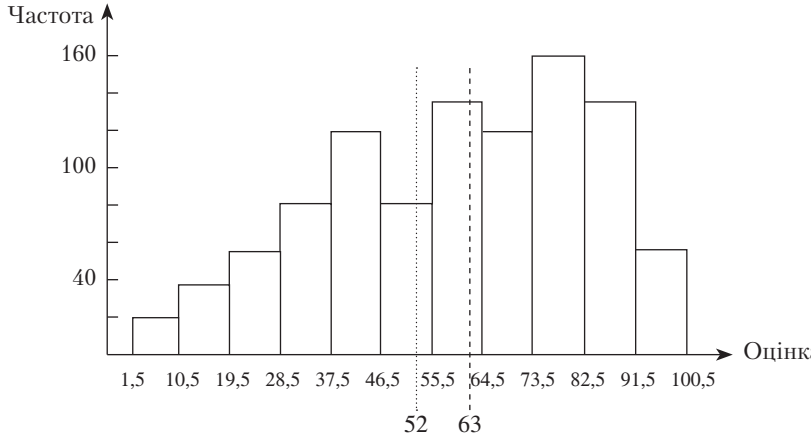


Рис. 4.1. **Гістограма ряду розподілу оцінок за тест**

Знайдемо суму частот для шести і семи нижніх інтервалів: 380 — для шести і 520 — для семи. Інакше кажучи, 380 оцінок розташовані нижче значення 55,5 і 520 оцінок розташовані нижче значення 64,5. Медіана повинна задовольняти нерівності $55{,}5 < Me < 64{,}5$, але все-таки бути ближчою до 64,5.

Площа прямокутника, побудованого над інтервалом 55,5÷64,5, відповідає 140 оцінкам, які були згруповані в даному інтервалі шляхом округлювання їх значень до 60. Для того щоб знайти медіану, необхідно визначити на горизонтальній осі точку, через яку можна провести вертикальну лінію, зліва від якої залишиться 120 із 140 оцінок даного інтервалу. У результаті 500 оцінок будуть знаходитися зліва від цієї лінії, а знайдена точка перетину з віссю вкаже на значення медіани. Таким чином, 120/140 або 6/7 площі прямокутника залишиться зліва, а 1/7 — справа. Проведемо, відповідно, вертикальну лінію, яка перетне горизонтальну вісь у точці приблизно 63.

Оскільки верхня сторона кожного стовпчика розташована горизонтально, це означає, що при побудові гістограми і визначенні медіани припускалось, що 140 оцінок (округлених до 60) мають рівномірний розподіл у даному інтервалі. Зазначимо, що половина загальної площі прямокутників гістограми розташована зліва і справа від пунктирної лінії.

Аналогічно, гістограму можна використати для *визначення процентилів* ряду. *Нагадаємо, що процентиль може бути окремим значен-*

ням ряду або ж представляти собою величину, яка лежить між двома значеннями. Так, 49-й процентиль ряду оцінок за тест дорівнює оцінці, яка перевищує 49 % оцінок + половина значення (у %) даної оцінки. Для того щоб визначити 49-й процентиль, необхідно знайти на горизонтальній осі таке число, яке було б більшим 49 % оцінок. Інакше кажучи, необхідно так провести вертикальну лінію на графіку, щоб 49 % площі знаходилося зліва від неї і 51 % — справа.

Зазначимо, що 49 % від 1000 складає 490 оцінок. У шести нижніх інтервалах знаходиться 380 оцінок, а в шести верхніх — 520. Це означає, що 49-й процентиль знаходиться між значеннями 55,5÷64,5. Очевидно, що 110/140 або 11/14 площі прямокутника над цим інтервалом необхідно розташувати зліва від лінії і 3/14 — справа. Ця лінія показана на рис. 4.1 крапками. Вона перетинає горизонтальну вісь приблизно в точці 62,57. Таким чином, 49 % загальної площі гістограми лежить зліва від цієї лінії і 51 % — справа. Оскільки медіана є 50-м процентилем, то її можна знайти аналогічно розглянутій процедурі.

4.2. Знаходження медіани і процентилів

Знаходження значення процентиля для інтервального ряду розподілу є просто чисельною версією розглянутої вище процедури. Розглянемо для початку смисл показників останнього стовпчика табл. 4.2.

Таблиця 4.2

Накопичені частоти для ряду оцінок за тест

Границя	Частота	Накопичена частота
91,5–100,5	60	1000
82,5–91,5	140	940
73,5–82,5	160	800
64,5–73,5	120	640
55,5–64,5	140	520
46,5–55,5	80	380
37,5–46,5	119	300
28,5–37,5	81	181
19,5–28,5	50	100
10,5–19,5	32	50
1,5–10,5	18	18

В останньому стовпчику наведено так звані накопичені частоти, які є сумою частот сусідніх інтервалів, починаючи знизу. Накопичені частоти інтерпретують наступним чином: є 18 оцінок, менше ніж 10,5; 50 оцінок, менше ніж 19,5; 800 оцінок, менше ніж 82,5. Тобто відношення "*менше ніж*" стосується верхньої границі інтервалу.

Щоб знайти медіану, необхідно визначити точку, в якій сума накопичених частот буде дорівнювати 500. Ця точка знаходиться в інтервалі 55,5÷64,5, в якому 140 оцінок. З цього числа необхідно 500 – 380 = 120 і додати до попереднього значення накопиченої частоти. Очевидно, що точку, яка відповідає медіані, знайдемо наступним чином згідно наведеної раніше формули:

$$\text{Me} = 55{,}5 + \frac{120}{140} \cdot 9 = 55{,}5 + \frac{6}{7} \cdot 9 = 63{,}2.$$

Для того щоб знайти 35-й процентиль, необхідно визначити точку, в якій сума накопичених частот складає 35 % від 1000, тобто 350. Із стовпчика накопичених частот табл. 4.2 видно, що вона знаходиться в інтервалі 46,5÷55,5. Менше значення 46,5 знаходиться 300 оцінок. Таким чином, до цього значення необхідно додати ще 50 (з 80-ти оцінок, що знаходяться в інтервалі 46,5÷55,5). Тепер 35-й процентиль знайдемо як

$$\text{Пр}(35) = 46{,}5 + \frac{50}{80} \cdot 9 = 52{,}1.$$

Загальна формула для обчислення процентилів має вигляд:

$$\text{Пр}(n) = L + \frac{s}{f} I, \qquad (4.2.1)$$

де L — нижня границя інтервалу, в який попадає значення Пр(n); f — число елементів у даному інтервалі; s — число елементів, яке необхідно для попадання в точку на горизонтальній осі, що відповідає даному процентилю; I — ширина інтервалу.

4.3. Квартилі і децилі

Крім процентиля використовують такі поняття, як *квартиль* і *дециль*. Поняття квартиля відноситься до четвертої частини сукупності даних, а дециля — до десятої частини.

Для попереднього прикладу з оцінками тестування перший квартиль, або Q_1, дорівнює значенню оцінок, нижче якого розташовано 25 % оцінок, а вище — 75 %. Інакше кажучи, це ще одна назва для

Пр(25). Другий квартиль (Q_2) збігається з медіаною і Пр(50), а третій, Q_3 — це Пр(75).

Таким чином, оцінка точно попадає в *перший* (нижній) *квартиль*, якщо вона *не перевищує* 25 % оцінок. Аналогічно, оцінка попадає у *верхній квартиль*, Q_3, якщо вона *перевищує рівно* 75 % оцінок. І, насамкінець, оцінка знаходиться у "верхній чверті", якщо її процентильний ранг попадає в інтервал 75÷100.

Поняття *децила* відноситься, відповідно, до десятих часток сукупності даних. Перший дециль, або D_1, відповідає значенню оцінки, менше якого знаходяться рівно 10 % оцінок. Очевидно, що D_1 = = Пр(10), D_2 = Пр(20) і т. д.

4.4. Кумулятивна крива

Якщо по осі абсцис відкласти значення інтервалів оцінок, а по осі ординат — значення накопичених частот, то отримаємо графік кумулятивних (накопичених) частот. Графік кумулятивних частот для розглянутого вище прикладу з оцінками за тест наведено на рис. 4.2.

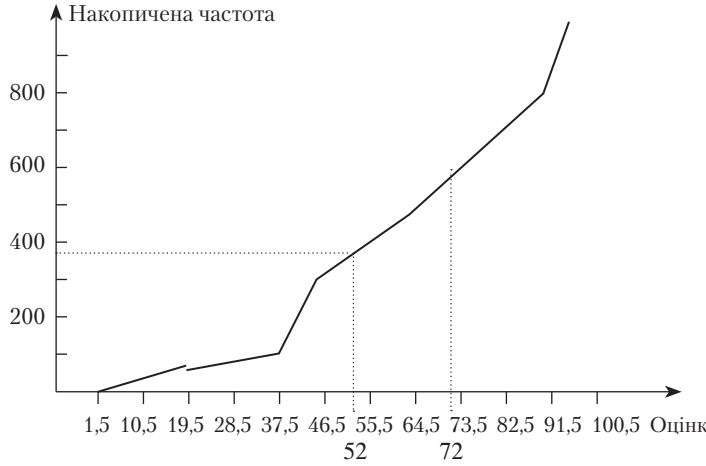


Рис. 4.2. **Кумулятивна крива для оцінок за тест**

Значення накопичених частот строго відповідають верхнім границям відповідних інтервалів. Щоб показати, що немає менших оцінок ніж 1,5, нанесемо відповідну точку на горизонтальній осі.

Кумулятивну криву накопичених частот можна використати для визначення процентилів і процентильних рангів. Нехай необхідно знайти 60-й процентиль Пр(60). На вертикальній осі знаходимо точку, яка відповідає 60 % від 1000, тобто 600 (оскільки всього було 1000 оцінок). З графіка кумулятивної кривої знаходимо, що Пр(60) = 70,5.

Для того щоб визначити процентильний ранг, необхідно йти у зворотному порядку. Так, для оцінок 70 і 71 частоти дорівнюють приблизно 600, що становить 60 %, тобто процентильний ранг становить 60.

4.5. Визначення моди згрупованих даних

Розглянемо задачу визначення моди спостережень, представлених у вигляді інтервального ряду розподілу. Раніше було встановлено, що ряд може мати дві і більше мод. Така сама ситуація може мати місце і для згрупованих даних. Спочатку необхідно вирішити, яке число ми будемо розглядати як моду.

У випадку згрупованих даних *інтервал, який характеризується найбільшою частотою, будемо називати модальним інтервалом, а його середнє значення — наближеною модою.*

Так, у випадку з оцінками за тест (див. табл. 4.1) модальним інтервалом є інтервал 74÷82, оскільки його частота 160 є найбільшою. Наближеним значенням моди є (74 + 82)/2 = 78. Якби було два модальних інтервали, то було б дві наближених моди. Зазначимо, що наближеною модою у подальшому ми не будемо користуватись.

4.6. Визначення середнього згрупованих даних

Раніше було встановлено, що віднімання константи від кожного значення ряду зменшує середнє на цю ж величину, а ділення кожного значення на константу веде до зменшення середнього в таке ж число разів. У даному параграфі розглянемо розрахунок середнього на основі таблиці згрупованих частот.

При розрахунку середнього інтервального ряду необхідно взяти тільки одне значення, яке представляє всі оцінки інтервалу. При рівномірному розподілі оцінок у рамках інтервалу таким значенням є центр інтервалу. Частоти інтервалів стають частотами їх центрів.

Розрахуємо середнє розподілу методом групування для ряду оцінок за тест, який складається з 1000 елементів. Результати попередніх розрахунків, необхідних для визначення середнього, наведені в

табл. 4.3. Таблиця містить добутки частот на середні значення кожного інтервалу, які необхідні для визначення середнього згрупованих даних.

Таблиця 4.3

До розрахунку середнього

Інтервал	Центр, x_i	Частота, f_i	Добуток, $f_i x_i$
92–100	96	60	5760
83–91	87	140	12180
74–82	78	160	12480
65–73	69	120	8280
56–64	60	140	8400
47–55	51	80	4080
38–46	42	119	4998
29–37	33	81	2673
20–28	24	50	1200
11–19	15	32	480
2–10	6	18	108
		$\Sigma f_i = 1000$	$\Sigma f_i x_i = 60639$

Таким чином, формула для середнього ряду розподілу при розрахунку методом групування має вигляд:

$$\mu = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}, \qquad (4.6.1)$$

де x_i — центри інтервалів; f_i — значення частот інтервалів; n — число інтервалів; очевидно, що $\Sigma f_i = 1000$ — загальне число оцінок за тест. Для даних, наведених у табл. 4.3, середнє для згрупованих даних: $\mu = 60639/1000 = 60{,}639$ або 60,6.

Якщо *центри інтервалів приймають дуже великі значення*, то середнє для згрупованих даних можна розрахувати також за формулою:

$$\mu = m\frac{\sum_{i=1}^{n} f_i u_i}{\sum_{i=1}^{n} f_i} + R, \qquad (4.6.2)$$

де $u_i = (x_i - R)/m$ (їх називають ще *u — оцінками або умовними відхиленнями*); R — константа, яку віднімають від кожного x_i з метою зменшення їх значень (її вибирають, як правило, рівною одному із серединних значень інтервалів x_i); m — масштабний коефіцієнт, який вибирають рівним відстані між сусідніми серединними значеннями x_i (тобто, ширині інтервалу). Наприклад, якщо користуватись цією формулою для даних, наведених у табл. 4.3, то для цієї мети можна вибрати $R = 60$, а ширина $m = 9$.

4.7. Знаходження дисперсії і стандартного відхилення для згрупованих даних

Раніше ми отримали формулу для знаходження дисперсії не згрупованих даних у вигляді:

$$\sigma_x^2 = \frac{1}{N-1}\sum_{i=1}^{N} x_i^2 - \frac{1}{(N-1)^2}\left(\sum_{i=1}^{N} x_i\right)^2. \qquad (4.7.1)$$

Вище було показано, що віднімання константи від кожного елемента ряду не впливає на величину дисперсії і стандартного відхилення. Друга властивість полягає в тому, що ділення кожного елемента ряду на константу приводить до зменшення стандартного відхилення (за абсолютною величиною) в таке саме число разів. Для ілюстрації застосування модифікованої формули (4.7.1) скористаємось прикладом аналізу оцінок за тест (табл. 4.4).

Таблиця 4.4

Розрахунок дисперсії

Інтервал	Центр, x_i	Частота, f_i	Добуток, $f_i x_i$	Квадрат, x_i^2	Добуток, $f_i x_i^2$
1	2	3	4	5	6
92–100	96	60	5760	9216	552960
83–91	87	140	12180	7569	1059660
74–82	78	160	12480	6084	973440
65–73	69	120	8280	4761	571320
56–64	60	140	8400	3600	504000
47–55	61	80	4080	2601	208080
38–46	42	119	4998	1764	209915
29–37	33	81	2673	1089	88209

Закінчення табл. 4.4

1	2	3	4	5	6
20–28	24	50	1200	576	28800
11–19	15	32	480	225	7200
2–10	6	18	108	36	648
		$\sum f_i = 1000$	$\sum f_i x_i =$ 60639		$\sum f_i x_i^2 =$ 4204233

Дисперсію для даних, згрупованих так, як показано в табл. 4.4, розраховують за формулою:

$$\sigma_x^2 = \frac{1}{N-1}\sum_{i=1}^{n} f_i x_i^2 - \frac{1}{(N-1)^2}\left(\sum_{i=1}^{n} f_i x_i\right)^2, \qquad (4.7.2)$$

де $N = \sum_{i=1}^{n} f_i$ — загальне число елементів розподілу; n — число груп елементів (інтервалів).

Для наведених у табл. 4.4 даних дисперсія складає:

$$\sigma_x^2 = \frac{4204233}{999} - \frac{1}{999^2}(60639)^2 = 524{,}0,$$

а стандартне відхилення дорівнює: $\sigma_x = \sqrt{524} = 22{,}96 \approx 23$.

Для знаходження дисперсії групових даних можна скористатися також *u*-оцінками або *умовними відхиленнями*:

$$\sigma_u^2 = \frac{1}{N-1}\sum_{i=1}^{n} f_i u_i^2 - \frac{1}{(N-1)^2}\left(\sum_{i=1}^{n} f_i u_i\right)^2, \qquad (4.7.3)$$

де u_i — оцінки. Значення *u*-оцінок для прикладу з оцінками тестування наведені в табл. 4.5.

Таблиця 4.5

Розрахунок дисперсії з використанням умовних відхилень

Центр, x_i	Частота, f_i	Значення, u_i	Добуток, $f_i u_i$	Квадрат, u_i^2	Добуток, $f_i u_i^2$
1	2	3	4	5	6
96	60	4	240	16	960
87	140	3	420	9	1260
78	160	2	320	4	640

Закінчення табл. 4.5

1	2	3	4	5	6
69	120	1	120	1	120
60	140	0	0	0	0
61	80	–1	–80	1	80
42	119	–2	–238	4	476
33	81	–3	–243	9	729
24	50	–4	–200	16	800
15	32	–5	–160	25	800
6	18	–6	–108	36	648
	$\sum f_i = 1000$		$\sum f_i u_i = 71$		$\sum f_i u_i^2 = 6513$

Значення u_i формуються наступним чином:

- центри x_i зменшують на деяке значення, вибране з середини стовпчика значень x_i (у даному випадку вибрано 60);
- значення u_7 (сьомий стовпчик знизу), оскільки центр $x_7 = 60$, а $x_7' = 60 - 60 = 0$;
- для інтервалів, серединні значення яких перевищують 60, у стовпчик u_i послідовно записують числа 1, 2, 3 і т. д., а для інтервалів, серединні значення яких є меншими 60, записують числа –1, –2, –3 і т. д.

Фактично, u-оцінки отримано шляхом віднімання 60 від серединних значень інтервалів з наступним діленням результату на 9, тобто на ширину інтервалів.

Оскільки при отриманні u-оцінок всі елементи початкового ряду були поділені на 9 (ширину інтервалу), то для того, щоб обчислити дисперсію початкового ряду на основі σ_u^2, необхідно помножити σ_u^2 на $9^2 = 81$. Підставимо дані з табл. 4.5 в (4.7.3):

$$\sigma_u^2 = \frac{6513}{999} - \left(\frac{71}{999}\right)^2 = 6,465 - (0,071)^2 = 6,46,$$

а стандартне відхилення: $\sigma_u = 2,54$. Статистичні характеристики для початкового ряду:

$\sigma_x = 9\sigma_u = 9 \cdot 2,54 = 22,89 \approx 23$;
$\sigma_x^2 = (22,89)^2 = 524$.

Таким чином, остаточна формула для розрахунку стандартного відхилення ряду згрупованих оцінок з використанням умовних відхилень має вигляд:

$$\sigma_x = m\sqrt{\frac{1}{N-1}\sum_{i=1}^{n} f_i u_i^2 - \frac{1}{(N-1)^2}\left(\sum_{i=1}^{n} f_i u_i\right)^2}, \qquad (4.7.4)$$

де m — довжина інтервалу. Зазначимо, що використання умовного відхилення значно прискорює і спрощує всі обчислення.

4.8. Контрольні питання і вправи

1. Поясніть, чи обов'язково медіана повинна збігатись із конкретним значенням ряду розподілу.
2. Як визначити медіану за гістограмою?
3. Яким чином визначаються процентилі за гістограмою?
4. Поясніть такі терміни: 50-й процентиль і 88-й процентиль.
5. Запишіть і поясніть формулу для обчислення процентилів.
6. Поясніть терміни *квартиль* і *дециль*. Поясніть на числовому прикладі.
7. Яким чином будується кумулятивна крива частот? Поясніть це на прикладі отриманих балів за тест.
8. Що таке модальний інтервал і як він визначається для згрупованих даних?
9. Поясніть на числовому прикладі обчислення середнього для згрупованих даних.
10. Поясніть термін *u*-оцінка і наведіть приклад практичного використання цих оцінок.
11. Запишіть і поясніть вираз для обчислення дисперсії і стандартного відхилення для згрупованих даних.

Розділ 5

НОРМАЛЬНИЙ РОЗПОДІЛ

5.1. Теоретичні розподіли

Вже встановлено, що розширення ряду розподілу все більшою кількістю значень дає можливість точніше відтворити форму його графіка. Однак, після досягнення деякої границі додавання нових значень до ряду розподілу випадкової величини (ВВ) перестає суттєво впливати на покращання форми графіка. Після того, як форма графіка перестала суттєво змінюватись, можна досліджувати властивості даного розподілу, не враховуючи кількість його елементів. У такому випадку кажуть, що кількість спостережень є "досить великою".

У подальшому будемо використовувати вираз "*нескінченна кількість спостережень*" тоді, коли ми хочемо сказати, що додавання додаткових спостережень до ряду не може надати нам нову інформацію щодо його властивостей.

Теоретичний розподіл — це розподіл, який складається з нескінченної кількості спостережень.

Інакше кажучи, це розподіл, характеристики якого вже сформувались і про його властивості можна вже сказати, що вони стали фіксованими. У теоретичному розподілі на кожне окреме значення змінної припадає такий незначний процент від загальної кількості спостережень, що процентильний ранг цього значення можна розглядати просто як процент спостережень, що знаходяться нижче нього у розподілі.

5.2. Нормальний розподіл і його графік

Одним із конкретних видів розподілів є *нормальний* або *Гаусів розподіл* (НР), який відіграє дуже велику роль у статистиці, теорії оцінювання та багатьох інших науках. Більша частина досягнень статистики базується на дослідженні властивостей нормального розподілу.

Крива нормального (гаусового) розподілу представлена на рис. 5.1. На горизонтальній осі абсцис відкладено значення нормованих відхилень (z-оцінок). Форма кривої нормального розподілу залежить

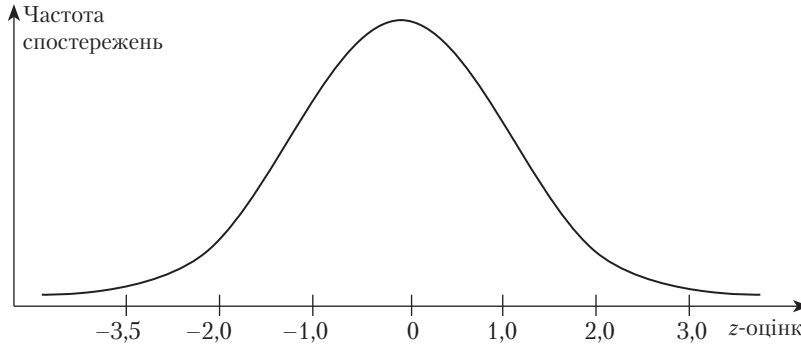


Рис. 5.1. **Крива нормального розподілу**

від масштабу, вибраного для *z*-оцінок. Наприклад, крива буде мати іншу ("гострішу") форму, якщо зменшити відстань між значеннями *z*-оцінок при постійному масштабі для частот спостережень на вертикальній осі. Крива НР є *симетричною*, а її найвища точка знаходиться над значенням $z = 0$.

З попереднього розділу відомо, що таблиця, яка містить значення *z*-оцінок і відповідних їм процентильних рангів, є числовим описом розподілу. Така таблиця дає можливість накреслити графік нормального розподілу без додаткової інформації.

У таблиці для нормального розподілу (вона, як правило, наводиться у додатках до підручників або у довідниках) наведено значення *z*-оцінок від –4 до +4. Якщо відома *z*-оцінка деякого спостереження, яке належить нормальному розподілу, то з цієї таблиці можна знайти частку спостережень, які за своєю величиною є меншими даного спостереження. Наприклад, з таблиці випливає, що спостереження із $z_i = -1{,}0$ перевищує приблизно 0,159 або 15,9 % процентів спостережень розподілу. Інакше кажучи, спостереження із $z_i = -1{,}0$ є наближено 16-м процентилем. Спостереження, яке має $z_i = 0$, перевищує приблизно половину спостережень нормального розподілу і є, таким чином, 50-м процентилем. Спостереження із $z_i = 1{,}5$ є приблизно 93-м процентилем, оскільки воно перевищує 93,3 % всіх спостережень. Чим детальнішою є така таблиця, тим точніше вона визначає нормальний розподіл.

Таким чином, твердження, що розподіл є нормальним, означає, що залежності між z-оцінками спостережень та їх процентильними рангами відповідає таблиці нормального розподілу.

5.3. Чотири властивості нормального розподілу

1. *Значення змінної мають властивість концентруватись навколо точки* $z = 0$.

2. *Нормальна крива є симетричною відносно вертикальної осі.* Звідси випливає, що половина спостережень розташовується зліва від вертикальної осі при $z = 0$ і має від'ємні z-оцінки.

3. *Значення спостережень, які належать нормальному розподілу, не обмежені за своєю величиною.* Теоретично спостереження можуть мати будь-які z-оцінки, наприклад, 1000. Ця властивість пояснюється тим, що крива нормального розподілу не перетинається з віссю абсцис.

4. *Середнє, медіана і мода нормального розподілу мають одне і те саме значення.* Спостереження, яке має $z_i = 0$, якраз і є цим значенням. Крива є симетричною відносно прямої, що проходить через цю точку і в ній крива має максимальну висоту.

5.4. Функція розподілу і числові характеристики неперервних величин

Функцією розподілу (або *кумулятивною функцією розподілу*) називають функцію $F(x)$, яка визначає ймовірність того, що випадкова величина X в результаті виконання деякого досліду прийме значення менше x, тобто

$$F(x) = p(X \le x).$$

Властивості функції розподілу

Властивість 1. *Значення функції розподілу належить відрізку* $[0, 1]$:

$$0 \le F(x) \le 1. \qquad (5.4.1)$$

Ця властивість випливає із визначення функції розподілу як ймовірності: *ймовірність* — це невід'ємне число, яке не перевищує одиниці.

Властивість 2. $F(x)$ — *неспадна функція, тобто*

$$F(x_2) \ge F(x_1), \text{ якщо } x_2 > x_1. \qquad (5.4.2)$$

Наслідок 1. *Ймовірність того, що випадкова величина прийме значення, яке знаходиться в інтервалі* (a, b), *дорівнює приросту функції розподілу на цьому інтервалі*:

$$P(a \le X < b) = F(b) - F(a). \qquad (5.4.3)$$

Наслідок 2. *Ймовірність того, що неперервна випадкова величина x прийме одне конкретне значення, дорівнює нулю.*

Похідну від функції розподілу називають *щільністю* або *густиною розподілу*, тобто

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

Функції розподілів випадкових величин називають *статистичними моделями* випадкових процесів. Нижче будуть розглянуті статистичні моделі нормального та деяких інших розподілів.

Ймовірність попадання неперервної випадкової величини у заданий інтервал

Теорема 5.1. *Ймовірність того, що неперервна випадкова величина прийме значення, яке належить інтервалу* (a, b), *дорівнює визначеному інтегралу від густини розподілу, взятому в межах від a до b*:

$$P(a < X < b) = \int_a^b f(x)\,dx.$$

Доведення. Скористаємось відомим співвідношенням:

$$P(a \le X < b) = F(b) - F(a).$$

За формулою Лейбніца: $F(b) - F(a) = \int_a^b F'(x)\,dx = \int_a^b f(x)\,dx$. Таким чином,

$$P(a \le X < b) = \int_a^b f(x)\,dx.$$

Оскільки $P(a \le X < b) = P(a < X < b)$, то остаточно отримаємо:

$$P(a < X < b) = \int_a^b f(x)\,dx.$$

Отриманий результат геометрично можна трактувати так: ймовірність того, що неперервна випадкова величина прийме значення, яке належить інтервалу (a, b), дорівнює площі криволінійної трапеції, обмеженої віссю $0x$, кривою розподілу $f(x)$ і вертикальними прямими $x = a$ і $x = b$.

Статистичні характеристики неперервних випадкових величин

При необмеженому збільшенні числа спостережень для дослідження розподілів випадкових величин зручно застосовувати харак-

теристики неперервних випадкових величин. Почнемо з математичного сподівання.

Нехай неперервна випадкова величина X задана щільністю розподілу $f(x)$. Припустимо, що X визначена на відрізку $[a, b]$. Розіб'ємо цей відрізок на n менших відрізків довжиною $\Delta x_1, \Delta x_2, ..., \Delta x_n$ і виберемо в кожному з них довільну точку x_i, $i = 1, ..., n$. Необхідно визначити математичне сподівання неперервної величини. За *аналогією з розподілом дискретних величин* складемо суму добутків можливих значень x_i на ймовірності їх попадання в інтервал Δx_i (нагадаємо, що добуток $f(x)\Delta x$ наближено дорівнює ймовірності попадання значення X в інтервал Δx):

$$\sum_{i=1}^{n} x_i f(x_i)\Delta x_i .$$

Якщо спрямувати довжину найбільшого із часткових відрізків до нуля, то

$$\lim_{\Delta x_i \to 0} \sum_{i=1}^{n} x_i f(x_i)\Delta x_i = \int_a^b xf(x)dx .$$

Означення 5.1. *Математичним сподіванням неперервної випадкової величини X, визначеної на інтервалі* $[a, b]$, *називають визначений інтеграл*

$$E[X] = \int_a^b xf(x)dx . \qquad (5.4.4)$$

Якщо X визначена на всій осі $0x$, то

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx . \qquad (5.4.5)$$

У даному випадку припускається, що невласний інтеграл збігається абсолютно, тобто існує інтеграл $\int_{-\infty}^{\infty} |x| f(x)dx$. Якби ця вимога не виконувалась, то значення інтегралу залежало б від швидкості наближення (окремо) нижньої границі до $-\infty$, а верхньої — до $+\infty$.

Означення 5.2. *Дисперсією неперервної випадкової величини називають математичне сподівання квадрату її відхилення.*

Якщо $X \in [a, b]$, то

$$\mathrm{var}[X] = \int_a^b \left[x - E(X)\right]^2 f(x)dx . \qquad (5.4.6)$$

Якщо ж $X \in [-\infty, \infty]$, то

$$\mathrm{var}[X] = \int_{-\infty}^{\infty} \left[x - E(X)\right]^2 f(x)dx . \qquad (5.4.7)$$

Примітка 1. Можна показати, що властивості математичного сподівання і дисперсії дискретних величин зберігаються і для неперервних величин.

Примітка 2. Для обчислення дисперсії можна легко отримати наступні формули:

$$\operatorname{var}[X]=\int_a^b x^2 f(x)\,dx-\left[E(X)\right]^2; \tag{5.4.8}$$

$$\operatorname{var}[X]=\int_{-\infty}^{\infty} x^2 f(x)\,dx-\left[E(X)\right]^2. \tag{5.4.9}$$

Приклад 5.1. Визначити математичне сподівання і дисперсію випадкової величини X, заданої функцією розподілу:

$$F(x)=\begin{cases}0, & x\le 0,\\ x, & 0<x\le 1,\\ 1, & x>1.\end{cases}$$

Розв'язок: Знайдемо щільність розподілу:

$$f(x)=F'(x)=\begin{cases}0, & x\le 0,\\ 1, & 0<x\le 1,\\ 0, & x>1.\end{cases}$$

Математичне сподівання:

$$E[X]=\int_0^1 x\cdot 1\cdot dx=\left.\frac{x^2}{2}\right|_0^1=\frac{1}{2}.$$

Дисперсія:

$$\operatorname{var}[X]=\int_0^1 x^2\cdot 1\cdot dx-\left(\frac{1}{2}\right)^2=\left.\frac{x^3}{3}\right|_0^1-\frac{1}{4}=\frac{1}{12}.$$

Приклад 5.2. Визначити математичне сподівання і дисперсію неперервної випадкової величини X, розподіленої рівномірно в інтервалі (a, b).

Розв'язок. Враховуючи, що густина рівномірного розподілу $f(x)=\dfrac{1}{b-a}$, знайдемо математичне сподівання:

$$E[X]=\int_a^b x\cdot f(x)\,dx=\frac{1}{b-a}\int_a^b x\,dx=\frac{1}{b-a}\cdot\left.\frac{x^2}{2}\right|_a^b=\frac{\left(b^2-a^2\right)}{2(b-a)}=\frac{a+b}{2}.$$

Дисперсія:

$$\mathrm{var}[X]=\int_a^b x^2\cdot f(x)\,dx-\left[E(X)\right]^2=\frac{1}{b-a}\int_a^b x^2dx-\left[\frac{a+b}{2}\right]^2=\frac{(b-a)^2}{12}.$$

5.5. Статистичні характеристики нормального розподілу

Статистична модель (щільність, густина) нормального розподілу, має вигляд:

$$f(x)=\frac{1}{\sigma_x\sqrt{2\pi}}\exp\left(-\frac{1}{2}\frac{(x-\mu_x)^2}{\sigma_x^2}\right), \qquad (5.5.1)$$

де σ_x^2 — дисперсія ряду розподілу змінної X; μ_x — середнє ряду; $f(x)$ — густина розподілу. Покажемо, що σ_x^2 і μ_x мають вказаний статистичний смисл.

1. За означенням математичного сподівання неперервної випадкової величини

$$E[X]=\int_{-\infty}^{\infty} xf(x)\,dx=\frac{1}{\sigma_x\sqrt{2\pi}}\int_{-\infty}^{\infty} x\cdot\exp\left(-(x-\mu_x)^2/2\sigma_x^2\right)dx. \qquad (5.5.2)$$

Введемо нову змінну $z=(x-\mu_x)/\sigma_x$, а звідси $x=\sigma_x z+\mu_x$ і $dx=\sigma_x dz$. Оскільки нові границі інтегрування залишаються незмінними, то

$$E[X]=\frac{\sigma_x}{\sigma_x\sqrt{2\pi}}\int_{-\infty}^{\infty}(\sigma_x z+\mu_x)e^{-z^2/2}dz=$$

$$=\frac{\sigma_x}{\sqrt{2\pi}}\int_{-\infty}^{\infty} ze^{-z^2/2}dz+\frac{\mu_x}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-z^2/2}dz.$$

Оскільки під знаком інтеграла в першому доданку *непарна функція* (z може бути меншим і більшим нуля), а *границі інтегрування симетричні стосовно початку координат*, то він дорівнює нулю. Другий доданок дорівнює μ_x, оскільки

$$\int_{-\infty}^{\infty}\exp\left(-z^2/2\right)dz=\sqrt{2\pi}$$

є інтегралом Пуассона. Таким чином, $E[X]=\mu_x$.

2. За означенням дисперсії неперервної випадкової величини маємо:

$$\mathrm{var}[X]=\frac{1}{\sigma_x\sqrt{2\pi}}\int_{-\infty}^{\infty}(x-\mu_x)^2\cdot\exp\left(-(x-\mu_x)^2/2\sigma_x^2\right)dx.$$

Знову введемо змінну $z = (x - \mu_x)/\sigma_x$, а звідси $x = \sigma_x z + \mu_x$ і $dx = \sigma_x dz$. І враховуючи, що границі інтегрування залишаються незмінними, отримаємо:

$$\mathrm{var}[X] = \frac{\sigma_x}{\sigma_x\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma_x z + \mu_x - \mu_x)^2 \cdot \exp\left(-z^2/2\right) dz =$$

$$= \frac{\sigma_x^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z \cdot z \cdot \exp\left(-z^2/2\right) dz = \frac{\sigma_x^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z d\left(-\exp\left(-z^2/2\right)\right).$$

Покладемо: $u = z$, $d\nu = ze^{-z^2/2}\,dz$ ($\nu = \int ze^{-z^2/2} dz = -e^{-z^2/2}$) і скористаємось методом інтегруванням за частинами: $\int u d\nu = u\nu - \int \nu du$:

$$\mathrm{var}[X] = \frac{\sigma_x^2}{\sqrt{2\pi}} \left[u\nu\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \nu du \right],$$

тобто

$$\mathrm{var}[X] = \frac{\sigma_x^2}{\sqrt{2\pi}} \left[-z \cdot \exp\left(-\frac{z^2}{2}\right) \right]_{-\infty}^{\infty} + \frac{\sigma_x^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{z^2}{2}\right) dz = \sigma_x^2,$$

оскільки $z \cdot \exp\left(-\frac{z^2}{2}\right) \to 0$ при z→∞, а $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{z^2}{2}\right) dz = 1$. Тобто перший доданок дорівнює нулю.

Примітка 1. *Загальним нормальним розподілом* називають розподіл з довільними значеннями параметрів μ_x і σ_x^2.

Нормованим називають нормальний розподіл з параметрами $\mu = 0$ і $\sigma^2 = 1$ (відповідно, стандартне відхилення $\sigma = 1$). Наприклад, якщо x — нормально розподілена величина з довільними параметрами μ_x і σ_x, то $u = (x - \mu_x)/\sigma_x$ — нормована величина з $\mu_u = 0$ і $\sigma_u = 1$.

Нормована густина нормального розподілу:

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-u^2/2\right). \qquad (5.5.3)$$

Ця функція затабульована, таблиці наведені у підручниках та довідниках.

Густина розподілу для довільної випадкової змінної X пов'язана з нормованою густиною виразом:

$$f(x) = \frac{1}{\sigma_x} \varphi(u) \quad \text{при} \quad u = \frac{x - \mu_x}{\sigma_x}.$$

Примітка 2. Функція $F(x)$ загального нормального розподілу визначається за формулою:

$$F(x)=\frac{1}{\sigma_x\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-(x-\mu_x)^2/2\sigma_x^2\right)dx, \qquad (5.5.4)$$

де x — змінна інтегрування. Нормована функція нормального розподілу визначається як

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-z^2/2\right)dz. \qquad (5.5.5)$$

Можна легко перевірити, що $F(x) = \Phi^*((x - \mu_x)/\sigma_x)$:

$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}e^{-(t-\mu)^2/2\sigma^2}dt=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{x-\mu}{\sigma}}e^{-u^2/2}du=\Phi^*\left(\frac{x-\mu}{\sigma}\right).$$

Примітка 3. Ймовірність попадання нормованої нормальної величини X в інтервал $(0, x)$ можна знайти за допомогою функції Лапласа (5.5.5).

Дійсно,

$$P(0<X<x)=\int_0^x\varphi(z)dz=\frac{1}{\sqrt{2\pi}}\int_0^x\exp\left(-z^2/2\right)dz=\Phi(x). \qquad (5.5.6)$$

Примітка 4. Враховуючи, що $\int_{-\infty}^{\infty}\varphi(x)dx=1$, а в силу симетрії $\varphi(x)$ відносно нуля

$$\int_{-\infty}^{0}\varphi(x)dx=0{,}5\text{ , а значить і }\quad P(-\infty<X<0)=0{,}5,$$

то легко визначити, що $F(x) = 0{,}5 + \Phi(x)$.

Дійсно,

$$F(x)=P(-\infty<X<x)=P(-\infty<X<0)+P(0<X<x)=0{,}5+\Phi(x).$$

5.6. Дослідження кривої нормального розподілу

Виконаємо дослідження функції

$$y(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-(x-a)^2/2\sigma^2\right]$$

методами диференціального числення. Для спрощення запису середнє μ_x позначено буквою a.

1. *Функція визначена на всій осі x.*

2. При всіх значеннях x *функція приймає додатні значення* (парна функція), тобто нормальна крива розташована над віссю $0x$.

3. *Границя функції* при необмеженому зростанні модуля x дорівнює нулю:

$$\lim_{|x|\to\infty} y = 0,$$

тобто вісь $0x$ служить горизонтальною асимптотою графіка нормального розподілу.

4. *Дослідимо функцію на екстремум*. Запишемо першу похідну:

$$y' = -\frac{x-a}{\sigma^3\sqrt{2\pi}}\exp\left[-(x-a)^2/2\sigma^2\right].$$

Видно, що $y' = 0$ при $x = a$; $y' > 0$ при $x < a$; $y' < 0$ при $x > a$. Тобто функція має максимум $\frac{1}{\sigma\sqrt{2\pi}}$ при $x = a$.

5. Різниця $x - a$ в квадраті міститься в аналітичному виразі функції, тобто *графік функції є симетричним стосовно прямої* $x = a$.

6. *Дослідимо функцію на точки перегину*. Знайдемо другу похідну:

$$y'' = -\frac{1}{\sigma^3\sqrt{2\pi}}e^{-(x-a)^2/2\sigma^2}\left[1-\frac{(x-a)^2}{\sigma^2}\right].$$

Видно, що друга похідна дорівнює нулю при $x = a + \sigma$ і $x = a - \sigma$, а при переході через ці точки вона змінює знак (в обох цих точках значення функції дорівнює $\frac{1}{\sigma\sqrt{2\pi e}}$). Таким чином, точками перегину графіка є наступні:

$$\left(a-\sigma, \frac{1}{\sigma\sqrt{2\pi e}}\right) \text{ і } \left(a+\sigma, \frac{1}{\sigma\sqrt{2\pi e}}\right).$$

5.7. Вплив параметрів нормального розподілу на форму нормальної кривої

Знову позначимо середнє розподілу $\mu_x = a$ і встановимо, як впливають параметри a і σ на форму і розташування нормальної кривої.

Відомо, що форма графіка нормального розподілу не залежить від значення середнього a. Тобто графіки функцій $f(x)$ і $f(x - a)$ мають однакову форму. При $a > 0$ вершина кривої буде розташована справа від осі ординат, а при $a < 0$ — буде розташована зліва від цієї осі. Звідси можна зробити такий висновок.

Зміна величини математичного сподівання не змінює форми нормального розподілу, а приводить лише до її зсуву вздовж осі $0x$ *— впра-*

во при зростанні значення середнього, і вліво — при зменшенні значення середнього.

Тепер розглянемо вплив стандартного відхилення. Максимум функції нормального розподілу дорівнює $1/(\sigma\sqrt{2\pi})$. Звідси випливає висновок.

Зростання σ *приводить до зменшення максимальної ординати нормальної кривої, а сама крива стає більш пологою; при зменшенні* σ *вершина нормальної кривої загострюється і розтягується в додатному напрямку осі* $0y$.

Зазначимо, що при будь-яких значеннях параметрів a і σ площа, обмежена нормальною кривою та віссю $0x$, залишається рівною одиниці (друга властивість густини розподілу).

5.8. Ймовірність попадання випадкової нормальної величини в заданий інтервал

З попереднього аналізу відомо, що ймовірність попадання значення випадкової величини X в інтервал (α, β) визначається за формулою:

$$P(\alpha < X < \beta) = \int_{\alpha}^{\beta} f(x)\,dx. \qquad (5.8.1)$$

Нехай випадкова величина X розподілена за нормальним законом. Ймовірність того, що X прийме значення, яке належить інтервалу (α, β), дорівнює

$$P(\alpha < X < \beta) = \frac{1}{\sigma\sqrt{2\pi}} \int_{\alpha}^{\beta} \exp\left[-(x-a)^2/2\sigma^2\right] dx. \qquad (5.8.2)$$

Виконаємо перетворення цієї формули так, щоб можна було скористатись готовими таблицями. Введемо нову змінну $z = (x - a)/\sigma$. Звідси маємо: $x = \sigma z + a$ і $dx = \sigma dz$. Нові границі інтегрування будуть такими: якщо $x = \alpha$, то $z = (\alpha - a)/\sigma$; а якщо $x = \beta$, то $z = (\beta - a)/\sigma$.

Таким чином, маємо:

$$P(\alpha < X < \beta) = \frac{1}{\sigma\sqrt{2\pi}} \int_{(\alpha-a)/\sigma}^{(\beta-a)/\sigma} \exp\left(-z^2/2\right) dz =$$

$$= \frac{1}{\sqrt{2\pi}} \int_{(\alpha-a)/\sigma}^{0} \exp\left(-z^2/2\right) dz + \frac{1}{\sqrt{2\pi}} \int_{0}^{(\beta-a)/\sigma} \exp\left(-z^2/2\right) dz =$$

$$= \frac{1}{\sqrt{2\pi}} \int_0^{(\beta-a)/\sigma} \exp\left(-z^2/2\right) dz - \frac{1}{\sqrt{2\pi}} \int_0^{(\alpha-a)/\sigma} \exp\left(-z^2/2\right) dz.$$

Користуючись функцією Лапласа

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-z^2/2} dz,$$

остаточно запишемо:

$$P(\alpha < X < \beta) = \Phi\left(\frac{\beta - a}{\sigma}\right) - \Phi\left(\frac{\alpha - a}{\sigma}\right). \qquad (5.8.3)$$

Приклад 5.3. Випадкова величина X має нормальний розподіл. Математичне сподівання і стандартне відхилення цієї величини дорівнюють 30 і 10, відповідно. Знайти ймовірність того, що X прийме значення в інтервалі (10, 50).

Розв'язок. Скористаємось формулою (5.8.3) за умови, що $\alpha = 10$, $\beta = 50$, $a = 30$ і $\sigma = 10$:

$$P(10 < X < 50) = \Phi\left(\frac{50-30}{10}\right) - \Phi\left(\frac{10-30}{10}\right) = 2\Phi(2).$$

З таблиці для функції Лапласа знайдемо, що $\Phi(2) = 0{,}4772$, а тому

$$P(10 < X < 50) = 2 \cdot 0{,}4772 = 0{,}9544.$$

5.9. Ймовірність отримання випадковою величиною заданого значення відхилення

У прикладних дослідженнях часто виникає задача знаходження ймовірності того, що відхилення випадкової нормальної величини X є меншим за модулем заданого додатного числа δ, тобто $|X - \mu_x| < \delta$.

Позначимо середнє $\mu_x = a$ і замінимо цю нерівність рівносильною нерівністю:

$$-\delta < X - a < \delta \text{ або } a - \delta < X < a + \delta.$$

Відповідно до формули (5.8.3) можна отримати:

$$P(|X - a| < \delta) = P(a - \delta < X < a + \delta) =$$
$$= \Phi\left[\frac{(a+\delta)-a}{\sigma}\right] - \Phi\left[\frac{(a-\delta)-a}{\sigma}\right] = \Phi\left(\frac{\delta}{\sigma}\right) - \Phi\left(-\frac{\delta}{\sigma}\right).$$

Оскільки функція Лапласа Ф — непарна, то $\Phi(-\delta/\sigma) = -\Phi(\delta/\sigma)$. Остаточно можна записати, що

$$P(|X - a| < \delta) = 2\Phi(\delta/\sigma). \qquad (5.9.1)$$

Зокрема, при $a = 0$:

$$P(|X| < \delta) = 2\Phi(\delta/\sigma). \qquad (5.9.2)$$

Нормальна крива стає "гострішою" при зменшенні значення σ. Звідси випливає, що для двох центрованих процесів ($a = 0$) ймовірність прийняття значення в інтервалі $(-\delta, \delta)$ є більшою у тієї величини, яка має менше σ. Це пояснюється тим, що частина графіка в інтервалі $(-\delta, \delta)$ буде мати більшу площу. Крім того, цей факт повністю відповідає ймовірнісному смислу параметра σ — чим меншим є розсіювання випадкової величини, тим більшою буде ймовірність попадання в інтервал $(-\delta, \delta)$.

Примітка. Очевидно, що події, які зв'язані з нерівностями $|X - a| < \delta$ і $|X - a| \geq \delta$, носять взаємно виключний характер. Тому, якщо ймовірність виконання нерівності $|X - a| < \delta$ дорівнює p, то ймовірність виконання нерівності $|X - a| \geq \delta$ буде $(1 - p)$.

Приклад 5.4. Випадкова нормальна величина X має $\mu_x = 20$ і $\sigma_x = 10$. Знайти ймовірність того, що абсолютне значення відхилення буде меншим 3, тобто $(\delta = 3)$.

Розв'язок. Скористаємось формулою (5.9.1):

$$P(|X - a| < \delta) = 2\Phi(\delta/\sigma) = p(|X - 20| < 3) = 2\Phi(3/10) = 2\Phi(0{,}3) = \\ = 2 \cdot 0{,}1179 = 0{,}2358.$$

5.10. Правило трьох сигм

У виразі (5.9.1), тобто $P(|X - a| < \delta) = 2\Phi(\delta/\sigma)$, покладемо $\delta = \sigma q$:

$$P(|X - a| < \sigma q) = 2\Phi(q).$$

Якщо $q = 3$, тобто $\sigma q = 3\sigma$, то

$$P(|X - a| < 3\sigma) = 2\Phi(3) = 2 \cdot 0{,}49865 = 0{,}9973, \qquad (5.10.1)$$

тобто ймовірність того, що абсолютне відхилення буде меншим трьох стандартних відхилень, дорівнює 0,9973.

Інакше кажучи, ймовірність того, що абсолютна величина відхилення випадкової величини *перевищить три стандартних відхилення* (СВ), є дуже малою, 0,0027. Це означає, що така ситуація може мати місце тільки у 0,27 % випадків. Якщо скористатись *принципом неможливості малоймовірних подій*, то можна сказати, що такі ситуації є практично неможливими. Таким чином, суть правила трьох сигм зводиться до такого:

Абсолютна величина відхилення нормальної випадкової величини від її математичного сподівання не перевищує трьох стандартних відхилень.

На практиці правило трьох сигм застосовують так: якщо розподіл випадкової величини не відомий, але умова (5.9.2) виконується, то можна припустити, що ця величина розподілена нормально, або вона має інший розподіл.

5.11. Оцінка відхилення теоретичного розподілу від нормального, асиметрія і ексцес

Емпіричним називають *розподіл відносних частот даних*. Його досліджують методами математичної статистики.

Теоретичним називають такий розподіл ВВ, подальше збільшення спостережень якої не впливає на форму кривої розподілу.

При дослідженні розподілів, які є відмінними від нормального, виникає необхідність знайти кількісну міру цієї відмінності. З цією метою введено спеціальні характеристики: *асиметрію, ексцес, статистику Жак-Бера* [11]. *Для нормального розподілу асиметрія і ексцес дорівнюють нулю.* Якщо асиметрія і ексцес мають для досліджуваного процесу невеликі значення, то можна припустити близькість цього розподілу до нормального.

Оцінювання асиметрії. Можна показати, що для симетричного розподілу кожний центральний момент непарного порядку дорівнює нулю.

Нагадаємо, що момент порядку p визначається як $m_p = E[X^p]$, а центральний момент порядку p — як $\mu_p = E[(X - E[X])^p]$.

Для несиметричних розподілів центральні моменти непарного порядку є відмінними від нуля. Тому будь-який з цих центральних моментів (крім моменту першого порядку, який дорівнює нулю для будь-якого розподілу) може служити для оцінки величини асиметрії.

Логічно вибрати для цієї мети самий простий з них, тобто момент третього порядку μ_3. Однак, приймати цей момент для оцінювання асиметрії незручно, тому що його величина залежить від одиниць, в яких вимірюється ВВ. Для того щоб позбутись цього недоліку, μ_3 ділять на σ^3 і, таким чином, отримують безрозмірну величину.

Асиметрією (*skewness*) розподілу називають відношення центрального моменту третього порядку до куба стандартного відхилення (він характеризується симетричністю хвостів розподілу):

$$Ac = \frac{\mu_3}{\sigma^3} \tag{5.11.1}$$

або

$$Ac = \frac{1}{N}\sum_{k=1}^{N}\left[\frac{y(k)-\bar{y}}{\sigma}\right]^3. \tag{5.11.2}$$

Якщо $Ac > 0$, то правий хвіст розподілу довший, а при $Ac < 0$ довшим є лівий хвіст розподілу; якщо $Ac = 0$, то розподіл симетричний.

Практично знак Ac визначають за розміщенням кривої розподілу відносно моди, тобто точки максимуму диференціальної функції: якщо "довга частина" кривої розташована правіше моди, то асиметрія додатна, а якщо зліва, то від'ємна.

Для оцінювання величини нахилу ("крутизни") кривої розподілу у порівнянні з кривою теоретичного розподілу користуються ексцесом.

Ексцес (*kurtosis*) — характеризує відмінність форми розподілу від нормального і розраховується за виразом:

$$E_k = \frac{\mu_4}{\sigma^4} - 3. \tag{5.11.3}$$

Для нормального розподілу $\mu_4/\sigma^4 = 3$, а тому ексцес $E_k = 0$. Якщо ексцес додатний, то крива має вищу і "гострішу" вершину ніж нормальний розподіл; якщо ексцес від'ємний, то досліджувана крива має нижчу і "плоскішу" вершину, ніж нормальна крива.

В англомовній літературі аналогом терміна *ексцес* є *куртозис* (*kurtosis*):

$$K = \frac{1}{N}\sum_{k=1}^{N}\left[\frac{y(k)-\bar{y}}{\sigma}\right]^4. \tag{5.11.4}$$

$K = 3$ для нормального розподілу; якщо $K > 3$, то форма розподілу буде "*гострішою*" від нормального; при $K < 3$ форма розподілу буде "*плоскішою*" від нормального.

Статистика Жак-Бера (*Jarque-Bera*) [4] — тестова статистика, яка показує, наскільки близьким є ряд до нормального розподілу.

$$JB = \frac{N-n}{6}\left[Ac^2 + \frac{1}{4}(K-3)^2\right], \qquad (5.11.5)$$

де n — число коефіцієнтів, використаних для побудови моделі ряду; при нуль-гіпотезі щодо нормальності розподілу статистика Жак-Бера має розподіл χ^2 з двома степенями вільності. Ймовірність, пов'язана із статистикою Жак-Бера, показує ймовірність справедливості нуль-гіпотези. *Мала ймовірність свідчить про те, що нуль-гіпотезу щодо нормальності розподілу необхідно відхилити.*

5.12. Використання таблиці відносних площ нормального розподілу

У таблиці для відносної площі нормального розподілу (додаток А) наведено частки площ, які знаходяться між оцінкою $z = 0$ та деяким іншим вибраним значенням z-оцінки.

Припустимо, що необхідно визначити, яка частина спостережень має z-оцінки, що знаходяться між $z = 0$ і $z = 0{,}7$. За допомогою лівої колонки при $z = 0{,}7$ та другого стовпчика $z = 0{,}0$ знайдемо значення 0,2580. Інакше кажучи, 25,8 % спостережень нормального розподілу мають z-оцінки в інтервалі 0÷0,7.

Оскільки нормальний розподіл симетричний, то в інтервалі $z = -0{,}7 \div 0{,}7$ буде знаходитись 51,6 % спостережень.

Нехай необхідно знайти частку спостережень нормального розподілу, які мають z-оцінки в інтервалі: $z = -1{,}96 \div 1{,}96$. Між $z = -1{,}96$ і $z = 0$ знаходяться 47,5 % спостережень. Таким чином, всього в інтервалі $z = -1{,}96 \div 1{,}96$ знаходиться 95 % спостережень.

Таким чином, всього в інтервалі $z = -1{,}96 \div 1{,}96$ **знаходиться 95 % спостережень**

Якщо необхідно знайти, яка частина спостережень має z-оцінки в інтервалі $z = 0{,}5 \div 1{,}0$, то необхідно спочатку знайти відсоток спостережень в інтервалі $z = 0 \div 1{,}0$, а потім з нього відняти відсоток спостережень, які знаходяться в інтервалі $z = 0 \div 0{,}5$. Тобто, отримаємо таке значення:

$$34{,}13\ \% - 19{,}15\ \% = 14{,}98\ \%.$$

Приклад 5.5. Розглянемо розподіл значень коефіцієнтів розумового розвитку (КРР) чоловіків у США. Його середнє становить 100, а стандартне відхилення — 15 пунктів за тестом Векслера [11], тобто $\mu = 100$, $\sigma = 15$.

Таким чином, якщо КРР = 115, то його $z = 1{,}0$, а якщо КРР = 130, то його $z = 2$. Процентильний ранг для КРР = 115 можна визначити так: в інтервалі $z = 0 \div 1{,}0$ знаходяться 34,13 % оцінок, а зліва від $z = 0$ знаходяться 50 % оцінок. Таким чином, процентильний ранг для КРР = 115 становить:

$$\text{Пр}(115) = 50{,}0\ \% + 34{,}13\ \% = 84{,}13\ \%.$$

Цей результат свідчить про те, що всі чоловіки, які мають КРР = = 115, перевищують оцінки приблизно 84 % інших індивідуумів.

Для КРР = 70 z-оцінка складає $z = -2$. Тобто ті чоловіки, які отримали 70 пунктів за тест, перевищують оцінки 2 % індивідуумів, тобто вони мають 2-й процентиль.

Як визначити, скільки чоловіків мають КРР в інтервалі КРР = = 90÷110? Фактично, це задача визначення частки відносних площ, z-оцінок, що лежать в інтервалі від $z = -0{,}67$ ($z_{90} = (90 - 100)/15 \approx \approx -0{,}67$) до $z = +0{,}67$ ($z_{110} = (110 - 100)/15 \approx 0{,}67$). Це буде

$$x_{90-110} = 24{,}86\ \% + 24{,}86\ \% = 49{,}72\ \%.$$

Визначення z-оцінок на основі процентильних рангів

Ця задача є оберненою до попередньої. Припустимо, що необхідно встановити, яка z-оцінка має процентильний ранг 35. Це означає, що 35 % площі під графіком розташовано зліва від шуканої оцінки і 15 % між цією оцінкою і $z = 0$. Найближчим до 0,15 табличним значенням є 0,1517, а відповідна йому z-оцінка становить: $z = 0{,}39$.

Практичне значення цієї задачі може бути наступним. Припустимо, що ми маємо інформацію тільки щодо 10 % спостережень, які є найбільшими за своєю величиною. Відомо, що ці 10 % спостережень належать нормальному розподілу із середнім $\mu = 80$ і стандартним відхиленням $\sigma = 5$. Необхідно визначити z-оцінку граничної (зліва) точки. Тобто, необхідно знайти z-оцінку для процентильного рангу 90. Між $z = 0$ та шуканою оцінкою знаходиться 40 % площі. Найближчим до 0,4 значенням є 0,3997. Йому відповідає $z = 1{,}28$, що становить 1,28 стандартного відхилення, а шуканим значенням граничної точки буде

$$x_{\text{гран}} = 80 + 1{,}28 \cdot 5 = 86{,}4.$$

Застосування таблиці відносних площ нормального розподілу

Розташування спостережень у нормальному розподілі може бути визначене за допомогою значень середнього та стандартного відхилень. Дійсно, можна дати повне описання нормально розподілених спостережень тільки за допомогою μ і σ.

Так, педагоги і психологи часто застосовують тести, результати яких потім узагальнюються на велике число спостережень. Розподіл отриманих оцінок дуже часто буває нормальним. Розраховані значення середнього та стандартного відхилень фіксують і в подальшому використовують для визначення точного розташування будь-якого значення оцінки в розподілі. Процедура полягає в перетворенні спостереження в z-оцінку з наступним визначенням положення цієї оцінки у нормальному розподілі за допомогою таблиці.

Якщо таблицю ввести в базу даних, то процес статистичного аналізу даних можна автоматизувати і, таким чином, суттєво прискорити його реалізацію. Таблиця може бути також частиною системи підтримки прийняття рішень при виконанні аналізу даних.

5.13. Поняття про теорему Ляпунова. Формулювання центральної граничної теореми (ЦГТ)

Широке розповсюдження нормально розподілених величин на практиці можна пояснити за допомогою теореми 5.2 (*теореми Ляпунова*):

Теорема 5.2. *Якщо випадкова величина X є сумою дуже великого числа взаємно незалежних випадкових величин і при цьому вплив кожної з них на всю суму є незначним, то X має розподіл, близький до нормального.*

Має місце теорема 5.3.

Теорема 5.3. *Нехай із нескінченної сукупності статистичних даних сформовано деяку множину випадкових вибірок і для кожної вибірки знайдено суму цих даних та з отриманих сум сформовано достатньо великий новий ряд розподілу. Тоді отриманий ряд розподілу буде близький до нормального.*

Приклад 5.6. На функціонування людського організму впливає багато факторів — температура навколишнього середовища, атмосферний тиск, хімічний склад їжі, час приймання їжі та її об'єм, харак-

тер спілкування з іншими індивідуумами, шум, ситуація з транспортом та ін. Кожний з цих факторів спричиняє деякі незначні порушення функціонування нервової системи та організму загалом. Однак, оскільки загальне число цих факторів є великим, то їх сукупна дія породжує помітний сумарний вплив. Цей вплив проявляється у втомі, дратівливості, зниженню працездатності і т. ін.

Таким чином, ми можемо розглядати сумарний негативний вплив на функціонування організму як суму великого числа взаємно незалежних часткових впливів. Це дає підставу стверджувати, що сумарний негативний вплив (якщо його коректно виміряти) має розподіл, близький до нормального.

Подібних прикладів утворення нормального розподілу можна навести безліч. Класичним прикладом є стрільба по мішені. Точність попадання залежить від освітленості місця, наявності вітру, стану зору, випадкових коливань карабіна, випадкових звуків, тіней від навколишніх дерев, будівель і т. ін.

Наведемо формулювання центральної граничної теореми, яка встановлює умови, при яких сума великого числа незалежних доданків має близький до нормального розподіл.

Теорема 5.4 (Центральна гранична теорема). Нехай X_1, X_2, ..., X_n — *послідовність незалежних випадкових величин, кожна з яких має скінченне математичне сподівання і дисперсію*:

$$E[X_k] = \mu_k, \ \text{var}\,[X_k] = \sigma_k^2$$

і

$$S_n = X_1 + X_2 + \ldots + X_n, \ \ A_n = \sum_{k=1}^{n} \mu_k, \ \ B_n^2 = \sum_{k=1}^{n} \sigma_k^2$$

та

$$F_n(x) = P\left(\frac{S_n - A_n}{B_n} < x \right)$$

є функція розподілу нормованої суми.

Тоді для будь-якого значення x функція розподілу нормованої суми $F_n(x)$ *прямує до функції нормального розподілу при* $n \to \infty$, *тобто*

$$P\left(\frac{S_n - A_n}{B_n} < x \right) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\left(-z^2/2\right)} dz \qquad (5.13.1)$$

при $n \to \infty$.

Має місце теорема 5.5.

Теорема 5.5. (ЦГТ для вибіркових середніх). *Середні значення деякої достатньо великої множини випадкових вибірок, сформованих з нескінченної сукупності спостережень, мають розподіл, близький до нормального.*

Нагадаємо, що кожна вибірка повинна складатися з однакового числа елементів. Вибірковий розподіл середніх є нормований розподіл сум, а тому він точно відображає їх вибірковий розподіл. Це можна довести наступним чином: відомо, що ділення кожного елемента ряду розподілу на одну й ту саму константу не впливає ні на *z*-оцінку, ні на процентильний ранг елементів. Таким чином, операція ділення всіх N елементів розподілу (в складі суми) на N залишає без зміни пари значень *z*-оцінок і процентильних рангів. З цього випливає, що вибірковий розподіл середніх також буде нормальним.

5.14. Властивості ряду розподілу середніх

Припустимо, що розподіл, сформований із середніх, є нескінченним рядом. Це означає, що із початкової (генеральної) сукупності було утворено нескінченне число вибірок.

Існують важливі зв'язки між значеннями *середнього і стандартного відхилення генеральної сукупності*, а також *середнім і стандартним відхиленням розподілу середніх*, взятих з цієї сукупності. При цьому середнє розподілу середніх дорівнює середньому генеральної сукупності.

Теорема 5.6. *Середнє ряду розподілу, утвореного із середніх значень великої кількості випадкових вибірок, взятих з однієї нескінченної сукупності, дорівнює середньому генеральної сукупності.*

Таким чином, $\mu_{рс} = \mu_{гс} = \mu$, де $\mu_{рс}$ — середнє ряду середніх; $\mu_{гс} = \mu$ — середнє генеральної сукупності.

Зв'язок стандартного відхилення середніх із стандартним відхиленням генеральної сукупності відображає наступна теорема.

Теорема 5.7. *Ряд розподілу, утвореного із середніх значень великої кількості випадкових вибірок обсягу N, взятих з однієї нескінченної сукупності, має стандартне відхилення, яке дорівнює стандартному відхиленню генеральної сукупності, поділеному на* $\sqrt{N}$, *тобто*

$$\sigma_{рс} = \frac{\sigma_{гс}}{\sqrt{N}} = \frac{\sigma}{\sqrt{N}}, \qquad (5.14.1)$$

де $\sigma_{гс} = \sigma$ *— стандартне відхилення генеральної сукупності, з якої взято велике число вибірок потужністю N.*

Розглянемо, наприклад, коефіцієнти розумового розвитку (КРР) чоловіків, які проживають в місті Києві. Середнє цієї сукупності $\mu = 100$, а стандартне відхилення $\sigma = 15$. Припустимо, що з цієї генеральної сукупності взято велике (нескінченне) число вибірок по 10 елементів у кожній. Сформуємо новий розподіл із середніх цих вибірок і знайдемо стандартне відхилення вибіркового розподілу середніх:

$$\sigma_{рс} = \frac{\sigma}{\sqrt{N}} = \frac{15}{\sqrt{10}} = 4{,}7.$$

Графік розподілу середніх є більш компактним (рис. 5.2), оскільки стандартне відхилення середніх $\sigma_{рс} = 4{,}7$, а стандартне відхилення генеральної сукупності $\sigma = 15$.

Інакше кажучи, середні значення вибірок сильніше групуються навколо точки рівноваги, ніж індивідуальні коефіцієнти розумового розвитку. Таким чином, середні значення вибірок у меншій мірі відрізняються одне від одного, ніж індивідуальні КРР.

Це явище можна пояснити тим, що завдяки усередненню відбувається процес "балансування" вибірки, в результаті якого середнє вибірки приймає значення, яке наближається до 100, тобто до середнього генеральної сукупності. При цьому, чим більшим буде обсяг вибірки, тим ближчими одне до одного будуть значення середніх, тобто зменшується стандартне відхилення розподілу середніх $\sigma_{рс}$.

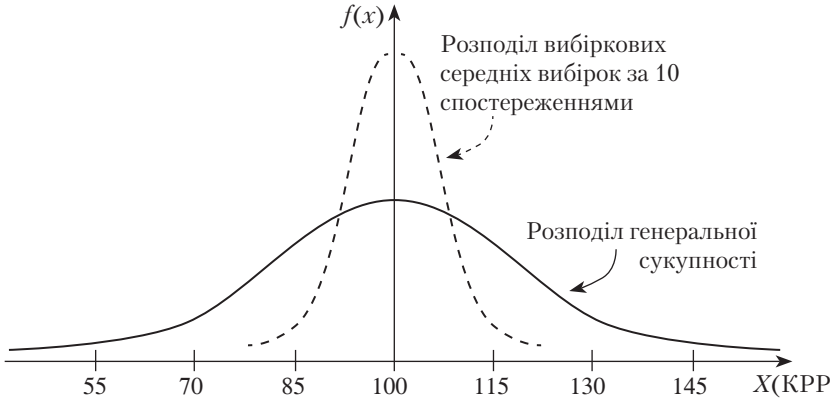


Рис. 5.2. **Криві розподілу генеральної сукупності і розподілу середніх**

На рис. 5.3 показано три криві розподілу. Перша крива (I) — це крива розподілу індивідуальних КРР із середнім 100 і стандартним відхиленням 15. Друга крива (II) є вибірковий розподіл середніх значень вибірок при $N = 100$, взятих з генеральної сукупності КРР. Цей розподіл має $\mu_{рс} = 100$ і $\sigma_{рс} = 4{,}5$.

Природно, що цей розподіл є компактнішим порівняно з розподілом індивідуальних КРР. Крива III ілюструє розподіл середніх значень вибірок обсягом $N = 100$. Стандартне відхилення становить $\sigma_{рс} = 1{,}5$. Цей розподіл самий "гострий", тобто середні значення вибірок обсягом $N = 100$, кожна, є дуже близькими одне до одного.

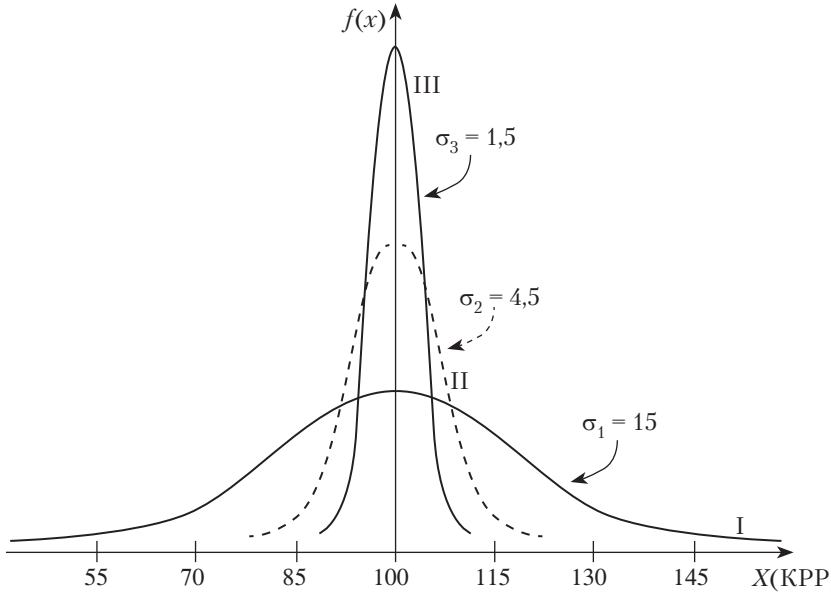


Рис. 5.3. **Криві розподілу при різних стандартних відхиленнях**

5.15. Процентильний ранг і *z*-оцінки вибіркового розподілу середніх

Як розраховуються *z*-оцінки і процентильний ранг окремого елемента нормального розподілу вже відомо. Розглянемо тепер, як визначити *z*-оцінку і процентильний ранг окремого значення нормального розподілу середніх. Будемо вважати, що обсяги вибірок є достатньо великими для того, щоб розподіл середніх був нормальним.

Також припустимо, що відомі середнє і стандартне відхилення (СВ) для генеральної сукупності, з якої сформовані вибірки. Доступною є тільки одна випадкова вибірка з генеральної сукупності, для якої можна розрахувати середнє.

Необхідно розрахувати z-оцінку і процентильний ранг значення середнього для доступної вибірки з вибіркового розподілу середніх, не розглядаючи інших вибірок.

Припускаємо, що вибірковий розподіл середніх є нормальним. На першому кроці можна розрахувати середнє і СВ вибіркового розподілу середніх, а потім можна перейти до розрахунку *z*-оцінки і процентильного рангу.

Повернемось до прикладу з розподілом КРР, для якого $\mu = 100$, $\sigma = 15$. Нехай з цієї сукупності отримано нескінченне число випадкових вибірок обсягом $N = 25$. Із середніх цих вибірок сформовано деякий розподіл. Але ми маємо інформацію тільки щодо одного розподілу і визначили, що її середнє $\mu_i = 106$. Необхідно знайти *z*-оцінку і процентильний ранг вибіркового середнього $\mu_i = 106$ у вибірковому розподілі середніх.

Знайдемо стандартне відхилення вибіркового розподілу:

$$\sigma_i = \frac{\sigma}{\sqrt{N}} = \frac{15}{\sqrt{25}} = 3.$$

Перший крок у знаходженні *z*-оцінки деякого вибіркового середнього полягає у знаходженні відстані від цього вибіркового середнього до точки рівноваги. Оскільки вибіркове середнє $\mu_i = 106$, то відхилення від точки рівноваги $\mu = 100$ складе: $\mu_i - \mu = 106 - 100 = 6$, а $\sigma_i = 3$. Отже, *z*-оцінка вибіркового середнього буде така:

$$z_i = \frac{\mu_i - \mu}{\sigma_i} = \frac{106 - 100}{3} = 2.$$

За допомогою *z*-оцінки можна визначити процентильний ранг вибіркового середнього, яке дорівнює 106. З таблиці відносних площ нормального розподілу знаходимо, що елемент із *z*-оцінкою $z_i = 2$ має процентильний ранг $\text{Пр}(z_i = 2) \approx 98$. Тобто, вибіркове середнє, $\mu_i = 106$, перевищує приблизно 98 % значень вибіркових середніх.

Наведемо ще одну ілюстрацію процедури визначення процентильного рангу. Нехай середнє і стандартне відхилення КРР для генеральної сукупності чоловіків призовного віку становлять: $\mu = 100$, $\sigma = 15$. Після чергового призову випадково сформовано множину підрозділів по 400 солдатів у кожному.

Середній КРР для деякого підрозділу *B* склав $\sigma_i = 99{,}25$. Необхідно знайти місце, яке займає за цим показником підрозділ *B* серед інших підрозділів. Інакше кажучи, який процент всіх інших підрозділів має середній КРР нижчий 99,25.

Стандартне відхилення середніх:

$$\sigma_{\text{рс}} = \frac{\sigma}{\sqrt{N}} = \frac{3}{4} = 0{,}75.$$

а його *z*-оцінка: $z_i = \dfrac{\mu_i - \mu}{\sigma} = \dfrac{99{,}25 - 100}{0{,}75} = \dfrac{-0{,}75}{0{,}75} = -1.$

З таблиці відносних площ нормального розподілу знаходимо, що елемент розподілу із $z_i = -1$ перевищує приблизно 16 % елементів цього розподілу. Таким чином, середній ККР підрозділу *B* має процентильний ранг 16. Фактично, ми виконали порівняння характеристик деякої випадкової вибірки з іншими вибірками за допомогою *z*-оцінки її середнього з наступним перетворенням середнього у процентильний ранг.

5.16. Ймовірність і нормальний розподіл

Знання характеристик нормального розподілу дає можливість зв'язати ймовірнісні судження з даним видом розподілу. Розглянемо, наприклад, ймовірність того, що елемент нормального розподілу буде мати *z*-оцінку більше нуля.

У нормальному розподілі ймовірність такого результату дорівнює частині елементів генеральної сукупності, які мають $z > 0$. Таким чином, ймовірність того, що вибраний елемент розподілу буде мати $z > 0$, буде дорівнювати 0,5, тобто $P(z_i > 0) = 0{,}5$.

Визначимо, яка буде ймовірність того, що *z*-оцінка вибраного *i*-го елемента буде знаходитись в інтервалі між –1 і +1, тобто $P(-1 < z_i < +1) = ?$ Для цього необхідно знайти частину елементів нормального розподілу, які знаходяться в інтервалі між –1 і +1. З таблиці відносних площ нормального розподілу знаходимо, що в цьому інтервалі знаходяться приблизно 68 % всіх елементів. Тобто ймовірність $P(-1 < z_i < +1) = 0{,}68$.

Відповідно, ймовірність того, що випадково вибраний елемент нормального розподілу не попадає у вказаний інтервал, становить 0,32.

Раніше було встановлено, що приблизно 95 % елементів нормального розподілу знаходяться в інтервалі між –2 і +2. Таким чином,

ймовірність того, що z-оцінка випадково вибраного елемента знаходиться в інтервалі між значеннями –2 і +2, становить 0,95.

Приблизно 2,5 % елементів розподілу мають оцінки, які перевищують +2 (вони знаходяться у правому хвості графіка розподілу). Так само, приблизно 2,5 % елементів розподілу мають z-оцінки, які є меншими, ніж –2 (вони знаходяться у лівому хвості графіка розподілу).

Розглянутий підхід до статистичного аналізу даних можна використати при розв'язуванні прикладних задач. Припустимо, що зріст чоловіків, які проживають у великому місті, має нормальний розподіл із середнім 175 см і стандартним відхиленням 10 см. Вибирається випадково чоловік із даної генеральної сукупності. Яка ймовірність того, що його зріст має значення в інтервалі: 175÷185 см?

Трансформуємо інтервал 175÷185 см в z-оцінки. Значенню 175 см відповідає $z = 0$. Значення 185 см має $z = 1$. Таким чином, необхідно визначити ймовірність попадання z-оцінки випадково вибраного елемента розподілу в інтервал $z = 0 \div 1$. У таблиці відносних площ нормального розподілу знаходимо, що ймовірність такого випадку становить 34/100 або 0,34, тобто

$$P(0 < z_i < 1) = 0{,}34.$$

Можна визначити також ймовірність того, що випадково вибраний чоловік матиме зріст більше ніж 185 см. Інакше кажучи, необхідно визначити ймовірність того, що його z-оцінка буде перевищувати 1. Оскільки ми встановили ймовірність попадання z-оцінки в інтервал $z = 0 \div 1$, то шукана ймовірність буде дорівнювати:

$$P(z_i > 1) = 0{,}5 - 0{,}34 = 0{,}16,$$

або ж значення 0,16 можна знайти безпосередньо з таблиці.

Останнє завдання полягає у визначенні ймовірності того, що зріст випадково вибраного чоловіка буде менш як 155 см чи більше ніж 195 см. Після трансформування значень в z-оцінки задача полягає у наступному: визначити ймовірність того, що z-оцінка випадково вибраного елемента буде меншою –2 або більшою +2. Ймовірність такої ситуації становить приблизно 5/100 або 0,05.

Задача визначення ймовірності середнього

При виконанні експериментів вибірка складається, як правило, з великого числа елементів, вибраних випадково з деякої сукупності.

У процесі аналізу даних знаходять середнє цієї сукупності. Як і раніше, будемо вважати, що при виконанні експерименту був визначений деякий інтервал, який є цікавим для дослідження. Задача полягає у тому, щоб визначити ймовірність попадання вибіркового середнього у визначений інтервал.

Розглянемо дані з попереднього прикладу. Середнє розподілу значень зросту чоловіків становить 175 см, а стандартне відхилення — 10 см. Нехай випадкова вибірка містить 100 спостережень і необхідно знайти середній зріст чоловіків, які попали в цю вибірку. Яка ймовірність того, що середній зріст чоловіків з цієї вибірки знаходиться між $171\frac{2}{3}$ см і $178\frac{1}{3}$ см?

Таким чином, необхідно визначити ймовірність того, що середнє вибірки із 100 спостережень у зазначений інтервал. Іншими словами, якщо сформувати велике число однакових за обсягом вибірок, то яка частина цих вибірок буде мати середнє, що знаходиться всередині вказаного інтервалу? Для того щоб відповісти на це питання, уявимо, що із деякої сукупності формується нескінченне число вибірок по 100 значень. Розподіл середніх цих вибірок буде нормальним, а його стандартне відхилення:

$$\sigma_{\text{рс}} = \frac{\sigma}{\sqrt{N}} = \frac{10}{\sqrt{100}} = 1 \text{ см}.$$

Необхідно визначити ймовірність того, що середнє деякої випадкової вибірки знаходиться у межах $171\frac{2}{3} \div 178\frac{1}{3}$. У розподілі середніх значенню $171\frac{2}{3}$ см відповідає $z(171\frac{2}{3}) = -3{,}3$, а значенню $178\frac{1}{3}$ см відповідає $z(178\frac{1}{3}) = 3{,}3$.

З таблиці відносних площ нормального розподілу знаходимо, що випадкова вибірка попаде у визначений інтервал з імовірністю 0,999, тобто майже з одиничною ймовірністю.

Зазначимо, що вибіркове середнє, яке дорівнює 176 см, має $z(176) = 1{,}0$. З таблиці знаходимо, що наближено 34 % вибіркових середніх даного розподілу знаходяться між значеннями 175 см і 176 см. Отже, ймовірність того, що середній зріст випадкової вибірки чоловіків розташований між 175 см і 176 см, становить приблизно 0,34.

Тепер розглянемо приклад, який стосується задачі середньої тривалості демонстрації фільмів. Припустимо, що середня тривалість де-

монстрації фільмів, випущених за останнє десятиліття, становить 100 хв, а стандартне відхилення $\sigma = 24$ хв.

Якщо вибрати випадково 36 різних фільмів, то якою буде ймовірність того, що середня тривалість фільму з цієї вибірки буде менш як 92 хв або більше ніж 108 хв?

Припустимо, що сформовано множину вибірок обсягом по 36 фільмів кожна і визначено середню тривалість фільму для кожної вибірки. Із середніх значень сформовано ряд розподілу. Отриманий розподіл середніх є нормальним із середнім $\mu_{pc} = 100$ і стандартним відхиленням

$$\sigma_{pc} = \frac{\sigma}{\sqrt{N}} = \frac{24}{\sqrt{36}} = 4.$$

Цей розподіл представлено на рис. 5.4.

Значенню 92 відповідає $z(92) = -2$, а значення 108 має $z(108) = +2$. Необхідно визначити ймовірність того, що взяте випадково вибіркове середнє буде менш як 92 або більше ніж 108. Інакше кажучи, необхідно визначити ймовірність того, що взяте випадково вибіркове середнє буде мати $z_i < -2$ або $z_i > +2$. Така ймовірність становить приблизно 5/100 або 0,05. Тобто тільки у 5 вибірках із 100 середнє буде відрізнятись від середнього генеральної сукупності більше ніж на 8 хв.

Розглянута задача може бути сформульована також наступним чином: яка ймовірність того, що середня тривалість фільму для деякої вибірки потужністю 36 елементів буде відрізнятись від середньої тривалості демонстрації всіх випущених фільмів (100 хв) щонайменше на 8 хв? Очевидно, що відповідь буде такою самою, як і в попередньому формулюванні задачі.

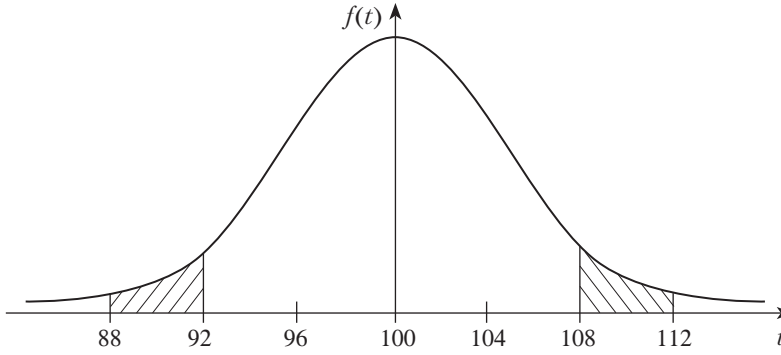


Рис. 5.4. **Крива розподілу для задачі з фільмами**

5.17. Квантилі нормованого розподілу

Квантилі — це ще одна назва z-оцінок випадкових змінних. На практиці квантилі використовують у вигляді зміщених z-оцінок, які розглядаються нижче. Квантилі нормованого нормального розподілу визначають з рівняння:

$$\Phi(z_p) = p,\ 0 \le p \le 1, \tag{5.17.1}$$

де Φ — функція Лапласа.

Значення z_p є аргументом функції Лапласа, який відповідає ймовірності p. Квантиль z_p зростає від $-\infty$ до $+\infty$ при збільшенні p від нуля до 1; при $p = 0{,}5$ значення $z_p = 0$.

Та обставина, що z_p приймає як додатні, так і від'ємні значення, часто веде до труднощів з обчисленнями. Цих труднощів можна уникнути, якщо вибрати іншу початкову точку відліку значень квантилів. Оскільки z надзвичайно рідко приймає значення менші –5, то можна встановити початкову точку відліку $\alpha = 5$. Таким чином, на практиці користуються зміщеними квантилями $z_p + 5$, тобто:

$$\text{Зміщений квантиль} = \text{квантиль} + 5. \tag{5.17.2}$$

Квантилі x_p випадкової величини X із середнім μ і стандартним відхиленням σ можна знайти із рівняння:

$$\Phi\left(\frac{x_p - \mu}{\sigma}\right) = p \tag{5.17.3}$$

так як

$$z_p = \frac{x_p - \mu}{\sigma}, \tag{5.17.4}$$

тоді

$$x_p = \mu + z_p\sigma. \tag{5.17.5}$$

5.18. Контрольні питання і вправи

1. Який розподіл називають теоретичним?
2. Що означає твердження, що розподіл є нормальним?
3. Яку функцію називають (кумулятивною) функцією розподілу?
4. Які властивості має функція розподілу?
5. Як називають похідну кумулятивної функції розподілу? Запишіть цю похідну для нормального розподілу.
6. Дайте означення математичного сподівання неперервної випадкової величини і запишіть вираз для його обчислення.

7. Дайте означення дисперсії неперервної випадкової величини і запишіть вираз для її знаходження.
8. Запишіть і поясніть статистичну модель (щільність) нормального розподілу.
9. Запишіть і поясніть застосування інтеграла Пуассона.
10. Запишіть і поясніть вираз для обчислення нормованої щільності нормального розподілу. У чому полягає перевага його використання?
11. Як обчислити ймовірність попадання нормованої нормальної величини X в інтервал $(0, x)$?
12. Зробіть дослідження кривої нормального розподілу на екстремум.
13. Як впливають зміни значення математичного сподівання випадкових величин, що утворюють розподіл, на форму нормального розподілу.
14. Як впливають зміни значення стандартного відхилення випадкових величин, що утворюють розподіл, на форму нормального розподілу?
15. За якою формулою знаходиться ймовірність попадання нормованої випадкової величини у заданий інтервал? Наведіть приклад обчислення цієї ймовірності.
16. Як обчислити ймовірність отримання випадковою величиною заданого значення відхилення?
17. Сформулюйте правило трьох сигм. Яка ймовірність того, що абсолютна величина відхилення випадкової величини від середнього перевищить три стандартних відхилення?
18. Який розподіл називають емпіричним? Поясніть на прикладі.
19. Який розподіл називають теоретичним і чому?
20. Поясніть такі статистичні характеристики нормального розподілу, як асиметрія і ексцес.
21. Запишіть і поясніть вираз для статистики Жак-Бера. Які значення може приймати ця статистика і як їх інтерпретують?
22. Скільки спостережень знаходиться в інтервалі значень z-оцінок: $-1{,}96 \le z \le 1{,}96$?
23. Як визначити z-оцінку за допомогою процентильних рангів?
24. Поясніть на прикладі можливість практичного використання площ нормального розподілу.
25. Сформулюйте центральну граничну теорему.

26. Сформулюйте центральну граничну теорему для вибіркових середніх.
27. Чому дорівнює середнє ряду розподілу, утвореного із середніх значень великої кількості випадкових вибірок, взятих з однієї нескінченної сукупності.
28. Як знайти стандартне відхилення ряду, утвореного із середніх значень великої кількості випадкових вибірок потужності N, взятих з однієї нескінченної сукупності?
29. Як розрахувати z-оцінки і процентильний ранг значення середнього для доступної вибірки з вибіркового розподілу середніх, не розглядаючи інші вибірки?
30. Як знайти ймовірність того, що елемент нормального розподілу буде мати z-оцінку більше ніж нуль?
31. Як визначаються квантилі нормального розподілу?

Розділ 6

АНАЛІЗ ПРОЦЕСУ ПРИЙНЯТТЯ РІШЕНЬ, РИЗИК І ПЕРЕВІРКА ГІПОТЕЗ

6.1. Вступ

Прийняття рішень та виконання дій відповідно до них — основний напрям нашої цілеспрямованої (розумної) діяльності. Причини, які спонукають прийняти те чи інше рішення, можуть бути навіть не завжди цілком зрозумілими і наші дії можуть здаватись рефлекторними. Це справедливо, наприклад, у випадку, коли водій чує позаду шум потужного автомобіля, що наздоганяє його, і повертає вправо, ближче до узбіччя. Така ситуація має місце у випадку, коли ми вибираємо вранці одяг або думаємо, де можна пообідати. Однак у більш серйозних ситуаціях виникає необхідність докладного вивчення проблеми, аналізу можливих варіантів розв'язку задачі та наслідків вибраних варіантів.

Важливу роль у прийнятті багатьох рішень відіграють *теорія ймовірностей* і *математична статистика*, які визначають процедури аналізу доступних даних та прийняття рішень на основі виконаного аналізу.

При прийнятті рішень необхідно приймати до уваги два таких моменти:

- необхідно визначити ймовірність того, що висновки, на основі яких ми збираємось виконати інші дії, є правильними;
- необхідно визначити можливі варіанти розв'язання задачі і зважити можливі наслідки наших дій.

Статистика дає можливість визначити ймовірність того, що наші висновки виявились правильними. Однак за її допомогою, як правило, не можна встановити ціну допущеної помилки, якщо прийняте рішення виявиться невірним.

Якщо існує ряд альтернатив щодо розв'язання задачі, то часто можна припустити, що помилка при прийнятті деякого рішення буде мати приблизно такі самі наслідки, як і при прийнятті будь-якого іншого рішення. У такому випадку розумніше всього зупинитись на рі-

шенні, яке має найбільшу ймовірність того, що воно виявиться правильним.

Іноді помилка у прийнятті деякого рішення може обійтись нам набагато дорожче, ніж наслідки помилки, допущеної при прийнятті інших рішень. У такому випадку дії, які потенційно можуть обійтись нам дуже дорого, повинні реалізовуватись тільки у тому випадку, якщо ймовірність того, що вони виявляться єдино правильними, є досить високою.

Звідси можна зробити висновок, що одне і те саме значення ймовірності дає змогу виконати деякі визначені дії в одній ситуації і не дає достатнього підґрунтя для їх виконання в іншій. Наприклад, гравець на біржі, який має солідний пакет акцій, може дозволити собі купити цінні папери, знаючи, що ймовірність підвищення їх ціни найближчим часом становить 0,7. Однак, уявімо собі, що деякий індивідуум обвинувачується в убивстві і свідоцтва проти нього такі, що в аналогічній ситуації 7 із 10 звинувачуваних виявлялись дійсно винними. Таким чином, ймовірність того, що даний звинувачений є також винним, становить 0,7. Однак присяжні в подібних ситуаціях проголосують, скоріш за все, за оправдання.

Причиною прийняття такого рішення є характер наслідків, до яких можуть призвести дії, які є можливими у цій ситуації. Помилка, яка є наслідком винесення несправедливого вироку, буде коштувати життя невинній людині, що значно перевищує негативні наслідки помилки, припущеної у випадку оправдання злочинця. Враховуючи подібну нерівноцінність наслідків, вирок буде, скоріш за все, буде виправдовувальним, хоча ймовірність справедливого звинувачення підсудного становить 0,7.

Наступний простий ілюстративний приклад дасть можливість нам ще більше зрозуміти важливість різних ставок у процесі прийняття рішень. Нехай десятирічний хлопчик попав у чуже місто і для того, щоб доїхати додому, йому потрібно 10 грн, але він має тільки 9. Він занадто гордий, щоб просити гривню у незнайомих людей, а тому сердитий та сумний прямує до передмістя. Хлопчикові завжди подобалось грати на невеликі суми грошей за допомогою монетки. Коли він зустрів по дорозі ровесника, то сказав йому: "Я підкину монету, а ти вгадай, що випаде — герб чи номінал. Хто виграє, той отримає гривню". Однак, зустрічний ровесник відмовився грати, так само як і всі інші.

Хлопчик вже зібрався шукати місце для ночівлі, але тут йому в голову прийшов сміливий план. Останній із ровесників, які зустрілись

йому, здавалось, трохи вагався щодо пропозиції зіграти, і хлопчик звернувся до нього ще раз.

На цей раз його пропозиція була такою: “Якщо ти виграєш, то отримаєш 9 грн, а якщо програєш, то віддаси мені тільки одну гривню.” Пропозиція була прийнята — 9 грн хлопчика і 1 грн його ровесника були покладені на камінь. Монетку підкинуто вверх і ровесник крикнув: “Герб!” Але монетка впала номіналом вгору і хлопчик швидко забрав усі 10 грн і побіг до автобуса.

Рішення хлопчика ризикнути своїми 9 грн було цілком вірним. У звичайних умовах воно було б авантюристичним, але життя наповнило його реальністю. Ймовірність того, що хлопчик виграє парі, становила 0,5, але, з його точки зору, цього було цілком достатньо. Втрата 9 грн ніяк не погіршувала ситуацію, в якій він опинився, але виграш гривні дозволив йому поїхати додому. Звичайно, що умови гри цілком влаштовували і його партнера, оскільки у випадку виграшу він отримував 9 грн, ризикуючи тільки однією.

У сфері бізнесу також часто виникають ситуації, коли ризикові дії є вигідними для обох партнерів, оскільки ставки, якими вони ризикують, мають різне значення. Аналіз проблем, що стосуються подібних аспектів, приводить у світ азартних ігор та фінансових операцій. *Наша мета у даному випадку полягає у тому, щоб підкреслити значення фактора величини та важливості ставки у конкретній ситуації.* Вплив цих факторів завжди великий, навіть у науковій галузі.

В експериментальній роботі необхідність врахування фактора ризику відчувається постійно. Так, великі ризики пов’язані з космічними польотами, але необхідність дослідження космосу і тяжіння людини до нових знань примушують ризикувати. Наприклад, можливі тяжкі наслідки лікування хворого деякими ліками, але, водночас, вони для нього необхідні. Ризик, з одного боку, полягає у тому, що лікар може припуститись помилки з тяжкими наслідками. З іншого боку, ризик виникає внаслідок того, що ми не використовуємо можливість вилікувати тяжку хворобу наявними ліками. Загалом, у процесі експериментування не можна виключити повністю появу помилок. Фактично, чим більше експериментатор зменшує ймовірність появи однієї помилки, тим більше підвищується ймовірність появи іншої. Необхідно докладно вивчати можливі наслідки всіх можливих рішень (наслідки помилок) і тільки після цього остаточно вибирати конкретні дії.

6.2. Основні принципи перевірки гіпотез

Вірогідність тієї чи іншої гіпотези необхідно оцінювати у будь-якій галузі науки. Гіпотези щодо вірогідності тих чи інших припущень підтверджують, як правило, експериментально.

При прийнятті рішень завжди розглядають дві можливості. Припустимо, що наше рішення полягає у тому, що ми відхиляємо гіпотезу.

1. Перша можливість полягає у тому, що наше рішення виявилось правильним. Це означає, що в дійсності висунута (сформульована) гіпотеза невірна і повинна бути відхилена.

2. Друга можливість полягає у тому, що наше рішення відхилити гіпотезу виявилось помилковим. У дійсності висунута гіпотеза виявилась вірною, але результат, отриманий на основі деякої експериментальної вибірки, ввів особу, яка приймає рішення (ОПР), в оману. Таку помилку називають *помилкою першого роду*.

Означення 6.1. *Якщо ОПР відхиляє вірну гіпотезу, то вона робить помилку першого роду*.

Тепер припустимо, що рішення полягає у прийнятті гіпотези.

1. Перша можливість полягає у тому, що наше рішення виявилось правильним. Це означає, що в дійсності висунута гіпотеза правильна і повинна бути прийнята.

2. Друга можливість полягає у тому, що наше рішення прийняти гіпотезу виявилось помилковим. Гіпотеза невірна і ОПР прийняв її помилково. Таку помилку називають *помилкою другого роду*.

Означення 6.2. *Якщо ОПР помилково приймає невірну гіпотезу, то вона робить помилку другого роду*.

Розглянемо приклад з підкиданням монети. У кожному досліді ми очікуємо один з двох можливих результатів, тобто маємо справу з *дихотомічною змінною*. Якщо монета вважається повноцінною, то ймовірність випадання герба і номіналу однакові: $P(\textit{герба}) = P(\textit{номіналу}) = 0{,}5$. Вид розподілу, що утворюється у результаті ряду послідовних експериментів з монетою, можна визначити за допомогою такої теореми.

Теорема 6.1. *Якщо у випадку дихотомічної змінної ймовірність появи події складає p і при цьому формується нескінченна кількість вибірок однакового обсягу N, то кожна вибірка X такого обсягу асимптотично наближається до нормального розподілу з параметрами*.

$$\mu = Np;\ \ \sigma = \sqrt{Np(1-p)}, \text{ тобто } \ \{X\} \leftrightarrow N\left(Np, Np(1-p)\right).$$

Нехай для прикладу з підкиданням монети $p = 0{,}5$ при $N = 100$. Теорема 6.1 стверджує, що розподіл отриманих результатів нормальний з параметрами:

$$\mu = Np = 100 \cdot 0{,}5 = 50; \quad \sigma = \sqrt{Np(1-p)} = \sqrt{100 \cdot 0{,}5 \cdot (1-0{,}5)} = 5.$$

Припустимо, що після 100 дослідів з підкиданням монети отримано такі результати:

герб — 55 разів; *номінал* — 45 разів.

Після цього ставиться питання: "Чи можна на основі цієї інформації прийняти або відхилити гіпотезу стосовно того, що монета повноцінна?" Тобто H_0: *монета повноцінна*.

Нехай перше правило прийняття рішення полягає в наступному: гіпотеза стосовно повноцінності монети вважається правильною, якщо кількість гербів, що випали в результаті експерименту, знаходиться в інтервалі 10÷90. Відповідно до цього правила гіпотеза стосовно повноцінності відхиляється тільки у тому випадку, коли кількість гербів буде менш як 10 або більше ніж 90. Це правило ілюструє рис. 6.1.

Таким чином, *якщо монета повноцінна, то результат експерименту практично завжди буде знаходитись в інтервалі* 10÷90. Оскільки зона, в якій даний критерій припускає істинність висунутої гіпотези, обмежена у кожний бік величиною 8-ми стандартних відхилень, то результат підкидання повноцінної монети майже ніколи не відхилиться від значення 50 настільки, щоб можна було відхилити гіпотезу відповідно до правила 1. Інакше кажучи, *використання правила* 1 *навряд чи призведе до помилки першого роду*.

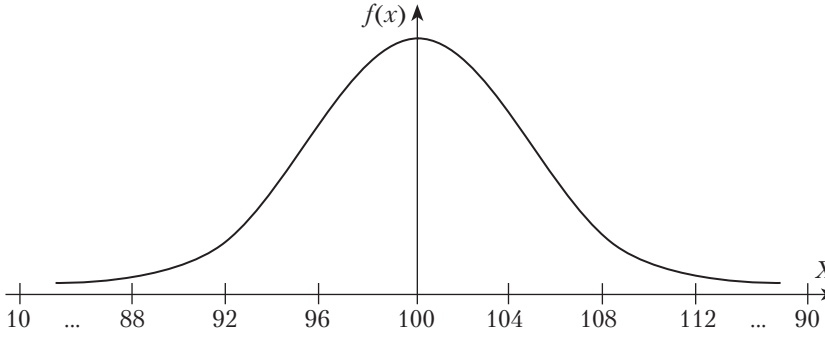


Рис. 6.1. **Ілюстрація першого правила прийняття рішень**

Припустимо, що сформульовано *друге правило*: *монета буде повноцінною, якщо кількість гербів попадає в інтервал* 45÷55. Це означає, що гіпотеза стосовно повноцінності монети буде прийнята, якщо це співвідношення буде лежати в дуже вузьких межах. На рис. 6.2 показано, що відбудеться, якщо при підкиданні монети рішення будуть прийматись на основі другого правила.

За допомогою теореми 6.1 можна розрахувати ймовірність того, що результати підкидання повноцінної монети попадуть у виділений нами інтервал. Цей інтервал обмежений значеннями 45 і 55. Кожне з цих значень відрізняється від середнього, що дорівнює 50, на одне стандартне відхилення. Використовуючи таблицю відносних площ нормального розподілу визначимо, що на область прийняття гіпотези приходиться у даному випадку приблизно 68 % загальної площі графіка. Відповідно, затінена область становить приблизно 32 %. Таким чином, *ймовірність того, що при підкиданні монети зустрінеться вибірка, в якій кількість гербів попаде у виділену область, становить* 0,68.

Важливо підкреслити те, що *ймовірність появи вибірки, в якій кількість гербів попаде в область, де гіпотеза стосовно повноцінності монети відхиляється, становить* 0,32. Тобто, це ймовірність того, що за правилом 2 буде помилково відхилена висунута гіпотеза, навіть якщо вона правильна — це ймовірність помилки першого роду.

Тепер можна зробити *висновок стосовно результатів застосування першого і другого правил* прийняття рішень. Згідно з правилом 1, навряд чи можна буде помилково відхилити висунуту гіпотезу, або зробити помилку першого роду. Згідно з правилом 2, у 32 % випадків гіпотеза щодо повноцінності монети буде відхилятись помилково і,

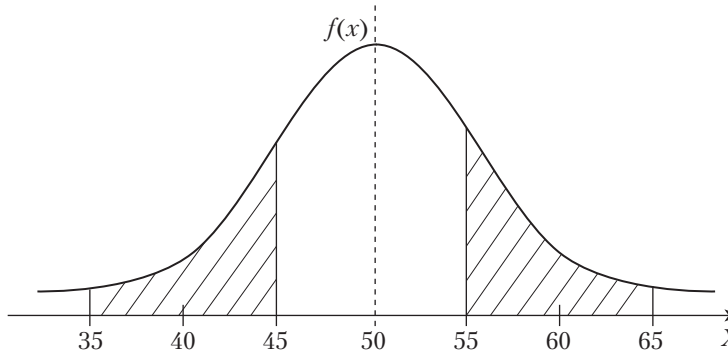


Рис. 6.2. **Ілюстрація другого правила прийняття рішень**

таким чином, ймовірність припущення помилки першого роду, у випадку повноцінності монети, становить приблизно 0,32.

Припустимо тепер, що для експерименту вибрали *неповноцінну монету*, тобто ймовірність випадання герба відрізняється від величини 0,5. Наприклад, ймовірність випадання орла при кожному окремому підкиданні може становити 0,2 або 0,6, або 0,9 (або іншу величину).

Нехай ймовірність випадання герба становить 0,6 для $N = 100$. Відповідно до теореми 6.1 розподіл результатів експерименту має такі середнє і стандартне відхилення:

$$\mu = Np = 100 \cdot 0,6 = 60;$$

$$\sigma = \sqrt{Np(1-p)} = \sqrt{100 \cdot 0,6(1-0,6)} = 4,9.$$

На рис. 6.3 показано розподіл результатів експерименту при використанні неповноцінної монети.

Розглядаючи експеримент з монетою, про неповноцінність якої відомо наперед, можна визначити наслідки застосування правил відносно відхилення гіпотези про повноцінність монети. Нагадаємо, що перше правило таке: *гіпотеза щодо повноцінності приймається, якщо кількість гербів попадає в інтервал* 10÷90. З рис. 6.3 видно, що навіть у випадку використання неповноцінної монети результати практично кожного експерименту будуть попадати в область прийняття гіпотези. Такий самий результат можна знайти за допомогою z-оцінок і таблиць відносних площ нормального розподілу. Інакше кажучи, *у випадку, коли ймовірність випадання герба становить* 0,6, *за першим*

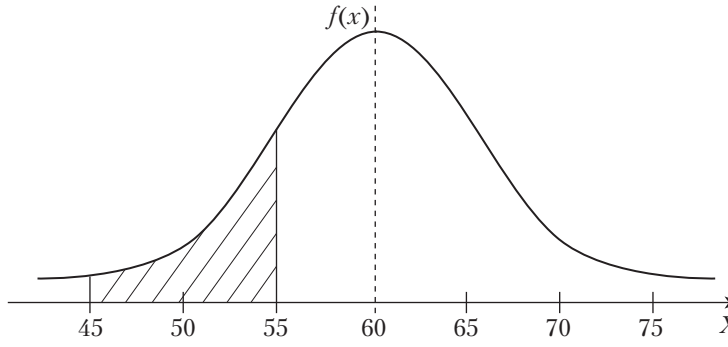


Рис. 6.3. **Розподіл результатів експерименту при використанні неповноцінної монети**

правилом практично завжди буде помилково прийматись висунута гіпотеза. Тим самим робиться помилка другого роду.

За *другим правилом* висунута гіпотеза приймається тільки у тому випадку, коли число гербів, що випадає, попадає в інтервал 45÷55. Прийнявши помилкову гіпотезу (що монета повноцінна), за другим правилом можна припуститись помилки другого роду у випадку, якщо кількість гербів при підкиданні неповноцінної монети буде попадати в указаний інтервал. Затінена площа на рис. 6.3 відображає кількість випадків, у яких буде припущена помилка другого роду за другим правилом. За допомогою *z*-оцінок для чисел 45 і 55 і таблиці відносних площ нормального розподілу знайдемо, що площа затіненої частини графіка становить 15 % від всієї площі під графіком.

Інакше кажучи, *у випадку використання неповноцінної монети* із ймовірністю випадання герба, p(*герба*) = 0,6, ймовірність того, що за другим правилом буде зроблена помилка другого роду, складає:

$$p(\text{помилки другого роду}) = 0{,}15.$$

Звідси можна зробити висновок, що при експериментуванні з неповноцінною монетою ймовірність того, що за другим правилом буде справедливо відхилено висунуту гіпотезу, становить 0,85, тобто:

$$p(\text{справедливого відхилення}) = 0{,}85.$$

Цей факт має важливе значення. Використовуючи одну і ту саму монету, за першим правилом ніколи не можна буде відхилити неправильну гіпотезу, а за другим правилом ймовірність справедливого відхилення неправильної гіпотези становить 0,85. Тобто *потужність другого правила є значною*.

Узагальнення отриманих результатів

Згідно правила 1 деяка гіпотеза відхиляється тільки в рідкісних випадках. Якщо гіпотеза правильна, то за правилом 1 вона ніколи помилково не буде відхилена, тобто не буде припущено помилки першого роду. У той самий час, як за правилом 2 помилку першого роду буде зроблено у 32 випадках із 100.

Якщо ж гіпотеза виявиться неправильною і ймовірність випадання герба при підкиданні монети становить 0,6, то за правилом 1 практично в усіх випадках буде робитись помилка другого роду, тобто прийматись неправильна гіпотеза. Критерій цього правила є значно меншим, ніж критерій правила 2, яке дає можливість зменшити ймо-

вірність припущення помилки другого роду до 0,15. Потужність цього критерію становить для описаного випадку 0,85.

Необхідно пам'ятати, що отримані числові результати справедливі для конкретного характеру неповноцінності монети. Вище було розглянуто випадок, коли ймовірність випадання герба становить 0,6. Якби розглядався інший випадок, то ми прийшли б до іншого результату, при якому ризик прийняття неправильної гіпотези мав би інше значення.

Можна очікувати, що чим більшою є неповноцінність монети, тим меншою буде ймовірність попадання результату експерименту в область прийняття гіпотези. Розглянемо крайню ситуацію: нехай ймовірність випадання герба становить 0,95. У такому випадку гіпотеза стосовно повноцінності буде відхилена за обома правилами практично завжди. Відповідний розподіл представлено на рис. 6.4.

Зазначимо, що ймовірність помилки другого роду залежить від конкретних умов виконання експерименту, тобто від ступеня неповноцінності монети. За обома правилами гіпотеза стосовно повноцінності монети буде відхилена скоріше у випадку, коли дійсність сильно відрізняється від припущень, ніж у випадку, коли дійсність незначно відрізняється від висунутої гіпотези.

Інакше кажучи, якщо неповноцінність монети є незначною, то ймовірність прийняття гіпотези щодо її повноцінності буде вищою за обома правилами. Тепер очевидно, що чим більше гіпотеза віддалена від дійсності, тим швидше вона буде відхилена.

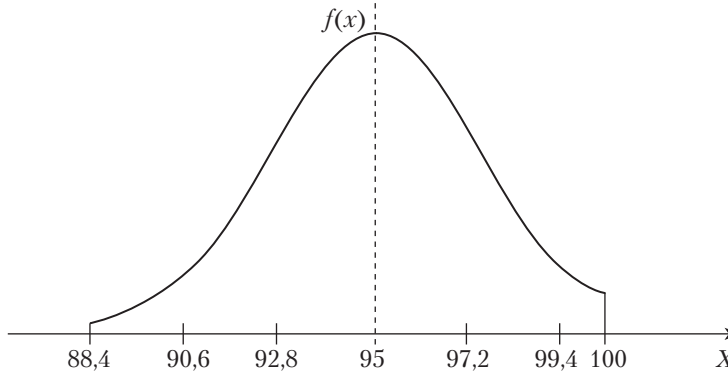


Рис. 6.4. **Випадок, коли ймовірність випадання герба становить 0,95**

У реальних ситуаціях при перевірці гіпотез експериментатор попадає в скрутне становище, якщо він не знає, чи правий, коли приймає або відхиляє гіпотезу. Звичайно, він завжди намагається прийняти правильну гіпотезу і відхилити неправильну. При цьому важливо знати ступінь ризику, який виникає в результаті прийняття того чи іншого рішення. Тобто *необхідно визначити деяку область прийняття сформульованої гіпотези і пам'ятати, що навіть у випадку правильної гіпотези може статись так, що отриманий результат не попаде у визначену область.*

Крім того, навіть, якщо гіпотеза є неправильною, то є деякі шанси, що отриманий результат попаде в область прийняття гіпотези. Виникає питання — як визначити границі області прийняття висунутих гіпотез?

Чим вужчою є ця область, тим більшим буде ризик помилки першого роду. Збільшуючи область прийняття гіпотези, ми тим самим зменшуємо ймовірність помилки першого роду і одночасно збільшуємо ризик помилки другого роду.

Очевидно, що *ми не можемо одночасно зменшувати ризик помилки одного роду без одночасного збільшення ризику помилки іншого роду.*

Рішення щодо величини області прийняття висунутої гіпотези залежить від того, наслідки якої помилки будуть гіршими. У випадку з монетою критерії ризику відрізнялись суттєво. *Перше правило* було спрямоване на те, щоб не забракувати повноцінну монету. Велика область прийняття гіпотези свідчить про те, що воно практично ніколи не допоможе виявити неповноцінну монету.

За *другим правилом* область прийняття гіпотези була досить вузькою, що попереджало прийняття монети з дефектом за повноцінну. Очевидно, що за другим правилом багато повноцінних монет можуть бути забракованими.

Загалом, при визначенні області прийняття висунутої гіпотези кожний експериментатор керується власними критеріями відносно можливості помилкового відхилення правильної гіпотези, а також, у деякій мірі, помилкового прийняття неправильної гіпотези. Величина вибраної області фактично відображає компроміс між цими двома вимогами.

Отже, при прийнятті рішень завжди існує ризик. У наступному розділі буде показано, як на практиці визначаються границі допустимого ризику при перевірці гіпотез. Вище ми розглянули принципи,

які лежать в основі процедури перевірки гіпотез, а тепер необхідно перейти безпосередньо до самої процедури перевірки гіпотез.

6.3. Статистична перевірка гіпотез

Існує стандартна процедура перевірки гіпотез, якою користуються в техніці, психології, освіті, суспільних науках та багатьох інших областях. Ця процедура буде розглянута у цій главі.

Необхідно підкреслити той важливий факт, що результат експерименту, який підтверджує справедливість висунутої гіпотези, майже ніколи не може бути основою для прийняття цієї гіпотези. Водночас *результат, несумісний з висунутою гіпотезою, є цілком достатнім для її відхилення як неправильної.* Очевидно, що наведене твердження потребує доведення.

Причиною того, що результат, який підтверджує висунуту гіпотезу, не обов'язково може бути основою для її прийняття, полягає в тому, що отриманий результат може бути сумісним і з іншими гіпотезами. Тобто він не обов'язково може служити доведенням справедливості даної гіпотези проти інших сформульованих альтернатив. Наприклад, випадання 51 герба при 100 підкиданнях монети є сумісним з гіпотезою про те, що монета є повноцінною. Разом з тим, цей результат є сумісним також з припущенням, що монета має деякий дефект і при її підкиданні герб випадає дещо частіше. Таким чином, результат, сумісний з гіпотезою, висунутою на початку, не може бути 100 %-м доведенням її правильності. Навіть випадання 50 гербів при підкиданні монети не є підґрунтям для висновку, що монета не має хоча б незначного дефекту.

У той же час випадання 80 гербів при 100 підкиданнях монети могло б служити для відхилення гіпотези щодо її повноцінності. Оскільки існує дуже низька ймовірність отримати такий результат, що може мати місце, то цілком обґрунтовано, і з мінімальним ризиком зробити помилку, його можна використати як основу для висновку щодо неповноцінності монети.

Приклад 6.1. Припустимо, що середній коефіцієнт розумового розвитку (КРР) деякої генеральної сукупності людей $\mu = 100$. Наслідок випадкової вибірки дає результат $\mu_i = 102$, який є сумісним з висунутою гіпотезою. Однак цей результат є також сумісним з припущенням, що $\mu = 101$ або $\mu = 99$ і, звичайно, є сумісним з гіпотезою, що

$\mu = 102$. Таким чином, на основі даного результату не можна віддати перевагу гіпотезі, що $\mu = 100$.

Припустимо тепер, що інша випадкова вибірка дала середнє $\mu = 135$. Якщо об'єм вибірки був достатньо великим, то можна показати таке: якщо початкова гіпотеза правильна, то ми практично ніколи не отримали б подібного результату. На основі цього висновку отриманий результат цілком обґрунтовано можна використати як підтвердження неправильності висунутої гіпотези і ризик помилки при цьому буде мінімальним.

Усе сказане ґрунтується на тому факті, що результат експерименту, який є сумісним з висунутою гіпотезою, виявляється також сумісним з іншими гіпотезами. Це веде до того, що подібний результат не може бути використано для обґрунтування вибору деякої гіпотези порівняно з іншими. Однак ми завжди можемо отримати результат, який не збігається з висунутою гіпотезою і може спричинити значні сумніви щодо її правильності або достовірності.

Гіпотезу можна порівняти із свідченнями звинуваченого в суді. Він не може довести істинність своїх слів. Разом з тим, деякі факти, наведені ним, залишають відкритою можливість для висування припущення, що він міг діяти не так, як говорить. Тому прокурор може піддати сумніву правдивість його розповіді і надати іншу інтерпретацію наведених фактів.

Нульова або нуль-гіпотеза

Відмова відхилити гіпотезу означає, що вона може бути правильною. *Відхилення сформульованої гіпотези означає, що ми робимо висновок щодо її неправильності*. Це положення є дуже важливим.

Таким чином, при перевірці гіпотез остаточний висновок можна зробити тільки у тому випадку, якщо ми можемо відхилити висунуту гіпотезу. Отже, мета експерименту повинна полягати у тому, щоб відхилити сформульовану гіпотезу.

Мета експерименту повинна полягати у тому, щоб відхилити сформульовану гіпотезу

А це означає, що початкова гіпотеза повинна формулюватись як альтернатива тому, у що ми віримо.

Якщо ми зможемо відхилити висунуту гіпотезу (довести її неправильність), то тим самим продемонструємо справедливість того твердження, у яке дійсно віримо.

Наприклад, якщо ми хочемо показати, що зріст чоловіків є більшим ніж зріст жінок, то висуваємо гіпотезу щодо відсутності відмінностей в їх зрості. Потім намагаємося відхилити цю гіпотезу. Для того, щоб довести існування відмінностей між різними партіями, необхідно перевірити гіпотезу, яка стверджує, що між партіями немає відмінностей. Відхиляючи цю гіпотезу, ми встановлюємо істинність вихідного припущення.

Означення 6.3. *Гіпотезу, відповідно до якої немає відмінностей між різними сукупностями (або варіантами), називають нуль-гіпотезою.*

Тобто у нуль-гіпотезі формулюється результат (висновок), протилежний до очікуваного. Відкидання нуль-гіпотези після проведення експерименту свідчить, що ми отримали очікуваний результат.

Гіпотеза, яку ми зможемо перевірити, *не може бути сформульована на основі будь-якого судження*. Так, судження, що монета неповноцінна, є недостатньо визначеним для того, щоб на його основі можна було сформулювати конкретну визначену гіпотезу. І подібні випадки зустрічаються досить часто. Із сказаного випливає наступне: *експериментатор повинен формулювати альтернативу тому, що він намагається довести, у вигляді чітко визначеної гіпотези.*

Експериментатор повинен формулювати альтернативу тому, що він намагається довести, у вигляді чітко визначеної гіпотези

Тільки у випадку, коли це можливо, він може спробувати відхилити її з метою доведення істинності початкових припущень.

Таким чином, перший крок експериментатора повинен полягати у формулюванні статистичної гіпотези, яку він сподівається *відхилити з метою доведення істинності свого вихідного (початкового) припущення*. Після цього він може застосувати процедуру перевірки гіпотези.

Відхилення гіпотези та рівень значущості

Припустимо, що експериментатор вже сформулював гіпотезу, яку він сподівається відхилити з метою довести істинність свого вихідного припущення. У прикладі з монетою це може відповідати тому, що він вирішив перевірити гіпотезу щодо повноцінності монети, сподіваючись її при цьому відхилити. Експериментатор приступає до формування вибірки і розрахунку на її основі числа гербів при підкиданні монети (наприклад, 100 разів) та інших параметрів.

Наступним кроком експериментатора є формулювання такого питання:

"*Якщо моя нульова гіпотеза вірна, то яка ймовірність того, що мною буде зроблена вибірка, показник якої відрізняється від очікуваного результату так само, як і отримане мною значення?*"

Припустимо, що монета дійсно є повноцінною і герби повинні випадати при її підкиданні у середньому в половині зробленого числа дослідів. Наскільки ймовірним є те, що може випасти число гербів, яке відрізняється від очікуваного значення 50 настільки, наскільки відрізняється результат, отриманий у дійсності?

Зверніть увагу на сформульоване тут запитання — воно є ключем для розуміння подальшого матеріалу даного розділу.

На наступному кроці виясняється, наскільки ймовірним буде отриманий результат вибірки за умови, що висунута гіпотеза є правильною. Ймовірність отримання вибірки з характеристиками, які відповідають сформульованій гіпотезі, може бути високою або низькою і цей факт визначає рішення щодо прийняття або відхилення гіпотези.

Припустимо, розрахунки показали, що у випадку правильності сформульованої гіпотези ймовірність отримання вибірки з характеристиками, які відповідають цій гіпотезі, є високою. У такому випадку отримана *вибірка повинна розглядатись як представницька* сукупність (наприклад, при випаданні 52 орлів при підкиданні монети). Це означає, що отримане значення має надто високу ймовірність появи, щоб дозволити нам відхилити висунуту гіпотезу.

Уявімо тепер, що отримане при проведенні досліду значення результату так сильно відрізняється від очікуваного (відповідно до висунутої гіпотези), що ймовірність його появи є дуже малою. (Наприклад, такий випадок буде мати місце, якщо при 100 підкиданнях монети випаде 93 герби.) Інакше кажучи, залишається припустити, що отримане вибіркове значення має таке велике відхилення (як свідчать відповідні розрахунки), що практично неможливо отримати подібний випадковий результат при справедливості нульової гіпотези. У такому разі рішення буде полягати у відхиленні висунутої нами гіпотези. Тобто швидше ми припустимо, що висунута гіпотеза виявилась неправильною, ніж допустимо ймовірність появи надзвичайно неправдоподібного результату.

Для експериментатора велике значення має мінімізація можливостей появи ситуацій, коли випадкове співпадіння низки чинників

може привести до відхилення правильної гіпотези. Тому перед тим, як відхилити гіпотезу, він вимагає, щоб ймовірність отримання відповідного вибіркового значення була дуже малою.

У деяких областях науки прийнято відхиляти гіпотезу тільки у тих випадках, коли випадкове вибіркове значення може зустрічатись не частіше, ніж 5 разів на 100 експериментів. В інших областях гіпотези відхиляються, якщо ймовірність появи відповідного вибіркового значення не перевищує 0,01. Очевидно, що експериментатор повинен прямувати до того, щоб ймовірність появи вибіркового значення, яке вказує на неправильність висунутої гіпотези, була досить малою.

Ймовірність появи вибіркового значення, яке вказує на неправильність висунутої гіпотези, вибирається експериментатором і називається *рівнем значущості* експерименту

Рівень значущості повинен визначатися до збору експериментальних даних, оскільки результати експерименту не повинні впливати на величину вибраного критерію. Після вибору рівня значущості і обробки даних експериментатор припускає, що його гіпотеза є правильною і визначає — більшою чи меншою вибраного рівня значущості буде ймовірність отриманого результату.

Якщо ймовірність отриманого результату перевищить рівень значущості, то експериментатор не зможе відхилити висунуту гіпотезу, справедливо вважаючи, що даний результат є в достатній мірі сумісним з нею. Наприклад, припустимо, що вибрано рівень значущості 0,05, а в результаті експерименту отримано 52 герби при 100 підкиданнях монети. Якщо експериментатор встановить, що відхилення у два герби від очікуваного значення зустрічається більше, ніж у 5 % випадків, то він не зможе відхилити висунуту гіпотезу.

Якщо розраховане значення ймовірності деякого результату виявилось меншим рівня значущості, то висунуту гіпотезу можна відхилити. Наприклад, припустимо, що при використанні рівня значущості 0,05 підкидання монети привело до 96 гербів, а розрахунок показав, що відхилення цього значення від очікуваного має ймовірність появи менше 0,05. Логіка, якою керується експериментатор, полягає у такому: "*Після формулювання висунутої гіпотези я отримав такий неймовірний результат, що не можу в нього повірити*". Після цього експериментатор може вважати свою гіпотезу доведеною і, таким чином, може сподіватись, що завжди будуть отримуватись результати, ймо-

вірність появи яких при даних обставинах буде меншою рівня значущості.

Вибір рівня значущості означає, що введено певне правило прийняття і неприйняття гіпотез. *Рівень значущості показує ймовірність відхилення деякого показника від його очікуваного значення, при якому дослідник може відхилити висунуту гіпотезу.*

Вибраний *рівень значущості вказує на точне значення ймовірності помилки першого роду* у випадку, якщо гіпотеза дійсно правильна. Наприклад, якщо рівень значущості дорівнює 0,01, то це означає, що у випадку правильності гіпотези в одному випадку із 100 буде робитись помилка першого роду на основі отриманого результату.

Інакше кажучи, рівень значущості означає величину ризику помилки першого роду. Чим меншим є рівень значущості, тим меншою є ймовірність припущення помилки першого роду. Однак, чим меншим є рівень значущості, тим більшою є ймовірність помилки другого роду, якщо гіпотеза виявиться помилковою. Таким чином, вибір рівня значущості означає вибір правила прийняття рішення при перевірці гіпотези.

Сформулюємо процедуру перевірки гіпотез у вигляді наступних етапів:

1. Формулювання нульової гіпотези, яку необхідно перевірити. Відхилення сформульованої гіпотези дає можливість вважати протилежну гіпотезу правильною. (У прикладі з монетою метою може бути встановлення факту неповноцінності монети. Таким чином, необхідно перевірити гіпотезу про те, що монета є повноцінною.)
2. Вибір рівня значущості. (Наприклад, рівень значущості 5 % означає, що ймовірність помилки першого роду становить 0,05.)
3. Виконання експерименту і обчислення необхідного статистичного параметра.
4. Припускаючи, що сформульована гіпотеза є правильною, необхідно визначити ймовірність відхилення обчисленого значення статистичного параметра від його очікуваного значення.
5. Якщо, в припущенні істинності гіпотези, розрахунки показують, що ймовірність відхилення отриманого вибіркового статистичного показника від очікуваного значення перевищує рівень значущості, то *відхилити висунуту гіпотезу неможливо*. Якщо, в припущенні істинності гіпотези, розрахунки показують, що ймовірність відхилення отриманого вибіркового

статистичного показника від очікуваного значення є меншим рівня значущості, то *висунута гіпотеза відхиляється*.

Нехай в експерименті з монетою випало 55 гербів. Гіпотеза щодо повноцінності монети буде перевірятись на рівні значущості 0,05. Припустимо, що монета є дійсно повноцінною, а отримана вибірка є однією з великої множини випадкових вибірок об'ємом 100 підкидань кожна.

Розподіл результатів великої множини експериментів з повноцінною монетою є нормальним. Цей розподіл представлено на рис. 6.5 ($\mu = 50$, $\sigma = 5$). Місце, де знаходиться оцінка 55, позначено хрестиком. Виникає наступне питання: якщо гіпотеза правильна, то *яка ймовірність відхилення числа гербів, які випали у випадковій вибірці із 100 спостережень, більше від очікуваного значення на 5 одиниць*? У розподілі, представленому на рис. 6.5, значення 55 знаходиться на відстані 5 одиниць від очікуваного значення. Таким чином, z-оцінка значення 55 дорівнює: $z(55) = (55 - 50)/5 = 1$.

Тепер можна записати формулу для визначення z-оцінки деякого результату експерименту для вибірки дихотомічної змінної при $N = 100$:

$$z = \frac{\text{Оцінка відхилення}}{\text{Стандартне відхилення}} = \frac{r - Np}{\sqrt{Np(1-p)}}, \qquad (6.3.1)$$

де r — значення отриманого результату експерименту (тобто, скільки разів дана подія зустрічається у вибірці); N — об'єм вибірки; p — ймовірність того, що дана подія буде мати місце.

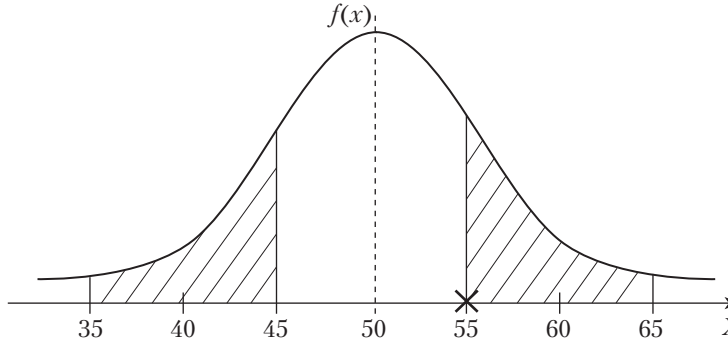


Рис. 6.5. **Розподіл результатів експерименту з повноцінною монетою ($\mu = 50$, $\sigma = 5$)**

Зазначимо, що Np — величина гіпотетичного очікування, яке дорівнює теоретичній ймовірності появи події, що визначається висунутою гіпотезою. Очевидно, що навіть у випадку, коли гіпотеза є правильною, можна очікувати деякого відхилення отриманого результату від його теоретичного значення. Однак вирішальну роль при прийнятті рішення відіграє ступінь цього відхилення, виражена через z-оцінку.

Для експерименту з повноцінною монетою $r = 55$, $n = 100$, $p = 0{,}5$. Таким чином:

$$z = \frac{r - Np}{\sqrt{Np(1-p)}} = \frac{55 - 100 \cdot 0{,}5}{\sqrt{100 \cdot 0{,}5 \cdot 0{,}5}} = 1{,}0.$$

Суть цієї формули полягає у такому: відповідно до нульової гіпотези ми припускаємо, що $p = 0{,}5$, і при цих умовах отриманий нами результат має $z = 1{,}0$ у розподілі результатів великої множини ідентичних вибірок.

Нагадаємо запитання, яке поставив перед собою експериментатор: *якщо гіпотеза правильна, то яка ймовірність отримання результату, який відхиляється від очікуваного значення більше ніж на* 5 *одиниць*? Це ж запитання, але записане за допомогою формули, звучить так: *яка ймовірність отримання випадкового результату, z-оцінка якого є на одну одиницю більшою від нульового значення*? За допомогою таблиці відносних площ нормального розподілу знаходимо, що приблизно 32 % спостережень нормального розподілу мають z-оцінки, що відрізняються від очікуваного значення щонайменше на одну одиницю. Це означає, що всі спостереження, які належать заштрихованим областям графіка, представленого на рис. 6.5, або 32 % всіх спостережень, відхиляються від очікуваного значення більше, ніж отриманий нами результат — 55 гербів.

Таким чином, результат 55 гербів означає наступне: якщо гіпотеза вірна, то ймовірність появи подібного результату або результату, який ще більше відрізняється від очікуваного значення, становить 32 %. Оскільки дане значення ймовірності перевищує вибраний рівень значущості 0,05, то ми не можемо відхилити гіпотезу щодо повноцінності монети. Можна сказати, що отриманий результат не є настільки неправдоподібним, щоб дати нам підґрунтя відкинути висунуту гіпотезу.

Припустимо далі, що при підкиданні монети випало 65 гербів. На цей раз відхилення від очікуваного значення становить 15 одиниць і можна припустити, що ймовірність отримання подібного результату

в експерименті з повноцінною монетою є значно меншою. З рис. 6.5 можна зробити висновок: *якщо висунута гіпотеза є вірною, то отримане на цей раз значення буде відрізнятись від очікуваного результату на три одиниці стандартного відхилення.*

За формулою (6.3.1) знайдемо:

$$z = \frac{r - Np}{\sqrt{Np(1-p)}} = \frac{65 - 100 \cdot 0{,}5}{\sqrt{100 \cdot 0{,}5 \cdot 0{,}5}} = 3{,}0,$$

що є формальним підтвердженням сформульованого висновку. З таблиці відносних площ нормального розподілу випливає, що ймовірність отримання результату, *z*-оцінка якого так суттєво відрізняється від очікуваного значення, приблизно дорівнює 0,001.

Таким чином, отриманий нами результат експерименту є настільки рідкісним, що ймовірність його появи становить приблизно 0,001. Оскільки це значення є меншим рівня значущості, то наше рішення буде полягати, у даному випадку, у відхиленні гіпотези щодо повноцінності монети, а це означає підтвердження початкового припущення щодо її неповноцінності.

Висновки до параграфа

Отже, розглянуто ідею методу перевірки гіпотез, послідовність реалізації якого полягає у такому:

- точне формулювання гіпотези, яка буде перевірятись;
- вибір рівня значущості і визначення результату (параметра), який нас цікавить, на основі вибірки;
- у припущенні істинності висунутої гіпотези визначити ймовірність того, що отриманий результат буде відрізнятись від очікуваного значення на величину відхилення, отриманого в експерименті;
- якщо ймовірність такого відхилення буде меншою рівня значущості, то висунута гіпотеза відхиляється;
- якщо ймовірність такого відхилення перевищить рівень значущості, то висунута гіпотеза приймається.

Логічна основа розглянутого методу залишається однаковою при розв'язуванні багатьох задач такого типу. Гіпотеза стосовно середнього, медіани, стандартного відхилення, а також багато інших гіпотез перевіряються за допомогою ідентичної процедури. Очевидно, що розрахунки у кожному окремому випадку будуть мати свої особливості.

6.4. Перевірка гіпотез у задачах трьох типів

Розглянемо три типи задач та шляхи їх розв'язку. Перший тип задач відноситься до дихотомічних змінних (приклад розглянуто у попередньому параграфі). У двох інших задачах буде сформульовано гіпотезу щодо невідомого середнього деякої сукупності.

Задача першого типу

Припустимо, що виборці деякого округу не збираються в цьому році голосувати за кандидатів тих партій, яким вони віддавали свої голоси в минулому. Як правило, 2/3 з них голосували за зелених, а 1/3 — за оранжевих. Однак мали місце події, які дають підґрунтя припускати, що виборці змінять свій вибір. При цьому визначити точно величину цього впливу неможливо. Це означає, що відповідно до нашого припущення зазначені події будуть мати велике значення для результатів голосування, але вгадати, яка з партій від цього виграє, неможливо.

Відповідно, можна висунути гіпотезу, що звичне співвідношення між голосами, поданими за обидві партії, не зміниться, і необхідно перевірити цю гіпотезу на рівні значущості 0,05. При цьому ми сподіваємось, що зможемо цю гіпотезу відкинути. У дійсності, ми просто сподіваємось показати, що звичне для нас співвідношення 2/3 до 1/3 буде змінено.

Нехай вибірка складається з 450 спостережень і встановлено, що 225 виборців будуть голосувати за зелених. Таким чином, маємо:

$$N = 450, p = 2/3, r = 225.$$

Звідси отримаємо: $N \cdot p = 300$. Це значення відповідає очікуваному результату експерименту за умови, що висунута гіпотеза є вірною. Інакше кажучи, якщо гіпотеза, відповідно до якої 2/3 виборців будуть голосувати за зелених, є вірною, то очікується, що 2/3 членів вибірки віддадуть свої голоси за зелених.

Припускаючи, що дана гіпотеза вірна, отримаємо:

$$z = \frac{r - Np}{\sqrt{Np(1-p)}} = \frac{225 - 300}{\sqrt{450 \cdot (2/3) \cdot (1/3)}} = \frac{-75}{10} = -7{,}5.$$

Оцінку $z = -7{,}5$ отримано у припущенні, що гіпотеза щодо збереження співвідношення між голосами виборців, поданими за кандидатів різних партій, є вірною.

Можна припустити, що отримано нескінченне число випадкових вибірок, потужністю $N = 450$ кожна, і на цій основі сформульовано новий розподіл, утворений із показників, які представляють собою число голосів, поданих за зелених у кожній вибірці.

Вибірка, яку ми фактично зібрали, містить тільки 225 голосів, поданих за зелених. Якщо розглядати її як одну із великої множини незалежних вибірок, то ми отримаємо для неї значення $z = -7{,}5$. На рис. 6.6 показано, який вигляд буде мати графік нашого розподілу.

Із сказаного випливає таке: якщо гіпотеза вірна, то в отриманій вибірці число голосів, поданих за зелених, відрізняється від очікуваного значення на 7,5 одиниць стандартного відхилення. Ймовірність отримання такого результату визначається частиною спостережень, для яких $z > 7{,}5$ або $z < -7{,}5$. З таблиці відносних площ знаходимо, що це значення ймовірності є таким: $p < 0{,}00001$.

Отже, гіпотеза відносно того, що співвідношення між голосами, поданими за зелених та оранжевих, дійсно становить 2/3 до 1/3 повинна бути відхилена. Хоча у даному випадку і не було доведено, які саме події стали вирішальним фактором впливу на результати голосування, все таки можна вважати, що співвідношення між голосами змінилось. При цьому необхідно розуміти, що розраховані значення z-оцінок відповідають z-оцінкам, отриманим за умови, що сформульована гіпотеза є вірною. *Логіка тут така: якщо значення z-оцінки буде відповідати події, ймовірність появи якої є незначною, то рішенням буде відхилення гіпотези.*

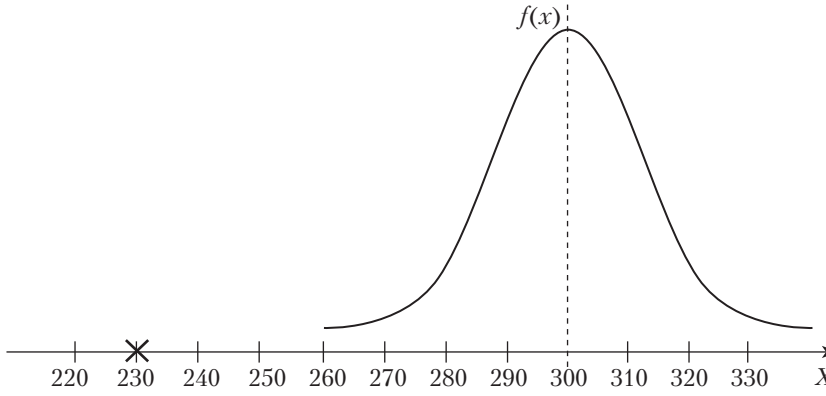


Рис. 6.6. **Графік розподілу кількості виборців**

Аналогічна процедура буде застосована також у випадку перевірки гіпотези щодо середнього значення деякої сукупності. У цьому випадку результатом експерименту буде значення вибіркового середнього, а не просто число повторень результату експерименту, який нас цікавить.

Задача другого типу: перевірка гіпотези щодо середнього значення деякої сукупності даних

Припустимо, що відоме стандартне відхилення деякої сукупності, яка є однією із великої множини випадкових вибірок однакового об'єму. Середні цих вибірок утворюють новий розподіл. Як і раніше, починаємо з нульової гіпотези і визначення рівня значущості. Необхідно визначити величину розбіжності між фактичним вибірковим середнім і величиною, яка очікується відповідно до нашої гіпотези. Оскільки ми припускаємо, що висунута гіпотеза є правильною, то за допомогою центральної граничної теореми для середніх можна встановити вигляд даного розподілу середніх.

Після формування розподілу визначимо *z*-оцінку обчисленого значення вибіркового середнього. На основі цієї оцінки встановимо, наскільки близькою є величина отриманого середнього до очікуваного значення. Якщо ймовірність розбіжності між теоретичним і фактичним значеннями середнього перевищує рівень значущості, то гіпотеза приймається. Якщо ж вона буде меншою рівня значущості, то гіпотеза відхиляється.

Приклад 6.2. Чи буде середній коефіцієнт розумового розвитку (КРР) *десятилітніх хлопчиків, які збираються стати водіями автомобілів*, відрізнятись від середнього генеральної сукупності. Нехай дослідження виконується в місті, в якому середній КРР десятилітніх хлопчиків $\mu = 100$, а стандартне відхилення $\sigma = 20$.

Необхідно показати, що десятилітні хлопчики, які збираються стати водіями автомобілів, відрізняються за рівнем свого інтелектуального розвитку від середнього генеральної сукупності. За нульгіпотезу прийнято:

H_0: вибрана група — типовий представник генеральної сукупності;

H_1: вибрана група відрізняється від генеральної сукупності.

Виберемо за рівень значущості 0,05 і будемо сподіватись, що зможемо відхилити висунуту гіпотезу. Нехай із генеральної сукупності

вибрано 64 хлопчики, які збираються стати водіями. Встановлено, що середній КРР для них становить: $\mu_{гр} = 108$.

Ми вважаємо, що нуль-гіпотеза є вірною, тобто їхні 64 КРР можна розглядати як випадкову вибірку із великої сукупності спостережень і представляє собою одну з численних вибірок, потужністю 64 кожна. З теорем 5.4 і 5.5 (для середнього ряду розподілу, утвореного із середніх значень) випливає, що при таких умовах розподіл середніх значень цих вибірок буде мати середнє $\mu = 100$, а стандартне відхилення

$$\sigma_{гр} = \frac{\sigma}{\sqrt{N}} = \frac{20}{\sqrt{64}} = 2{,}5.$$

Цей розподіл представлено на рис. 6.7, на якому середнє $\mu_{гр} = 108$ позначено хрестиком.

З рис. 6.7 видно, що в припущенні істинності нульової гіпотези ми отримали вибіркове середнє, величина якого на 8 пунктів перевищує точку рівноваги нормального розподілу, а стандартне відхилення випадкової вибірки становить 2,5. Інакше кажучи, припускаючи істинність нульової гіпотези, ми повинні зробити висновок, що отримане нами вибіркове середнє на 3,2 стандартного відхилення перевищує очікуване значення середнього, тобто 100.

З таблиці для відносних площ нормального розподілу знайдемо, що ймовірність отримання значення, *z*-оцінка якого відрізняється більше ніж на 3 одиниці (в обидва боки) від точки рівноваги, становить < 0,001. Таким чином, припустивши істинність нульової гіпоте-

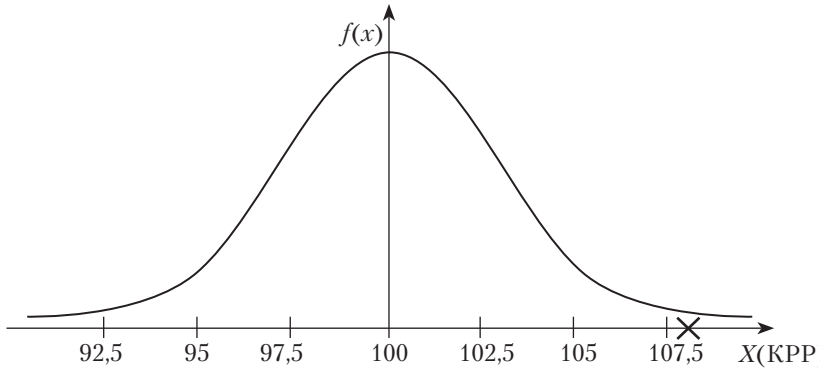


Рис. 6.7. **Форма розподілу для КРР при** $\mu_{гр} = 108$

зи, ми отримали вибіркове значення, яке настільки сильно відрізняється від очікуваної величини, що ймовірність його появи є меншою ніж 0,001. Очевидно, що такий результат свідчить про те, що ми повинні відхилити нульову гіпотезу і прийняти альтернативну. Тобто хлопчики, які у майбутньому збираються працювати водіями, в середньому відрізняються за рівнем свого інтелектуального розвитку від інших хлопчиків їхнього віку.

У такому випадку було б корисно застосувати формулу знаходження *z*-оцінки вибіркового середнього розподілу вибіркових середніх. Ця формула дає можливість швидше розв'язувати подібні задачі. Так, *z*-оцінка вибіркового середнього знаходиться за формулою:

$$z = \frac{\text{Оцінка відхилення середнього}}{\text{Стандартне відхилення середніх}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{N}}, \qquad (6.4.1)$$

де $\bar{x}$ — значення вибіркового середнього (у нашому прикладі $\bar{x} = 108$); μ — середнє розподілу вибіркових середніх (ми припускали, що $\mu = 100$); σ — стандартне відхилення КРР для генеральної сукупності (у нашому прикладі $\sigma = 20$).

Підставляючи конкретні значення у формулу (6.4.1), отримаємо:

$$z = \frac{\bar{x} - \mu}{\sigma\sqrt{N}} = \frac{108 - 100}{20\sqrt{64}} = 3{,}2.$$

Звертаючись тепер до таблиці відносних площ нормального розподілу (при $z = 3{,}2$), відхиляємо нуль-гіпотезу на рівні значущості 0,05.

Задача третього типу

Нехай *необхідно перевірити нуль-гіпотезу стосовно того, що студенти п'ятого курсу ІПСА НТУ КПІ є типовими представниками всіх студентів п'ятого курсу за вмінням виконувати курсові проекти* з проектування комп'ютерних інформаційних систем. Перевіримо нульову гіпотезу для даного випадку на рівні значущості 0,01. За основу візьмемо результати виконання курсових проектів у всіх університетах Києва. Середня оцінка за курсовий проект по місту дорівнює $\mu = 72$ (при максимальному значенні 100); стандартне відхилення $\sigma = 12$. У групі ІПСА, яка аналізувалась, налічувалось $N = 36$ студентів при середньому значення оцінки $\mu_{гр} = 74$. Тобто необхідно визначити, чи суттєво відрізняється середнє для вибраної групи від середнього генеральної сукупності.

Припустимо, що наша гіпотеза є вірною і що отримані 36 оцінок за курсові проекти можна розглядати як випадкову вибірку із всієї сукупності оцінок. Це означає, що ми можемо розглядати вибіркове середнє для групи студентів, $\mu_{гр} = 74$, як одне із значень теоретичного розподілу середніх різних вибірок з об'ємом $N = 36$ кожної. Відповідно до теореми 5.4 розподіл таких вибіркових середніх є нормальним. При цьому середнє розподілу середніх дорівнює 72 (таке саме значення має середнє генеральної сукупності), а стандартне відхилення:

$$\sigma_{рс} = \frac{\sigma}{\sqrt{N}} = \frac{12}{\sqrt{36}} = 2.$$

Таким чином, отримане вибіркове середнє для вибраної групи студентів, $\mu_{гр} = 74$, на $z = 1$ перевищує теоретично очікуване значення, тобто

$$z_{гр} = \frac{\overline{x} - \mu}{\sigma/\sqrt{N}} = \frac{74 - 72}{12\sqrt{36}} = 1{,}0.$$

Відповідно до таблиці відносних площ нормального розподілу, *ймовірність розбіжності між середнім генеральної сукупності і середнім вибраної групи студентів* є більшою вибраного нами рівня значущості, а тому *нуль-гіпотеза приймається*. Тобто середнє генеральної сукупності і середнє вибраної групи відрізняються несуттєво.

Отже, на основі виконаного експерименту не можна зробити висновок, що студенти ІПСА НТУ КПІ відрізняються за рівнем виконання курсових проектів з проектування інформаційних систем від студентів інших університетів.

6.5. Зауваження стосовно термінології

Необхідно строго розрізняти показники генеральної сукупності і показники, отримані для деякої конкретної вибірки. Так, наприклад, середній КРР генеральної сукупності хлопчиків може дорівнювати 100 при $\sigma = 12$, а значення вибіркового середнього конкретної вибірки, взятої з цієї генеральної сукупності, може становити 95, 110 чи 132. Середнє генеральної сукупності, $\mu = 100$, і її стандартне відхилення, $\sigma = 12$, називають *параметрами*, а показник, розрахований для окремої випадкової вибірки, називають *статистикою*.

Зазначимо, що будь-яке судження, сформульоване у вигляді гіпотези, відноситься до параметра. Для перевірки гіпотези формується конкретна вибірка і на її основі обчислюється статистика. Знайдене

значення статистики використовується для прийняття або відхилення висунутої гіпотези.

У процесі перевірки гіпотез ми часто звертаємося до нормального розподілу і легко запам'ятовуємо основні відповідності між значеннями *z*-оцінок та ймовірностями. Наприклад, ймовірність отримання нормального випадкового спостереження, *z*-оцінка якого відрізняється від середнього більше, ніж на 1,96, дорівнює 0,05. Таким чином, при перевірці гіпотез на рівні значущості $\alpha = 0{,}05$ можна автоматично зробити висновок, що гіпотезу не можна відхилити, якщо отримане значення оцінки попадає в інтервал $-1{,}96 \le z_i \le 1{,}96$. Якщо ж воно є меншим ніж −1,96 або більшим ніж +1,96, то гіпотеза відхиляється.

6.6. Односторонні критерії

Нехай наша мета полягає у тому, щоб відхилити гіпотезу відносно того, що деякий параметр, наприклад, середнє генеральної сукупності, дорівнює деякому значенню. Таким чином, необхідно перевірити гіпотезу, *що дане генеральне середнє дорівнює цьому значенню* і спробувати відхилити її на основі вибіркового значення.

Наприклад, припустимо, що КРР восьмилітніх хлопчиків, які збираються стати пожежниками, відрізняються від генеральної сукупності за рівнем свого інтелекту. Висувається гіпотеза, що ці хлопчики мають середній КРР = 100. На основі результатів деякої вибірки хлопчиків, які планують стати пожежниками, ми спробуємо відхилити гіпотезу, відповідно до якої їхній середній КРР виявиться значно меншим або значно більшим 100.

Розглянемо тепер ситуацію, коли необхідно показати, що члени деякої окремої групи відрізняються від генеральної сукупності. Це означає, що робиться припущення не тільки щодо відмінності вибіркового показника від параметра, але і щодо характеру цієї відмінності. Наприклад, припустимо, що хлопчики, які збираються стати пожежниками, в цілому повинні мати вищий рівень інтелектуального розвитку, ніж представники великої групи хлопчиків їхнього віку. Відповідно, ми прогнозуємо, що середній КРР вибірки хлопчиків, які збираються стати пожежниками, буде перевищувати 100. Тепер ми маємо справу *не просто із встановленням деякої відмінності, а й з прогнозуванням того, що в статистиці називають "напрямом" даної відмінності*. Отримання вибіркового середнього, яке виявиться на-

віть на 0,1 стандартного відхилення меншим від параметра, буде означати, що наше припущення було помилковим.

Сформулюємо нульову гіпотезу так: середній КРР групи з 64 хлопчиків, які збираються стати пожежниками, дорівнює відповідному показнику генеральної сукупності, тобто $\mu_{гр} = \mu = 100$ при $\sigma = 20$. Перевіримо цю гіпотезу на рівні значущості 0,05. Будемо також вважати, що 64 хлопчики вибрані випадково і таких груп існує велика множина. Таким чином, середнє цієї групи можна вважати одним із множини вибіркових середніх. Якщо сформульована гіпотеза правильна, то розподіл вибіркових середніх буде мати вигляд, як показано на рис. 6.8.

Тепер питання полягає у тому, яке значення отриманого нами вибіркового середнього дозволить відхилити висунуту гіпотезу. Це дасть можливість підтвердити наше початкове припущення. Оскільки вибрано рівень значущості 0,05, це означає, що у випадку справедливості висунутої гіпотези ми ризикуємо зробити помилку першого роду тільки у 5 випадках із 100. Тепер необхідно визначити такий інтервал, в який попаде тільки 5/100 вибіркових середніх, якщо висунута нами гіпотеза правильна.

Нашим початковим припущенням було те, що середній КРР хлопчиків, які збираються стати пожежниками, перевищує 100. Сформулюємо і перевіримо нуль-гіпотезу, що КРР = 100. Таким чином, ми не зможемо відхилити дану гіпотезу, якщо отримаємо будь-яке значення середнього, яке є менше ніж 100. Необхідно вибрати такий інтервал значень КРР, щоб шанси на успіх були максимальними, якщо наше

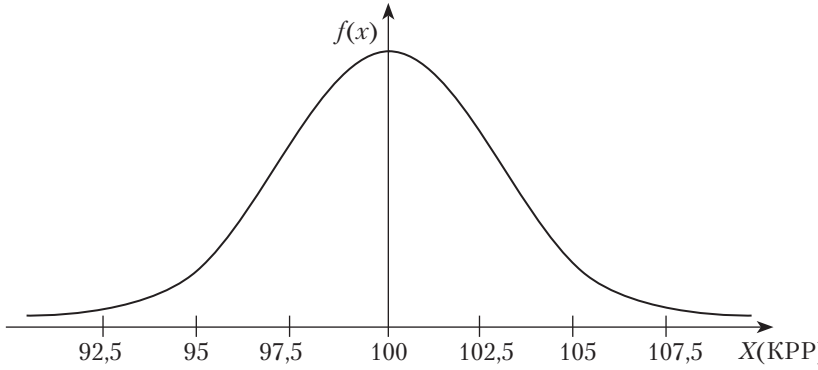


Рис. 6.8. **Розподіл вибіркових середніх**

початкове припущення виявиться правильним. Тобто необхідно вибрати 5 %-й інтервал, в якому будуть міститись найбільші значення вибіркових середніх.

Відповідно до таблиці відносних площ нормального розподілу, 5 % найбільших спостережень нормального розподілу повинні щонайменше на 1,64 стандартного відхилення перевищувати середнє генеральної сукупності. Для розподілу вибіркових середніх, представленого на рис. 6.9, 1,64 стандартного відхилення (тобто для вибіркових середніх $z = 1{,}64$) становлять 4,1:

$$z \cdot \frac{\sigma}{\sqrt{N}} = 1{,}64 \cdot \frac{20}{\sqrt{64}} = 1{,}64 \cdot 2{,}5 = 4{,}1.$$

Таким чином, 5% найбільших значень вибіркових середніх, які були б отримані при умові, що висунута гіпотеза є вірною, повинні перевищувати більше, ніж на 4,1 середнє генеральної сукупності. Інакше кажучи, 5% найбільших вибіркових середніх повинні мати величину, яка перевищує 104,1. Отриманий результат представлено на рис. 6.9 заштрихованою областю.

Нульова гіпотеза буде відхилена, якщо отримане значення вибіркового середнього перевищить 104,1. Тим самим ми підтвердимо початкове припущення, що середній КРР хлопчиків, які збираються стати пожежниками, перевищує 100. Інакше, гіпотеза буде прийнята. Вважається, що прийнятий рівень значущості 0,05 найкраще відповідає вимогам початкового припущення.

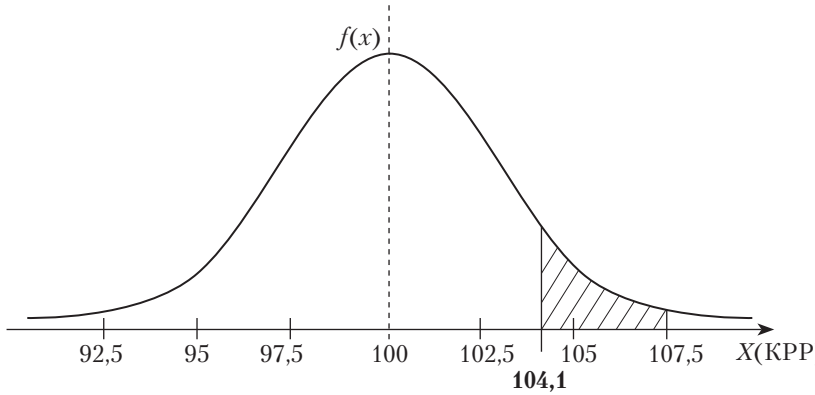


Рис. 6.9. **Форма розподілу вибіркових середніх для КРР; 5/100 найбільших значень перевищують рівень 104,1**

Такий метод перевірки називають *одностороннім тестом* (*критерієм*), тому що область відхилення гіпотези повністю міститься в одному з хвостів розподілу

Нехай необхідно довести, що середній КРР хлопчиків, які збираються стати пожежниками, є меншим ніж 115. У такому випадку потрібно відхилити гіпотезу, що середнє дорівнює 115. Для цього необхідно, щоб вибіркове середнє суттєво відрізнялось від цього значення, але тепер вже в іншому напрямку. Застосовуючи аналогічну аргументацію, можна встановити, що область відхилення гіпотези відповідає заштрихованій площі на рис. 6.10. У цьому прикладі ми також вважаємо, що $\sigma = 20$, а $N = 64$. Легко визначити, що на рівні значущості 0,05 граничне значення становить 110,9, тобто

$$\mu_{\text{гр}} - z \cdot \frac{\sigma}{\sqrt{N}} = 115 - 1{,}64 \cdot \frac{20}{\sqrt{64}} = 115 - 1{,}64 \cdot 2{,}5 = 110{,}9.$$

Коли необхідно зробити перевірку, що середнє просто відрізняється від деякого значення, використовується *двосторонній тест* (*критерій*). При цьому вибіркове середнє порівнюється з теоретично очікуваною величиною і визначається максимальне відхилення від неї в обидва боки. Це може дати нам можливість відхилити висунуту гіпотезу. Однак, якщо потрібно довести, що параметр відрізняється від очікуваного значення у *деякому визначеному напрямку*, то для цього використовується односторонній тест. Як і раніше, область відхилення гіпотези визначається рівнем значущості, але логіка задачі є та-

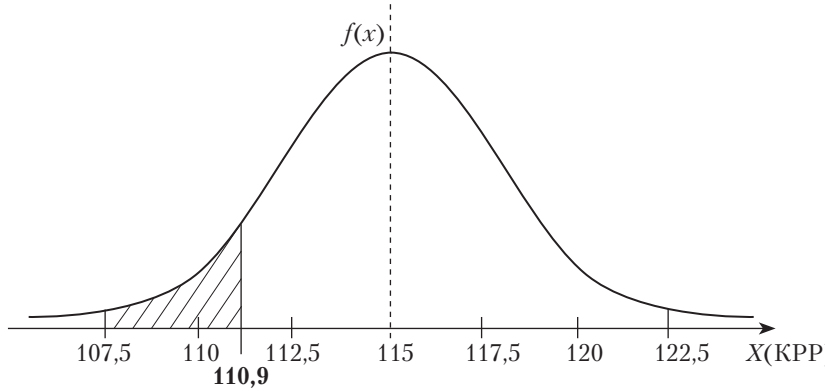


Рис. 6.10. **Приклад одностороннього тестування середнього**

кою, що у даному випадку необхідно приймати до уваги тільки одну область, що розташована в одному із хвостів розподілу.

Якщо необхідно довести, що середнє перевищує деяке значення, то область, в якій відхиляється гіпотеза, знаходиться у правому хвості. Якщо ж необхідно довести, що середнє вихідної (початкової) сукупності є меншим деякого значення, то така область буде розташована у лівому хвості.

Якщо гіпотеза припускає, *яким чином вибіркове середнє буде відрізнятись від очікуваної величини*, то необхідно користуватись одностороннім тестом. У такому випадку є більший шанс довести справедливість теорії, ніж при застосуванні двостороннього тесту. Так, якби у попередньому прикладі було застосовано двосторонній тест, то область відхилення гіпотези була б такою, як це показано на рис. 6.11.

Значення, які знаходяться між 104,1 і 104,9 у даному випадку будуть відрізнятись від очікуваного недостатньо, щоб довести вихідне припущення за допомогою двостороннього тесту. Однак, якщо отримане значення попаде в цей інтервал, то при застосуванні одностороннього тесту ми могли б відхилити висунуту гіпотезу.

По суті, у випадку застосування двостороннього тесту теорія просто прогнозує, що вибіркове середнє буде відрізнятись від очікуваного значення. Припущення вважається підтвердженим, якщо це середнє буде фактично відрізнятись від очікуваного значення. У випадку застосування одностороннього тесту гіпотеза припускає, що вибіркове середнє буде знаходитись з визначеного боку по відношенню до генерального середнього. Той факт, що таке припущення виявилось

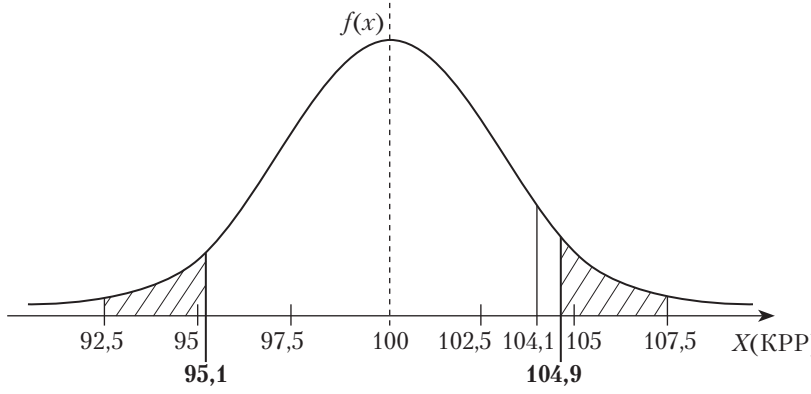


Рис. 6.11. **Приклад двостороннього тестування**

правильним, вже сам по собі є частковим підтвердженням висунутої гіпотези.

Інакше кажучи, цей факт також свідчить на користь даної гіпотези і, таким чином, величина розбіжності між очікуваним і фактичним показниками, яка необхідна для повного підтвердження висунутої гіпотези, буде меншою, ніж у випадку двостороннього тесту. Це зумовлено тим, що підтвердилось припущення щодо напрямку цієї відмінності.

Односторонній тест застосовують також при перевірці гіпотез стосовно інших параметрів. Наприклад, цей тест іноді застосовують для доведення того факту, що дисперсія вихідної сукупності перевищує деяке конкретне значення. У подальшому викладенні матеріалу будуть наведені інші приклади застосування одностороннього тесту.

6.7. Приклади застосування процедур перевірки гіпотез

Приклад 6.3. Термін функціонування (випадкова змінна X) дисплейних панелей, виготовлених за стандартною технологією, є нормальною випадковою величиною з параметрами розподілу: $\{X\} \sim N\{1000, (200)^2\}$. Після впровадження нової технології необхідно експериментально встановити, чи будуть мати нові панелі довший середній строк служби, ніж старі?

Сформулюємо *гіпотези*:

$$H_0: \mu = 1000 \text{ проти } H_1: \mu > 1000 \text{ годин функціонувань},$$

де μ — середнє генеральної сукупності.

Для перевірки гіпотези взяли випадкову вибірку обсягом $N = 25$ дисплейних панелей, виготовлених за новою технологією. Тобто вибірковий простір у даному випадку можна записати так:

$$\Omega = \{(x_1, x_2, ..., x_{25}) \mid 0 < x_i < \infty, i = 1, 2, ..., 25\},$$

де x_i — термін функціонування i-ї панелі.

Також припустимо, що ми відхилимо нуль-гіпотезу, якщо $\mu_1 > 1080$ годин, де μ_1 — середній термін функціонування нових панелей.

Приймемо наступні правила прийняття рішень (ППР):

Відхилити H_0 і прийняти H_1, якщо $\bar{x} > 1080$;
Відхилити H_1 і прийняти H_0, якщо $\bar{x} \le 1080$,

де $\bar{x}$ — середнє вибірки, взятої для дослідження. Таким чином, *область відхилення* H_0 визначається так:

$$\Omega_0 = \left\{ (x_1, x_2, \ldots, x_{25}) \middle| \bar{x} = \frac{x_1 + x_2 + \ldots + x_{25}}{25} > 1080 \right\}.$$

Будемо перевіряти гіпотези шляхом оцінювання ймовірностей припущення помилок 1-го та 2-го роду. Ймовірність помилки першого роду:

$$\begin{aligned}
\alpha &= P(\text{перевірка приведе до помилки 1-го роду}) = \\
&= P(\text{буде відхилена гіпотеза } H_0, \text{ коли } H_0 \text{ вірна}) = \\
&= P(\bar{x} > 1080 \text{ при } \mu = 1000) = \\
&= P\left(\frac{\bar{x} - \mu}{\sigma / \sqrt{N}} > \frac{1080 - 1000}{200 / \sqrt{25}} \right) = P\left(\frac{\bar{x} - \mu}{\sigma / \sqrt{N}} > 2 \right) = 0{,}02.
\end{aligned}$$

На рис. 6.12 наведено криву розподілу для цього випадку.

Обчислимо ймовірність помилки другого роду:

$$\begin{aligned}
\beta &= P(\text{перевірка приведе до помилки 2-го роду}) = \\
&= P(\text{буде прийнята гіпотеза } H_0, \text{ коли } H_0 \text{ невірна}) = \\
&= P(\bar{x} \le 1080 \text{ при } \mu > 1000) = \\
&= P\left(\frac{\bar{x} - \mu}{\sigma / \sqrt{N}} \le \frac{1080 - \mu}{200 / \sqrt{25}} \text{ при } \mu > 1000 \right).
\end{aligned}$$

Проте ми не можемо закінчити знаходження помилки другого роду, тому що не задано значення μ. У всіх прикладах, в яких альтер-

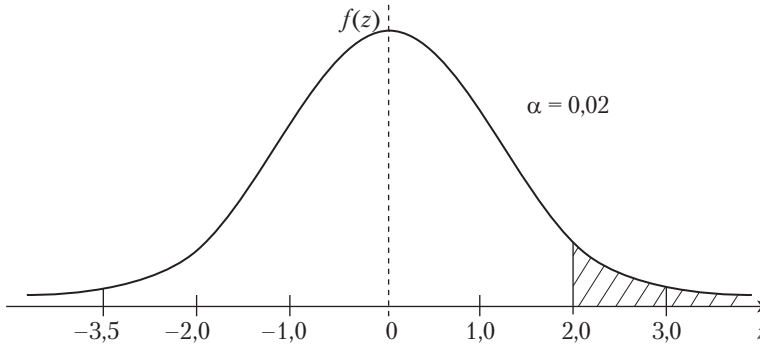


Рис. 6.12. **Крива нормального розподілу при перевірці гіпотез**

нативна гіпотеза не задає одне значення для параметра, обчислити β неможливо.

Повернемось до головного питання: чи будуть мати нові панелі довший середній термін функціонування, ніж старі? Припустимо, що для випадкової вибірки $N = 25$ нових панелей середній термін функціонування становив:

$$\bar{x} = 1100.$$

Відповідно до правила прийняття рішення ми повинні відхилити H_0, оскільки $\bar{x} > 1080$. Інакше кажучи, панелі, виготовлені за новою технологією, мають довший середній термін функціонування.

Якщо ми відхилимо H_0, то як ми дізнаємось, що прийняли правильне рішення? Пояснення наступне: у такому випадку єдино можливою помилкою є помилка 1-го роду (тобто відхилити H_0, коли H_0 вірна). Раніше ми визначили, що ймовірність помилки 1-го роду $\alpha = 0{,}02$. Така низька ймовірність помилки 1-го роду надає впевненості у правильності прийнятого рішення (відхиляючи її, ми помиляємось тільки у 2 % випадків).

6.8. Альтернативне формулювання процедури перевірки гіпотези

Припустимо, що необхідно сформулювати процедуру перевірки нуль-гіпотези, визначеної вище (для $N = 25$), але при цьому ймовірність помилки першого роду повинна дорівнювати 0,01 ($\alpha = 0{,}01$). Таку процедуру можна сформулювати наступним чином.

Скористаємось визначеними гіпотезами:

$$H_0 : \mu = 1000 \text{ проти } H_1 : \mu > 1000,$$

де μ — середнє нормальної генеральної сукупності при стандартному відхиленні $\sigma = 200$. Нехай $X = (x_1, x_2, ..., x_{25})$ — випадкова вибірка обсягом $N = 25$ з цієї генеральної сукупності; і нехай

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_{25}}{25},$$

де $\bar{x}$ — середній термін функціонування 25 панелей, виготовлених за новою технологією. Оскільки значення $\bar{x}$ очікується великим, якщо середнє генеральної сукупності μ велике, і значення $\bar{x}$ очікується малим, якщо середнє генеральної сукупності μ мале, то логічно відхили-

ти H_0 (і можливо прийняти H_1), якщо знайдене на основі експерименту $\bar{x}$ буде значно більшим 1000 ($\bar{x} > 1000$). Тепер можемо сформулювати правило прийняття рішення:

$$\text{відхилити } H_0, \text{ якщо } \bar{x} > c,$$

де c — деяка константа.

Визначимо граничне (критичне) значення константи c так, щоб ймовірність помилки 1-го роду дорівнювала 0,01 ($\alpha = 0{,}01$):

$$\begin{aligned}\alpha &= P(\text{перевірка приведе до помилки 1-го роду}) = \\ &= P(\text{буде відхилена гіпотеза } H_0, \text{ коли } H_0 \text{ вірна}) = \\ &= P(\bar{x} > c \text{ при } \mu = 1000) = \\ &= P\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{N}} > \frac{c-1000}{200/5}\right) = P\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{N}} > \frac{c-1000}{40}\right).\end{aligned}$$

З таблиці відносних площ нормального розподілу знайдемо, що при $\alpha = 0{,}01$ константу можна знайти з рівняння:

$$\frac{c-1000}{40} = 2{,}327.$$

Таким чином, $c = 1093$. Наведену рівність можна інтерпретувати так: z-оцінка $z = \dfrac{c-1000}{40} = 2{,}327$ відповідає лівій границі області нормального розподілу, яка займає 1 % від загальної відносної площі і знаходиться у правому хвості розподілу.

Тепер можна сформулювати правило прийняття рішень, яке базується на аналізі випадкової вибірки обсягом $N = 25$ при ймовірності припуститись помилки першого роду $\alpha = 0{,}01$:

$$\text{відхилити } H_0, \text{ якщо } \bar{x} > 1093.$$

На основі розглянутого прикладу можна сформулювати *альтернативний метод перевірки гіпотез*, який може бути застосований до всіх випадків з простою нуль-гіпотезою проти простої або композиційної альтернативної гіпотези.

Зазначимо, що гіпотеза називається **простою**, якщо вона задає одне значення для параметра, що аналізується (наприклад, $\mu = 1000$, $\sigma = 0{,}5$). Гіпотеза називається **композиційною**, якщо названий параметр може приймати більше одного можливого значення (наприклад, $\mu > 1000$).

6.9. Метод побудови процедури перевірки гіпотез

Крок 1. Знайти такий статистичний параметр (тобто функцію випадкової змінної), розподіл якої повністю визначений, якщо H_0 вірна. (Це дає змогу знайти α, якщо ми використовуємо даний статистичний параметр як тестову статистику.)

У попередньому прикладі цією статистикою було вибіркове середнє $\bar{x}$.

Крок 2. Скористатись статистикою, знайденою на попередньому кроці, як тестовою (перевірочною). Інакше кажучи, використати цю статистику для формулювання такого правила прийняття рішень, для якого ймовірність помилки 1-го роду становить α. (Тобто за допомогою вибраної статистики необхідно визначити критичну область відповідно до α.)

У попередньому прикладі ми визначили правило прийняття рішення після знаходження константи c з умови: $P(\bar{x} > c$ коли H_0 вірна$) = 0{,}01$.

Крок 3. Формалізувати правило прийняття рішення, сформульоване на попередньому кроці, із врахуванням конкретного експериментального значення тестової статистики.

У попередньому прикладі таким формальним правилом було:

$$\text{відхилити } H_0, \text{ якщо } \bar{x} > 1093.$$

Приклад 6.4. Припустимо, що середнє генеральної сукупності може приймати одне з двох значень: 1 або 2, але в обох випадках стандартне відхилення становить $\sigma = 0{,}5$. Використовуючи наведену трикрокову процедуру, необхідно побудувати процедуру перевірки гіпотез:

$$H_0: \mu = 1 \text{ проти } H_1: \mu = 2.$$

Для цього скористаємось випадковою вибіркою обсягом $N = 9$ та ймовірністю помилки першого роду $\alpha = 0{,}05$.

Крок 1. Знайти такий статистичний параметр (тобто функцію випадкової змінної), розподіл якої повністю визначений, якщо H_0 вірна.

У цьому випадку можна скористатись вибірковим середнім $\bar{x}$. Це можливо завдяки тому, що у випадку справедливості H_0 (тобто $\mu = 1$), $\bar{x}$ має нормальний розподіл із середнім $\mu = 1$ (при великому числі вибірок обсягом $N = 9$) і стандартним відхиленням $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{N}} = \dfrac{0{,}5}{\sqrt{9}} = \dfrac{0{,}5}{3}$.

Інакше кажучи, випадкову змінну $\bar{x}$ можна використати як тестову статистику, оскільки її розподіл (параметри розподілу) повністю відомий, якщо припустити, що H_0 вірна.

Крок 2. Використати цю статистику для формулювання такого правила прийняття рішень, для якого ймовірність помилки 1-го роду становить α.

Оскільки значення μ є більшим для H_1, ніж для H_0, то логічно очікувати більші значення $\bar{x}$ при справедливості гіпотези H_1, а не H_0. Таким чином, буде логічно відхилити гіпотезу H_0, якщо значення тестової статистики $\bar{x}$ будуть відносно великими. Тобто правило прийняття рішення можна сформулювати у вигляді:

$$\text{відхилити } H_0, \text{ якщо } \bar{x} > k,$$

де k — деяка константа. Знайдемо значення цієї константи, скориставшись визначенням ймовірність помилки 1-го роду (вона задана на рівні $\alpha = 0{,}05$):

$$\begin{aligned}\alpha &= P(\text{перевірка приведе до помилки 1-го роду}) = \\ &= P(\text{буде відхилена гіпотеза } H_0, \text{ коли } H_0 \text{ вірна}) = \\ &= P(\bar{x} > k \text{ при } \mu = 1) = \\ &= P\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{N}} > \frac{k-1}{0{,}5/\sqrt{9}}\right).\end{aligned}$$

З таблиці відносних площ нормального розподілу знайдемо, що значенню α відповідає ліва границя області в правому хвості розподілу, яка дорівнює 1,645, тобто:

$$\frac{k-1}{0{,}5/3} = 1{,}645.$$

Звідси маємо: $k = 1{,}2742$.

Крок 3. Формалізувати правило прийняття рішення, сформульоване на попередньому кроці, з урахуванням конкретного експериментального значення тестової статистики. Правило прийняття рішень, сформульоване на другому кроці, має вигляд:

$$\text{відхилити } H_0, \text{ якщо } \bar{x} > 1{,}2742.$$

На рис. 6.13 представлені елементи процедури перевірки гіпотез до прикладу 2. Зробимо резюме щодо основних елементів процедури перевірки гіпотез для розглянутого прикладу:

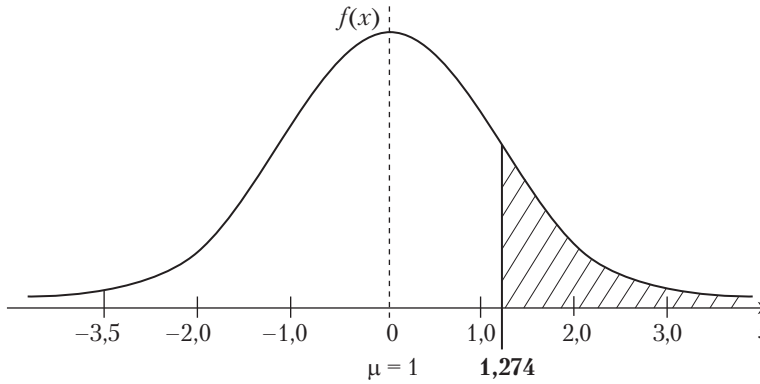


Рис. 6.13. **Крива нормального розподілу при перевірці гіпотез**

1. Гіпотези: відхилити $H_0: \mu = 1$ проти $H_1: \mu > 2$.
2. Суцільна крива нормального розподілу (рис. 6.13) відповідає щільності розподілу $\bar{x}$, якщо справедлива нуль-гіпотеза: $\mu = 1$.
3. Правило прийняття рішення: відхилити H_0, якщо $\bar{x} > 1{,}2742$. (Ймовірніше, що значення $\bar{x}$ будуть перевищувати 1,2742, якщо справедливою буде H_1, а не H_0.)
4. Площа правого хвоста під кривою розподілу (суцільна лінія) в межах від 1,2742 до $+\infty$ дорівнює $\alpha = P$(помилки 1-го роду) = = 0,05.

Приклад 6.5. Побудуйте трикрокову процедуру перевірки гіпотез (за аналогією попереднього прикладу) для перевірки таких гіпотез:

$$H_0: \mu = \mu_0 \text{ проти } H_1: \mu < \mu_0,$$

де μ — невідоме середнє нормальної генеральної сукупності із стандартним відхиленням σ; μ_0 — константа.

1. Побудуйте процедуру перевірки гіпотез, яка базується на випадковій вибірці розміром n при ймовірності зробити помилку 1-го роду $\alpha = 0{,}05$.

2. Побудуйте процедуру перевірки гіпотез, яка базується на випадковій вибірці обсягом n при ймовірності α зробити помилку 1-го роду.

Розв'язок

1. *Крок 1.* Знайти такий статистичний параметр, розподіл якого повністю визначений, якщо H_0 вірна.

Розглянемо випадкову змінну $\frac{\bar{x}-\mu_0}{\sigma/\sqrt{N}}$., де $\bar{x}$ — вибіркове середнє. Чи можна використати цю випадкову змінну як тестову статистику? Так, можна, тому що у випадку, коли H_0 вірна (тобто якщо $\mu = \mu_0$), то наведена випадкова змінна має стандартний (нормований) нормальний розподіл.

Крок 2. Використати статистику, вибрану на кроці 1, для формулювання такого правила прийняття рішень, для якого ймовірність помилки 1-го роду становить 0,05.

Оскільки спостереження $\bar{x}$ групуються навколо істинного значення середнього генеральної сукупності μ, то логічно відхилити H_0 у тому випадку (і, можливо, прийняти H_1), коли спостережуване значення $\bar{x}$ буде значно меншим μ_0.

Оскільки вибрана змінна $\frac{\bar{x}-\mu_0}{\sigma/\sqrt{N}}$ має нульове середнє, то логічно відхилити H_0, *коли спостережуване значення цієї змінної буде значно меншим нуля* (тобто значно меншим середнього). Інакше кажучи, правило прийняття рішення можна сформулювати так:

$$\text{відхилити } H_0, \text{ якщо } \frac{\bar{x}-\mu_0}{\sigma/\sqrt{N}} < c,$$

де $c < 0$ — константа. Знайдемо значення цієї константи за умови, що при застосуванні наведеного правила прийняття рішення можна зробити помилку 1-го роду з ймовірністю 0,05:

$$\begin{aligned}
\alpha &= P(\text{перевірка приведе до помилки 1-го роду}) = \\
&= P(\text{буде відхилена гіпотеза } H_0, \text{ коли } H_0 \text{ вірна}) = \\
&= P\left(\frac{\bar{x}-\mu_0}{\sigma/\sqrt{N}}. < c \text{ при } \mu = \mu_0\right) = \\
&= P\left(\frac{\bar{x}-\mu_0}{\sigma/\sqrt{N}} < c\right).
\end{aligned}$$

З таблиці відносних площ нормального розподілу знайдемо, що при $\alpha = 0{,}05$ значення константи c повинно дорівнювати: $c = -1{,}645$.

Крок 3. Формалізувати правило прийняття рішення, сформульоване на попередньому кроці, із врахуванням конкретного експериментального значення тестової статистики.

$$\text{відхилити } H_0, \text{ якщо } \frac{\bar{x}-\mu_0}{\sigma/\sqrt{N}} < -1{,}645.$$

2. Визначення правила прийняття рішення при умові, що ймовірність помилки 1-го роду становить α, зводиться до заміни числа 1,645 на загальну величину: $-z_\alpha$:

$$p\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{N}} < z_\alpha\right) = \alpha.$$

Примітка. Значення z_α визначається з рівняння:

$$\int_{z_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = \alpha.$$

Тобто, частина відносної площі нормального розподілу зліва від $-z_\alpha$ буде дорівнювати α. Таким чином, правило прийняття рішення приймає вигляд:

$$\text{відхилити } H_0, \text{ якщо } \frac{\bar{x}-\mu_0}{\sigma/\sqrt{N}} < -z_\alpha.$$

Основні елементи цієї процедури тестування гіпотези показано на рис. 6.14.

Приклад 6.6. Випадкова вибірка чашечок з кавою (N = 25), взята з кавоварного автомата, має середній вміст кави $\bar{x}$ = 93 г на чашку. Скористайтесь тестом, розробленим у попередньому прикладі, для перевірки нуль-гіпотези, що середнє генеральної сукупності для цього автомата становить μ = 95 проти альтернативної гіпотези: $\mu < 95$

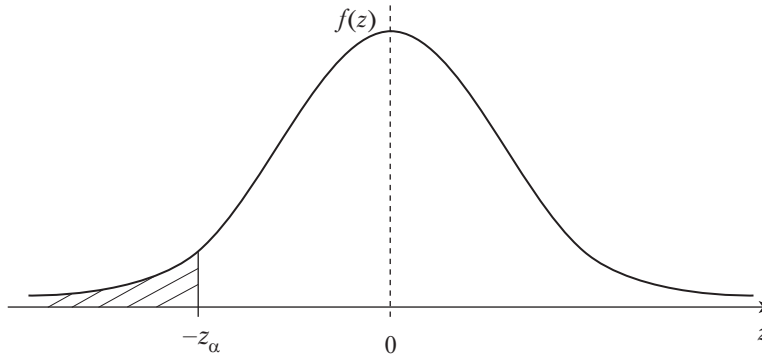


Рис. 6.14. **Елементи процедури тестування для прикладу 3.**
Площа під кривою від $-\infty$ до $-z_\alpha$ дорівнює
α = *P*(ймовірність помилки 1-го роду)

(рівень значущості $\alpha = 0{,}05$). Припустимо, що маса кави у чашечках має нормальний розподіл з дисперсією $\sigma^2 = 1$.

Розв'язок

Процедура тестування, розроблена у попередньому прикладі, визначена наступним правилом прийняття рішень:

$$\text{відхилити } H_0, \text{ якщо } \frac{\bar{x} - \mu_0}{\sigma/\sqrt{N}} < -1{,}645.$$

У цьому прикладі $\sigma = 1$, $N = 25$ і $\mu = 95$, а звідси значення тестової статистики

$$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{N}} = \frac{93 - 95}{1/\sqrt{25}} = -\frac{2}{0{,}2} = -10.$$

Оскільки -10 є меншим від $-1{,}645$, то нуль-гіпотеза повинна бути відхилена. Таким чином, середня маса кави, яка надається споживачеві автоматом, є суттєво меншою 95 г, тому що ми відхилили нуль-гіпотезу на рівні значущості 5 %.

Приклад 6.7. Використовуючи наведену вище трикрокову методику побудови гіпотез, побудуйте процедуру для тестування таких гіпотез:

$$H_0 : \mu = \mu_0 \text{ проти } H_1 : \mu \neq \mu_0,$$

де μ — невідоме середнє генеральної сукупності з нормальним розподілом; стандартне відхилення σ цієї сукупності відоме. Необхідно побудувати процедуру перевірки гіпотез з використанням випадкової вибірки обсягом N за припущення, що помилка першого роду можлива з ймовірністю α.

Розв'язок

Крок 1. Знайти такий статистичний параметр, розподіл якого повністю визначений, якщо H_0 вірна.

Оскільки генеральна сукупність та нуль-гіпотеза даного прикладу є ідентичними генеральній сукупності та нуль-гіпотезі прикладу 3, то в якості перевірочної статистики можна скористатись аналогічною випадковою змінною:

$$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{N}}.$$

Крок 2. Використати статистику, вибрану на кроці 1, для формулювання такого правила прийняття рішень, для якого ймовірність помилки 1-го роду становить α.

Оскільки цілком ймовірно, що спостереження $\bar{x}$ будуть концентруватись навколо μ, то буде логічно відхилити H_0 (і можливо прийняти H_1), коли спостережуване значення перевірочної статистики $(\bar{x} - \mu_0)/(\sigma/N^{1/2})$ суттєво відрізняється від нуля (тобто від його середнього).

Інакше кажучи, логічно скористатись наступним правилом прийняття рішення

$$\text{відхилити } H_0, \text{ якщо } \left|\frac{\bar{x}-\mu_0}{\sigma/\sqrt{N}}\right| > c,$$

де $c = \text{const} > 0$. Для визначення цієї константи скористаємось визначенням ймовірності помилки 1-го роду:

$$\begin{aligned}\alpha &= P(\text{перевірка приведе до помилки 1-го роду}) = \\ &= P(\text{буде відхилена гіпотеза } H_0, \text{ коли } H_0 \text{ вірна}) = \\ &= P\left(\left|\frac{\bar{x}-\mu_0}{\sigma/\sqrt{N}}\right| > c \text{ при } \mu = \mu_0\right) = P\left(\left|\frac{\bar{x}-\mu}{\sigma/\sqrt{N}}\right| > c\right).\end{aligned}$$

Оскільки змінна $(\bar{x} - \mu_0)/(\sigma/N^{1/2})$ розподілена за стандартним (нормованим) нормальним розподілом, то константа c задовольняє рівняння:

$$c = z_{\alpha/2}.$$

Крок 3. Формалізувати правило прийняття рішення, сформульоване на попередньому кроці, із врахуванням конкретного значення тестової статистики.

Правило прийняття рішення із врахуванням значення константи c:

$$\text{відхилити } H_0, \text{ якщо } \left|\frac{\bar{x}-\mu_0}{\sigma/\sqrt{N}}\right| > z_{\alpha/2}.$$

Основні елементи розглянутої процедури перевірки гіпотез наведено на рис. 6.15 (резюме до процедури перевірки).

1. Сформульовані гіпотези: $H_0: \mu = \mu_0$ проти $H_1: \mu \neq \mu_0$.
2. Крива щільності розподілу наведена для змінної $(\bar{x} - \mu_0)/(\sigma/N^{1/2})$ при $\mu = \mu_0$.
3. Правило прийняття рішень: відхилити H_0, якщо $\left|\frac{\bar{x}-\mu_0}{\sigma/\sqrt{N}}\right| > z_{\alpha/2}$. (Якщо вірна H_0, то ймовірніше, що значення $(\bar{x} - \mu_0)/(\sigma/N^{1/2})$ будуть близько 0.)
4. Площа під кривою від $-\infty$ до $-z_{\alpha/2}$ і від $z_{\alpha/2}$ до $+\infty$ дорівнює $\alpha = P$(помилки 1-го роду).

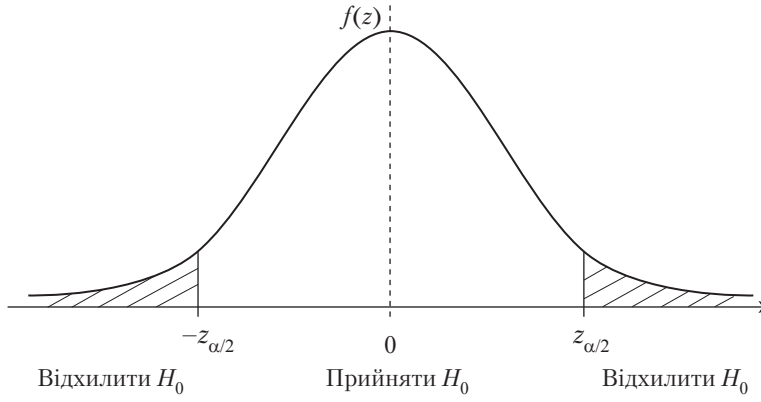


Рис. 6.15. **Елементи процедури тестування для прикладу 5**

6.10. Особливості односторонньої і двосторонньої перевірки гіпотез

У даному випадку перевірка гіпотез виконувалася із використанням обох хвостів розподілу. Таку перевірку називають **двосторонньою**, або перевіркою з двома хвостами розподілу. Двосторонні тести, як правило, застосовують для перевірки гіпотез наступного типу:

$$H_0: \theta = \theta_0 \text{ проти } H_1: \theta \neq \theta_0 .$$

Перевірка гіпотез, виконана у прикладах 1–3, називається **односторонньою** перевіркою (або перевіркою з використанням одного хвоста розподілу). При односторонній перевірці застосовують правила прийняття рішень, які передбачають відхилення нуль-гіпотези у випадку, коли тестова статистика приймає значення тільки зліва від деякого числа (порогового значення) або тільки справа від нього. Такі процедури перевірки, як правило, виконують перевірку гіпотез вигляду:

$$H_0: \theta = \theta_0 \text{ проти } H_1: \theta > \theta_0$$

або

$$H_0: \theta = \theta_0 \text{ проти } H_1: \theta < \theta_0 .$$

6.11. Оцінювання надійності правил перевірки статистичних гіпотез

1. Наскільки надійною є конкретна процедура перевірки статистичної гіпотези?

2. Чи є вона кращою від іншої процедури тестування тієї самої гіпотези?
3. Чи є вона найкращою серед множини всіх процедур перевірки гіпотез?
4. Які є стандартні критерії оцінювання якості (надійності) процедур перевірки гіпотез?

Відповіді на ці питання залежать від типу конкретної процедури, що аналізується. Процедури тестування *простих нуль-гіпотез проти простих альтернативних* гіпотез оцінюються їх *потужністю*. Процедури перевірки *простих гіпотез проти альтернативних композиційних* оцінюють за їх *функціями потужності* або за *операційними характеристичними кривими*.

Нижче розглянемо деякі з цих характеристик статистичних тестів і скористаємось ними для вибору *найкращих тестів* та *найбільш потужних тестів*.

Потужність тесту — це ступінь його здатності відхилити нуль-гіпотезу, якщо нуль-гіпотеза невірна.

Розглянемо процедуру перевірки простих гіпотез

$$H_0 : \theta = \theta_0 \text{ проти } H_1 : \theta = \theta_1 .$$

де θ — параметр розподілу генеральної сукупності.

Означення 6.4. *Потужністю* процедури (правила) перевірки називають ймовірність того, що тестування приведе до відхилення нуль-гіпотези $\theta = \theta_0$, якщо вірна альтернативна гіпотеза $\theta = \theta_1$. Тобто потужність тесту визначається величиною $1 - \beta$, де β — ймовірність помилки другого роду, тобто ймовірність прийняття нуль-гіпотези, якщо вірною є альтернативна.

Якщо виконується перевірка простої гіпотези $H_0 : \theta = \theta_0$ проти деякої композиційної гіпотези стосовно θ, то потужність тесту при $\theta = \theta_1$ визначається ймовірністю того, що тест приведе до відхилення нуль-гіпотези, $\theta = \theta_0$, якщо дійсним значенням параметра $\theta \in \theta_1$.

Приклад 6.8. Нехай $\mathfrak{R}$ — популяція, яка має нормальний розподіл з дисперсією $\sigma^2 = 0{,}25$, але її середнє μ — невідоме. Необхідно знайти α, β, а також потужності кожного з наведених нижче правил щодо перевірки гіпотез

$$H_0 : \mu = 1 \text{ проти } H_1 : \mu = 2.$$

Кожне правило базується на спостережуваному значенні середнього $\bar{x}$ випадкової вибірки (з генеральної сукупності) обсягом $N = 9$. Дано три правила [13]:

правило 1: відхилити H_0, якщо $\bar{x} > 1{,}3878$;

правило 2: відхилити H_0, якщо $\bar{x} < 0{,}6122$;

правило 3: відхилити H_0, якщо $\bar{x} > 2$ або якщо $\bar{x} < 0{,}6122$.

Розв'язок

Для правила 1:

$$\alpha = P(\text{відхилення } H_0, \text{ якщо } H_0 \text{ вірна}) =$$
$$= P(\bar{x} > 1{,}3878, \text{ якщо } \mu = 1) =$$
$$= P\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{N}} > \frac{1{,}3878-1}{0{,}5/\sqrt{9}}\right) = P\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{N}} > 2{,}327\right) = 0{,}01.$$

Примітка. При обчисленнях був використаний той факт, що змінна $(\bar{x}-\mu)/(\sigma/\sqrt{N})$ має стандартний нормальний розподіл.

$$\beta = P(\text{прийняття } H_0, \text{ якщо } H_0 \text{ невірна}) =$$
$$= P(\text{прийняття } H_0, \text{ якщо } H_1 \text{ вірна}) =$$
$$= P(\bar{x} \le 1{,}3878, \text{ якщо } \mu = 2) =$$
$$= P\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{N}} \le \frac{1{,}3878-2}{0{,}5/\sqrt{9}}\right) = P\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{N}} \le -3{,}6732\right) = 0{,}001.$$

Потужність 1-го правила $Tp1$:

$$Tp1 = P(\text{відхилення } H_0, \text{ якщо вірна } H_1) =$$
$$= P(\text{прийняття } H_0, \text{ якщо вірна } H_1) =$$
$$= 1 - \beta = 1 - 0{,}001 = 0{,}999.$$

Для правила 2:

$$\alpha = P(\text{відхилення } H_0, \text{ якщо } H_0 \text{ вірна}) =$$
$$= P(\bar{x} < 0{,}6122, \text{ якщо } \mu = 1) =$$
$$= P\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{N}} < \frac{0{,}6122-1}{0{,}5/\sqrt{9}}\right) = P\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{N}} < -2{,}3268\right) = 0{,}01.$$

$$\beta = P(\text{прийняття } H_0, \text{ якщо } H_1 \text{ вірна}) =$$
$$= P(\bar{x} \ge 0{,}6122, \text{ якщо } \mu = 2) =$$
$$= P\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{N}} \ge \frac{0{,}6122-2}{0{,}5/\sqrt{9}}\right) = P\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{N}} \ge -8{,}3268\right) = 1{,}0.$$

Потужність 2-го правила:

$$Tp2 = 1 - P(\text{прийняття } H_0, \text{ якщо вірна } H_1) = \\ = 1 - \beta = 1 - 1 = 0.$$

Для правила 3:

$$\alpha = P(\bar{x} > 2 \text{ або } \bar{x} < 0{,}6122, \text{ якщо } \mu = 1) = \\ = P(\bar{x} > 2, \text{ якщо } \mu = 1) + P(\bar{x} < 0{,}6122, \text{ якщо } \mu = 1) = 0{,}01;$$

$$\beta = P(0{,}6122 \le \bar{x} \le 2, \text{ якщо } \mu = 2) = 0{,}5\,.$$

Потужність 3-го правила: $Tp3 = 1 - \beta = 1 - 0{,}5 = 0{,}5$.

Таким чином, всі три правила прийняття рішень характеризуються однаковою ймовірністю відхилення нуль-гіпотези, коли вона вірна. Ймовірності прийняття нуль-гіпотези, якщо вірна альтернативна, становлять 0,001; 1,0 і 0,5, відповідно. Потужності ППР (*ймовірності відхилення нуль-гіпотез, якщо вірна альтернативна*) дорівнюють: 0,999; 0,0 і 0,5. Тобто з трьох правил перше характеризується самою високою ймовірністю прийняття правильного рішення; воно має найбільшу потужність і є, відповідно, найкращим з трьох наведених. Формальною мовою статистичного висновку найкраще правило (тест) визначається наступним чином.

Означення 6.5. При тестуванні простих нуль-гіпотез проти простих альтернативних гіпотез правило, що характеризується ймовірністю помилки 1-го роду α, буде *найкращим серед всіх правил*, які характеризуються ймовірністю помилки 1-го роду α, що має найбільшу потужність (інакше кажучи, воно має найменше β).

Примітка. Означення 6.5 передбачає, що всі правила прийняття рішень базуються на випадкових вибірках однакового обсягу *N*.

Порівняння правил прийняття рішень не завжди виконується просто. Розглянемо приклад перевірки гіпотез та порівняння ППР у випадку простої нуль-гіпотези і композиційної альтернативної.

Приклад 6.9. Нехай необхідно тестувати гіпотези

$$H_0 : \mu = 0 \text{ проти } H_1 : \mu \neq 0\,,$$

де μ — середнє нормальної генеральної сукупності з дисперсією $\sigma^2 = 4$. Для наведених нижче ППР необхідно знайти: α, потужність правила при $\mu = 0{,}5$ і потужність правила при $\mu = -0{,}5$. Позначимо ви-

біркове середнє через $\bar{x}$ для випадкової вибірки розміром $N = 25$ з генеральної сукупності:

правило 1: відхилити H_0, якщо $|\bar{x}| > 0{,}784$;
правило 2: відхилити H_0, якщо $\bar{x} > 0{,}658$;
правило 3: відхилити H_0, якщо $\bar{x} < -0{,}658$.

Розв'язок

Для правила 1:

$\alpha = 0{,}05$;
$Tp1(\text{при } \mu = 0{,}5) = 0{,}239$;
$Tp1(\text{при } \mu = -0{,}5) = 0{,}239$.

Для правила 2:

$\alpha = 0{,}05$;
$Tp2(\text{при } \mu = 0{,}5) = 0{,}347$;
$Tp1(\text{при } \mu = -0{,}5) = 0{,}002$.

Для правила 3:

$\alpha = 0{,}05$;
$Tp3(\text{при } \mu = 0{,}5) = 0{,}002$;
$Tp1(\text{при } \mu = -0{,}5) = 0{,}347$.

Усі три ППР мають однакову ймовірність відхилення нуль-гіпотези, коли вона вірна ($\alpha = 0{,}05$). При $\mu = 0{,}5$ потужності ППР мають такі значення: 0,239; 0,347 і 0,002. Таким чином, якби ми перевіряли нуль-гіпотезу $\mu = 0$ проти альтернативної $\mu = 0{,}5$, то найкращим було б правило 2, а на другому місці — правило 1.

Однак, якби ми перевіряли нуль-гіпотезу $\mu = 0$ проти альтернативної $\mu = -0{,}5$, то найкращим було б 3-тє ППР, на другому місці — 1-ше, а 2-ге було б найгіршим (відповідні потужності: 0,239; 0,002 і 0,347).

Оскільки конкретне значення потужності характеризує правило тільки при конкретному значенні тестової статистики, то для отримання повної оцінки ППР необхідно знайти його потужність для кожного можливого значення статистики. Це приводить до визначення *операційної характеристичної кривої* ППР та його *функції потужності*.

Розглянемо задачу перевірки простої нуль-гіпотези, що стосується параметра, проти альтернативної композиційної гіпотези стосовно

того самого параметра. Нехай $\beta(\theta)$ ймовірність прийняття нуль-гіпотези, якщо значення параметра дорівнює θ.

Означення 6.6. Значення $\beta(\theta)$ для кожного значення θ називають *операційною характеристикою* ППР, а криву, яка визначається функцією $\beta(\theta)$, називають *операційною характеристичною кривою* ППР.

Примітка. У випадку перевірки простої нуль-гіпотези ($\theta = \theta_0$) проти простої альтернативної гіпотези ($\theta = \theta_1$) функція $\beta(\theta)$ спрощується до $\beta(\theta_1) = \beta$, де β — ймовірність помилки 2-го роду.

Розглянемо перевірку простої нуль-гіпотези стосовно параметра проти композиційної альтернативної гіпотези стосовно того самого параметра.

Означення 6.7. Нехай $\pi(\theta)$ — ймовірність відхилення нуль-гіпотези, якщо значенням параметра є θ. Для кожного конкретного значення параметра θ значення $\pi(\theta)$ називають потужністю тесту. Множина значень $\pi(\theta)$ утворює функцію *потужності тесту*. Функція потужності та операційна характеристична крива тесту пов'язані співвідношенням

$$\pi(\theta) = 1 - \beta(\theta).$$

Примітка. Якщо нуль-гіпотезою є H_0: $\theta = \theta_0$, то $\pi(\theta_0) = \alpha$, де α — ймовірність помилки 1-го роду.

Операційні характеристичні криві та функції потужності використовують для оцінювання ППР щодо простих нуль-гіпотез проти композиційних альтернативних гіпотез. Наприклад, якщо перевіряють гіпотези

$$H_0 : \theta = \theta_0 \text{ проти } H_1 : \theta \neq \theta_1,$$

то бажано мати високу ймовірність прийняття нуль-гіпотези, якщо $\theta = \theta_0$, і низьку ймовірність прийняття гіпотези, якщо $\theta \neq \theta_0$. Тобто бажано мати велике значення для $\beta(\theta_0)$, але якщо $\theta \neq \theta_0$, то бажано мати малі значення для $\beta(\theta)$. Інакше кажучи, бажаним тестом (ППР) є такий, операційна характеристична крива якого має високе значення при $\theta = \theta_0$ і низькі значення при інших значеннях θ.

6.12. Контрольні питання і вправи

1. Поясніть помилки першого і другого роду, які можуть виникати при прийнятті рішень.

2. Чи можна одночасно зменшувати ризик припуститись помилки одного роду без одночасного збільшення помилки іншого роду?
3. Який результат експерименту є цілком достатнім для відхилення як неправильної висунутої гіпотези?
4. Яку мету стосовно перевірки гіпотез необхідно ставити при постановці експерименту?
5. Яку величину називають рівнем значущості експерименту?
6. Сформулюйте загальну багатокрокову процедуру перевірки гіпотез і продемонструйте її застосування на прикладі.
7. Поясніть величини, що входять у формулу для знаходження z-оцінок результату експерименту для вибірки дихотомічної змінної і наведіть приклад оцінювання:

$$z = \frac{\text{Оцінка відхилення}}{\text{Стандартне відхилення}} = \frac{r - Np}{\sqrt{Np(1-p)}}.$$

8. До змінних якого характеру відноситься перевірка гіпотез у задачах першого типу?
9. Які задачі відносять до другого типу при перевірці гіпотез? Наведіть ілюстративний приклад.
10. Наведіть приклад задачі третього типу при перевірці гіпотез.
11. Поясніть сутність використання одностороннього критерію при перевірці гіпотез. Скористайтесь графіком кривої нормального розподілу.
12. У чому полягає сутність використання двостороннього критерію при перевірці гіпотез? Поясніть на прикладі.
13. Термін функціонування (випадкова змінна X) мобільних телефонів, виготовлених за стандартною технологією, є нормальною випадковою величиною з параметрами розподілу: $\{X\} \sim N\{5000, (100)^2\}$. Після впровадження нової технології необхідно експериментально встановити, чи будуть мати нові панелі довший середній термін функціонування, ніж старі.
14. Які гіпотези називають простими, а які композиційними? Наведіть приклади обох видів гіпотез.
15. Сформулюйте трикроковий метод побудови процедури перевірки гіпотез.
16. Припустимо, що середнє генеральної сукупності може приймати одне з двох значень: 0,5 або 1, але в обох випадках стандартне відхилення становить $\sigma = 0{,}25$. Використовуючи наведену вище

трикрокову процедуру, необхідно побудувати процедуру перевірки гіпотез

$$H_0 : \mu = 0{,}5 \text{ проти } H_1 : \mu = 1.$$

Для цього скористайтесь випадковою вибіркою обсягом $N = 20$ та ймовірністю помилки першого роду $\alpha = 0{,}05$.

17. Побудуйте трикрокову процедуру перевірки гіпотез (за аналогією з попереднім прикладом) для перевірки таких гіпотез:

$$H_0 : \mu = \mu_0 \text{ проти } H_1 : \mu < \mu_0,$$

де μ — невідоме середнє нормальної генеральної сукупності із стандартним відхиленням σ; μ_0 — константа.

1. Побудуйте процедуру перевірки гіпотез, яка базується на випадковій вибірці обсягом 100 при ймовірності зробити помилку 1-го роду $\alpha = 0{,}05$.

2. Побудуйте процедуру перевірки гіпотез, яка базується на випадковій вибірці обсягом n при ймовірності зробити помилку 1-го роду α.

18. Випадкова вибірка стаканчиків із соком ($N = 50$), взята з торговельного автомата, має середній вміст соку $\bar{x} = 98$ г на чашку. Скористайтесь тестом, розробленим у попередньому прикладі, для перевірки нуль-гіпотези, що середнє генеральної сукупності для цього автомата складає $\mu = 100$ проти альтернативної гіпотези: $\mu < 100$ (рівень значущості $\alpha = 0{,}05$). Припустимо, що маса соку в стаканчиках має нормальний розподіл з дисперсією $\sigma^3 = 3$.
19. Сформулюйте означення потужності тесту.
20. Сформулюйте потужність правила перевірки гіпотез.
21. Яке правило, що характеризується ймовірністю помилки 1-го роду α, при тестуванні простих нуль-гіпотез проти простих альтернативних гіпотез, буде *найкращим серед всіх правил*, які характеризуються ймовірністю помилки 1-го роду α?
22. Дайте означення функції потужності тесту. Яким виразом пов'язані між собою функція потужності тесту і характеристична крива?
23. Наведіть приклади застосування функції потужності тесту і характеристичної кривої.

Розділ 7

ДЕЯКІ СТАНДАРТНІ ПРОЦЕДУРИ ПЕРЕВІРКИ ГІПОТЕЗ ТА ІНТЕРВАЛЬНЕ ОЦІНЮВАННЯ

7.1. Перевірка гіпотез щодо середніх

На практиці досить часто виникають задачі перевірки середніх у випадках, коли необхідно порівняти характеристики вибраної групи елементів із деякими прийнятими або стандартизованими характеристиками. Подібні приклади були розглянуті у попередньому розділі. Вони стосувалися середнього терміну функціонування телефонів, виготовлених за новою технологією; середнього об'єму соку, що розливається в стаканчики торговельним автоматом.

7.1.1. Перевірка гіпотези щодо середнього нормальної сукупності з відомою дисперсією

Нехай $\mathfrak{R}$ — генеральна сукупність з нормальним розподілом, який має невідоме середнє μ і відому дисперсію σ^2, тобто: $\{\mathfrak{R}\} \sim (\mu, \sigma^2)$. Стандартна процедура перевірки гіпотези

$$H_0 : \mu = \mu_0 \text{ проти } H_1 : \mu \neq \mu_0$$

базується на випадковій вибірці даних об'ємом N і на рівні значущості α. За тестову статистику вибирають випадкову змінну:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{N}},$$

яка має стандартний нормальний розподіл при $\mu = \mu_0$ (якщо $\mu \neq \mu_0$, то розподіл буде нормальний, але не стандартний). Правило прийняття рішень можна записати так:

$$\text{відхилити } H_0, \text{ якщо } |z| > z_{\alpha/2}.$$

Примітка. Значення z_α визначається з рівняння:

$$\frac{1}{\sqrt{2\pi}} \int_{z_\alpha}^{\infty} e^{-\left(x^2/2\right)} dx = \alpha.$$

Інакше кажучи, z_α — це значення на осі абсцис таке, що площа справа від нього (під кривою стандартного нормального розподілу) дорівнює α.

Якщо альтернативну гіпотезу змінити на $\mu > \mu_0$ або на $\mu < \mu_0$, то належним чином змінюється і правило прийняття рішення

відхилити H_0, якщо $z > z_\alpha$ і відхилити H_0, якщо $z < z_\alpha$,

відповідно.

Розглянуту процедуру перевірки гіпотез можна резюмувати наступним чином (табл. 7.1).

Таблиця 7.1

Стандартна процедура перевірки гіпотез щодо середнього генеральної сукупності з відомою дисперсією

Дана нормальна генеральна сукупність з відомою дисперсією σ^2.
$H_0 : \mu = \mu_0$; рівень значущості α.

Тестова статистика: $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{N}}$, (має стандартний нормальний розподіл, якщо $\mu = \mu_0$).

Альтернативні гіпотези H_1	Правило прийняття рішень: відхилити H_0, якщо
$\mu \neq \mu_0$	$\lvert z \rvert > z_{\alpha/2}$
$\mu > \mu_0$	$z > z_\alpha$
$\mu < \mu_0$	$z < -z_\alpha$

Приклад 7.1. Виробник корму для бройлерів запевняє керівника птахофабрики, що його новий корм дешевший, але не менш ефективний, ніж той, що використовується зараз. Керівник птахофабрики знає, що при використанні попереднього корму маса птахів віком до 3-х місяців має нормальний розподіл із середнім $\mu = 1{,}3$ кг і стандартним відхиленням $\sigma = 0{,}1$ кг.

Щоб перевірити запевнення виробника корму, взяли випадкову вибірку з 10 новонароджених курчат і годували їх новим кормом протягом 3-х місяців. Наприкінці 3-го місяця встановлено, що середня вибіркова маса 10 курчат становить: $\bar{x} = 1{,}2$ кг. Побудуйте процедуру перевірки гіпотез на рівні значущості $\alpha = 0{,}05$ з метою перевірки запевнення виробника корму за припущення, що вибіркове стандартне відхилення також становить 0,1 кг ($\sigma_в = 0{,}1$).

Послідовність розв'язання задачі

1. *Формулювання гіпотез.* Позначимо через μ середню масу птахів після 3-х місяців годування новим кормом. Оскільки для 10 птахів вибіркове середнє $\bar{x} = 1{,}2$ кг, то виникає сумнів, що середнє генеральної сукупності буде μ = 1,3 кг. Для перевірки запевнення виробника корму можна сформулювати наступні гіпотези

$$H_0 : \mu = 1{,}3 \text{ проти } H_1 : \mu < 1{,}3 .$$

2. *Вибір статистики і формулювання правила прийняття рішення.* Ми тестуємо середнє нормально розподіленої генеральної сукупності, яка має дисперсію $\sigma^2 = (0{,}1)^2 = 0{,}01$. Перевірочна статистика

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{N}} = \frac{\bar{x} - 1{,}3}{0{,}1 / \sqrt{10}} .$$

Правило прийняття рішення

$$\text{відхилити } H_0 \text{, якщо } z < -z_\alpha .$$

3. *Застосування ППР до даних.* Дані для цього прикладу

$$\bar{x} = 1{,}2;\ \mu_0 = 1{,}3;\ \sigma = 0{,}1;\ N = 10;\ \alpha = 0{,}05;\ -z_\alpha = -1{,}645.$$

Спостережуване значення тестової статистики z

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{N}} = \frac{1{,}2 - 1{,}3}{0{,}1 / \sqrt{10}} = -3{,}162.$$

Застосування правила прийняття рішень: оскільки

$$z = -3{,}162 < -1{,}645 = z_\alpha ,$$

то H_0 необхідно *відхилити*. Таким чином, на рівні значущості α = 0,05 керівник птахофабрики повинен відхилити запевнення виробника корму в тому, що новий корм є таким же ефективним, як і попередній, тобто він є менш ефективним.

Приклад 7.2. Розглянемо вихідні дані, наведені у попередньому прикладі. На основі отриманого результату не можна стверджувати, що стандартне відхилення σ маси бройлерів, відгодованих новим кормом, залишається таким самим, як і σ маси бройлерів, відгодованих старим кормом. Чи можна розв'язати цю задачу без зробленого припущення?

Розв'язок

Хоча на птахофабриці не знають стандартне відхилення для генеральної сукупності бройлерів, відгодованих новим кормом, але їм відомо, що $\sigma = 0{,}1$ у випадку, якщо запевнення виробника корму відповідають дійсності. Тому, якщо нуль-гіпотезу попереднього прикладу розширити таким чином, щоб врахувати запевнення виробника корму щодо σ, то H_0 приймає таку форму:

$$H_0 : \mu = 1{,}3 \; i \; \sigma = 0{,}1 \text{ проти } H_1 : \mu < 1{,}3.$$

У такому випадку, якщо нуль-гіпотеза вірна, то перевірочна статистика

$$z = \frac{\bar{x} - 1{,}3}{0{,}1/\sqrt{10}}$$

є стандартною нормальною випадковою змінною $\dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{N}}$, а це означає, що ми можемо скористатись у даному випадку тією самою тестовою статистикою, що і в попередньому прикладі, і розв'язок буде ідентичним попередньому прикладу.

Отже, можна сказати, що навіть у випадку, коли значення σ відоме тільки при вірній нуль-гіпотезі, можна скористатись табл. 7.1 за умови, що H_0 включає твердження відносно σ.

7.1.2. Перевірки гіпотези щодо середнього нормальної сукупності у випадку невідомої дисперсії

Нехай $\mathfrak{R}$ — генеральна сукупність з відомим середнім μ і невідомою дисперсією. Якщо сформулювати нуль-гіпотезу як $H_0 : \mu = \mu_0$, то стандартна процедура перевірки буде базуватись на використанні статистики

$$t = \frac{\bar{x} - \mu_0}{S/\sqrt{N}},$$

де випадкова змінна t буде мати t-розподіл з $N - 1$ ступенями свободи при умові, що $\mu = \mu_0$. Відомо, що при $N > 29$ t-розподіл з N ступенями вільності можна апроксимувати стандартним нормальним розподілом. Процедура перевірки гіпотези щодо середнього нормальної генеральної сукупності з невідомою дисперсією сформульована в табл. 7.2.

Приклад 7.3. Власник ресторану продає свій ресторан і запевняє потенційного покупця, що середнє число відвідувачів по суботах (без

Таблиця 7.2

Стандартна процедура перевірки гіпотез щодо середнього генеральної сукупності з невідомою дисперсією

Дана нормальна генеральна сукупність з відомою дисперсією σ^2.
$H_0 : \mu = \mu_0$; рівень значущості α.

Тестова статистика: $t = \dfrac{\bar{x} - \mu_0}{S/\sqrt{N}}$ (має t-розподіл з $N - 1$ ступенями вільності, якщо $\mu = \mu_0$).

Альтернативні гіпотези H_1	Правило прийняття рішень: відхилити H_0, якщо
$\mu \neq \mu_0$	$\lvert t \rvert > t_{\alpha/2,\, N-1}$
$\mu > \mu_0$	$t > t_{\alpha,\, N-1}$
$\mu < \mu_0$	$t < -t_{\alpha,\, N-1}$

врахування святкових днів) становить 100 ($\mu = 100$). Щоб перевірити запевнення власника ресторану, потенційний покупець підрахував число відвідувачів по 9-ти випадково вибраних суботах. У результаті він встановив, що вибіркове середнє $\bar{x} = 95$, а вибіркове стандартне відхилення складає $S = 10$. Необхідно встановити:

а) чи повинен відхилити потенційний покупець запевнення власника ресторану на рівні значущості $\alpha = 0{,}05$? (За припущення, що розподіл кількості відвідувачів є нормальним.)

б) яка буде відповідь при $S = 4$?

в) якщо на запитання а) і б) ви дали різні відповіді, то поясніть різницю.

Розв'язок

а)

1. *Формулювання гіпотез.* Оскільки $\bar{x} = 95$, то потенційний покупець має підозру, що фактичне середнє число відвідувачів $\mu < 100$ (для генеральної сукупності). Тому логічно сформулювати такі гіпотези:

$$H_0 : \mu_0 = 100 \text{ проти } H_1 : \mu < 100.$$

2. *Вибір перевірочної статистики і формулювання ППР.* У даному прикладі виконується перевірка середнього нормальної сукупності елементів з невідомою дисперсією, а тому необхідно скористатись табл. 7.2. Тестова статистика

$$t = \frac{\bar{x} - \mu_0}{S/\sqrt{N}},$$

і правило прийняття рішення

$$\text{відхилити } H_0, \text{ якщо } t < t_{\alpha, N-1}.$$

3. *Застосування ППР до даних.* У цьому випадку дані мають вигляд:

$$\bar{x} = 1{,}2;\ \mu_0 = 100;\ S = 10;\ N = 9;\ \alpha = 0{,}05;$$
$$-t_{\alpha, N-1} = -t_{0{,}05;\,8} = -1{,}860.$$

Спостережуване значення перевірочної статистики

$$t = \frac{\bar{x} - \mu_0}{S/\sqrt{N}} = \frac{95 - 100}{10/\sqrt{9}} = -1{,}5.$$

Оскільки $t = -1{,}5$, а $-t_{\alpha, N-1} = -1{,}86$, тобто $t > -t_{\alpha, N-1}$, то H_0 не може бути відхилена. Таким чином, потенційний покупець не може відхилити твердження власника ресторану на рівні значущості $\alpha = 0{,}05$ відносно того, що середнє число відвідувачів по суботах становить 100.

б) Якщо $S = 4$, то тестова статистика приймає значення:

$$t = \frac{\bar{x} - \mu_0}{S/\sqrt{N}} = \frac{95 - 100}{4/\sqrt{9}} = -3{,}75.$$

Оскільки $t = -3{,}75 < -1{,}86 = -t_{\alpha, N-1}$, то H_0 необхідно відхилити на рівні значимості $\alpha = 0{,}05$.

в) Вибіркові стандартні відхилення в пунктах а) і б) становили $S = 10$ і $S = 4$, відповідно. Це свідчить про вищу ступінь розсіювання числа відвідувачів у випадку а), ніж у випадку б). Тому, припускаючи, що $\mu = 100$, спостережуване значення $\bar{x} = 95$ ймовірніше буде отримане у випадку а).

7.1.3. Перевірка гіпотези щодо середнього генеральної сукупності на основі вибірок великого обсягу

Спочатку розглянемо випадок, коли дисперсія σ^2 відома. Нехай $\mathfrak{R}$ — генеральна сукупність з невідомим розподілом (тобто не обов'язково нормальним). Позначимо через μ і σ^2 середнє і дисперсію генеральної сукупності. Припустимо, що необхідно перевірити нульгіпотезу $H_0 : \mu = \mu_0$ проти деякої альтернативної. Якщо відоме σ^2,

і якщо обсяг вибірки N великий, то можна скористатись тестовою статистикою

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{N}}.$$

Зазначимо, що при $\mu = \mu_0$ і при великому N випадкова змінна z наближено має стандартний нормальний розподіл. Процедура перевірки гіпотез для даного випадку (велика вибірка з відомою дисперсією) наведена у табл. 7.3.

Таблиця 7.3

Процедура перевірки гіпотез щодо середнього генеральної сукупності з відомою дисперсією на основі великої вибірки

Дана будь-яка генеральна сукупність; відома дисперсія σ^2;
$H_0 : \mu = \mu_0$; велика вибірка даних N; рівень значущості α.

Тестова статистика: $z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{N}}$ (наближено має стандартний нормальний розподіл, якщо $\mu = \mu_0$).

Альтернативні гіпотези H_1	Правило прийняття рішень: відхилити H_0, якщо
$\mu \neq \mu_0$	$\lvert z \rvert > z_{\alpha/2}$
$\mu > \mu_0$	$z > z_\alpha$
$\mu < \mu_0$	$z < -z_\alpha$

Якщо дисперсія σ^2 генеральної сукупності невідома, то необхідно скористатись тестовою статистикою:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{N}}$$

при великому значенні N. Якщо N велике, то z має наближено стандартний нормальний розподіл. Процедура перевірки гіпотез для даного випадку (велика вибірка з невідомою дисперсією) наведена у табл. 7.4.

Приклад 7.4. В обласний відділ народної освіти надійшла інформація, що діти перших класів проводять біля телевізора 15 год на тиждень. Однак з іншого джерела надійшла інформація, що в деяких районах це число годин є меншим. Для перевірки цієї гіпотези були

Таблиця 7.4

Процедура перевірки гіпотез щодо середнього генеральної сукупності з невідомою дисперсією на основі великої вибірки

Дана будь-яка генеральна сукупність; дисперсія σ^2 невідома; $H_0 : \mu = \mu_0$; велика вибірка даних N; рівень значущості α.

Тестова статистика: $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{N}}$ (наближено має стандартний нормальний розподіл, якщо $\mu = \mu_0$).

Альтернативні гіпотези H_1	Правило прийняття рішень: відхилити H_0, якщо
$\mu \neq \mu_0$	$\lvert z \rvert > z_{\alpha/2}$
$\mu > \mu_0$	$z > z_\alpha$
$\mu < \mu_0$	$z < -z_\alpha$

випадково вибрані прізвища 49 дітей перших класів і їх батьки надали інформацію щодо того, скільки часу проводять їхні діти біля телевізорів.

а) Нехай вибіркове середнє значення часу, який діти проводять біля телевізорів, становить $\bar{x}$ = 10 год, а вибіркове стандартне відхилення — S = 5 год. Чи повинен відділ освіти відхилити $H_0 : \mu = 15$ на рівні значущості $\alpha = 0{,}05$ і прийняти альтернативну: $H_1 : \mu < 15$?

б) Чи можна на основі вказаних вибіркових даних (на рівні значущості $\alpha = 0{,}05$) стверджувати, що діти перших класів (для досліджуваного району) в середньому проводять біля телевізора менше 11 год на тиждень?

Розв'язок

1. *Формулювання гіпотез*

$$H_0 : \mu = 15 \text{ проти } H_1 : \mu < 15.$$

2. *Вибір тестової статистики і формулювання правила прийняття рішення.* Хоча дисперсія σ^2 невідома, але вибірка досить велика і тому можна скористатись процедурою перевірки, наведеною у табл. 7.4. Тестова статистика:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{N}}.$$

Правило прийняття рішення:

відхилити H_0, якщо $z < -z_\alpha$.

3. *Застосування правила прийняття рішення до даних*. Дані:

$$\bar{x} = 10;\ \mu_0 = 15;\ S = 5;\ N = 49;\ \alpha = 0{,}05;\ -z_\alpha = -1{,}645.$$

Спостережуване значення тестової статистики:

$$z = \frac{\bar{x} - \mu_0}{S/\sqrt{N}} = \frac{10 - 15}{5/\sqrt{49}} = -7.$$

Оскільки

$$z = -7 < -1{,}645 = -z_\alpha,$$

то згідно з правилом прийняття рішення H_0 необхідно відхилити, тобто приймається альтернативна гіпотеза: $\mu < 15$.

Гіпотези:

$$H_0: \mu = 11 \text{ проти } H_1: \mu < 11.$$

Правило прийняття рішення:

відхилити H_0, якщо $z < -z_\alpha$.

У даному випадку:

$$z = \frac{\bar{x} - \mu_0}{S/\sqrt{N}} = \frac{10 - 11}{5/\sqrt{49}} = -1{,}4 \quad \text{і} \quad -z_\alpha = -z_{0{,}5} = -1{,}645.$$

Оскільки $z = -1{,}4 > -1{,}645 = -z_\alpha$, то H_0 не може бути відхилена. Тобто не можна відхилити нуль-гіпотезу, що діти перших класів проводять (у вибраному районі) в середньому біля телевізорів 11 год на тиждень.

Приклад 7.5. Нехай $\Re$ — генеральна сукупність, яка має нормальний розподіл з невідомим середнім μ. Перевірте наступні гіпотези для значень α, обсягу вибірки N, вибіркового середнього $\bar{x}$ та вибіркового стандартного відхилення S, наведених нижче.

а) $\alpha = 0{,}05;\ N = 9;\ \bar{x} = 22{,}2;\ S = 3{,}3;$

$$H_0: \mu = 20 \text{ проти } H_1: \mu \neq 20.$$

б) $\alpha = 0{,}05;\ N = 9;\ \bar{x} = 22{,}2;\ S = 3{,}3;$

$$H_0: \mu = 20 \text{ проти } H_1: \mu > 20.$$

в) $\alpha = 0{,}01;\ N = 16;\ \bar{x} = 3{,}9;\ S = 1{,}6;$

$$H_0: \mu = 5 \text{ проти } H_1: \mu < 5.$$

г) $\alpha = 0{,}01$; $N = 16$; $\bar{x} = 3{,}9$; $S = 1{,}6$;

$$H_0: \mu = 5 \text{ проти } H_1: \mu \neq 5.$$

Розв'язок

а) Тестова статистика:

$$t = \frac{\bar{x} - \mu}{S/\sqrt{N}} = \frac{22{,}2 - 20}{3{,}3/\sqrt{9}} = \frac{2{,}2}{1{,}1} = 2{,}0.$$

Оскільки $\alpha = 0{,}05$, то $\alpha/2 = 0{,}025$; звідси $t_{0{,}025;\ 8} = 2{,}306$ при $N = 9$. Таким чином, $t = 2{,}0 < 2{,}306 = z_{\alpha/2}$, тобто H_0 не можна відхилити.

б) У даному випадку ми тестуємо гіпотезу:

$$H_0: \mu = 20 \text{ проти } H_1: \mu > 20.$$

На рівні значущості $\alpha = 0{,}05$ тестова статистика: $t_{0{,}05;\ 8} = 1{,}86$. Оскільки $z > z_{\alpha}$, то H_0 відхиляється і приймається альтернативна гіпотеза $H_1: \mu > 20$.

в) Тестова статистика:

$$t = \frac{\bar{x} - \mu}{\sigma/\sqrt{N}} = \frac{3{,}9 - 5}{1{,}6/\sqrt{16}} = -\frac{1{,}1}{0{,}4} = -2{,}75,$$

а критичне значення $t_{0{,}01;\ 15} = -2{,}602$. Оскільки $t = -2{,}75 < -2{,}602 = t_{\alpha,\ N-1}$, то H_0 необхідно відхилити і прийняти альтернативну.

г) У даному випадку $t = -2{,}75$; $t_{0{,}005;\ 15} = 2{,}947$, то H_0 приймається.

7.2. Поняття інтервальних оцінок

У третьому розділі було розглянуто, як за допомогою випадкової вибірки даних оцінити середнє і дисперсію вихідного (генерального) розподілу. Однак навіть найкраща оцінка може відрізнятись від істинного значення статистичного параметра. Тому на практиці бажано знати, наскільки знайдена оцінка є близькою до істинного значення параметра. У зв'язку з цим введемо деякі поняття, які дають можливість виносити судження щодо точності оцінок з ймовірнісної точки зору.

Оцінюючи статистичний параметр (наприклад, середнє розподілу), не можна говорити про ймовірність того, що оцінка є точною. Наприклад, нехай на основі деякої вибірки стверджують, що середній

зріст чоловіків в Україні становить 175 см. Очевидно, що таке твердження необхідно перевірити за допомогою даних щодо зросту чоловіків по всій країні. Тобто без такої перевірки ми не можемо говорити про ймовірність того, що оцінка є точною.

На практиці замість однієї (точкової) оцінки розглядають деякий інтервал і визначають ймовірність того, що значення невідомого параметра буде лежати між граничними точками цього інтервалу. Такі оцінки називають *інтервальними*. Як правило, спочатку вибирають бажану ймовірність попадання в інтервал (наприклад, 0,95 або 0,99), а потім знаходять два значення (точки), між якими буде знаходитись значення невідомого параметра із вибраною ймовірністю. Твердженням подібного типу може бути наступне: середній зріст дорослого чоловіка в США знаходиться в інтервалі 173–175 см з імовірністю 0,95.

Деякі статистики кажуть, що застосування слова "ймовірність" є в даному випадку некоректним, оскільки параметр — конкретне число і говорити про те, що воно міститься у деякому інтервалі немає смислу. Вони вважають, що поняття "ймовірність" можна застосовувати тільки до випадкових величин, а тому часто дають визначення ймовірності, яке відрізняється за смислом від того, що використовується тут. Нижче наведемо визначення ймовірності, яке є коректним не в усіх ситуаціях, але є цілком прийнятним для подальшого викладення.

Раніше ймовірність випадкової події визначалась як частота її появи, коли дослід повторюється багато разів. Інакше кажучи, ймовірність падіння монети визначеною стороною вверх є пропорціональною числу появи цієї сторони зверху з ростом числа підкидань.

Таким чином, що ж означає твердження, що зріст дорослих чоловіків коливається у межах 173–175 см з ймовірністю 0,95. Зокрема, що в даному випадку є тим результатом, який нас цікавить і які події можуть повторюватись велике число разів? Зазначимо, що ми маємо справу з однією випадковою вибіркою та інтервалом. Ми прогнозуємо, що невідомий параметр знаходиться у даному інтервалі.

Можна навести ще одну аналогію тієї ситуації, коли ймовірність відноситься до інтервальної оцінки. Наприклад, на борту військового літака є лише одна бомба, якою необхідно знищити ворожий об'єкт, місцезнаходження якого закрите хмарами. Припустимо, що ударна хвиля діє на відстані 10 км. Отже, льотчик знає, якщо бомба упаде у радіусі 10 км від місцезнаходження об'єкта, то він буде знищений.

Інакше мета не буде досягнута. У цьому випадку інтервальній оцінці параметра відповідає діаметр дії ударної хвилі вибуху, а місцеположення об'єкта є невідомий статистичний параметр. Вибух бомби, який охоплює об'єкт своєю ударною хвилею, подібний інтервальній оцінці, яка включає істинне значення параметра між своїми граничними точками.

Для визначення ймовірності, з якою бомба досягне цілі, потрібно повторити бомбування багато разів. Припустимо, що в результаті здійснення 100 вильотів льотчик зміг би 95 разів знищити ворожий об'єкт. У такому випадку буде справедливим твердження, що пілот досягне мети за допомогою лише однієї бомби з ймовірністю 0,95. Можна також сказати, що інтервальна оцінка включає невідомий параметр з ймовірністю 0,95.

7.3. Інтервальна оцінка середнього значення розподілу

Розглянемо процедуру визначення інтервальної оцінки середнього значення розподілу. Ця процедура повинна забезпечити включення в інтервал найкращої оцінки середнього значення розподілу. Водночас інтервал навколо цієї найкращої оцінки будується таким чином, щоб істинне середнє значення розподілу знаходилось у цьому інтервалі з визначеною ймовірністю. Наприклад, замість того, щоб казати, що середнє значення КРР даного розподілу дорівнює 115, ми можемо стверджувати, що середнє розподілу повинне бути в інтервалі між 111 і 119 з ймовірністю 0,95.

Звичайно, завжди бажано мати інтервальну оцінку по можливості вужчою. Ширина інтервальної оцінки — це одна з її характеристик. Наприклад, порівняємо дві наступні інтервальні оцінки:

1 — середня маса людей, які страждають хворобою L, знаходиться в інтервалі 50–80 кг з ймовірністю 0,95;

2 — середня маса людей, які страждають хворобою L, знаходиться в інтервалі 60–70 кг з ймовірністю 0,95.

У кожному припущенні приймає участь однакове ймовірнісне відношення. Однак у другому припущенні інтервал є вужчим, що робить інтервальну оцінку більш визначеною.

Також, чим більшою є ймовірність, тим точнішою буде інтервальна оцінка. Порівняємо два таких твердження:

1 — середня тривалість життя собак породи фокстер'єр становить 10–14 років з ймовірністю 0,9;

2 — середня тривалість життя собак породи фокстер'єр становить 10–14 років з ймовірністю 0,99.

Можна сказати, що друге припущення є більш чітким в ймовірнісному сенсі, ніж перше. Тобто другому твердженню можна довіряти більше, ніж першому.

Повернемось до процедури інтервального оцінювання середнього значення розподілу. Припустимо, що необхідно визначити інтервальну оцінку середнього коефіцієнтів розумового розвитку студентів з вибраної спеціальності. Необхідно визначити границі інтервалу, в який з ймовірністю 0,95 входить середнє значення КРР цих студентів. Нехай відоме стандартне відхилення КРР становить 16 одиниць. Припустимо, що середнє значення розподілу оцінено на основі вибірки, яка є в нас, і дорівнює 118.

У цьому випадку ми виходимо з того, що є багато дослідників, кожний з яких має вибірку з 64 елементів, на основі якої він оцінює середнє значення розподілу і знаходить його інтервальну оцінку. Розглянемо процедуру, яка у 95 випадках із 100 дасть для всіх дослідників інтервальні оцінки, які включають невідоме середнє значення розподілу.

Для розв'язання цієї задачі спочатку необхідно зробити деякі узагальнення. Середнє значення КРР невідоме. Стандартне відхилення розподілу відоме і становить 16 одиниць. Усі дослідники знаходять вибіркове середнє арифметичне на основі вибірок, які є у них. Відповідно до теореми про розподіл вибіркових оцінок середнього, вони мають нормальний розподіл. Їх середнє значення невідоме, але стандартне відхилення відоме: σ_M, тобто

$$\sigma_M = \frac{\sigma}{\sqrt{N}} = \frac{16}{\sqrt{64}} = 2,$$

де σ_M — стандартне відхилення для середніх або стандартна похибка оцінок вибіркових середніх.

У розподілі вибіркових середніх кожне значення є оцінкою середнього для 64 значень КРР, тобто кожне вибіркове середнє — це найкраща оцінка середнього для КРР даної групи студентів. Іноді найкращу оцінку називають *точковою оцінкою* для того, щоб відрізнити її від більш невизначеної інтервальної.

Оскільки вибіркові середні мають нормальний розподіл, а стандартне відхилення дорівнює 2, то за допомогою таблиці відносних площ під кривою нормального розподілу (див. Додаток) знайдемо,

що 68 із 100 вибіркових середніх відхиляються від істинного значення менше, ніж на 2 одиниці.

З цієї ж таблиці знайдемо, що для нормального розподілу 95 із 100 його вибіркових середніх відхиляються від середнього менше, ніж на 1,96 стандартного відхилення. Таким чином, 95 із 100 вибіркових середніх знаходяться в інтервалі $(-1{,}96\ \sigma;\ 1{,}96\ \sigma)$. Точніше можна сказати так:

Точкові оцінки, знайдені різними дослідниками, відхиляються від невідомого середнього значення розподілу в обидва боки не більше, ніж на 3,92 стандартного відхилення з ймовірністю 0,95

Подивимось тепер на результати з точки зору кожного дослідника окремо. Він може сказати: "Більшість моїх колег отримали вибіркове середнє, яке відхиляється від невідомого середнього не більше, ніж на $3{,}92\ \sigma$. Я можу бути одним із 95, які отримали такі оцінки, або одним із 5, чиї вибіркові середні відхиляються від невідомого середнього більше, ніж на $3{,}92\ \sigma$. Оскільки у 95 випадків із 100 дослідник попадає в першу категорію, то з ймовірністю 0,95 я також попаду в цю категорію. Інакше кажучи, моє вибіркове середнє буде відрізнятись від невідомого середнього значення розподілу не більше, ніж на $3{,}92\ \sigma$ з ймовірністю 0,95".

Припустимо далі, що за результатами всіх дослідників було знайдене середнє, яке склало 118 одиниць. Тепер можна сказати, що з ймовірністю 0,95 вибіркове середнє буде відрізнятись від середнього значення розподілу не більше, ніж на $3{,}92\ \sigma$. Усі вибіркові середні будуть знаходитися в області $118 \pm 3{,}92\ \sigma$ з ймовірністю 0,95. Таким чином, ми побудували інтервальну оцінку, яка містить всі числові значення в області $\pm\ 3{,}92\ \sigma$ від отриманої вибіркової оцінки середнього. Ця інтервальна оцінка називається *95 %-м довірчим інтервалом* для середнього.

Зазначимо, що процедуру оцінювання можна повторювати кожний день і отримувати при цьому нові вибіркові оцінки. Таким чином, центр інтервальної оцінки кожний день буде новим. Можна сказати лише те, що кожні 95 днів із 100 побудований нами інтервал буде містити невідоме середнє значення розподілу коефіцієнтів розумового розвитку для студентів з вибраної спеціальності.

Можна побудувати також таку інтервальну оцінку, яка буде містити невідоме середнє з ймовірністю 0,99. У такому випадку 99 із 100

елементів нормального розподілу відрізняються від його середнього не більше, ніж на 2,58 стандартного відхилення. Таким чином, знайдене вибіркове середнє буде знаходитись в інтервалі ± 2,58 стандартного відхилення від істинного значення середнього з ймовірністю 0,99. А це означає, що з імовірністю 0,99 вибіркове середнє, яке дорівнює 118, не більше, ніж на 5,16 відрізняється від невідомого генерального середнього значення розподілу. Насамкінець, можна стверджувати, що невідоме генеральне середнє значення КРР студентів вибраної спеціальності знаходиться в інтервалі між 112,84 і 123,16 з імовірністю 0,99. Цей інтервал називають *99 %-м довірчим інтервалом*.

Повторимо тепер всі етапи процесу побудови інтервальної оцінки середнього. Для цього необхідно виконати наступне:

1) знайти стандартне відхилення від середнього;

2) якщо вибрана ймовірність попадання оцінки в інтервал становить 0,95, то значення множника t для інтервальної оцінки становить 1,96 і вираз для 95 %-го довірчого інтервалу має вигляд:

$$\bar{x} - 1{,}96\frac{\sigma}{\sqrt{N}} < \mu < \bar{x} + 1{,}96\frac{\sigma}{\sqrt{N}},$$

де $\bar{x}$ — вибіркове середнє; σ — стандартне відхилення від середнього значення розподілу (для всієї вибірки); μ — невідоме генеральне середнє розподілу; N — кількість елементів у вибірці; $\sigma/\sqrt{N}$ — стандартне відхилення середнього значення. Для розглянутої вище задачі $\bar{x} = 118$, $\sigma = 16$, $N = 64$;

3) якщо вибрана ймовірність попадання оцінки в інтервал становить 0,99, то значення множника t для інтервальної оцінки становить 2,58 і вираз для 99 %-го довірчого інтервалу має вигляд:

$$\bar{x} - 2{,}58\frac{\sigma}{\sqrt{N}} < \mu < \bar{x} + 2{,}58\frac{\sigma}{\sqrt{N}}.$$

Отже, для розглянутої задачі дослідження КРР для студентів вибраної спеціальності можна знайти такі інтервальні оцінки:

• 95 %-й довірчий інтервал:

$$118 - 1{,}96\frac{16}{\sqrt{64}} < \mu < 118 + 1{,}96\frac{16}{\sqrt{64}},$$

$$114{,}08 < \mu < 121{,}92;$$

• 99 %-й довірчий інтервал:

$$118 - 2{,}58\frac{16}{\sqrt{64}} < \mu < 118 + 2{,}58\frac{16}{\sqrt{64}},$$

$$112{,}84 < \mu < 123{,}16.$$

Очевидно, що 95 %-й довірчий інтервал буде завжди меншим 99 %-го довірчого інтервалу.

7.4. Контрольні питання і вправи

1. Поясніть стандартну процедуру перевірки гіпотези стосовно середнього генеральної сукупності з відомою дисперсією.
2. Виробник мінеральних добрив стверджує, що його нове мінеральне добриво коштує менше, але не менш ефективне, ніж те, що вже використовується. Керівник агрофірми знає, що при використанні попереднього добрива урожайність з одного гектара має нормальний розподіл із середнім $\mu = 35$ ц і стандартним відхиленням $\sigma = 1{,}0$ ц. Щоб перевірити твердження виробника добрив, взяли випадкову вибірку із 20 га і застосували на ній нове добриво. Після збирання врожаю було встановлено, що середня врожайність на 20 га склала $\bar{x} = 34{,}5$ ц. Побудуйте процедуру перевірки гіпотез на рівні значущості $\alpha = 0{,}05$ з метою перевірки твердження виробника міндобрив за припущення, що вибіркове стандартне відхилення також складає 1,0 ц ($\sigma_в = 1{,}0$).
3. Опишіть стандартну процедуру перевірки гіпотези стосовно середнього генеральної сукупності з невідомою дисперсією.
4. Поясність процедуру перевірки гіпотези стосовно середнього генеральної сукупності з відомою дисперсією на основі великої вибірки.
5. Опишіть процедуру перевірки гіпотези стосовно середнього генеральної сукупності з невідомою дисперсією на основі великої вибірки.
6. Зібрані статистичні дані свідчать, що діти старших класів проводять біля комп'ютера по 20 год на тиждень. Однак з іншого джерела надійшла інформація, що в деяких районах це число годин є меншим. Для перевірки цієї гіпотези були випадково вибрані прізвища 60 дітей старших класів і їх батьки надали інформацію щодо того, скільки часу проводять їхні діти біля комп'ютерів.

а) Нехай вибіркове середнє значення часу, який діти проводять біля комп'ютерів, складає $\bar{x} = 16$ год, а вибіркове стандартне відхилення — $S = 6$ год. Чи можна відхилити $H_0 : \mu = 20$ на рівні значущості $\alpha = 0{,}05$ і прийняти альтернативну: $H_1 : \mu < 20$?

б) Чи можна на основі вказаних вибіркових даних (на рівні значущості $\alpha = 0{,}05$) стверджувати, що діти старших класів (для досліджуваного району) у середньому проводять біля телевізора менше 18 год на тиждень?

7. Які оцінки називають точковими, а які інтервальними? Наведіть числові приклади.
8. Опишіть процедуру і наведіть приклад знаходження інтервальної оцінки середнього значення вибірки даних.
9. Яким чином формуються довірчі інтервали? Які довірчі інтервали часто використовуються на практиці?

Розділ 8

ПОБУДОВА РЕГРЕСІЙНИХ МОДЕЛЕЙ

Розглянемо методику побудови математичних моделей часових рядів, яка більше орієнтована на побудову моделей фінансово-економічних, соціальних та екологічних процесів, для яких, як правило, набагато складніше поставити експеримент (або неможливо взагалі) та отримати інформативні експериментальні дані у достатньому обсязі. У наведеному вигляді ця методика також може бути успішно застосована до побудови моделей динаміки технічних систем і технологічних процесів при належному плануванні та реалізації експерименту з метою збору даних, необхідних для побудови високоякісних моделей.

Основи методики побудови моделей і аналізу часових рядів запропоновані Боксом і Дженкінсом [11]. Модифікована авторами методика побудови математичної моделі процесу з використанням даних у вигляді часового ряду і часового перерізу складається з таких кроків:

– Виконання аналізу процесу, для якого будується модель, на основі спеціальних літературних джерел, експертних оцінок перебігу процесу, візуального дослідження графіків вимірів вхідних і вихідних змінних, представлених часовими рядами або часовими перерізами, та іншої доступної інформації.

– Належна попередня обробка експериментальних даних з метою їх приведення до форми, найбільш придатної для оцінювання параметрів моделі.

– Аналіз часових рядів на можливу наявність нестаціонарності і нелінійності за допомогою множини статистичних критеріїв.

– Вибір структури моделей-кандидатів, для чого необхідно виконати такі дії: 1) знайти та виконати аналіз кореляційної матриці для часових рядів залежної та незалежних змінних з метою визначення тих екзогенних змінних, які необхідно включити в модель; 2) знайти автокореляційну (АКФ) та часткову автокореляційну функції (ЧАКФ) залежної змінної з метою визначення оцінки порядку авторегресійної частини моделі та ковзного середнього; 3) оцінити характеристики інших елементів структури математичної моделі, що буде розглянуто нижче.

– Вибір методу оцінювання параметрів математичних моделей вибраних структур і знаходження оцінки векторів їх параметрів. Найчастіше це метод найменших квадратів (МНК), метод максимальної правдоподібності (ММП) та їх модифікації (рекурсивні та нелінійні). В окремих випадках застосовують метод Монте-Карло для марковських ланцюгів (МКМЛ), який придатний для оцінювання нелінійних моделей.

– Вибір кращої з оцінених моделей-кандидатів за допомогою множини статистичних критеріїв адекватності (якості) моделі. Застосувати модель до розв'язання основної задачі — прогнозування, синтезу системи керування або поглибленого дослідження процесу і остаточно встановити її придатність.

Тепер розглянемо детальніше етапи побудови моделі.

8.1. Аналіз процесу

Аналіз процесу — це надзвичайно важливий етап, коректне виконання якого потребує досвіду дослідження реальних процесів різної природи. Ігнорування цього етапу приведе до неможливості побудови моделі високого ступеня адекватності процесу та її придатності для розв'язання задач, згаданих вище. Аналіз процесу спрямований на вирішення таких завдань:

- визначення кількості входів і виходів, тобто визначення вимірності процесу; як правило, вимірність визначається кількістю виходів процесу, кожний з яких описують окремим рівнянням;
- встановлення логічних зв'язків між змінними та аналіз можливостей їх спільного математичного опису (коректного об'єднання в одному математичному виразі); для цього необхідно використати всю наявну інформацію про процес із спеціальної літератури, наукових звітів та від експертів;
- визначення кількості зовнішніх збурень та їх типу (детермінованих чи стохастичних) та попереднє встановлення можливості їх статистичного опису за допомогою конкретних типів розподілів випадкових величин;
- встановлення можливості декомпозиції процесу на окремі підпроцеси, які є простішими як з точки зору їх функціонування, так і з точки зору математичного опису. Декомпозиція — це досить складний процес, який базується на спеціальних математичних методах;

- якщо процес має ієрархічну структуру (верхній та нижній рівень функціонування або більшу кількість рівнів), то необхідно чітко розмежувати ці рівні, визначити функції кожного з них і встановити, які типи зв'язків існують між ними. Наприклад, технологічні процеси часто можна розмежувати на два і більше рівнів, які пов'язані між собою інформаційними потоками або за логічними ознаками і т. ін.;
- використання знань зі спеціальної літератури стосовно особливостей функціонування процесу, відомих законів та закономірностей його перебігу, виявлення існуючих моделей процесу та досвіду його теоретичного чи експериментального дослідження;
- при наявності розроблених моделей досліджуваного процесу необхідно встановити їх недоліки та переваги, а також визначити можливість подальшого використання (модифікації); аналіз і використання існуючих моделей надає можливість дослідникам суттєво скоротити час та інші витрати на побудову і використання моделі.

Отриману інформацію максимально використовують для попереднього оцінювання структури моделі або кількох моделей-кандидатів, параметри яких оцінюють за допомогою експериментальних даних. У процесі виконання аналізу функціонування досліджуваного процесу доцільно використовувати та порівнювати інформацію з різних джерел. Це особливо стосується фінансово-економічних процесів, щодо яких може надходити інформація із суттєвими протиріччями, пропусками і похибками.

8.2. Попередня обробка даних

Процес попередньої обробки статистичних даних складається, зазвичай, з таких операцій:

- нормування та візуальна перевірка даних і, за необхідністю, їх коригування або приведення до зручного діапазону їх зміни, наприклад, від 0 до 1; від –1 до +1; від +10 до –10 і т. ін. Причому коригування даних полягає у заповненні пропусків та зменшенні рівнів викидів (екстремальних імпульсних значень), що виходять за основний діапазон значень змінних;
- бутстреп-аналіз з метою збільшення обсягів вибірок (розмноження вибірки);

- заміна некоректних вимірів інтерпольованими або усередненими значеннями;
- формування перших різниць та різниць вищих порядків, які необхідні для аналізу відповідних складових процесу, представленого часовим рядом;
- ортогональні перетворення і цифрова фільтрація даних з метою вилучення похибок вимірів, тобто шумових складових.

Як відомо, перша різниця є наближений дискретний аналог першої похідної, а друга різниця — другої похідної. Використання різниць дає можливість будувати моделі для швидкості та прискорення основної змінної. Часто із значень ряду віднімають його середнє для того, щоб отримати можливість працювати з відхиленнями, а не з повними значеннями змінних. Такий підхід застосовують, наприклад, при побудові моделей у просторі станів з їх подальшим використанням для оптимальної фільтрації або оптимального керування процесом.

Досить добрі результати нормування при оцінюванні параметрів множинної регресії

$$y(k) = \beta_0 + \beta_1 x_1(k) + \beta_2 x_2(k) + ... + \beta_p x_p(k) + \varepsilon(k) \qquad (8.2.1)$$

можна досягти завдяки одночасному нормуванню і центруванню даних наступним чином:

$$x_{i\text{Н}}(k) = \frac{x_i(k) - \bar{x}_i}{\sqrt{S_{x_i}}} = \frac{x_i(k) - \bar{x}_i}{\left(\frac{1}{N-1} \sum_{k=1}^{N} \left(x_i(k) - \bar{x}_i \right)^2 \right)^{1/2}}, \quad k = 1,\ldots,N; \quad i = 1,\ldots,p;$$

$$y_{\text{Н}}(k) = \frac{y(k) - \bar{y}}{\sqrt{S_y}} = \frac{y(k) - \bar{y}}{\left(\frac{1}{N-1} \sum_{k=1}^{N} \left(y(k) - \bar{y} \right)^2 \right)^{1/2}}, \qquad (8.2.2)$$

де $x_i(k)$ — елементи i-го стовпчика матриці значень незалежних змінних; $y(k)$ — значення залежної змінної; $x_{i\text{Н}}(k)$, $y_{i\text{Н}}(k)$ — нормовані значення змінних; $\bar{x}_i$, $\bar{y}$ — вибіркові середні значення незалежних змінних і залежної змінної, відповідно; N — кількість значень; p — кількість незалежних змінних (регресорів) x_i. Якщо ввести позначення для центрованих змінних

$$\tilde{x}_i(k) = x_i(k) - \bar{x}_i; \quad \tilde{y}(k) = y(k) - \bar{y}, \qquad (8.2.3)$$

то множинна регресія для центрованих змінних матиме вигляд

$$\tilde{y}(k) = \beta_1 \tilde{x}_1(k) + \beta_2 \tilde{x}_2(k) + \ldots + \beta_p \tilde{x}_p(k) + \varepsilon(k). \qquad (8.2.4)$$

Якщо підставити (8.2.2) у (8.2.4), то рівняння множинної регресії набуде вигляду

$$y_{\text{Н}}(k) S_y^{1/2} = \beta_1 S_1^{1/2} x_{1\text{Н}}(k) + \beta_2 S_2^{1/2} x_{2\text{Н}}(k) + \ldots + \beta_p S_p^{1/2} x_{p\text{Н}}(k) + \varepsilon'(k). \quad (8.2.5)$$

Тепер поділимо ліву і праву частини на $S_y^{1/2}$:

$$y_{\text{Н}}(k) = \alpha_1 x_{1\text{Н}}(k) + \alpha_2 x_{2\text{Н}}(k) + \ldots + \alpha_p x_{p\text{Н}}(k) + \varepsilon''(k), \qquad (8.2.6)$$

де $\alpha_1 = \beta_1 (S_1 / S_y)^{1/2}$, ..., $\alpha_p = \beta_p (S_p / S_y)^{1/2}$. Отримане рівняння (8.2.6) — це рівняння для нормованих значень.

У результаті центрування і нормування покращується ступінь зумовленості матриці значень, який вимірюється відношенням

$$\eta = \left| \frac{\lambda_{\max}}{\lambda_{\min}} \right|,$$

де $\lambda_{\max}$, $\lambda_{\min}$ — максимальне і мінімальне власні числа матриці значень. Для забезпечення належних умов оцінювання параметрів необхідно задовольнити умову: $\eta < 10$.

Застосування того чи іншого методу підготовки даних для моделювання визначається у кожному випадку по-різному.

Обробка екстремальних (аномальних) значень

Хоча виявлення та обробка екстремальних значень — це велика окрема тема для дослідження, розглянемо деякі можливості щодо розв'язання цієї проблеми. У подальшому будемо вважати дані *аномальними*, якщо вони виникли внаслідок впливу значних похибок вимірів або похибок, пов'язаних з некоректним збором статистичних даних. Якщо можна встановити факт наявності аномальних даних, то їх просто видаляють з розподілу.

Екстремальні значення — це правильно виміряні (зібрані) дані, які характеризують фактичні раптові (стрибкоподібні) зміни процесу. Підхід до розв'язання задачі дослідження екстремальних значень спостережень залежить від поставленої мети. Якщо дослідника цікавить тільки факт наявності таких значень (наприклад, з метою виявлення умов, що ведуть до появи екстремальних значень), то досить мати надійний критерій для виявлення таких спостережень.

Якщо ж ставиться завдання виявлення і виключення екстремальних значень (наприклад, з метою покращання оцінок статистичних

параметрів і моделей), то виникає задача — як правильно виконати обробку даних. Так, спираючись на критерій для визначення екстремальних значень, можна визначити величину зміщення оцінок параметрів.

Критерії аналізу екстремальних значень застосовують з метою:

- вирівняти спостереження перед аналізом (як правило, суттєво зменшити великі значення);
- переконатись, що дані містять аномальні значення, що свідчить про необхідність перегляду процедури отримання даних;
- виділити спостереження, які є цікавими з точки зору їх аномальності та, за можливістю, описати встановлений ефект математично.

Класичний підхід до виявлення аномальних спостережень полягає в тому, що вибіркові спостереження розглядають як випадкові, нормально розподілені величини. При цьому для аналізу (виявлення екстремальних значень) формується статистика (статистичний тест, який ґрунтується на статистичних даних), яка є чутливою до різких відхилень такого типу. Необхідно встановити розподіл цієї статистики при нульовій гіпотезі, що всі спостереження належать нормальній сукупності, а потім відхилити цю гіпотезу, якщо виявиться, що отримана статистика їй суперечить.

Розглянемо можливий критерій відкидання екстремальних значень. Нехай дана деяка вибірка $\{x_1, x_2, ..., x_N\}$, $N \geq 3$, яка, за припущенням, є випадковою для випадкової змінної X з нормальним розподілом: $\{X\} \sim (\mu_x, \sigma_x^2)$. Позначимо відхилення від середнього через

$$\tilde{x}_i = x_i - \bar{x}, \quad i = 1,2,\ldots,N, \quad \text{де} \quad \bar{x} = \frac{1}{N}\sum_{k=1}^{N} x(k).$$

Якщо вилучити одне значення із спостережень, то вибіркове середнє для спостережень, що залишились, визначається як

$$\sum_{\substack{k=1 \\ k \neq i}}^{N} \frac{x_k}{N-1} = \bar{x} - \frac{\tilde{x}_i}{N-1}. \tag{8.2.7}$$

Якщо вилучити кілька значень $x_1, x_2, ..., x_r$, то вибіркове середнє дорівнює

$$\bar{x} - \frac{\tilde{x}_1 + \tilde{x}_2 + \ldots + \tilde{x}_r}{N-r}. \tag{8.2.8}$$

Позначимо максимальне відхилення через $\tilde{x}_m = x_m - \bar{x}$. Тепер правило визначення екстремального значення можна сформулювати так:

при заданому значенні с спостереження x_m *відкидається, якщо* $|\tilde{x}_m| > cS_x$, де S_x — вибіркове стандартне відхилення змінної X.

Якщо вибірка має досить великий об'єм, то значення x_m видаляється і аналіз продовжується. Величина константи c може змінюватися зі зміною довжини вибірки; вона зв'язана неявно з t-статистикою [11]:

$$\sqrt{\frac{Nc^2(\nu+\nu_0-1)}{\nu\left(\nu+\nu_0-\dfrac{Nc^2}{\nu}\right)}} \approx t_{1-\alpha/2}^{\nu_0-\nu-1}, \qquad (8.2.9)$$

де $\nu = N - 1$; α — рівень значущості; ν_0 — будь-яке інше число додаткових ступенів вільності, яке зв'язане з оцінюванням σ_x^2 за вибіркою, об'єм якої не дорівнює N ($\nu_0 = 0$, якщо такої інформації немає). Також існує наближений вираз для c через розподіл F [11]:

$$c \approx \left(\frac{\nu}{N}\right)^{1/2}\left(\frac{3F_{1-q}}{1+\left[\left(3F_{1-q}-1\right)/\left(\nu+\nu_0\right)\right]}\right)^{1/2}, \qquad (8.2.10)$$

де $q = \Delta\hat{\sigma}_x^2 \dfrac{\nu}{N}$; $\Delta\hat{\sigma}_x^2$ — очікуваний приріст дисперсії внаслідок появи екстремальних значень. При використанні (8.2.10) значення c визначається додатним значенням квадратного кореня.

Виразом (8.2.10) можна скористатись наступним чином: якщо з ряду значення не видалялись, то допустимий (очікуваний) відносний приріст дисперсії ("премію") $\Delta\hat{\sigma}_x^2$ необхідно помножити на ν/N і таким чином отримаємо q. За його допомогою знайдемо відповідну верхню точку для відношення дисперсій F_{1-q} при 3-х і $\nu + \nu_0 - 1$ ступенях вільності. За виразом (8.2.10) знайдемо значення c і застосуємо критерій до x_m. Очікуваний відносний приріст дисперсії ("премія") залежить від того, наскільки ймовірною є поява екстремальних значень, наприклад, можна прийняти невеликий відносний приріст $\Delta\hat{\sigma}_x^2 = 0{,}01 \div 0{,}03$.

Наприклад, якщо $N = 4$, $\nu = 3$ і $\nu/N = 0{,}75$, то при $\Delta\hat{\sigma}_x^2 = 0{,}02$ маємо: $q = 0{,}02 \cdot 0{,}75 = 0{,}015$. При 3-х ступенях вільності $F_{1-q} = F_{1-0{,}015} = 9{,}28$ [16]. Тепер знайдемо значення c:

$$c = (0{,}75)^2\left(\frac{3F_{0{,}985}}{1+\left(3F_{0{,}985}-1\right)/3}\right)^{1/2} = 1{,}449.$$

Спостереження x_m необхідно видалити, якщо $|\tilde{x}_m| > 1{,}449\,S_x$. Можливі інші підходи до аналізу екстремальних значень.

Приклад 8.1. Знаходження критерію для виявлення екстремального значення. Є ряд значень: $X = \{23{,}2;\ 23{,}4;\ 23{,}5;\ 24{,}1;\ 25{,}5\}$. Встановити, чи можна вважати значення 25,5 екстремальним і чи необхідно його видалити з вибірки?

Розв'язок. Знаходимо: $\bar{x} = 23{,}9$; $\tilde{x}_5 = 25{,}5 - 23{,}9 = 1{,}6$; $S_x = 0{,}77$. Для $\alpha = 0{,}05$; $\nu = 4$ і $N = 5$ за формулою (8.2.9) маємо:

$$\left(\frac{15c^2}{16-5c^2}\right)^{1/2} = 2{,}776^3.$$

Звідси одержимо $c = 1{,}49$. Згідно з критерієм маємо:

$$|1{,}6| > 1{,}49 \cdot 0{,}77 = 1{,}05,$$

тобто спостереження x_5 видаляється.

Видалення похибок вимірів, тобто фільтрація шумів вимірів

При необхідності застосовують *фільтрацію даних від шумів*. Для розв'язання цієї важливої задачі використовують *цифрові* або *оптимальні* фільтри. Обидва методи фільтрації широко застосовуються у технічних системах, а також у системах обробки економічних, фінансових та інших типів даних. Розглянемо схематично ці можливості попередньої обробки даних.

Цифровий фільтр (ЦФ) можна представити, наприклад, рівнянням типу АР(p):

$$y(k) = a_1 y(k-1) + a_2 y(k-2) + \ldots + a_p y(k-p).$$

Такий фільтр має амплітудно-частотну характеристику (АЧХ), яка визначається значеннями коефіцієнтів рівняння. Мета застосування ЦФ: пропустити корисну частину і затримати шумову або просто непотрібну для аналізу складову. Сьогодні існують високорозвинені методи оптимізаційного проектування ЦФ, які дають можливість спроектувати ефективні структури фільтрів з частотними характеристиками заданої форми. Зокрема, корисний інструментарій для проектування ЦФ містить система Matlab.

Оптимальний фільтр потребує, як правило, модель процесу, представлену у просторі станів. На моделі такого типу базується фільтр Калмана.

Нехай нестаціонарна лінійна система описується у дискретному часі рівняннями з непостійними в часі коефіцієнтами (непостійність означає залежність від дискретного часу $k = 0, 1, 2, \ldots$):

$$\mathbf{x}(k) = \mathbf{A}(k, k-1)\,\mathbf{x}(k-1) + \mathbf{B}(k, k-1)\,\mathbf{u}(k-1) + \mathbf{w}(k),$$

де $\mathbf{x}(k)$ — n-вимірний вектор станів системи; $\mathbf{u}(k-1)$ — m-вимірний вектор детермінованих вхідних величин (сигнали керування); $\mathbf{w}(k)$ — n-вимірний вектор випадкових зовнішніх збурень; $\mathbf{A}(k, k-1)$ — $(n \times n)$ матриця динаміки системи (вона містить коефіцієнти, що характеризують динаміку, тобто швидкість зміни станів у часі); $\mathbf{B}(k, k-1)$ — $(n \times m)$ матриця коефіцієнтів керування. Подвійний часовий аргумент у вигляді $(k, k-1)$ означає, що величина з цим аргументом використовується в момент k, але її значення базується на попередніх даних, які відомі на момент $k-1$ включно. Далі будемо записувати для простоти матриці **A** і **B** з одним аргументом, тобто $\mathbf{A}(k)$ та $\mathbf{B}(k)$. Очевидно, що стаціонарна система описується матрицями з постійними коефіцієнтами, які записують просто **A** і **B**. Оскільки матриця **A** зв'язує поточний стан із попереднім, то її називають ще перехідною матрицею станів. Нагадаємо, що дискретний час k зв'язаний з неперервним часом t періодом дискретизації вимірів T_s: $t = kT_s$.

У класичній постановці задачі оптимальної фільтрації послідовність зовнішніх збурень $\mathbf{w}(k)$ задовольняє властивостям білого гаусового шуму з нульовим середнім значенням і коваріаційною матрицею **Q**, тобто статистики шуму мають вигляд:

$$E[\mathbf{w}(k)] = 0, \ \forall k;$$
$$E[\mathbf{w}(k)\,\mathbf{w}^T(j)] = \mathbf{Q}(k)\,\delta_{kj},$$

де δ_{kj} — дельта-функція Кронекера, що визначається так: $\delta_{kj} = \begin{cases} 0 & \text{для } k \neq j \\ 1 & \text{для } k = j \end{cases}$; $\mathbf{Q}(k)$ — додатно визначена коваріаційна матриця зовнішніх збурень стану розмірності $(n \times n)$. Діагональні елементи матриці є дисперсією компонент вектора збурень $\mathbf{w}(k)$.

Початковим станом системи $\mathbf{x}_0$ будемо вважати випадкові змінні з відомими статистиками:

$$E[\mathbf{x}_0] = \bar{\mathbf{x}}_0; \ E\left[\mathbf{x}_0\mathbf{x}_0^T\right] = \mathbf{M}; \ E\left[\mathbf{w}(k)\mathbf{x}_0^T\right] = 0, \ \forall k.$$

Нехай вектор вимірів $\mathbf{z}(k)$ вихідних змінних доступний у будь-який момент часу k, а його компоненти лінійно пов'язані з вектором стану і на них впливає шум вимірів, тобто

$$\mathbf{z}(k) = \mathbf{H}(k)\,\mathbf{x}(k) + \mathbf{v}(k),$$

де $\mathbf{H}(k)$ — матриця спостережень розмірності $(r \times n)$, $\mathbf{v}(k)$ — r-вимірний вектор випадкових величин шуму вимірів з відомими статистиками

$$E[\mathbf{v}(k)] = 0,\ E[\mathbf{v}(k)\,\mathbf{v}^T(j)] = \mathbf{R}(k)\,\delta_{kj},$$

де $\mathbf{R}(k)$ — додатно визначена коваріаційна матриця шумів вимірів розмірності $(r \times r)$, діагональні елементи якої є дисперсіями шуму у кожному каналі вимірів. Шум вимірів також задовольняє властивостям білого гаусового шуму. Тобто він має нормальний розподіл і вважається некорельованим із зовнішнім збуренням $\mathbf{w}(k)$ і початковим станом системи

$$E[\mathbf{v}(k)\,\mathbf{w}^T(j)] = 0, \quad \forall\, k, j\,;$$

$$E[\mathbf{v}(k)\,\mathbf{x}_0^T(j)] = 0, \quad \forall\, k\,.$$

Для визначеної вище системи з вектором стану $\mathbf{x}(k)$ необхідно знайти оцінку стану $\hat{\mathbf{x}}\,(k)$ в момент k як лінійну комбінацію оцінки $\hat{\mathbf{x}}\,(k-1)$ в момент $k-1$ і самого останнього виміру (статистичних даних) $\mathbf{z}(k)$.

Оцінка $\hat{\mathbf{x}}(k)$ повинна знаходитися як найкраща за мінімумом середнього значення суми квадратів похибок оцінок. Інакше кажучи, оцінка повинна бути такою, щоб

$$E\left[\left(\hat{\mathbf{x}}(k)-\mathbf{x}(k)\right)^T\left(\hat{\mathbf{x}}(k)-\mathbf{x}(k)\right)\right] = \min_{\mathbf{K}},$$

де $\mathbf{x}(k)$ — точне значення вектора стану, яке можна знаходити за допомогою детермінованої складової математичної моделі процесу (тобто без врахування випадкових складових — збурень стану і шумів вимірів); $\mathbf{K}$ — оптимальний матричний коефіцієнт фільтра, який необхідно знайти в результаті розв'язання оптимізаційної задачі.

Таким чином, фільтр потрібно будувати та використовувати для уточнення оцінок стану процесу в умовах впливу випадкових зовнішніх збурень та наявності шумів (похибок) вимірів. На сьогодні оптимальні фільтри — це практично обов'язкова складова комп'ютерних систем обробки експериментальних і статистичних даних. Докладніше оптимальний фільтр буде розглянуто нижче.

8.3. Аналіз наявності нелінійностей

Однією з проблем при визначенні структури моделі є встановлення факту наявності нелінійностей у досліджуваному процесі та їх

типу. Для розв'язання цієї проблеми обов'язково використовують *візуальний аналіз* даних та формальні тести на наявність нелінійностей. Досвідченому фахівцю з моделювання візуальний аналіз дає можливість оперативно виявити наявність ділянок з лінійним або нелінійним трендом, наявність гетероскедастичності та значних імпульсних викидів (екстремальних значень), які можуть суттєво впливати на якість моделі. Необхідно зазначити, що існують окремі навчальні курси з візуального аналізу даних. Це свідчить про те, що не варто нехтувати такою доступною, але ефективною можливістю дослідження даних, адже, на думку психологів, один інформативний рисунок може замінити до двох тисяч слів.

Існує також ряд формальних тестів на наявність нелінійності. Розглянемо простий тест для визначення наявності нелінійності [4]. Цей тест застосовується у випадку, коли можна набрати кілька груп (реалізацій) спостережень для одного і того самого процесу:

$$\hat{F} = \frac{\frac{1}{m-2}\sum_{i=1}^{m} n_i \left(\overline{y}_i - \hat{y}_i\right)^2}{\frac{1}{n-m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{ij} - \overline{y}_i\right)^2},$$

де $\overline{y}_i$ — середнє значення для i-ї реалізації (вибірки або групи) даних; $\hat{y}_i$ — середнє для лінійної апроксимації даних; m — кількість груп даних; n_i — кількість вимірів в i-й групі; n — загальна кількість вимірів. Фактично, наведена статистика — це таке співвідношення:

$$\hat{F} = \frac{\text{Відхилення середніх значень від прямої регресії}}{\text{Відхилення значень } y(k) \text{ від групових середніх}}.$$

Якщо статистика $\hat{F}$ з $\nu_1 = m - 2$ та $\nu_2 = n - m$ ступенями вільності досягає або перевищує рівень значущості, то гіпотезу стосовно лінійності процесу необхідно відхилити. Недоліком даного підходу є те, що для його застосування необхідно мати кілька (не менше трьох) груп (реалізацій) даних для одного і того самого процесу, які можна отримати в результаті виконання повторних експериментів. Очевидно, що це не завжди можливо.

Наявність нелінійності можна встановити також за допомогою вибіркових *нелінійних кореляційних функцій* (НКФ), тобто кореляційних функцій, розрахованих за вибірками експериментальних (статистичних) даних. Наприклад, якщо дискретна НКФ [10]

$$r_{yx^2}(s) = r_{y(k)x^2(k-s)} = \frac{1}{N} \frac{\sum_{k=s+1}^{N} \left\{ \left[y(k) - \bar{y} \right] \left[x(k-s) - \bar{x} \right]^2 \right\}}{\sigma_y \sigma_x^2}, \quad s = 0, 1, 2, 3, \ldots,$$

містить значення, які суттєво відрізняються від нуля у статистичному смислі, то процес містить квадратичну нелінійність відносно регресора x.

Наявність нелінійного детермінованого тренду у процесі можна визначити шляхом оцінювання рівняння

$$y(k) = a_0 + c_1 k + c_1 k^2 + \ldots + c_m k^m + \varepsilon(k),$$

яке є поліномом порядку m відносно часу; $\varepsilon(k)$ — випадковий процес, причини появи якого будуть розглянуті нижче у цьому розділі. Якщо хоча б один із коефіцієнтів моделі — c_i, $i = 1, \ldots, m$, є статистично значущим, то гіпотеза щодо відсутності тренду відхиляється. Якщо тренд відносно швидко змінює свій напрямок руху і для нього важко знайти адекватний функціональний опис, то застосовують моделі випадкових трендів, що базуються на комбінаціях випадкових величин. Наявність нелінійності стосовно регресора $x(k)$ можна встановити за допомогою відповідного полінома:

$$y(k) = a_0 + c_1 x(k) + c_1 x^2(k) + \ldots + c_m x^m(k) + \varepsilon(k).$$

Автоматично оцінює структуру математичної моделі *метод групового врахування аргументів* (МГВА), запропонований академіком О. Г. Івахненком (Інститут кібернетики АН України). Цей метод вже багаторазово застосовували до опису широкого класу процесів з метою оцінювання прогнозів та створення систем керування. Його успішно застосовують і сьогодні до моделювання процесів різної природи з нелінійностями та нестаціонарностями. Подальшим розвитком цього методу є нечіткий МГВА, що базується на нечіткому представленні параметрів оцінюваної моделі. Загалом задача встановлення наявності та визначення типу нелінійності залишається предметом дослідження.

8.4. Формування інших елементів структури моделі

На наступному етапі необхідно вибрати структури моделей-кандидатів. Поняття структури моделі було розглянуто вище, але нагадаємо, що воно включає *порядок моделі* (найвищий порядок рівнянь, що його утворюють); *вимірність* (кількість рівнянь моделі); час *запізнен-*

ня на вході (лаг) та його оцінка; *можливі нелінійності* та їх тип; *зовнішні збурення* та їх тип (детерміновані або випадкові; адитивні та мультиплікативні; імпульсні та неперервні); *обмеження* на параметри і змінні моделі.

Щоб визначити, які незалежні змінні (регресори) необхідно включити в праву частину рівняння, знаходять коефіцієнт кореляції між залежною та відповідною незалежною змінною.

Коефіцієнт кореляції, а в загальному випадку кореляційна функція, дає змогу встановити факт існування зв'язку між змінними. Кореляція може бути лінійною або нелінійною, залежно від типу функціональної залежності, яка фактично має місце між змінними. У більшості практичних випадків розглядають лінійну кореляцію (взаємозв'язок) між змінними, але більш глибокий аналіз потребує використання нелінійних залежностей. Складну нелінійну залежність можна спростити, але знати про її існування необхідно для того, щоб побудувати за необхідністю (складнішу за структурою) модель процесу з вищим ступенем адекватності.

Кореляційна матриця дає можливість встановити існування зв'язку між залежною (ендогенною) змінною та незалежними (екзогенними) змінними у правій частині. Розглянемо кореляційну матрицю **R** розмірності 3 × 3, яка будується для трьох змінних *x, y, z*

$$\mathbf{R} = \begin{bmatrix} r_{yy} & r_{xy} & r_{zy} \\ r_{yx} & r_{xx} & r_{zx} \\ r_{yz} & r_{xz} & r_{zz} \end{bmatrix}, \text{ де } r_{yx} = r_{xy},\ r_{yz} = r_{zy},\ r_{xz} = r_{zx}.$$

Нехай y — показник якості технологічного процесу; x, z — технологічні параметри, які, за припущенням, впливають на показник якості. Тобто ставиться задача встановлення існування залежності вигляду

$$y = f(x, z),$$

яка може бути представлена у формі регресії змінної y на незалежні змінні x, z :

$$y(k) = a_0 + a_1 x(k) + a_2 z(k) + \varepsilon(k),$$

де k — дискретний час (наприклад, у долях секунди, секундах, хвилинах, годинах, днях, тижнях, місяцях і т. ін.); $\varepsilon(k)$ — випадкова змінна, введення якої у модель пояснюється наступними причинами:

- часто буває неможливо встановити всі незалежні змінні, які впливають на залежну змінну, а тому наведене рівняння описує процес з похибкою;
- можуть існувати такі незалежні змінні, які неможливо виміряти і включити в модель, а тому їх розглядають як збурення і вважають, що їх спільний вплив на залежну змінну описується випадковою змінною $\varepsilon(k)$;
- у наведене вище регресійне рівняння можна ввести пояснювальні змінні, які формально корельовані із залежною змінною, але фактично не впливають на неї;
- для будь-якого методу оцінювання параметрів рівнянь властиві методичні обчислювальні похибки, які варто по можливості врахувати в моделі.

Вважається, що сукупний вплив усіх вказаних факторів можна описати певним чином за допомогою випадкової змінної $\varepsilon(k)$. Оскільки вона не вимірюється, то оцінити її значення (*похибку* моделі або *залишок*) можна тільки після оцінювання коефіцієнтів моделі, тобто

$$\hat{\varepsilon}(k) = y(k) - \hat{y}(k),$$

де $\hat{y}(k)$ — оцінка змінної $y(k)$, отримана за допомогою моделі; $y(k)$ — фактичний вимір.

Для знаходження елементів матриці **R** необхідно мати синхронізовані у часі вибірки значень всіх трьох змінних *y, x, z*. Формула для розрахунку вибіркових коефіцієнтів кореляції має вигляд [10]

$$r_{yx} = \frac{\dfrac{1}{N-1}\sum_{k=1}^{N}\left\{\left[x(k)-\overline{x}\right]\left[y(k)-\overline{y}\right]\right\}}{\sigma_x \sigma_y},$$

де $\overline{x}, \overline{y}$ — вибіркові середні значення змінних *x, y*; σ_x, σ_y — стандартні відхилення цих змінних, тобто корені квадратні з їх дисперсії. Наприклад,

$$\sigma_y = \sqrt{\sigma_y^2} = \left[\frac{1}{N-1}\sum_{k=1}^{N}\left[y(k)-\overline{y}\right]^2\right]^{1/2},$$

де N — число вимірів змінної y; $\overline{y}$ — вибіркове середнє значення ряду $\{y(k)\}$, яке обчислюється за відомою формулою

$$\overline{y} = \frac{1}{N}\sum_{k=1}^{N} y(k).$$

Очевидно, що перед формальним знаходженням коефіцієнтів кореляції, необхідно зробити аналіз процесу і визначити наявність (або відсутність) логічних зв'язків між змінними. Це дає змогу обмежитись розглядом тільки тих змінних, які дійсно впливають на залежну змінну, наприклад, на показник якості продукції. Для правильного вибору незалежних (екзогенних або пояснювальних) змінних необхідно досконало знати технологічний або інший процес, що моделюється.

На основі значень коефіцієнтів кореляції ухвалюється рішення про включення їх у рівняння регресії

$$y(k) = a_0 + b_1 x(k) + b_2 z(k) + \varepsilon(k).$$

У випадку багаточинникової регресії маємо

$$y(k) = a_0 + a_1 x_1(k) + a_2 x_2(k) + a_3 x_3(k) + ... + a_{p-1} x_{p-1}(k) + \varepsilon(k).$$

Можна показати, що між коефіцієнтами регресії b_1, b_2 і коефіцієнтами кореляції r_{yx}, r_{yz} існує однозначний аналітичний взаємозв'язок. Емпіричні дослідження свідчать, що незалежні змінні необхідно включати в рівняння регресії, якщо $| r_{yx} | \geq 0{,}4$. Однак необхідно пам'ятати, що ця рекомендація має досить наближений характер. Існують випадки, коли у модель слід включати регресори, які мають менші значення коефіцієнта кореляції із залежною змінною.

Останнє рівняння є рівнянням *лінійної регресії*, яке містить p параметрів (коефіцієнтів), але досить часто необхідно застосовувати складніші нелінійні моделі. Характерним представником нелінійної стосовно змінних регресії є поліноміальна регресія порядку $p - 1$:

$$y(k) = a_0 + a_1 x_1(k) + a_2 x^2(k) + a_3 x^3(k) + ... + a_{p-1} x^{p-1}(k) + \varepsilon(k).$$

Хоча в це рівняння включено тільки одну незалежну змінну $x(k)$, очевидно, що воно може бути розширене будь-якими іншими змінними.

Регресори, які включають у модель з фактичними значеннями їх лагів, називають *провідними індикаторами*. Наприклад, якщо результат (прибуток) від інвестицій з'являється через три квартали, то при використанні квартальних даних регресійне рівняння буде мати такий вигляд:

$$y(k) = a_0 + a_1 x(k - 3) + \varepsilon(k).$$

Таке рівняння дає можливість коректно прогнозувати залежну змінну на три кроки вперед.

Для визначення необхідності введення в рівняння регресії авторегресійної складової необхідно знайти і дослідити вибіркову *автокореляційну функцію* змінної $y(k)$. Рівняння з авторегресійною складовою має вигляд [9; 10]:

$$y(k) = a_0 + a_1 y(k-1) + a_2 y(k-2) + b_1 x(k) + b_2 z(k) + \varepsilon(k),$$

тобто в рівняння регресії введено авторегресійну (АР) складову другого порядку. Порядок авторегресії визначається за допомогою автокореляційної функції (АКФ) і часткової АКФ (ЧАКФ). Кількість значень автокореляційної функції, які відмінні від нуля у статистичному смислі й визначаються параметром зсуву s (що використовується при знаходженні кореляційних функцій), і буде становити оцінку порядку авторегресії.

Природа авторегресії пояснюється існуванням так званої "пам'яті" процесу, яка виявляється у тому, що його поточний стан може значною мірою визначатись попередніми станами. Наприклад, стан людини вранці залежить від того, яким було самопочуття увечері та в попередні дні. На формування ринкових цін суттєво впливають значення цін у попередні періоди часу. Очевидно, що поточний рівень валового внутрішнього продукту (ВВП) залежить від його попередніх значень, а поточний стан технічної системи або технологічного процесу також залежить від їх стану у попередні моменти часу.

АКФ та ЧАКФ використовують для визначення попередньої оцінки порядку авторегресійної частини моделі, тобто скільки затриманих у часі значень потрібно брати для опису процесу. При цьому необхідно врахувати, що ЧАКФ дає "чіткішу" оцінку порядку процесу, ніж АКФ. Наприклад, для процесу АР(1) значення основної змінної $y(k)$ та $y(k-2)$ будуть корельованими, незважаючи на те, що $y(k-2)$ не присутнє у моделі. Кореляція між $y(k)$ і $y(k-2)$, тобто ρ_2, дорівнює коефіцієнту кореляції між значеннями $y(k)$ і $y(k-1)$ помноженому на коефіцієнт кореляції між $y(k-1)$ і $y(k-2)$ або $\rho_2 = \rho_1\rho_1 = \rho_1^2$. Подібні "непрямі" кореляції наявні в АКФ будь-якого процесу авторегресії.

Вибіркова АКФ знаходиться за формулою [10]

$$r_y(s) = r_{y(k)y(k-s)} = \frac{1}{N-1} \frac{\sum_{k=s+1}^{N} \left\{ \left[y(k) - \bar{y} \right] \left[y(k-s) - \bar{y} \right] \right\}}{\sigma_y^2}, \quad s = 1, 2, 3, \ldots,$$

де σ_y^2 — вибіркова дисперсія змінної $y(k)$; $\bar{y}$ — середнє значення вибірки даних.

Кількість значень АКФ, відмінних від нуля, вказує на порядок авторегресійної частини моделі. Для стаціонарного процесу (це процес із постійними середнім, дисперсією та коваріацією) коефіцієнти $r_y(s)$ мають нормальний розподіл та нульове середнє.

На відміну від АКФ, часткова АКФ між значеннями $y(k)$ та $y(k - s)$ виключає вплив величин $y(k - 1)$ ... $y(k - s + 1)$, а це означає, що коефіцієнти ЧАКФ чіткіше відображають зв'язок між окремими значеннями основної змінної. Так, для процесу АР(1) ЧАКФ між $y(k)$ та $y(k - 2)$ дорівнює нулю за визначенням, що підтверджується знайденими значеннями ЧАКФ. Для того щоб знайти попередню оцінку порядку моделі, вибіркові коефіцієнти ЧАКФ (тобто коефіцієнти, знайдені за вибіркою даних) можна знайти також за допомогою простого методу, який полягає у наступному.

1. Знаходять додатковий часовий ряд з відхилень основної змінної:

$$\{y'(k)\} = \{y(k)\} - \mu,$$

де μ — середнє значення ряду.

2. Формують рівняння першого порядку:

$$y'(k) = \Phi_{11}\, y'(k - 1) + e(k),$$

де $e(k)$ — похибка моделі. У такому рівнянні Φ_{11} відіграє роль коефіцієнта АКФ та ЧАКФ між $y(k)$ та $y(k - 1)$. Для оцінювання двох коефіцієнтів можна записати рівняння другого порядку:

$$y'(k) = \Phi_{11}\, y'(k - 1) + \Phi_{22}\, y'(k - 2) + e(k),$$

де Φ_{22} — коефіцієнт ЧАКФ між $y(k)$ та $y(k - 2)$.

Таким чином, дискретні значення ЧАКФ можна знайти за допомогою значень АКФ, використовуючи такі вирази [14]:

$$\Phi_{11} = r_1, \qquad \Phi_{22} = \frac{r_2 - r_1^2}{1 - r_1^2};$$

$$\Phi_{ss} = \frac{r_s - \sum_{j=1}^{s-1} \Phi_{s-1,j} r_{s-j}}{1 - \sum_{j=1}^{s-1} \Phi_{s-1,j} r_j},$$

де $r_s = r_y(s)$.

У загальному випадку коефіцієнти ЧАКФ стаціонарного процесу АРКС (p, q) повинні прямувати до нуля, починаючи із p-го значення. АКФ процесу АРКС(p, q) починає прямувати до нуля при значеннях зміщення $s \geq q$.

Твердження, що значення автокореляційної функції повинні бути відмінними від нуля у статистичному смислі, означає, що існує вираз (формула), який дає змогу підтвердити або спростувати цей факт. Одним із загальноприйнятих підходів до встановлення факту, що значення АКФ суттєво відмінні від нуля у статистичному смислі, є знаходження та аналіз значущості статистичного параметра (або просто статистики) Люнга-Бокса $Q(r_k)$ за формулою [8]

$$Q(r_k) = N(N+2)\sum_{k=1}^{s} r_k^2 \big/ (N-k),$$

де N — обсяг вибірки даних змінної, для якої знайдено значення автокореляційної функції r_k ; s — кількість відліків АКФ, які досліджуються на суттєву відмінність від нуля. Якщо дані згенеровані процесом АР чи АРКС, то значення $Q(r_k)$ асимптотично мають розподіл χ^2 з s ступенями вільності, а тому для перевірки їх значущості необхідно користуватись відповідними статистичними таблицями. Очевидно, що більші значення вибіркової автокореляційної функції приводять до більших значень $Q(r_k)$.

Для ідеального процесу білого шуму $Q(r_k) = 0$. Якщо значення $Q(r_k)$, знайдене за наведеною формулою, перевищує критичне значення з розподілу χ^2 з s ступенями вільності, то існує щонайменше одне значення $r(k)$, яке є відмінним від нуля у статистичному смислі. Статистику Люнга-Бокса можна застосовувати також для встановлення ступеня близькості залишків моделі до білого шуму. Однак необхідно пам'ятати, що при знаходженні s значень кореляційної функції кількість ступенів вільності зменшується на кількість коефіцієнтів моделі. Таким чином, при аналізі залишків моделі АРКС (p, q) статистика $Q(r_k)$ має розподіл χ^2 з $s - p - q$ ступенями вільності, а із врахуванням константи $s - p - q - 1$.

Цей етап закінчується формуванням структур кількох моделей-кандидатів з векторами параметрів $\theta_1, ..., \theta_m$, де m — число кандидатів. Кандидатів може бути декілька, оскільки встановити структуру точно за один раз, як правило, неможливо. Загалом оцінювання структури моделі високого ступеня адекватності — це трудомісткий ітераційний процес, який вимагає поглиблених знань, досвіду моделювання

та значних зусиль. На наступному етапі оцінюють числові значення параметрів знайдених моделей-кандидатів.

8.5. Оцінювання параметрів моделей-кандидатів

Далі оцінюють параметри (коефіцієнти) моделей-кандидатів за умови відомої структури цих моделей, використовуючи принцип економії або збереження. Цей принцип означає, що *кількість коефіцієнтів, що оцінюються, не повинна перевищувати їх необхідну кількість* ("необхідність" можна визначити, наприклад, як необхідність збереження в моделі основних статистичних характеристик процесу — математичного сподівання, дисперсії та коваріації).

При моделюванні процесів будь-якої природи необхідно пам'ятати, що поведінку процесу варто коректно *апроксимувати* за допомогою моделей, а не намагатися описати її до найменших дрібниць. Необхідно враховувати також, що різні за структурою моделі можуть мати дуже схожі властивості.

Модель, що оцінюється, повинна задовольняти принцип інверсії, тобто щоб за допомогою отриманого рівняння можна було б згенерувати теоретичний ряд, на основі якого оцінювались коефіцієнти. Це означає, що хоча модель і спрощена, вона повинна співпадати з досліджуваним процесом за такими основними характеристиками, як середнє, дисперсія та коваріація.

У процедурі оцінювання часто використовують не абсолютні значення змінних, а їх відхилення від середнього, тобто

$$y(k) = Y(k) - \mu_y,$$

де $Y(k)$ — значення виміру, μ_y — середнє значення ряду. Якщо для оцінювання параметрів використовується рекурсивна процедура, то поточне середнє можна знаходити за формулою

$$\mu_y(k) = \mu_y(k-1) + \frac{1}{k}\left[y(k) - \mu_y(k-1)\right].$$

Найбільш поширеними методами оцінювання параметрів моделі є такі:

- метод найменших квадратів (МНК);
- метод максимальної правдоподібності (ММП);
- метод допоміжної (інструментальної) змінної (МДП);
- нелінійний метод найменших квадратів (НМНК);

- метод Монте-Карло для марковських ланцюгів (МКМЛ) та їх рекурсивні версії (РМНК, РММП, РМДП).

Деякі методи будуть розглянуті в розділі, присвяченому методам оцінювання. Оцінки (звичайного) МНК знаходять за допомогою формули

$$\hat{\theta} = \left[\mathbf{X}^T\mathbf{X}\right]^{-1}\mathbf{X}^T\mathbf{y},$$

де $\theta[p]$ — вектор оцінок параметрів вимірності p; $\mathbf{X}[N \times p]$ — матриця вимірів; $\mathbf{y}[N]$ — вектор вимірів залежної змінної. У квадратних дужках вказана вимірність векторів і матриці. Елементи матриці вимірів формують по-своєму для кожної конкретної моделі. Так, для моделі

$$y(k) = a_0 + a_1 x_1(k) + a_2 x_2(k) + a_3 x_3(k) + \varepsilon(k)$$

матриця вимірів має вигляд

$$\mathbf{X} = \begin{bmatrix} 1 & x_1(1) & x_2(1) & x_3(1) \\ 1 & x_1(2) & x_2(2) & x_3(2) \\ \dots & \dots & \dots & \dots \\ 1 & x_1(N) & x_2(N) & x_3(N) \end{bmatrix}.$$

Одиниці у першому стовпчику матриці $\mathbf{X}$ означають, що вимір при коефіцієнті a_0 завжди дорівнює одиниці.

Елементи матриці вимірів дещо ускладнюються у випадку використання поліноміальної моделі, але її також можна оцінювати за допомогою лінійних методів. Безпосереднє застосування методу мінімізації суми квадратів похибок до поліноміальної моделі порядку p приводить до формування такої матриці вимірів:

$$\mathbf{X}' = \begin{bmatrix} N & \sum_{k=1}^{N} x(k) & \sum_{k=1}^{N} x^2(k) & \dots & \sum_{k=1}^{N} x^p(k) \\ \sum_{k=1}^{N} x(k) & \sum_{k=1}^{N} x^2(k) & \sum_{k=1}^{N} x^3(k) & \dots & \sum_{k=1}^{N} x^{p+1}(k) \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{k=1}^{N} x^p(k) & \sum_{k=1}^{N} x^{p+1}(k) & \sum_{k=1}^{N} x^{p+2}(k) & \dots & \sum_{k=1}^{N} x^{2p}(k) \end{bmatrix}.$$

При такому представленні векторно-матричне рівняння для N вимірів залежної та незалежної змінних можна записати так: $\mathbf{y}' = \mathbf{X}'\,\theta$.

Звідси $\hat{\theta} = [\mathbf{X}']^{-1}\mathbf{y}'$, =

де $\mathbf{y}' = \left[\sum_1^N y(k);\ \sum_1^N x(k)y(k);\ \ldots;\sum_1^N x^p(k)y(k)\right]^T$; $\hat{\theta}$ — вектор оцінок параметрів моделі. Тобто оцінку вектора параметрів можна визначити шляхом знаходження розв'язку системи лінійних (нормальних) рівнянь.

Для отримання незміщених, консистентних та ефективних оцінок вектора параметрів θ лінійної регресійної математичної моделі, наприклад, моделі змішаної регресії

$$y(k) = a_0 + a_1 y(k-1) + a_2 y(k-2) + b_1 x(k) + b_2 z(k) + \varepsilon(k),$$

за допомогою методу найменших квадратів необхідно задовольнити наступні умови:

а) $\varepsilon(k)$ — некорельована послідовність однаково (нормально) розподілених випадкових чисел з нульовим середнім, тобто $E[\varepsilon(k)] = 0$,

$$\operatorname{cov}\left[\varepsilon(k)\right] = E\left[\varepsilon(k)\varepsilon(j)\right] = \begin{cases}\sigma_\varepsilon^2, & k = j;\\ 0, & k \neq j.\end{cases}$$

б) послідовності $\varepsilon(k)$ і $y(k)$ не повинні бути корельовані між собою.

Зазначимо, що перевірити виконання наведених умов ми можемо тільки після оцінювання коефіцієнтів моделі (апостеріорно), а до оцінювання (апріорно) можна тільки постулювати їх виконання. Тобто після оцінювання параметрів моделі оцінка значень випадкового процесу визначається похибками моделі

$$\hat{\varepsilon}(k) = e(k) = y(k) - \hat{y}(k),$$

що дає можливість виконати аналіз характеристик випадкового процесу $\{\varepsilon(k)\}$.

8.6. Діагностика моделей — вибір кращої з множини кандидатів

На цьому етапі аналізується якість моделі, тобто виконується перевірка оцінених кандидатів на адекватність процесу. Діагностика побудованих моделей складається з кроків, наведених нижче.

1. *Візуальне дослідження графіка похибок моделі* $e(k) = y(k) - \hat{y}(k)$, де $\hat{y}(k)$ — оцінка змінної, отримана за допомогою побудованої моделі.

На графіку не повинно бути значних викидів та довгих інтервалів, на яких похибка набуває великих значень (тобто довгих інтервалів суттєвої неадекватності). У випадку застосування рекурсивних методів оцінювання найбільші похибки будуть у перехідному процесі, коли інформаційна матриця ще не містить достатньо інформації щодо процесу. Однак для моделей низьких порядків (1–2) перехідний процес повинен закінчуватись через 20–30 кроків (не більше), а оцінки параметрів мають прямувати до точних значень.

Для аналізу характеристик похибок необхідно застосовувати статистичні характеристики (статистики), які свідчать про близькість цього випадкового процесу до білого шуму. Це коефіцієнт асиметрії, ексцес, статистика Жак-Бера.

2. *Похибки моделі не повинні бути корельовані* між собою. Для аналізу наявності кореляції між значеннями похибок необхідно знайти АКФ та ЧАКФ для ряду $\{e(k)\}$ і за допомогою Q-статистики визначити ступінь корельованості (наприклад, Q-статистика вважається несуттєвою до рівня 10 %).

Крім того, корельованість похибок визначають за допомогою статистики Дарбіна-Уотсона (DW), яка розраховується за формулою

$$DW = 2 - 2\rho,$$

де $\rho = E[e(k)\, e(k-1)]/\, \sigma_e^2$ — коефіцієнт кореляції між сусідніми значеннями похибки; σ_e^2 — дисперсія послідовності похибок $\{e(k)\}$. Таким чином, при повній відсутності кореляції між похибками $DW = 2$ — це ідеальне значення. Граничними значеннями для DW є 0 (при $\rho = 1$) та +4 (при $\rho = -1$).

Отримати формулу $DW = 2 - 2\rho$ можна досить просто. Автори цієї статистики (Дарбін та Уотсон) запропонували скористатись для перевірки корельованості похибок моделі такою формулою [10]

$$DW = \frac{\sum_{k=2}^{N}\left[e(k)-e(k-1)\right]^2}{\sum_{k=1}^{N} e^2(k)} = \frac{\sum_{k=2}^{N}\left[e(k)-e(k-1)\right]\left[e(k)-e(k-1)\right]}{\sum_{k=1}^{N} e^2(k)},$$

тобто DW можна, певним чином, трактувати як коефіцієнт автокореляції для (перших різниць) приростів похибок.

Розкриваючи квадрат різниці у чисельнику, отримаємо

$$DW = \frac{\sum_{k=2}^{N} e^2(k)}{\sum_{k=1}^{N} e^2(k)} + \frac{\sum_{k=2}^{N} e^2(k-1)}{\sum_{k=1}^{N} e^2(k)} - 2\frac{\sum_{k=2}^{N} e(k)e(k-1)}{\sum_{k=1}^{N} e^2(k)},$$

де $\frac{\sum_{k=2}^{N} e^2(k)}{\sum_{k=1}^{N} e^2(k)} \approx 1$; $\frac{\sum_{k=2}^{N} e^2(k-1)}{\sum_{k=1}^{N} e^2(k-1)} \approx 1$; а $\frac{\sum_{k=2}^{N} e(k)e(k-1)}{\sum_{k=1}^{N} e^2(k-1)} = \rho$.

Тому можна записати, що $DW = 2 - 2\rho$.

3. Для лінійної моделі 2–3 порядку *оцінки параметрів повинні прямувати до сталих значень* після 30–40 (не більше) ітерацій алгоритму оцінювання. Якщо кількість ітерацій набагато перевищує вказані числа, то це свідчить, що процес може бути нестаціонарним.

4. Перевірка статистичної значущості оцінок параметрів моделі. *Статистика Стьюдента або t-статистика* (випадкова величина, що має *t*-розподіл), яка використовується для визначення значущості оцінки кожного коефіцієнта у статистичному сенсі, визначається за формулою

$$t = \frac{\hat{a} - a^0}{SE_{\hat{a}}},$$

де $\hat{a}$ — оцінка коефіцієнта моделі; a^0 — нуль-гіпотеза (початкова гіпотеза) щодо цієї оцінки; $SE_{\hat{a}}$ — стандартна похибка оцінки. В якості нуль-гіпотези щодо значущості оцінки можна висувати будь-яку: що коефіцієнт значущий, тобто $H_0 : a^0 \neq 0$, або незначущий ($H_0 : a^0 = 0$). Статистична теорія перевірки гіпотез пропонує висувати нуль-гіпотезу, яка є протилежною бажаному результату. У даному випадку бажаним результатом є статистична значущість коефіцієнтів математичної моделі. Таким чином, необхідно висувати нульову гіпотезу, що коефіцієнт незначущий. Це дає можливість коректно підійти до визначення значущості оцінок коефіцієнтів та дещо спростити розрахунки.

Для того щоб встановити, чи оцінка коефіцієнта значуща, необхідно знати довжину вибірки даних N (обсяг вибірки); кількість ступенів вільності $f = N - n$, де n — число коефіцієнтів моделі, які оціню-

ються на основі ряду даних, і вибрати рівень значущості $\alpha = 0{,}01$, або $\alpha = 0{,}05$, або $\alpha = 0{,}10$ (для цих значень існують розраховані таблиці критичних значень статистики). Фактично рівень значущості означає ймовірність припуститись *помилки першого роду* при перевірці гіпотези. Згадаємо, що

$$\alpha = p\{X \in G/w \,|\, H_0\} = \int\limits_{n-m(G/w)} L_{H_0}(X)\,dx,$$

де $X = [x_1, ..., x_n] \in R^n$ — вся вибірка, яка розбивається на дві множини, що не перетинаються: w і G/w (w — область прийняття нуль-гіпотези); G/w — критична область: якщо $X \in G/w$, то H_0 відхиляється; $L_{H_0}(X)$ закон розподілу X. Помилка першого роду означає відхилення правильної гіпотези.

Користуючись значеннями N, f і α, з таблиць для t-розподілу знаходять критичне значення t-статистики, тобто $t_{кр}$. Для перевірки правильності висунутої гіпотези розраховане значення t порівнюють з критичним $t_{кр}$. Якщо

$$-t_{кр} < t < t_{кр} \quad \text{або} \quad |t| < |t_{кр}|,$$

то нуль-гіпотеза стосовно незначущості коефіцієнта приймається (його можна не враховувати в регресії). Звідси випливає, що чим більшим є значення t-статистики для оцінки коефіцієнта, тим імовірніше, що цей коефіцієнт є значущим.

Загалом послідовність дій при перевірці значущості оцінок коефіцієнтів побудованої моделі можна сформулювати так:

- сформулювати нуль-гіпотезу стосовно значущості коефіцієнта;
- обчислити значення t-статистики для кожного коефіцієнта регресії (це робить кожний пакет для математичного моделювання);
- за допомогою значень N, f і α знайти із таблиць для t-статистики її критичне значення;
- перевірити нуль-гіпотезу за наведеним вище простим правилом (аналіз виконання нерівності $-t_{кр} < t < t_{кр}$).

5. *Знаходимо коефіцієнт множинної детермінації* R^2 [1]

$$R^2 = \frac{\mathrm{var}(\hat{y})}{\mathrm{var}(y)} = 1 - \frac{SSE}{SST},$$

де $\mathrm{var}(\hat{y})$ — дисперсія залежної змінної, оціненої за допомогою побудованої моделі; var (y) — дисперсія вимірів залежної змінної; *SSE* = $= \sum_{k=1}^{N}\left[y(k)-\hat{y}(k)\right]^2$ — сума квадратів похибок (залишків) моделі (*sum of squared errors*); $SST = \sum_{k=1}^{N}\left[y(k)-\bar{y}\right]^2$ — загальна сума квадратів (*total sum of squares*); $\bar{y}$ — середнє значення; *SST* = *SSE* + *SSR*, де SSR = $= \sum_{k=1}^{N}\left[\hat{y}(k)-\bar{y}\right]^2$ — загальна сума квадратів для регресії (*sum of squares for regression*).

Очевидно, що найкращим значенням є $R^2 = 1$, коли дисперсії вимірів залежної змінної та оцінок цієї самої змінної, отриманих за моделлю, співпадають. Цей параметр можна трактувати також, як ступінь інформативності моделі, якщо за ступінь інформативності вибрати дисперсію. Таким чином, R^2 показує рівень інформативності моделі по відношенню до інформативності вибірки даних, за допомогою якої вона була оцінена.

6. *Сума квадратів похибок* для вибраної моделі повинна бути мінімальною, тобто

$$\sum_{k=1}^{N} e^2(k) = \sum_{k=1}^{N}\left[\hat{y}(k)-y(k)\right]^2 \to \min_{\hat{\theta}}$$

у порівнянні з усіма іншими моделями.

7. Для оцінки адекватності моделі використовують *інформаційний критерій Акайке* [1]

$$AIC = N\ln\left(\sum_{k=1}^{N} e^2(k)\right) + 2n,$$

та *критерій Байєса-Шварца* [10]

$$BSC = N\ln\left(\sum_{k=1}^{N} e^2(k)\right) + n\ln(N),$$

де $n = p + q + 1$ — кількість параметрів моделі, які оцінюються за допомогою статистичних даних (p — число параметрів авторегресійної частини моделі; q — кількість параметрів ковзного середнього; “1” з’являється тоді, коли оцінюється зміщення (або *перетин*, тобто a_0).

Критерії Акайке і Байєса-Шварца містять у правій частині суму квадратів похибок, а тому за цими критеріями вибирають ту модель, для якої ці критерії набувають найменших значень. Введення нового регресора приводить до збільшення критерію (при цьому збільшується n), але одночасно зменшується сума квадратів похибок і критерій загалом зменшується. Якщо регресор не покращує модель, то критерій збільшується. Необхідно також зазначити, що асимптотичні властивості для великих вибірок кращі у критерію Байєса-Шварца, тобто його рекомендують застосовувати при відносно великих значеннях N ($N > 100$).

8. Окрім згаданих параметрів, для визначення адекватності моделі в цілому використовують *F-статистику Фішера*, яка пропорційна відношенню

$$F \sim \frac{R^2}{1-R^2},$$

а для множинної (багатофакторної) регресії вона визначається за формулою

$$F = \frac{R^2}{1-R^2} \cdot \frac{(N-p-1)}{p},$$

де, як і раніше, N — число значень ряду; p — число параметрів моделі без врахування a_0.

Таким чином, якщо $R^2 \to 1$, то $F \to \infty$. Порядок застосування F-статистики такий самий, як і t-статистики. Нуль-гіпотезою у даному випадку є припущення про те, що модель неадекватна в цілому, тобто

$$H_0 : a_1 = a_2 = ... = a_p = 0,$$

проти альтернативної гіпотези:

H_1 : хоча б одне значення a_i відмінне від нуля у статистичному смислі.

Значення $F_{кр}$ знаходять із таблиць для F-розподілу. Послідовність застосування цієї статистики можна представити так:

1. Сформулювати нуль-гіпотезу стосовно адекватності моделі в цілому.

   Наприклад, H_0 : модель неадекватна в цілому (або

   $$H_0 : a_1 = a_2 = ... = a_p = 0).$$

2. Розрахувати значення F для оціненої моделі.
3. Вибрати рівень значущості $\alpha = 0{,}01$, або $\alpha = 0{,}05$, або $\alpha = 0{,}10$.
4. Користуючись значеннями N, f і α, знайти критичне значення $F_{кр}$ з таблиць для F-розподілу при $(p, N - p - 1)$ ступенях вільності.
5. Перевірити нуль-гіпотезу:
   якщо $F > F_{кр}$, то нуль-гіпотеза щодо неадекватності моделі в цілому відкидається на вибраному рівні значущості.

Коректне застосування запропонованої методики забезпечує побудову адекватної математичної моделі процесу, якщо статистичні (експериментальні) дані відповідають *вимогам представництва* та *інформативності*. Перша вимога означає, що вибірка даних повинна охоплювати досить великий проміжок часу, щоб повністю відображати поведінку того режиму функціонування процесу, для якого будується модель.

Вимога *інформативності* означає, що вибірка повинна містити в собі обсяг інформації, достатній для оцінювання коефіцієнтів моделі. Наприклад, якщо моделюється процес другого порядку, то вибірка повинна забезпечувати коректне знаходження першої та другої похідної. Іноді інформативність формально оцінюють за допомогою величини дисперсії процесу.

Умову інформативності даних пов'язують з *умовою достатнього збудження* процесу. Достатнє збудження означає, що вхідний сигнал повинен охоплювати всю смугу частот, які може пропускати на вихід процес (об'єкт). Тобто вхідний вплив (сигнал) повинен охоплювати всю амплітудно-частотну характеристику процесу. Ця вимога залишається справедливою для процесів будь-якої природи.

8.7. Приклади побудови моделей за статистичними даними

Приклад 8.2. Розглянемо можливості побудови математичних моделей динаміки для таких макроекономічних процесів України: формування внутрішнього валового продукту (ВВП), індекс споживчих цін (ІСЦ) і грошовий агрегат (М3). Для побудови використано фактичні щомісячні дані з січня 1996 по січень 2005 р., всього 109 значень. Кореляційна матриця для цих змінних [14]:

$$R = \begin{array}{c} \begin{matrix} \text{ІСЦ} & \text{ВВП} & \text{М3} \end{matrix} \\ \begin{pmatrix} 1 & -0{,}3043610 & -0{,}2491123 \\ -0{,}3043610 & 1 & 0{,}9317529 \\ -0{,}2491123 & 0{,}9317529 & 1 \end{pmatrix} \end{array}.$$

Корельованість ІСЦ з ВВП та агрегатом М3 незначна; у подальшому ця інформація буде використана при побудові альтернативних варіантів математичних моделей процесів, що розглядаються. Корельованість між ВВП і агрегатом М3 становить приблизно 0,932 — це велике значення, яке може негативно вплинути на якість оцінок моделі при використанні цих змінних у моделі, що будується.

8.7.1. Модель індексу споживчих цін

Авторегресійна модель індексу споживчих цін

Оскільки моделі парної регресії та авторегресійні моделі прості за своєю структурою, але досить часто вони дають можливість досягти високого ступеня адекватності досліджуваному процесу, то для опису індексу споживчих цін використаємо модель авторегресії з ковзним середнім [1–4]

$$y(k) = a_0 + \sum_{i=1}^{p} a_i y(k-i) + \sum_{j=1}^{q} b_j \varepsilon(k-j) + \varepsilon(k).$$

Значення автокореляційної функції процесу наведено в табл. 8.1. При побудові моделі індекс споживчих цін позначимо через *isc*. У табл. 8.1 прийнято такі скорочення: АКФ — автокореляційна функція, а ЧАКФ — часткова АКФ.

Результати оцінювання моделі авторегресії першого порядку по методу найменших квадратів наведені у табл. 8.2.

Отримана модель АР(1):

$$isc(k) = a_0 + a_1\, isc(k-1) + \varepsilon(k) = 43{,}294 + 0{,}57\, isc(k-1) + e(k),$$

де $e(k)$ — залишки (похибки) моделі, значення яких можна знайти у відповідному файлі пакету програм, що застосовується для побудови моделі. Основні статистичні характеристики якості моделі такі:

$$R^2 = 0{,}415;\ J = \text{СКП} = 141{,}99;\ DW = 1{,}931.$$

Таблиця 8.1

Автокореляційна функція процесу формування індексу споживчих цін (ІСЦ)

Часові дані: 1996:01 2005:01
Всього спостережень: 109

№ пор.	АКФ	ЧАКФ	Q-стат.	Ймовірність
1	0,570	0,570	36,373	0,000
2	0,230	–0,140	42,378	0,000
3	0,129	0,089	44,284	0,000
4	0,076	–0,022	44,957	0,000
5	0,064	0,043	45,433	0,000
6	0,120	0,101	47,137	0,000
7	0,180	0,088	50,960	0,000
8	0,068	–0,136	51,515	0,000
9	0,052	0,105	51,840	0,000
10	0,072	–0,001	52,468	0,000

Таблиця 8.2

Результати оцінювання моделі АР(1) для ІСЦ

Метод оцінювання: метод найменших квадратів
Скоригована часова вибірка даних: 1996:02 2005:01
Використано спостережень: 108 після коригування
Модель АР(1): $I = C(1) + C(2) \cdot I(-1)$

Параметр	Оцінка	Станд. похибка	*t*-статистика	Ймов.
C(1)	43,29386	6,624742	6,535178	0,0000
C(2)	0,569826	0,065703	8,672790	0,0000

R-квадрат	0,415067	Середнє залеж. змінної	100,7407
Скоригований *R*-квадрат	0,409549	Станд. відхил. залеж. змінної	1,506189
Станд. похибка регресії	1,157368	Інформ. критерій Акайке	3,148519
Сума квадратів похибок	141,9871	Критерій Шварца	3,198188
		Стат. Дарбіна-Уотсона	1,931805

Коефіцієнт детермінації має низьке значення (0,415), сума квадратів похибок досить висока (141,99), а статистика Дарбіна-Уотсона (1,931) наближається до найкращого значення. Таким чином, загалом адекватність моделі АР(1) досить низька, а тому структура моделі потребує уточнення. Характеристики якості (історичного, на навчальній вибірці) однокрокового прогнозу будуть такі [1]:

$$\text{СеКП} = 1{,}36;\ \text{САП} = 1{,}02;\ \text{САПП} = 1{,}008;\ U = 0{,}0067,$$

тобто середньоквадратична похибка (СеКП), середня абсолютна похибка (САП), середня абсолютна похибка у процентах (САПП) і коефіцієнт Тейла U [1], який свідчить про загальну придатність моделі для прогнозування (його ідеальне значення — нуль).

У табл. 8.3 наведено результати оцінювання авторегресійної моделі АР(3). Усі коефіцієнти моделі статистично значущі.

Таблиця 8.3

Результати оцінювання моделі АР(3)

Метод оцінювання: метод найменших квадратів
Скоригована часова вибірка даних: 1996:04 2005:01
Всього спостережень після коригування крайніх значень: 106
Модель: $I = C(1) + C(2) \cdot I(-1) + C(3) \cdot I(-2) + C(4) \cdot I(-3)$

Параметр	Оцінка	Станд. похибка	*t*-статистика	Ймов.
C(1)	46,96556	9,330171	5,033729	0,0000
C(2)	0,613458	0,098563	6,223995	0,0000
C(3)	–0,157670	0,113613	–1,387787	0,1682
C(4)	0,077517	0,086493	0,896229	0,3722

R-квадрат	0,316655	Середнє залеж. змінної	100,6604
Скоригований *R*-квадрат	0,296556	Станд. відхил. залеж. змінної	1,372428
Станд. похибка регресії	1,151076	Інформ. критерій Акайке	3,156277
Сума квадратів похибок	135.1476	Критерій Шварца	3,256785
		Стат. Дарбіна-Уотсона	1,992389

Отже, можна записати таку модель для індексу споживчих цін (*isc*):

$$isc(k) = a_0 + a_1\, isc(k-1) + a_2\, isc(k-2) + a_3\, isc(k-3) + \varepsilon(k) =$$
$$= 46{,}9 + 0{,}61\, isc(k-1) - 0{,}15\, isc(k-2) + 0{,}08\, isc(k-3) + e(k).$$

Для цієї моделі спостерігається зменшення коефіцієнта детермінації від 0,415 до 0,317 (деяке погіршення); зменшення суми квадратів похибок від 141,99 до 135,148 і покращання значення статистики Дарбіна-Уотсона від 1,931 до 1,992 (похибки моделі можна вважати практично некорельованими). Отримані значення характеристик моделі:

$$R^2 = 0{,}317;\ J = \text{СКП} = 135{,}148;\ DW = 1{,}992.$$

Характеристики однокрокового прогнозу для даної моделі:

$$\text{СеКП} = 1{,}36;\ \text{САП} = 1{,}02;\ \text{САПП} = 1{,}01;\ U = 0{,}0068,$$

тобто модель загалом придатна для прогнозування (за коефіцієнтом Тейла, який наближається до ідеального нульового значення), а три інших показники свідчать про високу точність прогнозу. Необхідно зазначити, що показники якості прогнозу для моделей АР(1) і АР(3) є практично однаковими.

Розглянемо характеристики моделі вищого порядку. У табл. 8.4 наведені результати оцінювання моделі АР(7).

Таблиця 8.4

Результати оцінювання моделі АР(7) для ІСЦ

Метод оцінювання: метод найменших квадратів
Скоригована часова вибірка даних: 1996:08 2005:01
Всього використано спостережень після коригування крайніх значень: 102
Модель: $I = C(1) + C(2) \cdot I(-1) + C(3) \cdot I(-2) + C(4) \cdot I(-3) + C(5) \cdot I(-5) + C(6) \cdot I(-6) + C(7) \cdot I(-7)$

Параметр	Оцінка	Станд. похибка	*t*-статистика	Ймов.
C(1)	33,13446	12,79240	2,590168	0,0111
C(2)	0,603211	0,101625	5,935625	0,0000
C(3)	–0,184907	0,117804	–1,569623	0,1198
C(4)	0,125878	0,101891	1,235415	0,2197
C(5)	–0,031530	0,102954	–0,306254	0,7601
C(6)	0,093246	0,117601	0,792904	0,4298
C(7)	0,064850	0,089527	0,724365	0,4706

R-квадрат	0,346348	Середнє залеж. змінної	100,6667
Скоригований *R*-квадрат	0,305064	Станд. відхил. залеж. змінної	1,388306
Станд. похибка регресії	1,157330	Інформ. критерій Акайке	3,196268
Сума квадратів похибок	127,2443	Критерій Шварца	3,376413
		Стат. Дарбіна-Уотсона	1,811821

Порівнюючи моделі АР(3) і АР(7), можна сказати, що коефіцієнт детермінації збільшився від 0,317 до 0,346; сума квадратів похибок моделі зменшилась від 135,15 до 127,24, а статистика Дарбіна-Уотсона зменшилась від 1,992 до 1,811. Характеристики однокрокового прогнозу для АР(7):

$$\text{СеКП} = 1{,}36;\ \text{САП} = 1{,}012;\ \text{САПП} = 1{,}002;\ U = 0{,}0067,$$

тобто на 0,008 зменшились середня абсолютна похибка і середня абсолютна похибка у процентах.

Результати оцінювання моделі 12-го порядку наведено у табл. 8.5. Модель АР(12) має кращі характеристики, ніж попередні моделі:

$$R^2 = 0{,}435;\ J = \text{СКП} = 97{,}80;\ DW = 1{,}94.$$

Значно зменшилась сума квадратів похибок, підвищилось значення R^2, а значення статистики Дарбіна-Уотсона майже таке саме, як для моделі АР(3).

Таблиця 8.5

Результати оцінювання моделі АР(12) для ІСЦ

Метод оцінювання: метод найменших квадратів
Скоригована часова вибірка даних: 1997:01 2005:01
Всього використано спостережень після коригування крайніх значень: 97
Модель: I = C(1) + C(2) · I(–1) + C(3) · I(–2) + C(4) · I(–3) + C(5) · I(–6) + +C(6) · I(–7) + C(7) · I(–8) + C(8) · I(–9) + C(9) · I(–11) + C(10) · I(–12)

Параметр	Оцінка	Станд. похибка	*t*-статистика	Ймов.
C(1)	26,23923	14,28763	1,836500	0,0697
C(2)	0,666410	0,102198	6,520741	0,0000
C(3)	–0,247996	0,122488	–2,024654	0,0460
C(4)	0,179378	0,103248	1,737345	0,0859
C(5)	–0,038240	0,096715	–0,395392	0,6935
C(6)	0,016650	0,112406	0,148120	0,8826
C(7)	–0,037042	0,111025	–0,333639	0,7395
C(8)	0,030055	0,095744	0,313915	0,7543
C(9)	0,020445	0,093392	0,218911	0,8272
C(10)	0,149194	0,081977	1,819964	0,0722

R-квадрат	0,435002	Середнє залеж. змінної	100,6082
Скоригований *R*-квадрат	0,376553	Станд. відхил. залеж. змінної	1,342857
Станд. похибка регресії	1,060301	Інформ. критерій Акайке	3,052366
Сума квадратів похибок	97,80880	Критерій Шварца	3,317800
		Стат. Дарбіна-Уотсона	1,940130

Характеристики однокрокового прогнозу для цієї моделі такі:

$$\text{СеКП} = 1{,}337;\ \text{САП} = 1{,}02;\ \text{САПП} = 1{,}013;\ U = 0{,}0066.$$

Таким чином, характеристики однокрокового прогнозу також найкращі для моделі АР(7). Можна зробити висновок, що процес формування індексу оптових цін може бути описаний моделлю авторегресії АР(7) з високим ступенем адекватності. Ця модель забезпечує також отримання кращого однокрокового прогнозу.

Модель авторегресії для відхилень ІСЦ від середнього

Якщо з вихідних (фактичних) значень ряду ІСЦ відняти середнє ($isc(k)$ – 100,66), k = 1, ..., 109, то отримаємо ряд відхилень від середнього. Автокореляційна функція залишається фактично незмінною. Результати оцінювання моделі АР(1) наведені у табл. 8.6.

Таблиця 8.6

Результати оцінювання моделі АР(1) для відхилень ІСЦ

Метод оцінювання: метод найменших квадратів
Скоригована часова вибірка даних: 1996:02 2005:01
Всього використано спостережень після коригування крайніх значень: 109
Модель: IV = C(1) + C(2) · IV(–1)

Параметр	Оцінка	Станд. похибка	*t*-статистика	Ймов.
C(1)	–0,007477	0,111831	–0,066857	0,9468
C(2)	0,569826	0,065703	8,672790	0,0000

R-квадрат	0,415067	Середнє залеж. змінної	0,080741
Скоригований *R*-квадрат	0,409549	Станд. відхил. залеж. змінної	1,506189
Станд. похибка регресії	1,157368	Інформ. критерій Акайке	3,148519
Сума квадратів похибок	141,9871	Критерій Шварца	3,198188
		Стат. Дарбіна-Уотсона	1,931805

Три вибрані статистичні характеристики адекватності цієї моделі:

$$R^2 = 0{,}415;\ J = \text{СКП} = 141{,}99;\ DW = 1{,}931.$$

Характеристики якості однокрокового прогнозу:

$$\text{СеКП} = 1{,}36;\ \text{САП} = 1{,}02;\ \text{САПП} = 100{,}40;\ U = 0{,}662,$$

У порівнянні з моделлю АР(1) для повних значень (без відніманння середнього) середньоквадратична похибка і середня абсолютна похибка не змінились, але в 100 разів збільшилась САПП і коефіцієнт Тейла: від 0,0067 до 0,662. Таким чином, модель АР(1) для відхилень не придатна для прогнозування. Це можна пояснити тим, що відхилення мають різні знаки, що затрудняє побудову моделі.

Авторегресія 13-го порядку для відхилень ІСЦ від середнього наведена у табл. 8.7.

Три вибрані статистичні характеристики адекватності цієї моделі:

$$R^2 = 0{,}440;\ J = \text{СКП} = 95{,}69;\ DW = 1{,}965.$$

Таблиця 8.7

Результати оцінювання моделі АР(13) для відхилень ІСЦ

Метод оцінювання: метод найменших квадратів
Скоригована часова вибірка даних: 1997:02 2005:01
Всього використано спостережень після коригування крайніх значень: 96
Модель: IV = C(1) + C(2) · IV(–1) + C(3) · IV(–2) + C(4) · IV(–3) + C(5) × × IV(–5) + C(6) · IV(–6) + C(7) · IV(–7) + C(8) · IV(–8) + C(9) × × IV(–9) + C(10) · IV(–11) + C(11) · IV(–12) + C(12) · IV(–13)

Параметр	Оцінка	Станд. похибка	*t*-статистика	Ймов.
C(1)	–0,045818	0,110205	–0,415756	0,6786
C(2)	0,706976	0,107583	6,571429	0,0000
C(3)	–0,261301	0,123875	–2,109395	0,0379
C(4)	0,194879	0,105046	1,855172	0,0671
C(5)	–0,110121	0,107924	–1,020352	0,3105
C(6)	0,043907	0,120182	0,365334	0,7158
C(7)	0,002351	0,115561	0,020347	0,9838
C(8)	–0,023657	0,114078	–0,207375	0,8362
C(9)	0,033030	0,096947	0,340705	0,7342
C(10)	0,002612	0,096331	0,027112	0,9784
C(11)	0,187169	0,109248	1,713242	0,0904
C(12)	–0,074932	0,085883	–0,872494	0,3854

R-квадрат	0,440922	Середнє залеж. змінної	–0,066250
Скоригований *R*-квадрат	0,367709	Станд. відхил. залеж. змінної	1,342254
Станд. похибка регресії	1,067316	Інформ. критерій Акайке	3,084639
Сума квадратів похибок	95,68969	Критерій Шварца	3,405183
Логарифм правдоподіб.	–136,0627	Стат. Дарбіна-Уотсона	1,965193

Характеристики якості однокрокового прогнозу:

$$\text{СеКП} = 1{,}346;\ \text{САП} = 1{,}02;\ \text{САПП} = 109{,}09;\ U = 0{,}848.$$

Спостерігається погіршення характеристик прогнозу, особливо погіршилось значення САПП і коефіцієнта Тейла: від 0,662 для моделі АР(1) до 0,848 для моделі АР(13). За цим параметром модель не придатна для прогнозування. Коефіцієнти C(6) – C(10) є статистично незначущими, а тому їх можна вилучити з моделі.

Після вилучення цих коефіцієнтів із моделі та відповідних їм складових процесу отримаємо модель, параметри якої наведені у табл. 8.8.

Таблиця 8.8

Результати оцінювання моделі АР(13) для відхилень

Метод оцінювання: метод найменших квадратів
Скоригована часова вибірка даних: 1997:02 2005:01
Всього використано спостережень після коригування крайніх значень: 96
Модель: IV = C(1) + C(2) · IV(–1) + C(3) · IV(–2) + C(4) · IV(–3) +
+ C(5) · IV(–5) + C(6) · IV(–12) + C(7) · IV(–13)

Параметр	Оцінка	Станд. похибка	*t*-статистика	Ймов.
C(1)	–0,046129	0,107257	–0,430077	0,6682
C(2)	0,700062	0,103640	6,754728	0,0000
C(3)	–0,256660	0,120078	–2,137440	0,0353
C(4)	0,195599	0,101703	1,923247	0,0576
C(5)	–0,081877	0,083177	–0,984368	0,3276
C(6)	0,190417	0,089745	2,121760	0,0366
C(7)	–0,067455	0,081058	–0,832180	0,4075

R-квадрат	0,438765	Середнє залеж. змінної	–0,066250
Скоригований *R*-квадрат	0,400929	Станд. відхил. залеж. змінної	1,342254
Станд. похибка регресії	1,038900	Інформ. критерій Акайке	2,984323
Сума квадратів похибок	96,05881	Критерій Шварца	3,171307
Логарифм правдоподіб.	–136,2475	Стат. Дарбіна-Уотсона	1,959533

Вибрані статистичні характеристики адекватності цієї моделі:

$$R^2 = 0{,}439;\ J = \text{СКП} = 96{,}05;\ DW = 1{,}96.$$

Характеристики якості однокрокового прогнозу:

$$\text{СеКП} = 1{,}346;\ \text{САП} = 1{,}017;\ \text{САПП} = 108{,}2;\ U = 0{,}854.$$

Таким чином, після видалення з моделі несуттєвих коефіцієнтів її характеристики залишились практично незмінними. Можна зробити висновок, що побудовані авторегресійні моделі для відхилень ІСЦ від середнього не придатні для прогнозування. Це свідчить про те, що побудувати модель для процесу, який має різні знаки вимірів у різні моменти часу (тобто розвиток відбувається у двох квадрантах), складніше, ніж для процесу, який змінюється у межах одного квадранту.

Авторегресія з ковзним середнім для ІСЦ

Розглянемо можливість описання ІСЦ за допомогою моделі АРКС. Характеристики моделі АРКС(1,1) наведені у табл. 8.9.

Таблиця 8.9

Результати оцінювання моделі АРКС(1, 1) для ІСЦ

Метод оцінювання: метод найменших квадратів
Скоригована часова вибірка даних: 1996:02 2005:01
Всього використано спостережень після коригування крайніх значень: 108

Параметр	Оцінка	Станд. похибка	*t*-статистика	Ймов.
C	100,6529	0,263645	381,7742	0,0000
AR(1)	0,546826	0,093191	5,867793	0,0000
КС(1)	0,060696	0,133153	0,455838	0,6494

R-квадрат	0,415932	Середнє залеж. змінної	100,7407
Скоригований *R*-квадрат	0,404807	Станд. відхил. залеж. змінної	1,506189
Станд. похибка регресії	1,162006	Інформ. критерій Акайке	3,165558
Сума квадратів похибок	141,7772	Критерій Шварца	3,240062
Логарифм правдоподіб.	–167,9401	F-статистика	37,38676
Статист. Дарбіна-Уотсона	1,996202	Ймовірність (F-стат.)	0,000000

Інвертовані АР корені	.55
Інвертовані КС корені	–.06

Вибрані статистичні характеристики адекватності цієї моделі:

$$R^2 = 0{,}416;\ J = \text{СКП} = 141{,}78;\ DW = 1{,}996.$$

Характеристики якості однокрокового прогнозу:

$$\text{СеКП} = 1{,}362;\ \text{САП} = 1{,}016;\ \text{САПП} = 1{,}005;\ U = 0{,}0067.$$

Моделі АР(1) і АРКС(1,1) мають практично однакові характеристики адекватності та якості однокрокового прогнозу, а тому перевагу (при виборі з цих двох моделей) можна надати моделі АР(1), яка є простішою. Нижче наведена порівняльна таблиця для всіх побудованих моделей.

Врахування впливу на ІСЦ агрегату М3

Коефіцієнт кореляції між ІСЦ та агрегатом М3 від'ємний: –0,249, тобто формальний взаємозв'язок між цими змінними незначний, але цікаво розглянути вплив М3 на ІСЦ за допомогою моделі. Врахування регресора може покращити деякі характеристики моделі, а також врахувати причинний зв'язок між вибраними змінними. Характеристики змішаної моделі: авторегресія АР(1) + парна регресія наведені у табл. 8.10.

Таблиця 8.10

Результати оцінювання моделі ІСЦ: АР(1) + регресор М3

Метод оцінювання: метод найменших квадратів
Скоригована часова вибірка даних: 1996:02 2005:01
Всього використано спостережень після коригування крайніх значень: 108
Модель: I = C(1) + C(2) · I(–1) + C(3) · M

Параметр	Оцінка	Станд. похибка	*t*-статистика	Ймов.
C(1)	44,97869	6,920057	6,499757	0,0000
C(2)	0,554264	0,068261	8,119764	0,0000
C(3)	–2,82E-06	3,30E-06	–0,854463	0,3948

R-квадрат	0,419106	Середнє залеж. змінної	100,7407
Скоригований *R*-квадрат	0,408041	Станд. відхил. залеж. змінної	1,506189
Станд. похибка регресії	1,158844	Інформ. критерій Акайке	3,160108
Сума квадратів похибок	141,0066	Критерій Шварца	3,234612
Логарифм правдоподіб.	–167,6458	Стат. Дарбіна-Уотсона	1,919823

Спостерігається незначне покращання характеристик моделі та якості прогнозу у порівнянні з АР(1), але коефіцієнт при М3 дуже малий (–2,82E-06). Однак формально він є значимим. Таким чином, на розглянутому часовому інтервалі вплив агрегату М3 на індекс споживчих цін незначний і ним можна знехтувати.

Характеристики однокрокового прогнозу:

$$\text{СеКП} = 1{,}340;\ \text{САП} = 1{,}004;\ \text{САПП} = 0{,}994;\ U = 0{,}0066.$$

Можна припустити, що в обігу було недостатньо грошової маси для того, щоб її вплив на ІСЦ був істотним. З іншого боку, недостатній обсяг грошової маси у національній валюті компенсувався (і продовжує компенсуватись) "твердою" іноземною валютою, зокрема доларами США і, деякою мірою, євро. Таким чином, слабка українська економіка позитивно впливає на курс долара США, завдяки фактичному введенню його у частковий обіг на відносно великій території України. Встановити фактичне співвідношення між обсягами національної та іноземної валют в обороті можливо, але для цього необхідно отримати додаткові статистичні дані і виконати спеціальне дослідження. Зокрема, необхідно мати дані стосовно обсягів реалізації торговельних операцій у валюті підприємствами всіх форм власності. Очевидно, що отримати такі дані для тіньового обороту досить складно.

Визначення впливу ВВП на ІСЦ

Як було показано на початку цього параграфа, коефіцієнт кореляції між ІСЦ та ВВП становить –0,304, тобто формально це невелике значення. Також логічно припустити, що зростання ВВП має приводити до зменшення ІСЦ (про це свідчить також знак коефіцієнта кореляції між цими змінними).

Характеристики змішаної моделі: авторегресія АР(1) + парна регресія наведені у табл. 8.11.

Таблиця 8.11

Результати оцінювання змішаної регресії для ІСЦ і ВВП

Метод оцінювання: метод найменших квадратів
Скоригована часова вибірка даних: 1996:02 2005:01
Всього використано спостережень після коригування крайніх значень: 108
Модель: I = C(1) + C(2) · I(–1) + C(3) · V

Параметр	Оцінка	Станд. похибка	*t*-статистика	Ймов.
C(1)	45,53125	7,156626	6,362112	0,0000
C(2)	0,549611	0,070124	7,837741	0,0000
C(3)	–1,37E-05	1,64E-05	–0,833627	0,4064

R-квадрат	0,418913	Середнє залеж. змінної	100,7407
Скоригований *R*-квадрат	0,407844	Станд. відхил. залеж. змінної	1,506189
Станд. похибка регресії	1,159037	Інформ. критерій Акайке	3,160441
Сума квадратів похибок	141,0536	Критерій Шварца	3,234945
Логарифм правдоподіб.	–167,6638	Стат. Дарбіна-Уотсона	1,916289

Отримано таке рівняння:

$$isc(k) = 45{,}53 + 0{,}55\ isc(k-1) - (1{,}37E{-}05)\ \nu\nu p(k) + \varepsilon(k)$$

з характеристиками:

$$R^2 = 0{,}419;\ J = \text{СКП} = 141{,}05;\ DW = 1{,}92.$$

Характеристики якості однокрокового прогнозу:

$$\text{СеКП} = 1{,}335;\ \text{САП} = 1{,}004;\ \text{САПП} = 0{,}993;\ U = 0{,}0066.$$

Коефіцієнт при змінній ВВП невеликий і знаходиться на межі статистичної значимості, тобто вплив ВВП на ІСЦ незначний. Причиною такого незначного впливу може бути некоректний розподіл ВВП, що веде до того, що більшість населення проживає на межі або нижче межі бідності. Одночасно інша (менша) частина населення привлас-

нює більшість благ і користується ними, але очевидно, що це користування не веде до позитивного впливу на ІСЦ. Звідси маємо, що ВВП зростає значними темпами, а добробут населення помітно від нього відстає. Можна зробити висновок, що на часовому періоді, який розглядається у цьому прикладі, відбувався несправедливий розподіл суспільних благ, що проявилось у незначному впливі ВВП на споживчі ціни для більшості населення України. Побудована математична модель є формальним (об'єктивним) підтвердженням відомого факту нерівномірного розподілу матеріальних благ, вироблених в Україні.

У табл. 8.12 зведені характеристики математичних моделей, побудованих для індексу споживчих цін, і характеристики однокрокових прогнозів, обчислених на основі цих моделей. Ця таблиця дає можливість оперативно порівняти результати моделювання та прогнозування, отримані за допомогою методики Бокса-Дженкінса, а також встановити можливість практичного використання цих результатів.

Таблиця 8.12

Результати моделювання і однокрокового прогнозування індексу оптових цін

Тип моделі	Характеристики моделі			Характеристики однокрокового прогнозу			
	R^2	$\sum e^2(k)$	DW	СеКП	САП	САПП	Коефіцієнт Тейла
АР(1)	0,415	141,99	1,931	1,360	1,020	1,008	0,0067
АР(3)	0,317	135,148	1,992	1,360	1,020	1,011	0,0068
АР(7)	0,346	127,244	1,811	1,360	1,012	1,002	0,0067
АР(12)	0,435	97,80	1,941	1,337	1,020	1,013	0,0066
АРКС(1,1)	0,416	141,78	1,996	1,362	1,016	1,005	0,0067
АР(1)+М3	0,419	141,007	1,919	1,340	1,004	0,994	0,0066
АР(1)+ВВП	0,419	141,054	1,916	1,335	1,004	0,993	0,0066
Моделі для відхилень ІСЦ від середнього							
АР(1)	0,415	141,99	1,931	1,360	1,020	100,40	0,662
АР(13)	0,440	95,69	1,965	1,346	1,020	109,09	0,848

Результати моделювання, наведені у табл. 8.12, свідчать про те, що практично всі моделі, побудовані для ІСЦ (перші сім моделей), є придатними для прогнозування, оскільки коефіцієнт Тейла вимірюється тисячними частками. Характеристики однокрокового прогнозу для

моделей різного порядку відрізняються несуттєво. Найкращі характеристики щодо прогнозування має модель АР(1)+ВВП. Для моделі АР(12) отримано найменше значення суми квадратів похибок моделі, але характеристики прогнозів, отриманих за цією моделлю, не кращі від інших.

Дві моделі, які побудовані для відхилень ІСЦ від середнього, малопридатні для прогнозування. Про це свідчать високі значення коефіцієнта Тейла та середньої абсолютної похибки у процентах. Цей факт можна пояснити тим, що описати за допомогою АРКС процес, динаміка якого спостерігається у двох квадрантах, складніше, ніж процес, який спостерігається в одному квадранті. Тому для прогнозування відхилень від середнього необхідно знайти іншу структуру математичної моделі, придатну для опису різнознакових величин часового ряду.

8.8. Контрольні питання і вправи

1. Назвіть етапи побудови математичних моделей за методикою Бокса-Дженкінса. Що забезпечує коректне використання цієї методики на практиці?
2. Яка мета аналізу функціонування процесу? Які елементи структури моделі можна встановити за допомогою попереднього аналізу процесу на основі відомої інформації?
3. Яка мета попередньої обробки даних? Назвіть основні операції, які виконують у процесі попередньої обробки даних. До чого веде визначення значень змінних у великому числовому діапазоні?
4. Які два основних типи нелінійностей зустрічаються в аналізі часових рядів? Який з них ускладнює процедуру оцінювання параметрів моделі? Поясніть на прикладах.
5. Яким чином можна встановити наявність нелінійностей у процесі? В яких випадках нелінійний процес можна описати лінійною моделлю?
6. Який метод дає можливість автоматизувати процес визначення та врахування нелінійностей процесу?
7. Яку інформацію можна отримати на основі візуального аналізу даних? Як можна нею скористатись?
8. Яким чином можна знайти оцінку порядку авторегресійної частини моделі?

9. У чому полягає відмінність між автокореляційною та частковою автокореляційною функціями процесу? Чи існує необхідність розрахунку обох функцій у процесі аналізу даних?
10. На чому ґрунтується відбір незалежних змінних (регресорів, екзогенних змінних) для включення у праву частину математичної моделі?
11. Назвіть три умови коректного застосування методу найменших квадратів до оцінювання параметрів математичної моделі. Як можна перевірити виконання цих умов?
12. Для чого призначена статистика Люнга-Бокса і як вона обчислюється? Скористайтесь наявним пакетом програм для статистичного аналізу даних для обчислення цього статистичного параметра і поясніть його значущість (чи незначущість) на прикладі.
13. У чому полягає принципова різниця між методом найменших квадратів (МНК), призначеним для оцінювання лінійних моделей, та МНК для оцінювання моделей, нелінійних відносно параметрів? Які критерії якості мінімізують ці методи?
14. Виведіть формулу МНК для поліноміальної моделі.
15. Сформуйте матрицю вимірів для математичної моделі такого вигляду:

$$y(k) = a_0 + a_1 y(k-1) + b_1 x(k) + b_2 z(k) + \varepsilon(k).$$

16. Про що свідчить корельованість похибок моделі між собою? За допомогою якого статистичного параметра якості моделі можна визначити ступінь корельованості похибок? Яке значення приймає ця статистика в ідеальному випадку?
17. Що означає значущість оцінки параметра (коефіцієнта) моделі у статистичному смислі? За допомогою якої статистики можна встановити значущість оцінки параметра моделі?
18. Що означають помилки першого і другого роду при перевірці статистичних гіпотез?
19. Поясніть фізичну сутність коефіцієнта (множинної) детермінації. Яке його ідеальне значення? Чому R^2 коригують із врахуванням кількості ступенів вільності?
20. У чому полягає різниця між критеріями Акайке та Байєса-Шварца? Чи можна обмежитись використанням тільки одного з цих критеріїв?

21. Яка мета розрахунку статистики Фішера? Наведіть послідовність розрахунку та аналізу цієї статистики.
22. Сформулюйте правила перевірки гіпотез для статистики Стьюдента і статистики Фішера. Яким чином можна визначити критичні (порогові) значення цих статистик?
23. Поясніть фізичну сутність умови достатнього збудження процесу. До чого призводить невиконання цієї умови?
24. Яким чином досягають достатнього збудження процесів (об'єктів) на практиці? Які сигнали використовують для досягнення цієї мети? Які властивості цих сигналів використовують у даному випадку?

Розділ 9

ЗАСТОСУВАННЯ РІЗНИЦЕВИХ РІВНЯНЬ У РЕГРЕСІЙНОМУ МОДЕЛЮВАННІ

9.1. Загальні відомості про різницеві рівняння

При використанні дискретних рівнянь (моделей) незалежну змінну, час t, замінюють дискретним часом, тобто покладають $t = kT_s$, де T_s — період дискретизації вимірів, який у технічних системах набуває значення від десятків мікросекунд до десятків секунд і навіть хвилин, а при моделюванні фінансово-економічних процесів — від кількох хвилин до одного року залежно від того, які статистичні дані можна отримати. У моделях, які описують процеси у дискретному часі, період дискретизації вимірів, як правило, нормують до одиниці і незалежною змінною залишається k (дискретний час), яка набуває цілих значень від 0 до ∞. При цьому для кожної прикладної задачі дискретна одиниця часу має відповідне фактичне значення.

Завдяки простоті структури та наявності надійних методів оцінювання параметрів, різницеві рівняння (РР) знайшли надзвичайно широке застосування при створенні моделей процесів у технічних системах, економіці та фінансах, екології, біології та інших прикладних і наукових галузях. Простим прикладом різницевого рівняння є стохастичне рівняння авторегресії першого порядку з одиничним коефіцієнтом (окремий випадок) при затриманому в часі значенні основної змінної:

$$y(k) = y(k-1) + \varepsilon(k), \qquad (9.1.1)$$

де $y(k)$ — основна змінна; $\varepsilon(k)$ — випадкова величина, яка відображає вплив різноманітних невимірюваних факторів на основну змінну. У першу чергу, це випадкові збурення, що діють на процес. Частіше за все припускають, що випадкова величина має нормальний розподіл: $\{\varepsilon(k)\} \sim N_n(0, \sigma_\varepsilon^2)$, тобто це некорельований процес із нульовим середнім та скінченною дисперсією σ_ε^2. Наприклад, якщо $y(k)$ — ціни на деякі біржові акції в k-й день, то $\varepsilon(k)$ характеризує коливання ціни під впливом багатьох випадкових факторів, які неможливо ввести у

модель, оскільки неможливо отримати їх виміри, або ж ці впливи носять якісний характер. Як правило, у наведеній моделі процес $\{\varepsilon(k)\}$ має такі обмеження: $E[\varepsilon(k)] = 0$, $E\left[\varepsilon(k)\varepsilon(l)\right] = \begin{cases}\sigma_\varepsilon^2, & k = l,\\ 0, & k \neq l.\end{cases}$ Докладніше роль випадкової змінної буде розглядатись окремо у кожному конкретному випадку моделювання.

За допомогою рівняння (9.1.1) описують, наприклад, ціну акції на біржі в момент часу, що відповідає аргументу k. Його називають ще рівнянням, яке описує *процес випадкового кроку* (випадкове блукання або *random walk*). Таку назву воно отримало з тієї причини, що приріст значення основної змінної визначається, фактично, випадковим процесом. Воно може бути записане також у формі першої різниці

$$\Delta y(k) = \varepsilon(k),$$

де $\Delta y(k) = y(k) - y(k-1)$. Саме можливості застосування різниць перших і вищих порядків до опису часових рядів зумовило використання назви *різницеві рівняння*.

Для рівняння випадкового кроку можна записати таке однорідне рівняння:

$$y(k) - y(k-1) = 0,$$

яке має характеристичне рівняння: $\alpha - 1 = 0$ (як отримати характеристичне рівняння у загальному випадку, розглянемо нижче у цьому розділі). Таким чином, дане характеристичне рівняння має один корінь $\alpha = 1$. Якщо характеристичне рівняння має хоча б один одиничний корінь, то кажуть, що відповідне йому різницеве рівняння описує *процес з одиничним коренем* (нестаціонарний процес з трендом). Нижче покажемо, що процеси з одиничними коренями — це *процеси з трендами* або *інтегровані* процеси. У подальших викладках під трендом будемо розуміти загальний довгостроковий напрям розвитку процесу. Фактично він співпадає з поточним середнім значенням.

Більш загальною формою різницевого рівняння (9.1.1), тобто процесу авторегресії першого порядку — АР(1) є наступна:

$$y(k) = a_0 + a_1 y(k-1) + \varepsilon(k), \qquad (9.1.2)$$

але для того, щоб воно відповідало процесу випадкового кроку, необхідно покласти: $a_0 = 0$, $a_1 = 1$; інакше це рівняння вже не буде відпо-

відати своєму означенню. Зазначимо, що порядок авторегресійної частини різницевого рівняння визначається числом попередніх (затриманих) вимірів залежної змінної, які використовуються у правій частині для пояснення її зміни у часі.

Різницеві рівняння, в правій частині яких наявні попередні (затримані) виміри основної змінної, називають *авторегресійними* (АР), тобто регресія змінної на саму себе. Рівняння авторегресії *n*-го порядку має вигляд

$$y(k) = a_0 + \sum_{i=1}^{n} a_i y(k-i) + \varepsilon(k). \qquad (9.1.3)$$

Якщо на процес, що моделюється, впливає деяка вхідна змінна, то вона записується у правій частині:

$$y(k) = a_0 + \sum_{i=1}^{n} a_i y(k-i) + \sum_{j=1}^{q} b_j x(k-j) + \varepsilon(k). \qquad (9.1.4)$$

Якщо $x(k)$ випадковий процес, то таке рівняння називають авторегресією з ковзним середнім (АРКС). Формально у випадку ковзного середнього необхідно також виконати умову $\sum_{j=1}^{q} b_j = 1$.

Характеристичне рівняння, записане для (9.1.3), може мати одиничні корені, тобто один або більше коренів характеристичного рівняння можуть набувати значення "1". Такі процеси називають процесами авторегресії з інтегрованим ковзним середнім — АРІКС(n, d, q), де d — кількість одиничних коренів характеристичного рівняння. Процеси цього класу *нестаціонарні* — вони мають тренд, порядок якого визначається кількістю одиничних коренів. Якщо $d = 1$, то тренд лінійний; якщо $d = 2$, то тренд квадратичний і т. д. Таким чином, *процеси з трендом, інтегровані процеси і процеси з одиничними коренями* — це різні назви нестаціонарних процесів, що містять тренд.

Застосування перших різниць та різниць вищих порядків

Разом з основною змінною для опису процесів використовують *перші* та *другі різниці*, наприклад,

$$\Delta y(k) = y(k) - y(k-1),$$
$$\Delta y(k+1) = y(k+1) - y(k),$$
$$\Delta y(k+2) = y(k+2) - y(k+1).$$

Наведені перші різниці відображають швидкість зміни основної змінної, що відповідає першій похідній для рівнянь, які описують модельований процес у неперервному часі, тобто для диференціальних рівнянь.

Знаходження перших різниць веде до видалення лінійного тренду з процесу. Наприклад, нехай тренд описується лінійним рівнянням

$$y(k) = a_0 + a_1 k.$$

Перша різниця для цього процесу:

$$\Delta y(k) = y(k) - y(k-1) = a_0 + a_1 k - a_0 - a_1 (k-1) = a_1 .$$

Тобто після дискретного диференціювання залишилась константа.

Другі різниці відображають швидкість зміни у часі перших різниць (тобто прискорення) і записуються так:

$$\begin{aligned}\Delta^2 y(k) &= \Delta(\Delta\, y(k)) = \Delta[y(k) - y(k-1)] = \\ &= [y(k) - y(k-1)] - [y(k-1) - y(k-2)] = \\ &= y(k) - 2y(k-1) + y(k-2), \\ \Delta^2 y(k+1) &= \Delta(\Delta\, y(k+1)) = y(k+1) - 2y(k) + y(k-1).\end{aligned}$$

Останній вираз називають різницевою схемою знаходження другої похідної. На практиці другі різниці використовують досить рідко, а різниці вищого порядку не використовуються. Застосування других різниць до процесу викликає видалення квадратичного тренду, що можна легко проілюструвати на прикладі опису тренду поліномом другого порядку від k.

Наприклад, нехай

$$y(k) = a_0 + a_1 k + a_2 k^2 .$$

Перша різниця:

$$\begin{aligned}\Delta y(k) &= a_0 + a_1 k + a_2 k^2 - [a_0 + a_1 (k-1) + a_2 (k-1)^2] = \\ &= a_0 + a_1 k + a_2 k^2 - a_0 - a_1 k + a_1 - a_2 k^2 + 2a_2 k - a_2 = \\ &= a_1 + 2a_2 k - a_2 .\end{aligned}$$

Друга різниця:

$$\Delta^2 y(k) = a_1 + 2a_2 k - a_2 - [a_1 + 2a_2 (k-1) - a_2] = 2a_2 .$$

Редукована та структурована форми різницевого рівняння. Для ілюстрації рівнянь такого типу розглянемо модель Самуельсона для валового внутрішнього продукту (ВВП), споживання та інвестицій [7, 8]:

$$y(k) = c(k) + I(k); \tag{9.1.5}$$

$$c(k) = \alpha\, y(k-1) + \varepsilon_C(k),\ 0 < \alpha < 1; \tag{9.1.6}$$

$$I(k) = \beta\,[c(k) - c\,(k-1)] + \varepsilon_I(k),\ \beta > 0, \tag{9.1.7}$$

де $y(k)$, $c(k)$, $I(k)$ — ВВП, споживання та інвестиції в момент k відповідно. У цій моделі $y(k)$, $c(k)$, $I(k)$ — ендогенні змінні, а $\varepsilon_C(k)$, $\varepsilon_I(k)$ — стохастичні змінні з нульовим середнім, які характеризують вплив збурень на споживання та інвестиції, тобто вплив тих змінних, які не введені у модель. Третє рівняння ілюструє принцип прискорення (акселерації), оскільки в нього введена швидкість зміни споживання $[c(k) - c\,(k-1)]$.

Рівняння (9.1.7) називають **структурованим рівнянням** [8], оскільки в ньому ендогенна змінна $I(k)$ залежить від поточного (а не попереднього) значення іншої ендогенної змінної $c(k)$.

Редукованою формою різницевого рівняння називають таку, в якій основна змінна (у лівій частині) залежить від попередніх вимірів цієї самої змінної, попередніх вимірів інших ендогенних змінних, поточних та попередніх вимірів екзогенних змінних та збурення.

Таким чином, рівняння для споживання вже представлене у редукованій формі. Рівняння для інвестицій не відповідає редукованій формі, оскільки у правій частині наявне поточне значення споживання. Для того щоб отримати редуковану форму для інвестицій, підставимо (9.1.6) у рівняння для інвестицій і отримаємо:

$$\begin{aligned} I(k) &= \beta[\alpha\, y(k-1) + \varepsilon_C(k) - c\,(k-1)] + \varepsilon_I(k) = \\ &= \alpha\beta\, y(k-1) - \beta c(k-1) + \beta\varepsilon_C(k) + \varepsilon_I(k). \end{aligned}$$

Необхідно зазначити, що отримана редукована форма для інвестицій не єдина. Якщо дискретний час k зменшити на одиницю, то рівняння (9.1.6) можна переписати у вигляді $c(k-1) = \alpha y(k-2) + \varepsilon_C(k-1)$. За допомогою отриманого рівняння запишемо редуковану форму для інвестицій у вигляді

$$\begin{aligned} I(k) &= \alpha\beta\, y(k-1) - \beta[\alpha y(k-2) + \varepsilon_C(k-1)] + \beta\varepsilon_C(k) + \varepsilon_I(k) = \\ &= \alpha\beta\,[y(k-1) - y(k-2)] + \beta\,[\varepsilon_C(k) - \varepsilon_C(k-1)] + \varepsilon_I(k). \end{aligned} \tag{9.1.8}$$

За аналогією можна отримати редуковану форму для ВВП, якщо підставити (9.1.6) і (9.1.8) у рівняння (9.1.5):

$$y(k)=\alpha y(k-1)+\varepsilon_C(k)+\alpha\beta\left[y(k-1)-y(k-2)\right]+$$
$$+\beta\left[\varepsilon_C(k)-\varepsilon_C(k-1)\right]+\varepsilon_I(k)= \qquad (9.1.9)$$
$$=\alpha(1+\beta)y(k-1)-\alpha\beta y(k-2)+(1+\beta)\varepsilon_C(k)+\varepsilon_I(k)-\beta\varepsilon_C(k-1).$$

Рівняння (9.1.9) називають редукованою формою стосовно однієї змінної, оскільки у правій частині наявні тільки попередні виміри основної змінної та збурення. Така форма зручна для прогнозування, оскільки для цього необхідно мати виміри тільки однієї змінної. Нижче будуть розглянуті підходи до прогнозування, які ґрунтуються на розв'язках різницевих рівнянь. Вони дають можливість знаходити прогноз на довільну кількість кроків уперед.

Необхідно також зазначити, що різницеві рівняння, які описують фінансово-економічні процеси перехідного періоду, можна використати для створення систем оптимального управління цими процесами при розробці інформаційних систем підтримки прийняття оптимальних (чи субоптимальних) рішень, при створенні експертних систем.

Позитивною стороною використання РР є те, що їх параметри легко оновлюються при надходженні нових вимірів, також можна досить просто змінити порядок і структуру математичної моделі. При цьому під **структурою моделі** будемо розуміти *кількість рівнянь* моделі, *їх порядок, наявність нелінійностей* та їхній тип, *наявність запізнення* (*лагу*) відносно входу та його оцінка, а також *збурення процесу* і його тип. У поняття структури включають також обмеження на параметри й змінні моделі.

9.2. Ітераційний метод знаходження розв'язків різницевих рівнянь

Поняття розв'язку різницевого рівняння

Розв'язки різницевих рівнянь (так само, як і диференціальних) необхідні для формування на їх основі функцій прогнозування, а також для порівняння характеристик різних процесів між собою.

Розв'язок різницевого рівняння є *функцією часу, вхідної змінної, початкових умов, параметрів та збурення*. Існують різні підходи до знаходження розв'язків різницевих рівнянь.

Як простий приклад знаходження розв'язку розглянемо процес, який описується першою різницею

$$\Delta y(k) = 2{,}5, \tag{9.2.1}$$

тобто приріст змінної за один період дискретизації становить 2,5 одиниці у вибраному масштабі. Замінимо першу різницю відповідними значеннями змінної:

$$y(k) = y(k-1) + 2{,}5,$$

і задамо початкову умову $y(0) = y_0 = 0{,}25$. Тепер запишемо значення $y(k)$ для кількох моментів часу, починаючи з $k = 1$:

$$y(1) = y_0 + 2{,}5;$$
$$y(2) = y(1) + 2{,}5 = y_0 + 2{,}5 + 2{,}5;$$
$$y(3) = y(2) + 2{,}5 = y_0 + 2{,}5 + 2{,}5 + 2{,}5 = y_0 + 3 \cdot 2{,}5,$$

і для довільного k:

$$y(k) = y_0 + 2{,}5 \cdot k = 2{,}5k + 0{,}25. \tag{9.2.2}$$

Отриманий розв'язок для (9.2.1) справедливий для будь-якого значення k. Для того щоб це довести, підставимо отриманий розв'язок у рівняння (9.2.1):

$$2{,}5k + 0{,}25 = 2{,}5(k-1) + 0{,}25 + 2{,}5.$$

або

$$0{,}25 = 0{,}25.$$

Таким чином, знайдений розв'язок справедливий для $\forall k$, $k = 0, 1, 2, \ldots$. Він є рівнянням прямої, яка перетинає вісь ординат у точці $y_0 = 0{,}25$. У зв'язку з цим, значення $y_0 = a_0$ називають ще *перетином*.

Ітераційний підхід до знаходження розв'язку рівняння першого порядку

Розглянемо знаходження розв'язку РР першого порядку ітераційним методом. Рівняння першого порядку з ненульовим перетином має вигляд

$$y(k) = a_0 + a_1 y(k-1) + \varepsilon(k), \tag{9.2.3}$$

з початковою умовою y_0. Для моментів часу $k = 1, 2, 3$ можна записати, що

$$y(1) = a_0 + a_1 y(0) + \varepsilon(1);$$

$$y(2) = a_0 + a_1 y(1) + \varepsilon(2) = a_0 + a_1\left[a_0 + a_1 y(0) + \varepsilon(1)\right] + \varepsilon(2) =$$
$$= a_0 + a_0 a_1 + (a_1)^2 y(0) + a_1\varepsilon(1) + \varepsilon(2);$$
$$y(3) = a_0 + a_1 y(2) + \varepsilon(3) =$$
$$= a_0\left[1 + a_1 + (a_1)^2\right] + (a_1)^3 y(0) + a_1^2\varepsilon(1) + a_1\varepsilon(2) + \varepsilon(3).$$

Тепер можна записати розв'язок рівняння (9.2.3) для довільного моменту k:

$$y(k) = a_0 \sum_{i=0}^{k-1} a_1^i + a_1^k y_0 + \sum_{i=0}^{k-1} a_1^i \varepsilon(k-i). \qquad (9.2.4)$$

Розв'язок (9.2.4) також можна отримати, якщо розпочати ітерації у зворотному часі, тобто

$$y(k) = a_0 + a_1\left[a_0 + a_1 y(k-2) + \varepsilon(k-1)\right] + \varepsilon(k) =$$
$$= a_0 \cdot (1 + a_1) + a_1^2\left[a_0 + a_1 y(k-3) + \varepsilon(k-2)\right] + \varepsilon(k) + a_1\varepsilon(k-1).$$

Продовження ітераційного процесу приведе до розв'язку у формі (9.2.4).

Загальний розв'язок (9.2.4) складається з двох частин. Перша частина залежить від початкової умови $a_1^k y_0$, а друга — від наявності у рівнянні перетину a_0 і випадкової змінної $\varepsilon(k)$. Відповідно, першу частину називають однорідним розв'язком або розв'язком однорідного рівняння і позначають так:

$$y^h(k) = a_1^k y_0 = A\alpha^k,$$

де індекс h — *homogeneous* (однорідний). У більш загальному вигляді однорідний розв'язок записують так: $y^h(k) = A\alpha^k$, де A — довільна константа, яка визначається за допомогою початкових умов; α — корінь характеристичного рівняння, який дорівнює параметру a_0 у випадку рівняння АР(1).

Частинний розв'язок неоднорідного рівняння позначають так:

$$y^p(k) = a_0 \sum_{i=0}^{k-1} a_1^i + \sum_{i=0}^{k-1} a_1^i \varepsilon(k-i),$$

де верхній індекс p — *partial* (частинний). Загальний розв'язок є сумою загального розв'язку однорідного рівняння і частинного розв'язку неоднорідного рівняння, тобто

$$y(k) = y^h(k) + y^p(k).$$

Якщо початкова умова не задана, то замість y_0 у (9.2.4) можна підставити $y(0) = a_0 + a_1 y(-1) + \varepsilon(0)$ (випадок можливий, якщо існують дані про процес до моменту $k = 0$):

$$y(k) = a_0 \sum_{i=0}^{k-1} a_1^i + a_1^k \left[a_0 + a_1 y(-1) + \varepsilon(0) \right] + \sum_{i=0}^{k-1} a_1^i \varepsilon(k-i) = \\ = a_0 \sum_{i=0}^{k} a_1^i + \sum_{i=0}^{k} a_1^i \varepsilon(k-i) + a_1^{k+1} y(-1). \tag{9.2.5}$$

Альтернативний підхід до знаходження розв'язків

Існує ***альтернативний підхід*** до *знаходження розв'язків* різницевих рівнянь, який полягає у знаходженні окремо розв'язку однорідного рівняння та часткового розв'язку. Наприклад, розв'язок однорідного рівняння виду

$$y(k) = a_1 y(k-1) \quad \text{або} \quad y(k) = a_1 y(k-1) + a_2 y(k-2)$$

називають *однорідним розв'язком*. Тривіальним розв'язком однорідного рівняння є $y(1) = y(2) = \ldots = 0$. Нехай $a_0 = 0$ і всі значення випадкової змінної $\{\varepsilon(k)\} = 0$, то з (9.2.4) випливає, що розв'язком рівняння буде також $y(k) = a_1^k y_0$. Можна показати, що $y(k) = A a_1^k$, де A — довільна константа також буде розв'язком однорідного рівняння. Підставимо останнє значення в однорідне рівняння і отримаємо тотожність:

$$A(a_1)^k = a_1 A (a_1)^{k-1}. \tag{9.2.7}$$

Процес знаходження часткового розв'язку буде розглянуто нижче, а зараз наведемо загальну методику розв'язання різницевого рівняння, яка логічно випливає з отриманого ітераційним методом повного розв'язку рівняння АР(1).

Загальна методика знаходження розв'язку різницевого рівняння

Користуючись результатами знайденого розв'язку рівняння першого порядку, отриманими вище, можна сформулювати загальну методику знаходження розв'язків різницевих рівнянь типу АР та АРКС.

Крок 1. Для моделі АРКС (n, q) записати однорідне рівняння та знайти всі його однорідні розв'язки (їх кількість дорівнює n).

Крок 2. Знайти всі складові часткового розв'язку, зумовлені константами, детермінованими та випадковими функціями у правій частині рівняння.

Крок 3. Записати загальний розв'язок у вигляді суми однорідного та частинного розв'язків.

Крок 4. За допомогою початкових умов знайти значення довільних констант.

Розглянемо приклад знаходження повного розв'язку рівняння авторегресії другого порядку.

Приклад 9.1. Знайдемо загальний розв'язок РР другого порядку

$$y(k) = 2{,}1 + 0{,}8y(k-1) - 0{,}15y(k-2).$$

Запишемо однорідне рівняння:

$$y(k) - 0{,}8y(k-1) + 0{,}15y(k-2) = 0.$$

Оскільки це рівняння другого порядку, то воно має два однорідних розв'язки, які знаходять за допомогою розв'язку характеристичного рівняння, записаного для однорідного (більш докладно ця методика буде розглянута нижче). Характеристичне рівняння має вигляд

$$\lambda^2 - 0{,}8\lambda + 0{,}15 = 0.$$

Це квадратне рівняння має два корені: $\lambda_1 = 0{,}5$; $\lambda_2 = 0{,}3$. Таким чином, два однорідних розв'язки можна записати так: $y_1^h(k) = A_1(0{,}5)^k$; $y_2^h(k) = A_2(0{,}3)^k$.

Правильність знаходження однорідних розв'язків можна перевірити шляхом підстановки в однорідне рівняння. Підставимо в рівняння $y_2^h(k) = A_2(0{,}3)^k$. Тоді

$$A_2(0{,}3)^k - 0{,}8\,A_2(0{,}3)^{k-1} + 0{,}15\,A_2(0{,}3)^{k-2} = 0.$$

Якщо розділити всі члени на $(0{,}3)^{k-2}$, то отримаємо

$$(0{,}3)^2 - 0{,}8(0{,}3) + 0{,}15 = 0{,}09 - 0{,}24 + 0{,}15 = 0.$$

Другий однорідний розв'язок перевіряється аналогічно.

Частинний розв'язок для детермінованого збурення шукаємо у вигляді константи y^p:

$$y^p = 2{,}1 + 0{,}8\,y^p - 0{,}15\,y^p,$$

тобто під впливом константи на виході також буде константа y^{p} для всіх моментів часу. Звідси $y^{p} = 6{,}0$, де верхній індекс p = *partial* (частинний). Загальна методика визначення частинного розв'язку буде наведена нижче. Тепер об'єднаємо однорідний і частинний розв'язки:

$$y(k) = 6{,}0 + A_1(0{,}5)^k + A_2(0{,}3)^k,$$

де A_1, A_2 — довільні константи, які знайдемо за допомогою початкових умов. Оскільки є дві невідомі константи, то необхідно мати два значення для початкових умов. Нехай $y(0) = 1{,}0$; $y(1) = 2{,}0$. Тепер можемо записати систему із двох рівнянь для знаходження констант:

$$k = 0; \quad 1{,}0 = 6{,}0 + A_1 + A_2;$$
$$k = 1; \quad 2{,}0 = 6{,}0 + 0{,}5\,A_1 + 0{,}3\,A_2.$$

Розв'язуючи систему, знайдемо, що $A_1 = -12{,}5$; $A_2 = 7{,}5$, і розв'язок набуває вигляду

$$y(k) = 6{,}0 - 12{,}5\,(0{,}5)^k + 7{,}5\,(0{,}3)^k.$$

9.3. Знаходження загальних розв'язків однорідних рівнянь та частинних розв'язків неоднорідних

9.3.1. Розв'язування однорідних рівнянь

Однорідне рівняння другого порядку

При моделюванні економічних процесів найчастіше використовують різницеві рівняння 1–3 порядку, але зовсім не виключена можливість використання рівнянь вищих порядків. У зв'язку з цим виникає необхідність знаходження розв'язків однорідних рівнянь різних порядків. Оскільки розв'язок рівняння другого порядку дає можливість зробити деякі узагальнення на рівняння вищих порядків, то розглянемо спочатку рівняння другого порядку:

$$y(k) - a_1\,y(k-1) - a_2\,y(k-2) = 0. \qquad (9.3.1)$$

За аналогією з рівнянням першого порядку виберемо шуканий однорідний розв'язок у загальному вигляді $y^h(k) = A\alpha^k$. Підставимо шуканий розв'язок у рівняння (9.3.1):

$$A\alpha^k - a_1 A\alpha^{k-1} - a_2 A\alpha^{k-2} = 0 \qquad (9.3.2)$$

і поділимо обидві частини отриманого рівняння на $A\alpha^{k-2}$. У результаті отримаємо так зване *характеристичне рівняння* відносно α:

$$\alpha^2 - a_1\alpha - a_2 = 0. \tag{9.3.3}$$

Запишемо корені цього квадратного рівняння:

$$\alpha_{1,2} = \frac{a_1 \pm \sqrt{a_1^2 + 4a_2}}{2} = \frac{a_1 \pm \sqrt{d}}{2}, \tag{9.3.4}$$

де дискримінант $d = \sqrt{a_1^2 + 4a_2}$. Кожний корінь характеристичного рівняння дає один однорідний розв'язок. Розв'язком є також лінійна комбінація

$$A_1\alpha_1^k + A_2\alpha_2^k, \tag{9.3.5}$$

що легко перевірити підстановкою (9.3.5) у ліву частину (9.3.1). Тоді одержимо

$$A_1\alpha_1^k + A_2\alpha_2^k - a_1(A_1\alpha_1^{k-1} + A_2\alpha_2^{k-1}) - a_2(A_1\alpha_1^{k-2} + A_2\alpha_2^{k-2}).$$

Згрупуємо члени і маємо

$$A_1(\alpha_1^k - a_1\alpha_1^{k-1} - a_2\alpha_1^{k-2}) + A_2(\alpha_2^k - a_1\alpha_2^{k-1} - a_2\alpha_2^{k-2}) = 0,$$

оскільки α_1, α_2 — це корені характеристичного рівняння (9.3.3), і через це обидва члени в дужках дорівнюють нулю. Таким чином, повний однорідний розв'язок РР другого порядку має вигляд:

$$y^h(k) = A_1\alpha_1^k + A_2\alpha_2^k. \tag{9.3.6}$$

Оскільки *рівняння другого порядку* — це основа для аналізу розв'язків рівнянь вищих порядків, у літературі розглядають три можливих випадки знаходження їх розв'язку, залежно від значення дискримінанта: $d > 0$, $d = 0$ і $d < 0$. Крім того, рівняння другого порядку — досить поширена модель економічних, фінансових, технічних та інших процесів. Так, у технічних системах рівняннями другого порядку можна описати гармонійні коливання елементів конструкцій.

А. *Дискримінант додатний*: $a_1^2 + 4\,a_2 > 0$. У такому випадку корені будуть різними дійсними числами. Загальний розв'язок однорідного рівняння має вигляд:

$$y^h(k) = A_1\,\alpha_1^k + A_2\,\alpha_2^k.$$

Приклад 9.2. Розглянемо процес другого порядку:

$$y(k) = 0{,}95\, y(k-1) + 0{,}35\, y(k-2),$$

тобто $a_1 = 0{,}95$; $a_2 = 0{,}35$. Для однорідного рівняння $y(k) - a_1\, y(k-1) - a_2\, y(k-2) = 0$ характеристичним є наступне: $\alpha^2 - 0{,}95\,\alpha - 0{,}35 = 0$. Дискримінант $d = a_1{}^2 + + 4\, a_2 = 0{,}95^2 + 4 \cdot 0{,}35 = 2{,}3025$. Корені характеристичного рівняння:

$$\alpha_1 = \frac{1}{2}\left(a_1 + \sqrt{d}\right) = \frac{1}{2}(0{,}95 + 1{,}517) = 1{,}233; \quad \alpha_2 = \frac{1}{2}(0{,}95 - 1{,}517) = -0{,}284.$$

Таким чином, однорідний розв'язок має вигляд

$$y^h(k) = A_1\, 1{,}233^k + A_2\, (-1{,}284)^k.$$

Б. *Дискримінант* $a_1^2 + 4a_2 = 0$ (*випадок кратних коренів*). При $d = 0$ отримаємо два однакових корені: $\alpha_1 = \alpha_2 = a_1/2$. Таким чином, однорідним розв'язком є $A \cdot (a_1/2)^k$. Однак при $d = 0$ існує також однорідний розв'язок такого вигляду: $A \cdot k(a_1/2)^k$ (тобто дискретний час k виступає множником, що веде до нелінійної поведінки однорідного розв'язку). Покажемо, що це так, шляхом підстановки другого розв'язку в ліву частину рівняння (9.3.1)

$$k(a_1/2)^k - a_1[(k-1)(a_1/2)^{k-1}] - a_2[(k-2)(a_1/2)^{k-2}].$$

Для зручності аналізу винесемо за дужки $(a_1/2)^{k-2}$ і отримаємо в дужках:

$$-[(a_1^2/4) + a_2]k + [(a_1^2/2) + 2a_2].$$

Оскільки $a_1^2 + 4a_2 = 0$, то обидва вирази в дужках будуть дорівнювати нулю і, таким чином, $k(a_1/2)^k$ є розв'язком рівняння другого порядку при $d = 0$. Тепер можемо записати загальний розв'язок однорідного рівняння для даного випадку у вигляді лінійної комбінації двох розв'язків:

$$y^h(k) = A_1\left(\frac{a_1}{2}\right)^k + A_2 k\left(\frac{a_1}{2}\right)^k.$$

У випадку, якщо характеристичне рівняння має m-кратний корінь, загальний розв'язок однорідного рівняння матиме такий вигляд:

$$y^h(k) = A_1\,\alpha^k + A_2\, k\,\alpha^k + A_3\, k^2\,\alpha^k + \ldots + A_m\, k^{m-1}\,\alpha^k.$$

В. *Дискримінант* $a_1^2 + 4\,a_2 < 0$. У даному випадку характеристичне рівняння має два комплексно-спряжених корені. Оскільки $a_1^2 \geq 0$ завжди, то комплексні корені можуть з'явитись тільки у випадку, коли $a_2 < 0$.

Таким чином, корені характеристичного рівняння можна записати у вигляді:

$$\alpha_1 = \frac{1}{2}\left(a_1 + i\sqrt{d}\right),\ \ \alpha_2 = \frac{1}{2}\left(a_1 - i\sqrt{d}\right),\ \text{де}\quad i = \sqrt{-1}.$$

При знаходженні розв'язку однорідного рівняння скористаємось такими тригонометричними тотожностями:

$$\sin(\theta_1 + \theta_2) = \sin(\theta_1)\cos(\theta_2) + \cos(\theta_1)\sin(\theta_2);$$
$$\cos(\theta_1 + \theta_2) = \cos(\theta_1)\cos(\theta_2) - \sin(\theta_1)\sin(\theta_2),$$

які при $\theta_1 = \theta_2$ спрощуються до вигляду:

$$\sin(2\theta) = 2\sin(\theta)\cos(\theta);$$
$$\cos(2\theta) = \cos^2(\theta) - \sin^2(\theta).$$

Нагадаємо, що комплексне число $a + ib$ можна представити точкою на комплексній площині (рис. 9.1) і перейти до представлення коренів характеристичного рівняння у полярних координатах.

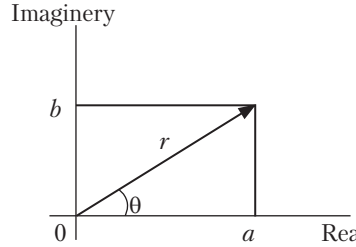


Рис. 9.1. **Представлення комплексного числа на площині**

Запишемо комплексне число через тригонометричні функції кута θ і отримаємо його представлення у полярних координатах:

$$a = r\cos(\theta),\ b = r\sin(\theta).$$

По відношенню до виразів для коренів однорідного рівняння α_1, α_2 можна записати, що $a = a_1/2$, $b = \sqrt{d}/2$.

Таким чином, для коренів α_1, α_2 можна записати вирази у полярних координатах:

$$\alpha_1 = a + ib = r[\cos(\theta) + i\sin(\theta)];$$
$$\alpha_2 = a - ib = r[\cos(\theta) - i\sin(\theta)].$$

Щоб побудувати розв'язок, необхідно отримати вирази для α_1^k і α_2^k. Спочатку запишемо вираз для α_1^k з урахуванням того, що $i^2 = -1$:

$$\alpha_1^2 = \left\{r\left[\cos(\theta) + i\sin(\theta)\right]\right\}\left\{r\left[\cos(\theta) + i\sin(\theta)\right]\right\} =$$
$$= r^2\left[\cos(\theta)\cos(\theta) - \sin(\theta)\sin(\theta) + 2i\sin(\theta)\cos(\theta)\right] =$$
$$= r^2\left[\cos(2\theta) + i\sin(2\theta)\right].$$

Якщо продовжити процес піднесення до степеня, то отримаємо такі вирази для коренів:

$$\alpha_1^k = r^k[\cos(k\theta) + i\sin(k\theta)], \quad \alpha_2^k = r^k[\cos(k\theta) - i\sin(k\theta)].$$

Оскільки однорідний розв'язок $y^h(k)$ — це дійсне число, а α_1, α_2 — комплексні числа, то довільні константи A_1, A_2 — повинні бути спряженими комплексними числами у наступній формі:

$$A_1 = B[\cos(\varphi) + i\sin(\varphi)], \quad A_2 = B[\cos(\varphi) - i\sin(\varphi)],$$

де B, φ — довільні дійсні числа.

Тепер можна знайти добутки $A_1\alpha_1^k$ і $A_2\alpha_2^k$. Так,

$$A_1\alpha_1^k = B[\cos(\varphi) + i\sin(\varphi)]\, r^k[\cos(k\theta) + i\sin(k\theta)] =$$
$$= Br^k[\cos(\varphi)\cos(k\theta) - \sin(\varphi)\sin(k\theta) + i\cos(k\theta)\sin(\varphi) +$$
$$+ i\sin(k\theta)\cos(\varphi)].$$

Використовуючи наведені тригонометричні тотожності, отримаємо:

$$A_1\alpha_1^k = Br^k[\cos(k\theta + \varphi) + i\sin(k\theta + \varphi)].$$

За аналогією можна отримати вираз для другого добутку:

$$A_2\alpha_2^k = Br^k[\cos(k\theta + \varphi) - i\sin(k\theta + \varphi)].$$

Тепер запишемо загальний розв'язок однорідного рівняння як суму двох розв'язків:

$$y^h(k) = Br^k[\cos(k\theta + \varphi) + i\sin(k\theta + \varphi)] + Br^k[\cos(k\theta + \varphi) - i\sin(k\theta + \varphi)] =$$
$$= 2\,Br^k\cos(k\theta + \varphi).$$

Якщо покласти $2B = \beta_1$ і $\theta = w$, то розв'язок набирає вигляду

$$y^h(k) = \beta_1 r^k\cos(wk + \varphi),$$

де β_1, φ — довільні константи; $r = \sqrt{(-a_2)}$; значення w вибирається з умови:

$$\cos(w) = \frac{a_1}{2(-a_2)^{1/2}} = \frac{a_1}{2r}.$$

З погляду фізики величина w має смисл кругової частоти, тобто $w = 2\pi f$ (f — лінійна частота у герцах), а φ — початкова фаза коливань розв'язку. Тобто однорідний розв'язок має у цьому випадку форму затухаючих або розбіжних коливань залежно від значення r.

Наведений вище вираз для r отримується так:

$$r = \sqrt{a^2 + b^2} = \sqrt{\left(\frac{a_1}{2}\right)^2 + \left(\frac{\sqrt{|d|}}{2}\right)^2} = \sqrt{\frac{a_1^2 + \left[-\left(a_1^2 + 4a_2\right)\right]}{4}} = \sqrt{-a_2}\,.$$

Розглянемо *приклад знаходження однорідного розв'язку* у випадку комплексних коренів.

Приклад 9.3. Нехай рівняння другого порядку має вигляд

$$y(k) = 1{,}6\, y(k-1) - 0{,}8\, y(k-2).$$

Характеристичне рівняння для нього:

$$\alpha^2 - 1{,}6\,\alpha + 0{,}8 = 0.$$

Дискримінант $d = a_1^2 + 4\, a_2 = -0{,}64$. Корені характеристичного рівняння і параметри однорідного розв'язку:

$$\alpha_1 = 0{,}5\left(1{,}6 + i\,0{,}8\right), \quad \alpha_2 = 0{,}5\left(1{,}6 - i\,0{,}8\right);$$

$$r = \left(-a_2\right)^{1/2} = \left(0{,}80\right)^{1/2} = 0{,}894;$$

$$\cos(w) = \frac{a_1}{2(-a_2)^{1/2}} = \frac{1{,}6}{2 \cdot 0{,}894} = 0{,}894.$$

Загальний розв'язок однорідного рівняння буде:

$$y(k) = \beta_1 (0{,}894)^k \cos(k \arccos 0{,}894 + \varphi).$$

Структура розв'язку однорідного рівняння вищого порядку

Розв'язки однорідних рівнянь вищих порядків знаходять за аналогією з рівнянням другого порядку. Розглянемо однорідне рівняння n-го порядку:

$$y(k)-\sum_{i=1}^{n} a_i y(k-i)=0.$$

Підставимо в нього однорідний розв'язок загального виду $y^h(k) = A\alpha^k$:

$$A\alpha^k - a_1 A\alpha^{k-1} - a_2 A\alpha^{k-2} - ... - a_n A\alpha^{k-n} = 0$$

і розділимо отримане рівняння на $A\alpha^{k-n}$. У результаті отримаємо характеристичне рівняння n-го порядку:

$$\alpha^n - a_1 \alpha^{n-1} - a_2 \alpha^{n-2} - ... - a_n = 0.$$

Воно має n розв'язків, тобто n коренів α_i, $i = 1, ..., n$. Згідно з отриманими вище результатами, лінійна комбінація n коренів також буде розв'язком. Якщо серед знайдених коренів буде корінь кратності m, то розв'язок матиме вигляд

$$y^h(k) = A_1\alpha^k + A_2 k\,\alpha^k + ... + A_m k^{m-1}\alpha^k + A_{m+1}\alpha^k{}_{m+1} + ... + A_n \alpha^k{}_n.$$

Очевидно, що у випадку наявності комплексно-спряженого кореня розв'язок буде мати відповідну гармонійну складову вигляду $y^h(k) = \beta_1 r^k \cos(wk + \varphi)$.

9.3.2. Знаходження частинних розв'язків різницевих рівнянь

Частинний розв'язок неоднорідного рівняння зумовлений наявністю у рівнянні констант, а також детермінованих і випадкових змінних і функцій у правій частині. Форма частинного розв'язку залежить від конкретного типу збурюючої функції у правій частині різницевого рівняння. Методика знаходження частинних розв'язків складається з таких кроків:

- вибір шуканої функції;
- підстановка шуканої функції у рівняння і визначення невідомих коефіцієнтів шуканої функції (наприклад, за методом невизначених коефіцієнтів);
- аналіз характеру отриманого розв'язку.

Розглянемо кілька характерних випадків знаходження частинних розв'язків.

Випадок 1

У правій частині різницевого рівняння наявні тільки постійна складова (зміщення) a_0 і авторегресійна (АР) частина. Тобто рівняння має вигляд

$$y(k) = a_0 + a_1 y(k-1) + a_2 y(k-2) + \ldots + a_n y(k-n). \qquad (9.3.7)$$

При подачі константи на вхід лінійного процесу на виході також повинна бути константа. Тобто логічно припустити, що частинним розв'язком цього рівняння повинна бути деяка константа, що можна записати так: $y(k) = y(k-1) = \ldots = c$. Підставимо цей шуканий розв'язок у (9.3.7) і отримаємо:

$$c = a_0 + a_1 c + a_2 c + \ldots + a_n c.$$

Звідси у випадку $\sum_{i=1}^{n} a_i \neq 1$ маємо

$$c = \frac{a_0}{1 - a_1 - a_2 - \ldots - a_n}. \qquad (9.3.8)$$

Константа є частинним розв'язком і можна записати:

$$y^p = \frac{a_0}{1 - a_1 - a_2 - \ldots - a_n} = \frac{a_0}{1 - \sum_{i=1}^{n} a_i}. \qquad (9.3.9)$$

Якщо $\sum_{i=1}^{n} a_i = 1$, то маємо окремий випадок, коли процес $\{y(k)\}$ називають *процесом з одиничним коренем*. У простому випадку (наявний один одиничний корінь характеристичного рівняння) можна скористатися шуканим розв'язком виду $y^p(k) = ck$. Підставимо цей шуканий розв'язок у рівняння (9.3.7) і отримаємо:

$$ck = a_0 + (k-1)\, a_1 c + (k-2)\, a_2 c + \ldots + (k-n)\, a_n c;$$

або

$$(1 - a_1 - a_2 - \ldots - a_n)\, ck = a_0 - (a_1 + 2a_2 + 3a_3 + \ldots + na_n) \cdot c.$$

Оскільки $1 - a_1 - a_2 - \ldots - a_n = 0$, то

$$c = \frac{a_0}{a_1 + 2a_2 + 3a_3 + \ldots + na_n}.$$

Наприклад, для рівняння

$$y(k) = 2{,}5 + 0{,}75\, y(k-1) + 0{,}25\, y(k-2),$$

$a_1 + a_2 = 0{,}75 + 0{,}25 = 1$, і частинний розв'язок має вигляд ck, де $c = 2{,}5/(0{,}75 + 2 \cdot 0{,}25) = 2{,}0$. Якщо шукана функція ck не приводить до успішного знаходження розв'язку, то необхідно взяти іншу шукану функцію. Наприклад, можна скористатися нелінійними функціями виду $y^p(k) = ck^2, ck^3, \ldots, ck^n$. Для рівняння довільного порядку одна з цих функцій обов'язково буде частинним розв'язком.

Випадок 2

Лінійний процес збурюється *експонентою* $x(k) = bd^{rk}$, де b, d, r — константи. Наприклад, для економічних процесів коефіцієнт r можна розглядати як фактор зростання. Розглянемо послідовність знаходження частинного розв'язку на прикладі рівняння першого порядку:

$$y(k) = a_0 + a_1 y(k-1) + bd^{rk}. \qquad (9.3.10)$$

У правій частині цього рівняння є константа a_0. Тому можна очікувати, що у частинному розв'язку також буде наявна константа. Для лінійного процесу можна також передбачити, що при подачі на його вхід сигналу bd^{rk} на виході з'явиться реакція такої самої форми, але з деяким коефіцієнтом передачі. Тому логічно записати шуканий розв'язок у формі:

$$y^p(k) = c_0 + c_1 d^{rk},$$

де c_0, c_1 — невідомі константи, які знайдемо за методом невизначених коефіцієнтів. Підставимо шуканий розв'язок у рівняння (9.3.10) і в результаті отримаємо:

$$c_0 + c_1 d^{rk} = a_0 + a_1 (c_0 + c_1 d^{r(k-1)}) + bd^{rk}. \qquad (9.3.11)$$

Перенесемо в ліву частину члени, які містять невідомі константи c_0, c_1, і об'єднаємо їх таким чином:

$$c_0 + c_1 \frac{d^{rk}}{1-a_1}\left(\frac{d^r - a_1}{d^r}\right) = \frac{a_0}{1-a_1} + \frac{b}{1-a_1} d^{rk}.$$

Звідси знайдемо, що

$$c_0 = \frac{a_0}{1-a_1}; \qquad c_1 \frac{d^{rk}}{1-a_1}\frac{d^r - a_1}{d^r} = \frac{b}{1-a_1} d^{rk},$$

або

$$c_1 = \frac{bd^r}{d^r - a_1}.$$

Тепер запишемо частинний розв'язок рівняння (9.3.10):

$$y^p(k) = \frac{a_0}{1-a_1} + \frac{bd^r}{d^r - a_1} d^{rk}. \qquad (9.3.12)$$

Якщо $a_1 = 1$, то при знаходженні частинного розв'язку необхідно взяти шукану функцію у вигляді $y^p(k) = ck$. При $a_1 = d^r$ шукана функція має вигляд:

$$c_1 = \frac{kbd^r}{d^r - a_1}.$$

Методика знаходження частинного розв'язку для рівнянь вищих порядків залишається такою самою. Спочатку необхідно коректно вибрати шукану функцію відповідно до структури рівняння, а потім застосувати метод невизначених коефіцієнтів для знаходження невідомих параметрів.

Випадок 3

На вхід процесу подається *функція* у формі $x(k) = bk^d$, де b — константа; d — ціле додатне число. Різницеве рівняння має вигляд:

$$y(k) = a_0 + \sum_{i=1}^{n} a_i y(k-i) + bk^d. \qquad (9.3.13)$$

У даному випадку $y(k)$ залежить від k^d, значення $y(k-1)$ залежить від $(k-1)^d$, значення $y(k-1)$ залежить від $(k-2)^d$ і т. д. Тому шукану функцію для частинного розв'язку можна вибрати у вигляді:

$$y^p(k) = c_0 + c_1 k + c_1 k^2 + ... + c_d k^d.$$

Наприклад, щоб знайти невідомі коефіцієнти c_i, підставимо шуканий розв'язок $y = c_0 + c_1 k$ у рівняння другого порядку $y(k) = a_0 + a_1 y(k-1) + a_2 y(k-2)$, тоді при $d = 1$:

$$c_0 + c_1 k = a_0 + a_1 [c_0 + (k-1)c_1] + a_2[c_0 + (k-2)c_1] + bk. \qquad (9.3.14)$$

Значення коефіцієнтів c_0, c_1 виберемо такими, щоб (9.3.14) залишалось тотожністю для всіх можливих значень k. Згрупуємо всі константи і всі члени, які мають множник k:

$$c_0 - a_0 - a_1 c_0 + a_1 c_1 - a_2 c_0 + 2\, a_2 c_1 = a_1 k c_1 + a_2 k c_1 + bk - c_1 k.$$

Звідси отримаємо значення невідомих констант за *методом невизначених коефіцієнтів*:

$$c_1 = b\,/(1 - a_1 - a_2);$$

$$c_0 = [a_0 - (2a_2 + a_1)c_1]/(1 - a_1 - a_2)$$

або

$$c_0 = \frac{a_0}{1 - a_1 - a_2} - \frac{b(2a_2 + a_1)}{(1 - a_1 - a_2)^2}.$$

Таким чином, частинний розв'язок включає в себе лінійний часовий тренд. Якщо $a_1 + a_2 = 1$, то шуканий розв'язок необхідно помно-

жити на k. Методика залишається незмінною для рівнянь вищих порядків.

Приклад 9.4. Розглянемо детермінований процес авторегресії другого порядку:

$$y(k) = 2{,}5 + a_1 y(k-1) + a_2 y(k-2);$$

$$a_1 = 0{,}75;\ a_2 = 0{,}25;\ y(0) = 1;\ y(1) = 2.$$

Оскільки сума коефіцієнтів моделі $\sum_{i=1}^{2} a_i = 0{,}75 + 0{,}25 = 1$, то за шуканий частинний розв'язок вибираємо $y^p_{\text{ш}}(k) = ck$, де

$$c = \frac{a_0}{a_1 + 2a_2} = \frac{2{,}5}{0{,}75 + 2 \cdot 0{,}25} = \frac{2{,}5}{1{,}25} = 2{,}0.$$

Таким чином, частинний розв'язок має вигляд: $y^p(k) = 2{,}0k$.

Характеристичне рівняння $\alpha^2 - 0{,}75\alpha - 0{,}25 = 0$ має корені: $\alpha_1 = 1$, $\alpha_2 = -0{,}25$. Запишемо загальний розв'язок:

$$y(k) = A_1 + A_2(-0{,}25)^k + 2{,}0k.$$

Використаємо початкові умови для знаходження довільних констант:

$$k = 0 : 1 = A_1 + A_2;$$
$$k = 1 : 2 = A_1 - 0{,}25A_2 + 2{,}0.$$

Розв'язуючи цю систему двох рівнянь, знайдемо: $A_1 = 0{,}2$; $A_2 = 0{,}8$.

Приклад 9.5. Необхідно знайти загальний розв'язок рівняння

$$y(k) = a_0 + 2y(k-1) - y(k-2).$$

Характеристичне рівняння $\alpha^2 - 2\alpha + 1 = 0$ має корені: $\alpha_1 = \alpha_2 = 1$, тобто маємо випадок кратних коренів. Оскільки шукана функція $y^p_{\text{ш}}(k) = ck$ не дає необхідного результату, то вибираємо за шукану функцію $y^p_{\text{ш}}(k) = ck^2$. У результаті підстановки шуканого розв'язку у рівняння знайдемо:

$$c = \frac{a_0}{k^2 - 2(k-1)^2 + (k-2)^2} = \frac{a_0}{2}.$$

Тепер запишемо загальний розв'язок:

$$y(k) = A_1 + A_2 k + \frac{a_0}{2} k^2.$$

Випадок 4

Застосування методу невизначених коефіцієнтів для знаходження частинного розв'язку *стохастичного* різницевого рівняння (у правій частині наявний випадковий процес).

Процес авторегресії першого порядку

Як уже зазначалося, стохастична авторегресія першого порядку описується рівнянням

$$y(k) = a_0 + a_1 y(k-1) + \varepsilon(k).$$

Природа процесу $\{y(k)\}$ у даному випадку така, що частинний розв'язок залежить тільки від постійного члена a_0, часу k та послідовності $\{\varepsilon(k)\}$. Враховуючи, що часова змінна у збурюючій функції задана неявно, то змінна k може входити у частинний розв'язок тільки у тому випадку, коли корінь характеристичного рівняння дорівнює 1.

Для ілюстрації методу скористаємось такою шуканою функцією:

$$y(k) = b_0 + b_1 k + \sum_{i=0}^{k} \alpha_i \varepsilon(k-i), \tag{9.3.15}$$

де b_0, b_1, α_i — невідомі коефіцієнти. Для того щоб знайти значення цих коефіцієнтів, підставимо шуканий розв'язок (9.3.15) у РР першого порядку:

$$b_0 + b_1 k + \alpha_0 \varepsilon(k) + \alpha_1 \varepsilon(k-1) + \alpha_2 \varepsilon(k-2) + \ldots =$$
$$= a_0 + a_1 \left[b_0 + b_1 (k-1) + \alpha_0 \varepsilon(k-1) + \alpha_1 \varepsilon(k-2) + \ldots \right] + \varepsilon(k).$$

Після групування членів отримаємо:

$$(b_0 - a_0 - a_1 b_0 + a_1 b_1) + b_1 (1 - a_1) k + (\alpha_0 - 1)\varepsilon(k) + (\alpha_1 - a_1 \alpha_0)\varepsilon(k-1) +$$
$$+ (\alpha_2 - a_1 \alpha_1)\varepsilon(k-2) + (\alpha_3 - a_1 \alpha_2)\varepsilon(k-3) + \ldots = 0. \tag{9.3.16}$$

Отримане рівняння (9.3.16) повинне бути справедливим при всіх можливих значеннях k і всіх можливих значеннях $\varepsilon(k-i)$, $i = 0, 1, 2, \ldots$; тобто мають виконуватись такі умови:

$$\begin{aligned}
&\alpha_0 - 1 = 0;\\
&\alpha_1 - a_1 \alpha_0 = 0;\\
&\alpha_2 - a_1 \alpha_1 = 0;\\
&\quad\vdots\\
&b_0 - a_0 - a_1 b_0 + a_1 b_1 = 0;\\
&b_1 - a_1 b_1 = 0.
\end{aligned} \tag{9.3.17}$$

Зазначимо, що першою групою умов (перші три рядки) можна скористатись для знаходження розв'язку відносно α_i у рекурсивній формі. Використовуючи умови (9.3.17), знайдемо:

$$\alpha_0 = 1, \quad \alpha_1 = a_1, \quad \alpha_2 = a_1\alpha_1 = a_1^2, \quad \alpha_i = a_1^i.$$

Якщо $a_1 \neq 1$ (немає одиничного кореня характеристичного рівняння), то $b_1 = 0$, $b_0 = a_0/(1 - a_1)$, і частинний розв'язок має вигляд

$$y(k) = \frac{a_0}{1-a_1} + \sum_{i=0}^{k} a_1^i \varepsilon(k-i).$$

Такий самий розв'язок був отриманий вище для РР першого порядку за методом ітерацій. Загальний розв'язок отримаємо у результаті підсумовування частинного та загального розв'язків однорідного рівняння

$$y(k) = \frac{a_0}{1-a_1} + Aa_1^k + \sum_{i=0}^{k} a_1^i \varepsilon(k-i). \qquad (9.3.18)$$

Якщо відома початкова умова $y(0) = y_0$, то можна знайти значення довільної константи з рівняння

$$y_0 = \frac{a_0}{1-a_1} + A + \varepsilon(0).$$

Таким чином, частинний розв'язок неоднорідного рівняння можна записати у вигляді

$$y(k) = \frac{a_0}{1-a_1} + a_1^k\left[y_0 - \frac{a_0}{1-a_1} - \varepsilon(0)\right] + \sum_{i=0}^{k} a_1^i \varepsilon(k-i). \qquad (9.3.19)$$

Якщо $a_1 = 1$ (корінь характеристичного рівняння також дорівнює одиниці), то b_0 може бути довільною константою, а $b_1 = a_0$. У цьому випадку можливою формою частинного розв'язку може бути функція, яка включає добуток $a_0 k$, оскільки з'являється одиничний корінь. Така функція має вигляд

$$y(k) = b_0 + a_0 k + \sum_{i=0}^{k} \varepsilon(k-i). \qquad (9.3.20)$$

Стохастичний процес АРКС(1,1)

Як розширення попереднього випадку, розглянемо тепер стохастичне рівняння АРКС(1,1):

$$y(k) = a_0 + a_1 y(k-1) + \varepsilon(k) + \beta_1 \varepsilon(k-1). \qquad (9.3.21)$$

Розв'язок цього рівняння буде включати константу, послідовність $\{\varepsilon(k)\}$ та час k у першому степені (тому що рівняння першого порядку). Як і в попередньому прикладі, k потрібно включати у шуканий розв'язок, якщо корінь характеристичного рівняння дорівнює одиниці. Отже, скористаємось таким шуканим розв'язком:

$$y_{\text{ш}}(k) = b_0 + b_1 k + \sum_{i=0}^{k} \alpha_i \varepsilon(k-i).$$

Підставимо шуканий розв'язок у (9.3.21) і отримаємо:

$$b_0 + b_1 k + \sum_{i=0}^{k} \alpha_i \varepsilon(k-i) = a_0 + a_1 \left[b_0 + b_1 (k-1) + \sum_{i=0}^{k} \alpha_i \varepsilon(k-i-1) \right] +$$
$$+ \varepsilon(k) + \beta_1 \varepsilon(k-1)$$

або

$$b_0 + b_1 k + \sum_{i=0}^{k} \alpha_i \varepsilon(k-i) = a_0 + a_1 b_0 + a_1 b_1 k - a_1 b_1 + a_1 \sum_{i=0}^{k} \alpha_i \varepsilon(k-i-1) +$$
$$+ \varepsilon(k) + \beta_1 \varepsilon(k-1).$$

Прирівнюючи коефіцієнти при однакових змінних $\varepsilon(k)$, $\varepsilon(k-1)$, $\varepsilon(k-2)$, ..., отримаємо такі значення невідомих коефіцієнтів розв'язку:

$$\alpha_0 = 1;$$
$$\alpha_1 = a_1 \alpha_0 + \beta_1 = a_1 + \beta_1;$$
$$\alpha_2 = a_1 \alpha_1 = a_1 (a_1 + \beta_1);$$
$$\alpha_3 = a_1 \alpha_2 = a_1^2 (a_1 + \beta_1);$$
$$\vdots$$
$$\alpha_i = a_1 \alpha_{i-1} \quad \text{або} \quad \alpha_i = a_1^{i-1} (a_1 + \beta_1).$$

Прирівнюючи константи та члени, які містять час k, отримаємо:

$$b_0 = a_0 + a_1 b_0 - a_1 b_1;$$
$$b_1 = a_1 b_1.$$

Для даного прикладу також необхідно розглянути два випадки: $a_1 \neq 1$ і $a_1 = 1$. Якщо $a_1 \neq 1$, то це означає, що тренду немає і $b_1 = 0$. Звідси випливає, що

$$b_0 = \frac{a_0}{1 - a_1},$$

а частковий розв'язок має вигляд

$$y(k)=\frac{a_0}{1-a_1}+\varepsilon(k)+(a_1+\beta_1)\sum_{i=1}^{k}a_1^{i-1}\varepsilon(k-i).$$

Загальний розв'язок включає ще однорідний:

$$y(k)=\frac{a_0}{1-a_1}+Aa_1^k+\varepsilon(k)+(a_1+\beta_1)\sum_{i=1}^{k}a_1^{i-1}\varepsilon(k-i).$$

Якщо $a_1 = 1$, то з виразу $b_0 = a_0 + a_1b_0 - a_1b_1$ випливає, що $b_1 = a_0$, а коефіцієнт b_0 є довільною константою, яка може бути визначена за допомогою початкових умов. Таким чином, розв'язок для даного випадку має вигляд

$$y(k)=b_0+a_0k+\varepsilon(k)+(1+\beta_1)\sum_{i=1}^{k}\varepsilon(k-i). \qquad (9.3.22)$$

Стохастичний процес АР(2)

Оскільки на практиці досить часто зустрічаються моделі другого порядку, то розглянемо знаходження розв'язку для стохастичної авторегресії АР(2). Це рівняння має вигляд [11, 12]

$$y(k) = a_0 + a_1y(k-1) + a_2y(k-2) + \varepsilon(k). \qquad (9.3.23)$$

Оскільки це рівняння другого порядку, то його характеристичне рівняння може мати два одиничні корені, а тому максимальний порядок полінома для опису можливого тренду повинен становити 2. Тобто шуканий розв'язок запишемо так:

$$y_{ш}(k)=b_0+b_1k+b_2k^2+\sum_{i=0}^{k}\alpha_i\,\varepsilon(k-i),$$

де b_0, b_1, b_2, α_i — невідомі коефіцієнти розв'язку рівняння (9.3.23). Підстановка шуканого розв'язку в рівняння (9.3.23) приводить до такого виразу:

$$\begin{aligned}&b_0+b_1k+b_2k^2+\alpha_0\varepsilon(k)+\alpha_1\varepsilon(k-1)+\alpha_2\varepsilon(k-2)+\ldots= \qquad (9.3.24)\\ &=a_0+a_1\left[b_0+b_1(k-1)+b_2(k-1)^2+\alpha_0\varepsilon(k-1)+\alpha_1\varepsilon(k-2)+\alpha_2\varepsilon(k-3)+\ldots\right]+\\ &+a_2\left[b_0+b_1(k-2)+b_2(k-2)^2+\alpha_0\varepsilon(k-2)+\alpha_1\varepsilon(k-3)+\alpha_2\varepsilon(k-4)+\ldots\right]+\varepsilon(k).\end{aligned}$$

З умов, які приводять наведений вираз до тотожності при довільних реалізаціях процесу $\{\varepsilon(k)\}$, отримуємо такі значення невизначених коефіцієнтів:

$$\begin{aligned}
&\alpha_0 = 1;\\
&\alpha_1 = a_1\alpha_0 = a_1;\\
&\alpha_2 = a_1\alpha_1 + a_2\alpha_0 = a_1^2 + a_2;\\
&\alpha_3 = a_1\alpha_2 + a_2\alpha_1 = a_1^3 + 2a_1a_2;\\
&\quad\vdots
\end{aligned} \qquad (*)$$

З наведених виразів для обчислення коефіцієнтів α_i видно, що починаючи з α_2, всі коефіцієнти можна визначити за допомогою різницевого рівняння другого порядку:

$$\alpha_j = a_1\,\alpha_{j-1} + a_2\,\alpha_{j-2}. \qquad (**)$$

Оскільки α_0 і α_1 відомі, то всі наступні коефіцієнти можна визначити рекурсивно.

Примітка. Якщо складові збурюючої функції у правій частині рівняння, тобто $\varepsilon(k)$, мають такий самий вигляд, як і складові $y^h(k)$, то шукана функція часткового розв'язку змінюється. Всі складові $y^p(k)$, які відповідають таким самим складовим $y^h(k)$, необхідно помножити на k у такому найменшому степені, щоб їх тотожність не порушувалась. Відповідний приклад буде розглянуто у наступному параграфі.

Процес АР(2) з випадковою та детермінованою функціями на вході

Розглянемо процес з випадковою та детермінованою функціями на вході:

$$y(k) = a_0 + a_1y(k-1) + a_2y(k-2) + x(k) + \varepsilon(k), \qquad (9.3.26)$$

де $a_0 = 0{,}3$; $a_1 = -0{,}5$; $a_2 = 1{,}5$; $x(k) = c_1\, e^{c_2(k+1)}$; $\varepsilon(k)$ — центрований випадковий процес типу білого шуму; c_1, c_2 — константи.

Характеристичне рівняння $\alpha^2 + 0{,}5\alpha - 1{,}5 = 0$ має корені: $\alpha_1 = 1$ і $\alpha_2 = -1{,}5$. Тобто це процес з одиничним коренем, який має тренд. Таким чином, однорідний розв'язок має вигляд

$$y^h(k) = A_1 + A_2(-1{,}5)^k.$$

Підставимо шукану функцію виду

$$y_{\text{ш}}^{p}(k) = b_0 + b_1 k + b_2 k^2 + b_3 e^{c_2(k+1)} + \sum_{i=0}^{k} \alpha_i\, \varepsilon(k-i)$$

у рівняння (9.3.26)

$$b_0 + b_1 k + b_2 k^2 + b_3 e^{c_2(k+1)} + \sum_{i=0}^{k} \alpha_i\, \varepsilon(k-i) = a_0 +$$

$$+a_1\left[b_0 + b_1(k-1) + b_2(k-1)^2 + b_3 e^{c_2 k} + \alpha_0 \varepsilon(k-1) + \alpha_1 \varepsilon(k-2) + \ldots\right] +$$

$$+a_2\left[b_0 + b_1(k-2) + b_2(k-2)^2 + b_3 e^{c_2(k-1)} + \alpha_0 \varepsilon(k-2) + \alpha_2 \varepsilon(k-3) + \ldots\right] +$$

$$+c_1 e^{c_2(k+1)} + \varepsilon(k).$$

Прирівняємо коефіцієнти у лівій і правій частинах при однакових ступенях k:

$$b_0 = a_0 + a_1 b_0 - a_1 b_1 + a_1 b_2 + a_2 b_0 - 2a_2 b_1 + 4\, a_2 b_2 . \qquad (9.3.27)$$

Коефіцієнти при k :

$b_1 = a_1 b_1 - 2a_1 b_2 + a_2 b_1 - 4a_2 b_2$ або $b_1(1 - a_1 - a_2) = -2\, b_2(a_1 + 2\, a_2)$.

Оскільки $1 - a_1 - a_2 = 1 + 0{,}5 - 1{,}5 = 0$, а $a_1 + 2\, a_2 = -0{,}5 + 3 = 2{,}5$, то необхідно покласти $b_2 = 0$.

При $b_2 = 0$ із (9.3.27) можна записати, що

$$b_0 = a_0 + a_1 b_0 - a_1 b_1 + a_2 b_0 - 2a_2 b_1,$$

або

$$b_0(1 - a_1 - a_2) = a_0 - b_1(a_1 + 2a_2).$$

Звідси $b_1 = \dfrac{a_0}{a_1 + 2a_2} = \dfrac{0{,}3}{-0{,}5 + 3} = \dfrac{0{,}3}{2{,}5} = 0{,}12.$

Прирівняємо ще коефіцієнти при $e^{c_2 k}$:

$$b_3 e^{c_2} = a_1 b_3 + a_2 b_3 e^{-c_2} + c_1 e^{c_2},$$

і визначимо константу b_3 :

$$b_3 = \frac{c_1 e^{c_2}}{e^{c_2} - a_1 - a_2 e^{-c_2}} \quad \text{або } b_3 = \frac{c_1}{1 - a_1 e^{-c_2} - a_2 e^{-2c_2}} = \frac{c_1}{1 + 0{,}5 e^{-c_2} - 1{,}5 e^{-2c_2}} .$$

Залишився невизначеним коефіцієнт b_0 , який можна знайти за допомогою початкових умов.

Запишемо повний розв'язок:

$$y(k) = b_0 + A_1 + A_2(-1{,}5)^k + 0{,}12k + \frac{c_1}{1 + 0{,}5 e^{-c_2} - 1{,}5 e^{-2c_2}} e^{c_2(k+1)} + \sum_{i=0}^{k} \alpha_i\, \varepsilon(k-i) ,$$

причому α_i (i = 0, 1, 2, ...) визначаються як і в стохастичному процесі АР(2) згідно формул (*) та (**). Тоді $\alpha_0 = 1$; $\alpha_1 = a_1 = -0{,}5$; $a_2 = 1{,}5$ та при i = (2, 3, ...) $\alpha_i = -0{,}5\alpha_{i-1} + 1{,}5\alpha_{i-2}$.

9.4. Приклади знаходження повних розв'язків різницевих рівнянь

Приклад 9.6. *Модель процесу інфляції.*

Для моделювання процесу інфляції запропоновано різницеве рівняння такого вигляду [14]

$$m(k) = \alpha + p(k) - \beta[p^e(k+1) - p(k)],\ \beta > 0,$$

де $m(k)$ — логарифм номінального забезпечення грошовою масою в момент k; $p(k)$ — логарифм поточного рівня цін; $p^e(k+1)$ — логарифм рівня цін, які очікуються в момент часу $k+1$, α — константа, що враховує вплив неврахованих факторів на інфляцію, β — коефіцієнт, що характеризує вплив приросту цін. Очевидно, що наведене рівняння може бути використане тільки для нормованих безрозмірних змінних.

Перепишемо наведену модель у вигляді

$$m(k) - p(k) = \alpha - \beta[p^e(k+1) - p(k)].$$

Тепер можна стверджувати, що потреба в грошах $[m(k) - p(k)]$ знаходиться у від'ємному взаємозв'язку із приростом цін $[p^e(k) - p(k)]$, тобто відносне перевищення цін над об'ємом грошової маси в лівій частині рівняння вимагає знаку мінус перед приростом цін у правій частині.

Звичайно, що наведена модель є дуже спрощеною для економіки перехідного періоду, тому що для її побудови використано тільки дві змінні. Фактично, на інфляцію у перехідний період впливає ряд інших факторів, а саме: значне падіння рівня виробництва, перекачування валюти за кордон (у той час, як вона повинна використовуватися для збільшення внутрішніх інвестицій), рівень тіньової економіки і т. ін. Така модель буде набагато складнішою за своєю структурою, але адекватнішою процесу інфляції.

Для визначення впливу на процес інфляції майбутніх збурень отримаємо розв'язок рівняння відносно цих значень. Нехай забезпечення грошовою масою має такий простий вигляд:

$$m(k) = m_a + \varepsilon(k),$$

де m_a — середнє значення логарифма номінального забезпечення грошовою масою; $\varepsilon(k)$ — некорельований випадковий процес з нульовим середнім, $E[\varepsilon(k)] = 0$. Покладемо також, що

$$p^e(k + 1) = p(k + 1),$$

тобто прогнозовані ціни відповідають дійсним значенням цін у момент $k + 1$.

Таким чином, рівняння рівноваги грошової маси та рівня цін набуває вигляду

$$m_a + \varepsilon(k) - p(k) = \alpha - \beta[p(k + 1) - p(k)]$$

або

$$p(k+1)-\left(1+\frac{1}{\beta}\right)p(k)=-(m_a-\alpha)/\beta-\varepsilon(k)/\beta.$$

Знайдемо розв'язок цього рівняння. Для знаходження частинного розв'язку скористаємось методом невизначених коефіцієнтів. Оскільки перед змінною $p(k)$ стоїть коефіцієнт $a_1 = 1 + 1/\beta > 1$, то будемо шукати розв'язок відносно майбутніх відліків збурень. Шуканий розв'язок запишемо у вигляді:

$$p_{\text{ш}}(k)=b_0+\sum_{i=0}^{n}\gamma_i\,\varepsilon(k+i),$$

де n — велике число.

Підставимо його у наведене рівняння інфляції для того, щоб знайти невідомі коефіцієнти b_0, γ_i:

$$b_0+\sum_{i=0}^{k}\gamma_i\,\varepsilon(k+1+i)-(1+1/\beta)\left[b_0+\sum_{i=0}^{n}\gamma_i\,\varepsilon(k+i)\right]=\left[\alpha-m_a-\varepsilon(k)\right]/\beta.$$

Для того щоб останнє рівняння було тотожністю для всіх можливих ненульових значень $\{\varepsilon(k)\}$, необхідно прирівняти коефіцієнти при однакових значеннях $\varepsilon(k + i)$, $i = 0, 1, 2, ...$:

– константи: $b_0 - b_0(1 + \beta)/\beta = (\alpha - m_a)/\beta \Rightarrow b_0 = m_a - \alpha$;

– коефіцієнти при $\varepsilon(k)$: $-\gamma_0(1 + \beta)/\beta = -1/\beta \Rightarrow \gamma_0 = 1/(1 + \beta)$;

– коефіцієнти при $\varepsilon(k + 1)$: $\gamma_0 - \gamma_1(1 + \beta)/\beta = 0 \Rightarrow \gamma_1 = \beta/(1 + \beta)^2$;

– коефіцієнти при $\varepsilon(k + 2)$: $\gamma_1 - \gamma_2(1 + \beta)/\beta = 0 \Rightarrow \gamma_2 = \beta^2/(1 + \beta)^3$;

...

– коефіцієнти при $\varepsilon(k + i)$: $\gamma_{i-1} - \gamma_i(1 + \beta)/\beta = 0 \Rightarrow \gamma_i = \beta^i/(1 + \beta)^{i+1}$.

Частинний розв'язок у компактній формі матиме вигляд

$$p^p(k) = m_a - \alpha + \frac{1}{\beta}\sum_{i=0}^{n}\left(\frac{\beta}{1+\beta}\right)^{i+1}\varepsilon(k+i).$$

Для однорідного рівняння першого порядку

$$p(k+1) - \left(1+\frac{1}{\beta}\right)p(k) = 0,$$

загальний розв'язок однорідного рівняння запишемо у вигляді

$$p^h(k) = A\left(1+\frac{1}{\beta}\right)^k.$$

Таким чином, загальний розв'язок, як сума частинного та загального розв'язку однорідного рівняння, набуває наступного вигляду

$$p(k) = m_a - \alpha + \frac{1}{\beta}\sum_{i=0}^{n}\left(\frac{\beta}{1+\beta}\right)^{i+1}\varepsilon(k+i) + A\left(1+\frac{1}{\beta}\right)^k.$$

Скористаємось початковими умовами для визначення сталої A:

$$p_0 = m_a - \alpha + \frac{1}{\beta}\sum_{i=0}^{n}\left(\frac{\beta}{1+\beta}\right)^{i+1}\varepsilon(i) + A,$$

або

$$A = p_0 - m_a + \alpha - \frac{1}{\beta}\sum_{i=0}^{n}\left(\frac{\beta}{1+\beta}\right)^{i+1}\varepsilon(i).$$

Тоді, для того щоб $A = 0$, випливає:

$$p_0 = m_a - \alpha + \frac{1}{\beta}\sum_{i=0}^{n}\left(\frac{\beta}{1+\beta}\right)^{i+1}\varepsilon(i).$$

У цьому рівнянні $(m_a - \alpha)$ — детермінована умова довгострокової "*рівноваги*", яка зустрічається також в інших моделях макроекономіки. Якщо розв'язок початкового рівняння (моделі інфляції) існує, то він повинен прямувати до цієї детермінованої частини. Складова розв'язку

$$\frac{1}{\beta}\sum_{i=0}^{n}\left(\frac{\beta}{1+\beta}\right)^{i+1}\varepsilon(k+1)$$

ілюструє вплив збурень на величину грошової маси в обороті. Якщо прийняти до уваги те, що

$$\left|\frac{\beta}{1+\beta}\right|<1,$$

то вплив цієї складової на поведінку розв'язку загалом буде помітним тільки на короткому початковому проміжку часу (при відносно невеликих значеннях k), оскільки коефіцієнти зменшуватимуться з ростом значення степеня $i + 1$.

Зі сказаного можна зробити висновок, що частинний розв'язок характеризує поведінку (рівновагу) головної змінної $p(k)$ на короткому та більш значному інтервалах часу. Однорідний розв'язок можна розглядати як порушення рівноваги у початковий період часу.

Оскільки частинний розв'язок

$$p^p(k)=m_a-\alpha+\frac{1}{\beta}\sum_{i=0}^{n}\left(\frac{\beta}{1+\beta}\right)^{i+1}\varepsilon(k+i)$$

відображає рівновагу між ціною та грошовою масою на k-й момент часу, то початкове значення

$$p_0=m_a-\alpha+\frac{1}{\beta}\sum_{i=0}^{n}\left(\frac{\beta}{1+\beta}\right)^{i+1}\varepsilon(i)$$

є рівноважним значенням ціни для нульового моменту. Таким чином, умова

$$A(1 + 1/\beta)^k = 0$$

є умовою нульового відхилення від рівноваги у початковий момент часу.

Якщо на послідовність $\{p(k)\}$ накласти обмеження, що вона не повинна розбігатися, то загальний розв'язок матиме вигляд

$$p(k)=m_a-\alpha+\frac{1}{\beta}\sum_{i=0}^{n}\left(\frac{\beta}{1+\beta}\right)^{i+1}\varepsilon(k+i).$$

Зазначимо, що у кожний момент часу k ціна пропорційна середньому значенню грошової маси в обороті. Це можна легко показати, оскільки всі змінні задані в логарифмічному масштабі і частинна похідна

$$\frac{\partial p(k)}{\partial m}=1.$$

Короткочасні зміни надходжень грошової маси в оборот ведуть себе так, що наслідком останнього розв'язку є зростання грошової

маси в майбутньому, яке представлене значеннями $\varepsilon(k+1)$ і викликає зростання цін у поточний період часу. Але основна ідея полягає в тому, що збільшення випуску грошової маси у майбутньому буде сприяти також збільшенню цін у майбутньому, а звідси випливає, що гроші невигідно тримати без руху, вони повинні бути в обороті.

Приклад 9.7. Знайти загальний розв'язок рівняння

$$y(k) = -2y(k-1) - y(k-2) + k(-1)^k. \qquad (9.4.1)$$

Запишемо характеристичне рівняння

$$\lambda^2 + 2\lambda + 1 = 0$$

і знайдемо його корені: $\lambda_{1,2} = \frac{1}{2}\left(-2 \pm \sqrt{4-4}\right)$, тобто $\lambda_1 = \lambda_2 = -1$. Запишемо однорідний розв'язок із врахуванням наявності кратного кореня:

$$y^h(k) = A_1(-1)^k + A_2 k(-1)^k.$$

Якщо складові збурюючої функції у правій частині рівняння (9.4.1), тобто $x(k) = k(-1)^k$, мають такий самий вигляд, як і складові $y^h(k)$, то шукана функція частинного розв'язку змінюється. Усі складові $y^p(k)$, які відповідають таким самим складовим $y^h(k)$, необхідно помножити на k у такому найменшому степені, щоб їх тотожність не порушувалась. Таким чином, шукана функція для частинного розв'язку буде:

$$y^p_{\text{ш}}(k) = \left[c_2 k + c_3 k^2\right] k(-1)^k = \left[c_2 k^2 + c_3 k^3\right](-1)^k. \qquad (9.4.2)$$

Підставимо (9.4.2) у (9.4.1) і знайдемо невідомі коефіцієнти c_2, c_3 за допомогою методу невизначених коефіцієнтів. Тоді отримаємо:

$$\left[c_2 k^2 + c_3 k^3\right](-1)^k = -2\left[c_2(k-1)^2 + c_3(k-1)^3\right](-1)^{(k-1)} -$$
$$-\left[c_2(k-2)^2 + c_3(k-2)^3\right](-1)^{(k-2)} + k(-1)^k$$

або

$$\left[c_2 k^2 + c_3 k^3\right](-1)^k = 2\left[c_2(k-1)^2 + c_3(k-1)^3\right](-1)^k -$$
$$-\left[c_2(k-2)^2 + c_3(k-2)^3\right](-1)^k + k(-1)^k.$$

Звідси знайдемо значення невідомих коефіцієнтів $c_2 = 1/2$; $c_3 = 1/6$ і повний розв'язок:

$$y(k)=\left[c_0+c_1k+\frac{1}{2}k^2+\frac{1}{6}k^3\right](-1)^k.$$

Приклад 9.8. Розглянемо дискретне рівняння з гармонічним збуренням:

$$y(k) = a_0 + a_1y(k - 1) + A\sin(\omega k),$$

де A і ω — відомі значення амплітуди і кругової частоти вхідного сигналу.

Для знаходження частинного розв'язку виберемо його у вигляді:

$$y^p_{\text{ш}}(k) = c_0 + c_1k + c_2\sin(\omega k + \varphi),$$

де φ — зсув фази сигналу при проходженні через лінійну систему. Скористаємось тригонометричною тотожністю

$$\sin(\alpha + \beta) = \sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta)$$

і підставимо шукану функцію в рівняння процесу:

$$c_0+c_1k+c_2\left[\sin(\omega k)\cos(\varphi)+\cos(\omega k)\sin(\varphi)\right]=a_0+a_1\left\{c_0+c_1(k-1)+\right.$$
$$\left.+c_2\left[\sin\left(\omega(k-1)\right)\cos(\varphi)+\cos\left(\omega(k-1)\right)\sin(\varphi)\right]\right\}+A\sin(\omega k).$$

Введемо позначення: $\cos(\varphi) = b_1$, $\sin(\varphi) = b_2$, $\sin(\omega) = b_3$, $\cos(\omega) = b_4$ і в результаті отримаємо тотожності:

$$c_0 = a_0 + a_1c_0 - a_1c_1;\ c_1 = a_1c_1;$$
$$c_2b_1 = c_2a_1b_1b_4 + c_2a_1b_2b_3 + A;$$
$$c_2b_2 = -c_2a_1b_1b_3 + c_2a_1b_2b_4.$$

Якщо $a_1 \neq 1$, то $c_1 = 0$, і з першої тотожності знайдемо, що $c_0 = a_0/(1 - a_1)$. За допомогою третьої можна визначити коефіцієнт c_2:

$$c_2 = \frac{A}{b_1 - a_1b_1b_4 - a_1b_2b_3}.$$

Оскільки коефіцієнти b_3 і b_4 відомі, а $b_1^2 + b_2^2 = 1$, то з четвертої тотожності можна знайти b_1 і b_2, і фазовий кут φ.

Таким чином, при $a_1 \neq 1$ загальний розв'язок має вигляд:

$$y(k)=\frac{a_0}{1-a_1}+Ba_1^k+\frac{A}{b_1-a_1b_1b_4-a_1b_2b_3}\sin(\omega k+\varphi).$$

Приклад 9.9. Знайдемо загальний розв'язок рівняння третього порядку АРКС(3, 2):

$$y(k) = a_0 + a_1 y(k-1) + a_2 y(k-2) + a_3 y(k-3) + \varepsilon(k) + \\ + \beta_1 \varepsilon(k-1) + \beta_2 \varepsilon(k-2)$$

при таких значеннях параметрів $a_0 = 0{,}5$; $a_1 = 1$; $a_2 = 0{,}25$; $a_3 = -0{,}25$; $\beta_1 = -0{,}125$; $\beta_2 = 0{,}125$.

Запишемо характеристичне рівняння:

$$\lambda^3 - \lambda^2 - 0{,}25\lambda + 0{,}25 = 0$$

і знайдемо його корені: $\lambda_1 = 1$; $\lambda_2 = -0{,}5$; $\lambda_3 = 0{,}5$ (процес з одиничним коренем). Загальний розв'язок однорідного рівняння має вигляд

$$y^h(k) = A_1 + A_2(-0{,}5)^k + A_3(0{,}5)^k.$$

Для знаходження частинного розв'язку виберемо шукану функцію

$$y^p_{\text{ш}}(k) = b_0 + b_1 k + \sum_{i=0}^{k-1} \alpha_i \, \varepsilon(k-i)$$

і підставимо її у рівняння процесу, скориставшись загальними позначеннями для коефіцієнтів:

$$\begin{aligned} & b_0 + b_1 k + \alpha_0 \varepsilon(k) + \alpha_1 \varepsilon(k-1) + \alpha_2 \varepsilon(k-2) + \alpha_3 \varepsilon(k-3) + \ldots = \\ & = a_0 + a_1 \big[b_0 + b_1 (k-1) + \alpha_0 \varepsilon(k-1) + \alpha_1 \varepsilon(k-2) + \alpha_2 \varepsilon(k-3) + \\ & + \alpha_3 \varepsilon(k-4) + \ldots \big] + a_2 \big[b_0 + b_1 (k-2) + \alpha_0 \varepsilon(k-2) + \alpha_1 \varepsilon(k-3) + \\ & + \alpha_2 \varepsilon(k-4) + \alpha_3 \varepsilon(k-5) + \ldots \big] + a_3 \big[b_0 + b_1 (k-3) + \alpha_0 \varepsilon(k-3) + \\ & + \alpha_1 \varepsilon(k-4) + \alpha_2 \varepsilon(k-5) + \alpha_3 \varepsilon(k-6) + \ldots \big] + \varepsilon(k) + \beta_1 \varepsilon(k-1) + \beta_2 \varepsilon(k-2). \end{aligned}$$

Прирівняємо константи у лівій і правій частинах, а також коефіцієнти при однакових змінних:

$$b_0 = a_0 + a_1 b_0 - a_1 b_1 + a_2 b_0 - 2a_2 b_1 + a_3 b_0 - 3a_3 b_1;$$

при k: $b_1 = a_1 b_1 + a_2 b_1 + a_3 b_1$;

при $\varepsilon(k)$: $\alpha_0 = 1$;

при $\varepsilon(k-1)$: $\alpha_1 = a_1 \alpha_0 + \beta_1$;

при $\varepsilon(k-2)$: $\alpha_2 = a_1 \alpha_1 + a_2 \alpha_0$;

при $\varepsilon(k-3)$: $\alpha_3 = a_1 \alpha_2 + a_2 \alpha_1 + a_3 \alpha_0$.

Сталі b_0 і b_1 визначаються за допомогою початкових умов. Тоді загальний розв'язок має вигляд:

$$y(k) = A + A_2(-0{,}5)^k + A_3(0{,}5)^k + b_1 k + \sum_{i=0}^{k} \alpha_i \varepsilon(k-i),$$

де $A = A_1 + b_0$, $\alpha_i = \alpha_{i-1} + 0{,}25\,\alpha_{i-2} - 0{,}25\,\alpha_{i-3}$.

9.5. Контрольні питання і вправи

1. Поясніть, для чого необхідно знаходити розв'язки різницевих рівнянь. Чи можна скористатись моделлю АРКС(p, q) для аналізу поведінки процесу?
2. Який процес називають процесом з одиничними коренями? Які синоніми цієї назви? Дайте визначення тренду. Що означає нестаціонарність процесу?
3. На яку частину розв'язку впливають початкові умови? У чому проявляється вплив початкових умов? Яким чином можна задати початкові умови?
4. Що характеризують перші і другі різниці? Запишіть рівняння АР(2) через перші різниці; для розв'язання якої задачі можна скористатись цією моделлю?
5. Яким чином можна звести однорідний розв'язок до нуля? Що дасть на практиці зведення однорідного розв'язку до нуля?
6. У якому випадку неоднорідний розв'язок дорівнює нулю?
7. Знайдіть однорідні розв'язки рівнянь:

   $y(k) = 0{,}35 - 0{,}75y(k-1) + 0{,}25y(k-2) + \varepsilon(k)$;

   $y(k) = a_0 + 1{,}5y(k-1) - 0{,}5y(k-2) + \varepsilon(k)$;

   $y(k) = a_0 + y(k-2) + \varepsilon(k)$;

   $y(k) = a_0 + 2y(k-1) - y(k-2) + \varepsilon(k)$;

   $y(k) = a_0 + y(k-1) + 0{,}25y(k-2) - 0{,}25y(k-3) + \varepsilon(k)$.
8. Запишіть загальний вигляд однорідного рівняння для випадку, коли характеристичне рівняння містить трикратний корінь.
9. У яких випадках однорідний розв'язок містить гармонійну складову? Чим визначається амплітуда, частота і фаза коливань у даному випадку?
10. Що таке кругова частота? Чим зручне використання кругової частоти у порівнянні з лінійною при аналізі поведінки процесів?

11. У чому полягає подібність та відмінності розв'язків однорідних різницевих та диференціальних рівнянь? Покажіть це на прикладі рівняння другого порядку.
12. Запишіть у загальному вигляді розв'язок наведеного рівняння:
$$y(k) = a_0 + a_1 y(k-1) + a_2 y(k-2) + a_3 y(k-3) + b_1 \varepsilon(k-1) + \varepsilon(k).$$
Нехай характеристичне рівняння має у даному випадку один дійсний і два кратних корені. Виконайте аналіз отриманого розв'язку.
13. Накресліть графік однорідного розв'язку для випадку двократного кореня характеристичного рівняння.
14. Запишіть у загальному вигляді розв'язок однорідного рівняння довільного порядку з m-кратним коренем характеристичного рівняння. Який вигляд матиме графік цього розв'язку?
15. Запишіть у загальному вигляді розв'язок однорідного рівняння, якщо його характеристичне має два дійсних різних корені і комплексно спряжену пару коренів.

Розділ 10

ФАКТОРНИЙ АНАЛІЗ

10.1. Завдання факторного аналізу

Завдання факторного аналізу полягають у *виявленні ступеня впливу на залежну змінну інших змінних* (*факторів*).

Виконання факторного аналізу статистичних (експериментальних) даних потребує розв'язання наступних задач:

- задача робастного оцінювання параметрів регресійних рівнянь;
- узагальнення отриманих результатів;
- визначення інформативних факторів для моделі;
- задача повороту для вибору однієї матриці вагових коефіцієнтів **A** із множини можливих;
- оцінювання значень факторів;
- побудова моделей динаміки досліджуваних процесів.

Розглянемо матрицю вимірів розмірністю $p \times N$, де p — кількість ознак (незалежних змінних); N — загальна кількість вимірів. Елементи матриці x_{ij} є j-м значенням i-ї змінної. Наприклад, x_1 — продуктивність праці (обсяг продукції, яка виробляється одним робітником за одиницю часу); x_2 — енерговитрати на виробництво за таку саму одиницю часу.

У класичному факторному аналізі наступним етапом обробки даних є знаходження матриці *нормованих значень вихідних* (початкових) *даних*, яку отримують шляхом нормування елементів матриці вимірів **X**. Однак сучасні дослідження свідчать, що необхідно враховувати ще можливе зашумлення (забруднення) даних досить грубими похибками, що надзвичайно негативно впливає на остаточний результат аналізу. Тому на сьогодні ці дві матриці (вихідні виміри та їх нормовані значення) об'єднує загальна проблема — *проблема робастності оцінювання* параметрів моделей (**це перша задача факторного аналізу**). Технічно ця задача розв'язується завдяки застосуванню методу *робастного* (стійкого) оцінювання вибіркового середнього $\bar{x}$ та середньоквадратичного відхилення s.

При цьому до даних застосовується так званий тест на "забруднення". Якщо забруднення нема, то $\bar{x}$ і s розраховуються так само, як

і у класичному випадку. Якщо ж забруднення виявлено, то застосовують тест на симетричність розподілу.

До симетричних забруднених розподілів застосовують методи робастного оцінювання $\bar{x}$ і s. Для асиметричних розподілів цей підхід не придатний. У випадку асиметричного розподілу застосовують оцінювання джекнайф (jack-knife), яке враховує наявність асиметрії [2; 4].

Введемо такі позначення для подальшого використання:

- **Y** — матриця нормованих значень змінних (вихідних ознак) розмірності ($p \times N$), тобто її розмірність збігається з розмірністю вихідної матриці **X**, але значення її елементів вже нормовані, тобто безрозмірні;
- **R** ($p \times p$) — симетрична матриця парних кореляцій (кореляційна матриця); вона має одиничні елементи головної діагоналі, які відповідають одиничним дисперсіям нормованих змінних;
- $\mathbf{R}_h$ ($p \times p$) — симетрична редукована кореляційна матриця, яка є основою для факторного аналізу (**її визначення — це друга задача**); на її головній діагоналі замість одиниць знаходяться так звані узагальнюючі значення (узагальнення) $h_j^{\#}$, які необхідно визначити у процесі аналізу даних;
- **A** — матриця вагових коефіцієнтів, які є характеристиками стохастичного зв'язку між вихідними ознаками та узагальненими факторами.

Між знаходженням матриць $\mathbf{R}_h$ і **A** є **третя задача** — *знаходження факторів*, яка включає у себе задачі визначення кількості загальних факторів та їх вигляду. Значення вагових коефіцієнтів є координати ознак на нових осях координат, якими і є загальні фактори.

Вибір матриці **A** — **це четверта задача**. Загальні фактори займають довільне положення відносно вимірів ознак, які утворюють первісну конфігурацію векторів вимірів. Таким чином, матрицю $\mathbf{R}_h$ можна отримати за допомогою значної кількості матриць вагових коефіцієнтів **A**, тобто розв'язок задачі не є однозначним.

Який критерій застосовують для вибору цієї матриці? Існує декілька підходів до вибору **A**, але найбільш сучасним є *підхід на основі принципу простої структури Терстоуна* [7]. Цей підхід дає можливість визначити матрицю $\mathbf{A}^* = (a_{jr}^*)$ розмірністю ($p \times m$), елементи якої (тобто вагові коефіцієнти) отримані після виконання операції повороту.

- **F** — ($m \times N$) — матриця індивідуальних значень факторів для кожної змінної (об'єкта дослідження).

Для отримання цієї матриці використані нові інтегральні одиниці виміру для кожної змінної.

П'ята задача — оцінювання значень факторів, що включає в себе перехід від матриці $\mathbf{A}^*$ до матриці **F**.

Розв'язком згаданих п'яти задач *закінчується статичний варіант факторного аналізу.*

Якщо виконувати факторний аналіз протягом кількох років підряд і потім порівняти отримані результати, то отримаємо динаміку процесу факторного аналізу — **це шоста задача факторного аналізу**. Шоста задача стосується моделювання динаміки досліджуваних за таким методом процесів. За допомогою динамічних моделей можна виявити ті ознаки (змінні), вплив яких буде зростати або зменшуватись у майбутньому.

10.2. Основна модель факторного аналізу

Нехай є декілька змінних (ознак), які характеризують, наприклад, діяльність підприємства:

x_1 — продуктивність праці на підприємстві;

x_2 — фондовіддача;

⋮

x_n — собівартість продукції на підприємстві.

Значення змінних x_1 , x_2 , ..., x_n відомі для кожного з N підприємств. Вихідну інформацію можна представити у вигляді матриці $\mathbf{X} = (x_{ij})$ розмірності ($n \times N$). Кожний рядок матриці буде значенням одного показника для всіх підприємств. Припустимо, що кожний елемент цієї матриці (x_{ij}) є результатом впливу деякого числа m загальних гіпотетичних факторів і одного характерного фактора, тобто можна записати рівняння вигляду:

$$x_{ij} = a'_{j1} f'_{1i} + a'_{j2} f'_{2i} + ... + a'_{jm} f'_{mi} + d'_j \nu'_{ji}, \qquad (10.2.1)$$
$$j = 1, ..., n;\ i = 1, ..., N;\ r = 1, ..., m,$$

де x_{ij} — центроване значення j-го показника (змінної) для i-го об'єкта дослідження (тобто i-го підприємства); f'_{ri} — значення r-го загального фактора на i-му об'єкті дослідження (тобто на i-му підприємстві); ν'_{ji} — значення j-го характерного фактора для i-го об'єкта досліджен-

ня; a'_{jr} — ваговий коефіцієнт j-ї змінної для r-го загального фактора (або навантаження j-ї змінної з боку r-го загального фактора); d'_j — ваговий коефіцієнт (або навантаження) j-го характерного фактора у рівнянні для j-ї змінної.

Як правило, припускають, що характерні фактори не корельовані між собою, а також не корельовані із загальними факторами.

Фактори, які пов'язані значущими ваговими коефіцієнтами більше, ніж з однією змінною, називають *загальними*.

Загальний фактор, який пов'язаний значущими ваговими коефіцієнтами з усіма змінними, називають *генеральним*.

У рівнянні (11.2.1) змінні x_{ij} мають розмірність. Для того щоб перейти до безрозмірних змінних, значення центрують і нормують:

$$y_{ji} = \frac{x^*_{ji} - \bar{x}_j}{s_j} = \frac{x_{ji}}{s_j}, \qquad (10.2.2)$$

де y_{ji} — нормоване значення j-ї змінної для i-го об'єкта; x^*_{ij} — вихідне (початкове) значення змінної (тобто оригінальний вимір); $\bar{x}_j$ — середнє значення j-ї змінної; s_j — вибіркове середньоквадратичне відхилення j-ї змінної, яке розраховується за формулою

$$s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{N} x^2_{ji}}\,. \qquad (10.2.3)$$

Після перетворення вимірів (вихідних даних) за формулою (10.2.2) вибіркові дисперсії змінних будуть дорівнювати одиниці. Таким чином, замість рівняння (11.2.1) надалі буде використовуватись таке рівняння:

$$y_{ji} = a_{j1} f_{1i} + a_{j2} f_{2i} + ... + a_{jm} f_{mi} + d_j \nu_{ji}, \qquad (10.2.4)$$
$$j = 1, ..., n;\ i = 1, ..., N;\ r = 1, ..., m\,.$$

Середні значення змінних $\bar{y}_j = 0$, а їх вибіркові дисперсії $s^2_{y_j} = 1$. Вирази (10.2.1) і (10.2.4) нагадують регресійні рівняння. Дійсно, залежна змінна y_{ji} описується за допомогою m інших змінних (факторів) плюс залишок: $d_j \nu_{ji}$. Однак різниця між регресійним рівнянням і рівнянням (10.2.4) полягає у тому, що у регресійному аналізі незалежні змінні у правій частині безпосередньо вимірюються (або є статистичні дані).

Але вирази (10.2.1) і (10.2.4) призначені для розв'язання іншої задачі — у факторному аналізі загальні фактори f_r, а також індивіду-

альні (характерні) фактори ν_{ji} — це гіпотетичні змінні, які потрібно оцінити.

Із виразу (10.2.4) випливає, що спостережувані значення змінних — це лінійні комбінації неспостережуваних, гіпотетичних факторів, тобто їх вимірів нема. Таким чином, метою будь-якого методу факторного аналізу є така задача: представити елемент матриці нормованих вимірів **Y** у вигляді лінійної комбінації деякої кількості m, загальних факторів і одного характерного фактора.

Представимо тепер вираз (10.2.4) у матричній формі. У матричній формі вагові коефіцієнти a_{jr} утворюють прямокутну матрицю $\mathbf{A}[n \times (m + n)]$:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} & 0 & 0 & \dots & 0 \\ a_{21} & a_{22} & \dots & a_{2m} & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nm} & 0 & 0 & \dots & 0 \end{bmatrix}, \quad (10.2.5)$$

а коефіцієнти d_j утворюють прямокутну квазідіагональну матрицю $\mathbf{D}[n \times (m + n)]$ виду:

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & \dots & 0 & d_1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & d_2 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & d_n \end{bmatrix}. \quad (10.2.6)$$

Сума матриць (10.2.5) і (10.2.6) дає матрицю **M**:

$$\mathbf{M} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} & d_1 & 0 & \dots & 0 \\ a_{21} & a_{22} & \dots & a_{2m} & 0 & d_2 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nm} & 0 & 0 & \dots & d_n \end{bmatrix}. \quad (10.2.7)$$

Матриці значень загальних і характерних факторів мають вигляд:

$$\mathbf{F} = \begin{bmatrix} f_{11} & f_{12} & \dots & f_{1N} \\ f_{21} & f_{22} & \dots & f_{2N} \\ \dots & \dots & \dots & \dots \\ f_{m1} & f_{m2} & \dots & f_{mN} \end{bmatrix}; \quad \mathbf{V} = \begin{bmatrix} \nu_{11} & \nu_{12} & \dots & \nu_{1N} \\ \nu_{21} & \nu_{22} & \dots & \nu_{2N} \\ \dots & \dots & \dots & \dots \\ \nu_{n1} & \nu_{n2} & \dots & \nu_{nN} \end{bmatrix}. \quad (10.2.8)$$

Значення факторів у зведеній таблиці:

$$\mathbf{F}^* = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1N} \\ f_{21} & f_{22} & \cdots & f_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ f_{m1} & f_{m2} & \cdots & f_{mN} \\ \nu_{11} & \nu_{12} & \cdots & \nu_{1N} \\ \nu_{21} & \nu_{22} & \cdots & \nu_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ \nu_{n1} & \nu_{n2} & \cdots & \nu_{nN} \end{bmatrix}. \qquad (10.2.9)$$

Тепер модель (10.2.4) можна представити у вигляді:

$$\mathbf{Y} = \mathbf{M}\,\mathbf{F}^*, \qquad (10.2.10)$$

де $\mathbf{F}^*$ — матриця розмірності $[(m + n) \times N)]$; матриця $\mathbf{Y}[n \times N]$; $\mathbf{M}$ — матриця факторного відображення розмірністю $[n \times (m + n)]$, яка включає навантаження загальних і характерних факторів (її називають ще *повною факторною матрицею*); $\mathbf{F}^*$ $[(m + n) \times N)]$ — матриця значень загальних і характерних факторів для всіх об'єктів дослідження.

З трьох матриць рівняння (10.2.10) невідомі матриці $\mathbf{M}$ і $\mathbf{F}^*$.

10.3. Аналіз дисперсії вимірів у факторному аналізі

З теорії вимірів відомо, що вимірювана величина містить щонайменше дві компоненти: *істинне значення* і *випадкову похибку вимірів*:

$$x_{ji} = x^a_{ji} + \varepsilon_{ji}, \qquad (10.3.1)$$

де x_{ji} — вимір j-ї змінної i-го об'єкта; x^a_{ji} — фактичне (істинне) значення вимірюваної змінної; ε_{ji} — похибка виміру j-ї змінної i-го об'єкта.

Якщо ж виміри ведуться в біології, психології, економіці, медицині і т. ін., то у вимірах з'являється третя компонента, яка залежить від варіабельності досліджуваної змінної на об'єктах даного класу. Таким чином, вимір змінної стає сумою трьох складових:

$$x_{ji} = x^a_{ji} + x^s_{ji} + \varepsilon_{ji}, \qquad (10.3.2)$$

де x^s_{ji} — варіативне значення вимірюваної змінної для i-го об'єкта дослідження. Необхідно встановити, що є істинним значенням вимірюваної змінної і яка величина складової x^a_{ji}. Мова може йти про математичне сподівання, оскільки для випадкових величин x^a_{ji} — матема-

тичне сподівання досліджуваної змінної, а дві інші компоненти характеризують відхилення від математичного сподівання.

Похибка вимірів є, як правило, значно меншою варіативної компоненти, а тому їх часто об'єднують. Оскільки варіативна складова і похибки вимірів виникають незалежно одна від одної, то їх легко обчислити окремо.

Якщо перша компонента x_{ji}^{a} — це загальна статистична характеристика сукупності досліджуваних об'єктів, то друга і третя компоненти характеризують відхилення значень окремої змінної від середнього. Вони відображають особливості, які характерні для даного об'єкта та застосованого методу вимірювання.

Простою характеристикою цих особливостей є різниця:

$$x_{ji} - x_{ji}^{a} = x_{ji}^{s} + \varepsilon_{ji}. \tag{10.3.3}$$

Однак при досліджуванні множини об'єктів використовують такі узагальнені характеристики, як дисперсія σ_j^2 та/або середньоквадратичне (стандартне) відхилення σ_j.

Розглянемо компоненти дисперсії у факторному аналізі; для цього повернемось до виразу (10.2.4):

$$y_{ji} = a_{j1} f_{1i} + a_{j2} f_{2i} + ... + a_{jm} f_{mi} + d_j \nu_{ji},$$
$$j = 1, ..., n;\ i = 1, ..., N;\ r = 1, ..., m.$$

Оскільки y_{ji} — нормована випадкова велична, то її дисперсія дорівнює одиниці і є сумою квадратів значень показників (змінних) по всіх досліджуваних об'єктах, яка ділиться на кількість об'єктів N або $N - 1$ (із врахуванням зміщення статистичної характеристики):

$$\hat{s}_j^2 = 1 = \frac{1}{N}\sum_{i=1}^{N} y_{ji}^2 = \frac{1}{N}\Bigg[a_{j1}^2 \sum_{i=1}^{N} f_{1i}^2 + a_{j2}^2 \sum_{i=1}^{N} f_{2i}^2 + \ldots + a_{jm}^2 \sum_{i=1}^{N} f_{mi}^2 +$$
$$+ d_j^2 \sum_{i=1}^{N} \nu_{ji}^2 + 2\Bigg(a_{j1}a_{j2} \sum_{i=1}^{N} f_{1i}f_{2i} + a_{j1}a_{j3} \sum_{i=1}^{N} f_{1i}f_{3i} + \ldots + \tag{10.3.4}$$
$$+ a_{j(m-1)}a_{jm} \sum_{i=1}^{N} f_{(m-1)i}f_{mi} + a_{jm}d_j \sum_{i=1}^{N} f_{mi}\nu_{ji} \Bigg)\Bigg],$$
$$j = 1, \ldots, n;\ i = 1, \ldots, N;\ r = 1, \ldots, m.$$

Розглянемо доданки, які містять співмножник $\frac{1}{N}\sum_{i=1}^{N} f_{ri}^2$. У даному випадку величина

$$\frac{1}{N}\sum_{i=1}^{N} f_{ri}^2 = \hat{s}_{f_r}^2 = 1$$

є дисперсією нормованого загального фактора f_r і дорівнює одиниці завдяки його нормуванню, тобто

$$f_{ri} = \frac{f_{ri}^* - \overline{f}_r}{s_{f_r}},$$

де f_{ri}^* — ненормоване значення загального фактора; $\overline{f}_r$ — середнє значення ненормованого фактора; s_{f_r} — середньоквадратичне відхилення ненормованого загального фактора.

Тепер розглянемо в (10.3.4) доданки, які містять співмножник $\frac{1}{N}\sum_{i=1}^{N} f_{ri} f_{li}$; це коефіцієнт кореляції між двома факторами, тобто

$$\frac{1}{N}\sum_{i=1}^{N} f_{ri} f_{li} = r_{f_r f_l},$$

де $r = 1, ..., m$; $l = 1, ..., m$; $r \neq l$.

Очевидно, що $\frac{1}{N}\sum_{i=1}^{N} f_{mi}\nu_{ji} = r_{f_m \nu_j}$.

Таким чином, вираз (10.3.4) можна представити у вигляді:

$$\hat{s}_j^2 = a_{j1}^2 + a_{j2}^2 + \ldots + a_{jm}^2 + d_j^2 + 2\left(a_{j1}a_{j2}r_{f_1 f_2} + a_{j1}a_{j3}r_{f_1 f_3} + \ldots + a_{jm}d_j r_{f_m \nu_j}\right),$$

а звідси виходить, що

$$\hat{s}_j^2 = 1 = \sum_{r=1}^{m} a_{jr}^2 + d_j^2 + 2\sum_{l>r=1}^{m} a_{jr}a_{jl}r_{f_r f_l} + 2d_j\sum_{r=1}^{m} a_{jr}r_{f_r \nu_j}. \qquad (10.3.5)$$

Оскільки характерний фактор відноситься тільки до однієї j-ї змінної і завжди не корельований із загальними факторами, то $r_{f_r \nu_j} = 0$, а тому вираз (10.3.5) можна спростити:

$$\hat{s}_j^2 = \sum_{r=1}^{m} a_{jr}^2 + d_j^2 + 2\sum_{l>r=1}^{m} a_{jr}a_{jl}r_{f_r f_l}. \qquad (10.3.6)$$

Подальше спрощення досягається завдяки врахуванню некорельованості загальних факторів (вони повинні бути некорельованими), тобто при $r_{f_r f_l} = 0$ маємо:

$$\hat{s}_j^2 = d_j^2 + \sum_{r=1}^{m} a_{jr}^2. \qquad (10.3.7)$$

Таким чином, дисперсія змінної y_j визначається сумою відносних вкладів у дисперсію цієї змінної кожного із m загальних і одного характерного фактора.

Вираз

$$\sum_{r=1}^{m} a_{jr}^2 = h_j^2 \qquad (10.3.8)$$

називають *узагальненням показника* y_j, тобто узагальнення показника — це сума відносних вкладів усіх m загальних факторів у дисперсію змінної y_j. Вклад характерного фактора v_j у дисперсію змінної y_j визначається доданком d_j^2.

Дисперсія характерного фактора складається з двох компонент: компоненти, пов'язаної із специфікою цього фактора S_j, і компоненти, пов'язаної із похибками вимірів E_j. Якщо складова, пов'язана із специфікою S_j, некорельована із складовою E_j, то модель факторного аналізу приймає вигляд:

$$y_j = a_{j1}f_1 + a_{j2}f_2 + \ldots + a_{jm}f_m + b_jS_j + c_jE_j. \qquad (10.3.9)$$

Вклад характерного фактора у дисперсію ознаки (змінної) можна представити так:

$$d_j^2 = b_j^2 + c_j^2. \qquad (10.3.10)$$

Якщо вилучити із дисперсії ознаки складову похибки, то отримаємо характеристику, яку називають *надійністю*:

$$r_j^2 = h_j^2 + b_j^2. \qquad (10.3.11)$$

У табл. 10.1 наведено формули, які визначають вклади факторів у дисперсію ознаки.

Таблиця 10.1

Формули для складових дисперсії

Характеристика	Позначення	Вираз для розрахунку
Повна дисперсія	s_j^2	$h_j^2 + b_j^2 + c_j^2 = 1$
Надійність	r_j^2	$h_j^2 + b_j^2 = 1 - c_j^2$
Узагальнення	h_j^2	$1 - d_j^2$
Характерність	d_j^2	$b_j^2 + c_j^2 = 1 - h_j^2$
Специфічність	b_j^2	$d_j^2 - c_j^2$
Дисперсія похибки	c_j^2	$1 - r_j^2$

Якщо необхідно визначити вклад фактора f_r у сумарну дисперсію всіх змінних, то це можна записати так:

$$V_r = \sum_{j=1}^{n} a_{jr}^2, \qquad (10.3.12)$$

де n — кількість змінних.

Вклад усіх характерних факторів у сумарну дисперсію ознак розраховується таким чином:

$$V_0 = \sum_{r=1}^{m} \nu_r, \qquad (10.3.13)$$

де ν_r — вклад одного характерного фактора у сумарну дисперсію ознак.

Повнотою факторизації називають відношення

$$k = \frac{V_0}{n}. \qquad (10.3.14)$$

При виконанні аналізу отриманих результатів факторизації корисно побудувати діаграми, в яких відображаються частки дисперсії кожної змінної і вклади загальних факторів у дисперсії вихідних ознак.

Приклад 10.1. У результаті розв'язання задачі, яка має сім ознак (змінних), знайдено два загальних фактори; необхідно визначити:

1 — вклади загальних і характерних факторів у дисперсію ознак, %;

2 — вклад усіх семи ознак у кожний загальний фактор, %;

3 — вклад кожного загального фактора у сумарну дисперсію, %;

4 — скласти таблицю відносних вкладів факторів у сумарну дисперсію.

Нехай матриця вагових коефіцієнтів загальних факторів має вигляд [15]:

$$\mathbf{A} = \begin{bmatrix} 0{,}90 & -0{,}30 \\ 0{,}80 & -0{,}30 \\ 0{,}60 & 0{,}30 \\ 0{,}50 & 0{,}20 \\ 0{,}50 & 0{,}50 \\ -0{,}30 & 0{,}60 \\ 0{,}20 & 0{,}80 \end{bmatrix}. \qquad (10.3.15)$$

Перший стовпчик цієї матриці є вектором вагових коефіцієнтів $\mathbf{a}_1$ першого загального фактора. Другий стовпчик матриці **A** є вектором вагових коефіцієнтів $\mathbf{a}_2$ другого загального фактора. Наприклад, a_{31} — ваговий коефіцієнт, який встановлює зв'язок між змінною y_3 і першим загальним фактором, а $a^2_{31} = 0{,}36$ — вклад третьої змінної у дисперсію першого загального фактора. Вклад першої змінної у дисперсію другого загального фактора становить: $a^2_{12} = 0{,}09$.

Розв'язок

1. Визначимо вклади загальних і характерного факторів у дисперсію ознак. Вклад першої ознаки (змінної) у дисперсію першого фактора становить:

$$a^2_{11} = 0{,}9^2 = 0{,}81,$$

а її вклад у другий фактор: $a^2_{12} = 0{,}3^2 = 0{,}09$.

Таким чином: $h^2_1 = a^2_{11} + a^2_{12} = 0{,}81 + 0{,}09 = 0{,}90$, а $d^2_1 = 1 - 0{,}90 = 0{,}10$.

Результати таких розрахунків представлено у табл. 10.2.

Таблиця 10.2

Розрахункові значення h^2_j і d^2_j

№ змінної j	a^2_{j1}	a^2_{j2}	$h^2_j = a^2_{j1} + a^2_{j2}$	$d^2_j = 1 - h^2_j$
A	1	2	3	4
1	0,81	0,09	0,90	0,10
2	0,64	0,09	0,73	0,27
3	0,36	0,09	0,45	0,55
4	0,25	0,04	0,29	0,71
5	0,25	0,25	0,50	0,50
6	0,09	0,36	0,45	0,55
7	0,04	0,64	0,68	0,32

2. Визначення вкладів ознак у дисперсії факторів:

а) вклад *першої змінної у дисперсію першого* загального фактора. За 100 % приймаємо дисперсію першого загального фактора. Дисперсія першого загального фактора дорівнює сумі елементів другого стовпчика табл. 10.2:

$$\lambda_1 = V_1 = \sum_{j=1}^{7} a^2_{j1} = 2{,}44.$$

Вклад першої змінної у дисперсію першого фактора становить:

$$\frac{a_{11}^2}{V_1} = \frac{0,81}{2,44} = 0,332 \approx 0,33;$$

б) вклад *першої змінної у дисперсію другого* загального фактора. За 100 % приймаємо дисперсію другого загального фактора:

$$\lambda_2 = V_2 = \sum_{j=1}^{7} a_{j2}^2 = 1,56.$$

Вклад першої змінної у дисперсію другого фактора становить:

$$\frac{a_{12}^2}{V_2} = \frac{0,09}{1,56} = 0,0577 \approx 0,06;$$

в) складемо таблицю вкладів змінних у дисперсію загальних факторів (табл. 10.3).

Таблиця 10.3

Вклади змінних у дисперсії загальних факторів

№ фактора	Вклади змінних, %						
	y_1	y_2	y_3	y_4	y_5	y_6	y_7
1	33	26	15	10	10	4	2
2	6	6	6	2	16	23	41

3. Розрахунок вкладів загальних факторів у сумарний узагальнюючий параметр, тобто визначимо:

а) сумарний узагальнюючий параметр:

$$\sum_{j=1}^{7} h_j^2 = V_0 = \sum_{r=1}^{2} V_r = V_1 + V_2 = 2,44 + 1,56 = 4,00;$$

б) вклад першого фактора у сумарний узагальнюючий параметр:

$$\frac{V_1}{V_0} = \frac{2,44}{4,0} = 0,61;$$

в) вклад другого фактора у сумарний узагальнюючий параметр:

$$1 - \frac{V_1}{V_0} = 1 - 0,61 = 0,39;$$

г) вклади кожної змінної в узагальнення першого і другого факторів з точністю до 1 % (див. табл. 10.4). Для цього необхідно вклад

кожної змінної (див. табл. 10.3) помножити на ваговий коефіцієнт відповідного фактора у сумарному узагальненні процесу, або значення a_{j1}^2, a_{j2}^2 (див. табл. 10.2) розділити на сумарний узагальнюючий параметр (який дорівнює 4,0).

Можна побудувати графік вкладів змінних у кожний із загальних факторів.

4. Складання результуючої таблиці часток дисперсій факторів.

Таблиця 10.4

Вклади змінних у сумарний узагальнюючий параметр із врахуванням вкладів факторів

№ фактора r	Вклади змінних, %						
	y_1	y_2	y_3	y_4	y_5	y_6	y_7
1	20	16	9	6	6	2	1
2	2	2	2	1	6	9	16

Таблиця 10.5

Частки дисперсій факторів

№ фактора r	Вид дисперсії	Формула	Значення вкладу	Вклад, %
A	1	2	3	4
1	Дисперсія процесу (повна дисперсія)	S_{Σ}^2	7,00	100
2	Дисперсія першого фактора	$V_1 = \sum_{j=1}^{7} a_{j1}^2$	2,44	34,86
3	Дисперсія другого фактора	$V_2 = \sum_{j=1}^{7} a_{j2}^2$	1,56	22,29
4	Узагальнюючий параметр процесу (сумарний)	$V_0 = \sum_{j=1}^{7} h_j^2 = V_1 + V_2$	4,0	57,14
5	Сумарна характерна дисперсія	$V_x = \sum_{j=1}^{7} d_j^2$	3,0	42,86

Зазначимо, що дисперсія процесу дорівнює 7 і співпадає з кількістю ознак. Дисперсія кожної нормованої ознаки дорівнює 1, а тому повна дисперсія процесу для семи ознак дорівнює 7.

Природно, що

$$V_0 + V_x = 4{,}0 + 3{,}0 = 7.$$

Також необхідно зазначити, що аналіз дисперсій був виконаний тільки на основі заданої матриці вагових коефіцієнтів загальних факторів. Тобто значення a_{jr} загальних факторів однозначно визначає значення вагових коефіцієнтів характерних факторів.

10.4. Знаходження матриці коефіцієнтів парної кореляції та її перетворення у факторному аналізі

Вихідні (початкові) дані матриці **Y** дають можливість отримати матрицю **R** — це матриця коефіцієнтів парної кореляції або кореляційна матриця. Із записів матриць (10.2.5), (10.2.6) і (10.2.7) видно, що

$$\mathbf{M} = \mathbf{A} + \mathbf{D}.$$

Тобто для відтворення усіх зв'язків змінних у кореляційній матриці необхідно скористатись матрицею **M**.

Тепер кореляційну матрицю можна визначити так:

$$\begin{aligned}\mathbf{R} = \mathbf{M}\cdot\mathbf{M}^T = (\mathbf{A}+\mathbf{D})\cdot(\mathbf{A}+\mathbf{D})^T = (\mathbf{A}+\mathbf{D})\cdot\left(\mathbf{A}^T+\mathbf{D}^T\right) = \\ = \mathbf{A}\mathbf{A}^T + \mathbf{A}\mathbf{D}^T + \mathbf{D}\mathbf{A}^T + \mathbf{D}\mathbf{D}^T.\end{aligned} \tag{10.4.1}$$

Із записів матриць (10.2.5) і (10.2.6) випливає, що

$$\mathbf{A}\mathbf{D}^T = \mathbf{D}\mathbf{A}^T = 0, \tag{10.4.2}$$

а тому кореляційна матриця матиме вигляд:

$$\mathbf{R} = \mathbf{A}\mathbf{A}^T + \mathbf{D}\mathbf{D}^T. \tag{10.4.3}$$

Введемо позначення $\mathbf{A}\mathbf{A}^T = \mathbf{R}_h$; і оскільки **D** — квазідіагональна матриця, то $\mathbf{D}\mathbf{D}^T = \mathbf{D}^2$. Піднесення матриці **D** до квадрату означає, що ми отримаємо діагональну матрицю, на головній діагоналі якої будуть елементи цієї матриці у квадраті.

Таким чином, матриця значень парних кореляцій для вихідних (початкових) вимірів може бути отримана за допомогою матриці **M**:

$$\mathbf{R} = \mathbf{R}_h + \mathbf{D}^2, \tag{10.4.4}$$

або

$$\mathbf{R} = \mathbf{A}\mathbf{A}^T + \mathbf{D}^2. \qquad (10.4.5)$$

Матриця $\mathbf{R}$ — дійсна симетрична кореляційна матриця з одиницями на головній діагоналі, а матриця $\mathbf{R}_h$ — кореляційна матриця з узагальненими параметрами на головній діагоналі.

У симетричній матриці $\mathbf{R}$ діагональні елементи r_{ii} — це дисперсії досліджуваних ознак. Так, r_{11} — це дисперсія випадкової величини — ознаки y_1; r_{ii} — дисперсія ознаки y_i. Всі ці дисперсії дорівнюють одиниці. Таким чином, сумарна дисперсія всіх досліджуваних ознак дорівнює сумі діагональних елементів матриці $\mathbf{R}$, тобто сумі дисперсій ознак.

Представимо елементи матриці $\mathbf{R}$ у розгорнутому вигляді:

$$\begin{aligned}
&r_{11} = 1;\\
&r_{12} = \frac{1}{N}\left(y_{11}y_{21} + y_{12}y_{22} + \ldots + y_{1N}y_{2N}\right);\\
&\quad\vdots\\
&r_{jk} = \frac{1}{N}\left(y_{j1}y_{k1} + y_{j2}y_{k2} + \ldots + y_{jN}y_{kN}\right);\\
&\quad\vdots\\
&r_{nn} = 1,
\end{aligned} \qquad (10.4.6)$$

де $r_{jj} = r_{kk} = 1$; $r_{jk} = r_{kj}$.

На основі (10.4.6) можна записати вираз для $\mathbf{R}$ у матричній формі:

$$\mathbf{R} = \frac{1}{N}\mathbf{Y}\mathbf{Y}^T. \qquad (11.4.7)$$

Скористаємось отриманою раніше формулою $\mathbf{Y} = \mathbf{A}\mathbf{F}$ з метою перетворення матриці $\mathbf{R}$ у редуковану форму:

$$\mathbf{R}_h = \frac{1}{N}\mathbf{A}\mathbf{F}\left(\mathbf{A}\mathbf{F}\right)^T = \frac{1}{N}\mathbf{A}\mathbf{F}\mathbf{F}^T\mathbf{A}^T = \mathbf{A}\frac{1}{N}\mathbf{F}\mathbf{F}^T\mathbf{A}^T.$$

Вираз, який знаходиться між $\mathbf{A}$ і $\mathbf{A}^T$, за аналогією із (10.4.7) є кореляційною матрицею стохастичних зв'язків між загальними факторами. Позначимо її так:

$$\mathbf{C} = \frac{1}{N}\mathbf{F}\mathbf{F}^T,$$

що приводить до виразу для $\mathbf{R}_h$:

$$\mathbf{R}_h = \mathbf{ACA}^T. \qquad (10.4.8)$$

Якщо загальні фактори не корельовані між собою, то **C** буде одиничною матрицею $\mathbf{I}_n$ і вираз (10.4.8) спрощується до вигляду:

$$\mathbf{R}_h = \mathbf{AA}^T. \qquad (10.4.9)$$

Приклад 10.2. Наведемо ілюстрацію знаходження матриці $\mathbf{R}_h$ за умови відомої матриці вагових коефіцієнтів **A**.

Використовуючи матрицю **A** (приклад 10.1), необхідно знайти:

1 — редуковану матрицю $\mathbf{R}_h$;

2 — матрицю розсіювання $\mathbf{R}^1$;

3 — матрицю залишків: $\mathbf{R}_1 = \mathbf{R}_h - \mathbf{R}^1$.

Виконати аналіз вкладів у дисперсію загальних і характерних факторів.

Розв'язок

1. Знаходимо матрицю $\mathbf{R}_h$.

Для цього помножимо **A** на $\mathbf{A}^T$ і отримаємо редуковану кореляційну матрицю $\mathbf{R}_h$, тобто відновлену із моделі факторного аналізу за умови, що фактори некорельовані:

$$\mathbf{R}_h = \mathbf{A} \cdot \mathbf{A}^T = \begin{bmatrix} 0{,}90 & -0{,}30 \\ 0{,}80 & -0{,}30 \\ 0{,}60 & 0{,}30 \\ 0{,}50 & 0{,}20 \\ 0{,}50 & 0{,}50 \\ -0{,}30 & 0{,}60 \\ 0{,}20 & 0{,}80 \end{bmatrix} \cdot \begin{bmatrix} 0{,}90 & 0{,}80 & 0{,}60 & 0{,}50 & 0{,}50 & -0{,}30 & 0{,}20 \\ -0{,}30 & -0{,}30 & 0{,}30 & 0{,}20 & 0{,}50 & 0{,}60 & 0{,}80 \end{bmatrix};$$

$$\mathbf{R}_h = \begin{bmatrix} 0{,}90 & 0{,}81 & 0{,}45 & 0{,}39 & 0{,}30 & -0{,}45 & -0{,}06 \\ 0{,}81 & 0{,}73 & 0{,}39 & 0{,}34 & 0{,}25 & -0{,}42 & -0{,}08 \\ 0{,}45 & 0{,}39 & 0{,}45 & 0{,}36 & 0{,}45 & 0{,}00 & 0{,}36 \\ 0{,}39 & 0{,}34 & 0{,}36 & 0{,}29 & 0{,}35 & -0{,}03 & 0{,}26 \\ 0{,}30 & 0{,}25 & 0{,}45 & 0{,}35 & 0{,}50 & 0{,}15 & 0{,}50 \\ -0{,}45 & -0{,}42 & 0{,}00 & -0{,}03 & 0{,}15 & 0{,}45 & 0{,}42 \\ -0{,}06 & -0{,}08 & 0{,}36 & 0{,}26 & 0{,}50 & 0{,}42 & 0{,}68 \end{bmatrix}.$$

2. Знаходимо матрицю розсіювання $\mathbf{R}^1$.

Необхідно відповісти на такі питання: що буде, якщо не враховувати другий загальний фактор, тобто інтерпретацію виконаємо на

основі тільки першого загального фактора? Як буде відтворена у такому випадку матриця розсіювання?

Для цього необхідно знайти добуток векторів, як матриць:

$$\mathbf{R}^1 = \begin{bmatrix} 0{,}90 \\ 0{,}80 \\ 0{,}60 \\ 0{,}50 \\ 0{,}50 \\ -0{,}30 \\ 0{,}20 \end{bmatrix} \cdot \begin{bmatrix} 0{,}90 & 0{,}80 & 0{,}60 & 0{,}50 & 0{,}50 & -0{,}30 & 0{,}20 \end{bmatrix} =$$

$$= \begin{bmatrix} 0{,}81 & 0{,}72 & 0{,}54 & 0{,}45 & 0{,}45 & -0{,}27 & 0{,}18 \\ 0{,}72 & 0{,}64 & 0{,}48 & 0{,}40 & 0{,}40 & -0{,}24 & 0{,}16 \\ 0{,}54 & 0{,}48 & 0{,}36 & 0{,}30 & 0{,}30 & -0{,}18 & 0{,}12 \\ 0{,}45 & 0{,}40 & 0{,}30 & 0{,}25 & 0{,}25 & -0{,}15 & 0{,}10 \\ 0{,}45 & 0{,}40 & 0{,}30 & 0{,}25 & 0{,}25 & -0{,}15 & 0{,}10 \\ -0{,}27 & -0{,}24 & -0{,}18 & -0{,}15 & -0{,}15 & 0{,}09 & -0{,}06 \\ 0{,}18 & 0{,}16 & 0{,}12 & 0{,}10 & 0{,}10 & -0{,}06 & 0{,}04 \end{bmatrix}.$$

Відтворена або редукована за першим загальним фактором матриця відновлює зв'язки, які пояснюються першим вектором (фактором) — вектором матриці **A**. І перша, і друга відтворені матриці не відображають всієї інформації про процес.

При цьому друга матриця $\mathbf{R}_1$ відображає менше інформації, ніж перша $\mathbf{R}_1$. Це пояснюється тим, що $\mathbf{R}_1$ відтворює зв'язки, що відповідають $V_1 = 2{,}44$, а $\mathbf{R}_1$ — відтворює зв'язки при $V_2 = 1{,}56$.

Однак і повніша матриця $\mathbf{R}_h$ не відтворює зв'язків, які визначаються характерними факторами, оскільки ця матриця об'єднує вагові коефіцієнти тільки загальних факторів.

3. Знайдемо матрицю залишків $\mathbf{R}_1$.

Матриця $\mathbf{R}_1$ містить ту частину інформації, яка не пояснена матрицею $\mathbf{R}^1$ у порівнянні з інформацією, яка пояснюється матрицею $\mathbf{R}_h$. Виникає питання: де ж міститься ця частина інформації?

Матриця $\mathbf{R}_h$ пояснює всю інформацію, представлену матрицею загальних факторів **A**. Таким чином, не пояснена частина інформації повинна міститись у другому векторі матриці **A**, тобто в $\mathbf{a}_2$.

$$\mathbf{R}_1 = \mathbf{R}_h - \mathbf{R}^1 = \begin{bmatrix} 0{,}09 & 0{,}09 & -0{,}09 & -0{,}06 & -0{,}15 & -0{,}18 & -0{,}24 \\ 0{,}09 & 0{,}09 & -0{,}09 & 0{,}06 & -0{,}15 & -0{,}18 & -0{,}24 \\ -0{,}09 & -0{,}09 & 0{,}09 & 0{,}06 & 0{,}15 & 0{,}18 & 0{,}24 \\ -0{,}08 & -0{,}06 & 0{,}06 & 0{,}04 & 0{,}10 & 0{,}12 & 0{,}16 \\ -0{,}15 & -0{,}15 & 0{,}15 & 0{,}10 & 0{,}25 & 0{,}30 & 0{,}40 \\ -0{,}18 & -0{,}18 & 0{,}18 & 0{,}12 & 0{,}30 & 0{,}36 & 0{,}48 \\ -0{,}24 & -0{,}24 & 0{,}24 & 0{,}16 & 0{,}40 & 0{,}48 & 0{,}64 \end{bmatrix}.$$

Якщо добуток як матриць вектора $\mathbf{a}_2$ на $\mathbf{a}_2^T$ дасть матрицю розсіювання, яка співпадатиме з матрицею $\mathbf{R}_1$, то $\mathbf{R}_1$ відображає інформацію, яка міститься у векторі $\mathbf{a}_2$. (Необхідно самостійно отримати добуток $\mathbf{a}_2\,\mathbf{a}_2^T$ і переконатись, що $\mathbf{R}_1 = \mathbf{a}_2\,\mathbf{a}_2^T$.)

Частина інформації, яка не пояснена матрицями $\mathbf{R}_h$ і $\mathbf{A}$, припадає на характерні фактори.

4. Аналіз вкладів у дисперсію загальних і характерних факторів виконаємо на основі табл. 10.2, у якій наведено розрахунки вкладів ознак у загальні і характерні фактори.

За виконаними обчисленнями можна зробити такі висновки.

1. Вклад першого загального фактора $V_1 = 2{,}44$ і вклад другого загального фактора $V_2 = 1{,}56$ наведено у табл. 10.5 (стовпчик 3). Сумарний вклад у дисперсію процесу загальних факторів складає V_0. У даному прикладі сумарний вклад характерних факторів доповнює V_0 до значення $n = 7$.

2. Елементи головної діагоналі матриці залишків $\mathbf{R}_1$ збігаються з відповідними елементами табл. 10.2 (стовпчик 3). У цьому стовпчику наведено вклади ознак у дисперсію другого фактора $\mathbf{a}_{j2}^2$.

3. У матриці $\mathbf{R}^1$ на головній діагоналі знаходяться вклади у дисперсію першого фактора відповідних змінних $\mathbf{a}_{j2}^2$ (див. табл. 10.2, стовпчик 2).

4. На головній діагоналі матриці $\mathbf{R}_h$ знаходяться дисперсії, які є узагальнюючі параметри, а також сумарний вклад у змінні двох наявних загальних факторів (див. табл. 10.2, стовпчик 4).

5. Матриця $\mathbf{R}^1$ характеризує зв'язки між змінними, які пояснюються тільки першим загальним фактором, а матриця $\mathbf{R}_1$ характеризує зв'язки між змінними, які пояснюються другим загальним фактором.

Зв'язки між змінними, які пояснюються усіма загальними факторами, характеризує матриця $\mathbf{R}_h$, яка є сумою $\mathbf{R}_h = \mathbf{R}^1 + \mathbf{R}_1$.

Матриця $\mathbf{R}_h$ у даному прикладі, так само як і $\mathbf{R}^1$, називається редукованою матрицею.

Як правило, на практиці для отримання $\mathbf{R}_h$ замість одиниць на головній діагоналі матриці $\mathbf{R}$ ставлять оцінки узагальнюючих параметрів h_j^2.

Згідно з обчисленнями, наведеними у прикладі 10.2, $\mathbf{R}$ можна представити у вигляді:

$$\mathbf{R} = \left(\mathbf{R}_h + \mathbf{D}^2\right) = \mathbf{M}\begin{pmatrix} \mathbf{C} & 0 \\ 0 & \mathbf{E} \end{pmatrix}\mathbf{M}^{\mathbf{T}}. \qquad (10.4.10)$$

Матриця $\mathbf{R}_h$ відрізняється від матриці $\mathbf{R}$ на матрицю $\mathbf{D}^2$. Таким чином, потрібно розглядати перехід від матриці $\mathbf{R}$ з одиницями на головній діагоналі до матриці $\mathbf{R}_h$.

10.5. Контрольні питання і вправи

1. У чому полягає основна задача факторного аналізу?
2. Які задачі необхідно розв'язувати при виконанні факторного аналізу?
3. Поясність відмінність між статистичним і динамічним варіантами факторного аналізу.
4. Поясніть змінні регресійної моделі:

$$x_{ji} = a_{j1}f_{1i} + a_{j2}f_{2i} + ... + a_{jm}f_{mi} + d_j\,\nu_{ji},$$
$$j = 1, ..., n;\ \ i = 1, ..., N;\ \ r = 1, ..., m\ .$$

5. Поясніть операції центрування і нормування даних. Яка мета цих обчислювальних операцій?
6. Поясніть складові рівняння після нормування:

$$y_{ji} = a_{j1}f_{1i} + a_{j2}f_{2i} + ... + a_{jm}f_{mi} + d_j\,\nu_{ji},$$
$$j = 1, ..., n;\ \ i = 1, ..., N;\ \ r = 1, ..., m\ .$$

7. Які компоненти містять виміри даних?
8. Який фактор називають характерним? Як його вибирають?
9. З яких компонент складається дисперсія характерного фактора?
10. Поясніть послідовність обчислень при виконанні факторного аналізу.

Розділ 11

ДИСКРИМІНАНТНИЙ АНАЛІЗ

11.1. Завдання дискримінантного аналізу

Завданням *дискримінантного аналізу* є розділення досліджуваної множини індивідуумів або об'єктів на два або більше класів. Наприклад, у банківській системі розв'язується задача класифікації позичальників кредитів на дві групи — "надійні" клієнти, тобто ті, хто поверне кредит, і "ненадійні". Для цього використовують такі змінні, як *вік клієнта*, його *стать, середній дохід, стаж роботи, сімейний стан* і т. ін. Очевидно, що прикладів такого аналізу можна навести багато.

Дискримінантний аналіз розглядається у системі координат багатовимірних спостережень. Факторний аналіз застосовують в економіці, фінансах, соціальних та політичних дослідженнях, технічних системах та інших галузях. Наприклад:

- для виявлення "факторів" економічного розвитку;
- для дослідження стану валютного ринку з метою вироблення коректної валютної політики;
- при виконанні операцій на валютній біржі;
- для дослідження технічних задач, пов'язаних із застосуванням деяких методів оцінювання параметрів систем одночасних рівнянь.

11.2. Лінійна дискримінантна функція

Розглянемо задачу аналізу оцінок студентів однієї групи, що отримані в результаті усного опитування та письмового іспиту. У ході дискримінантного аналізу ставиться таке питання: чи можна знайти такий критерій, який дасть можливість віднести кожного студента цієї групи до підгрупи успішних або підгрупи неуспішних? Очевидно, що крім отриманих оцінок, можна скористатись деякою додатковою інформацією стосовно студентів (наприклад, їх середньою успішністю), але для спрощення аналізу обмежимось двома отриманими оцінками.

Отже, будемо будувати критерій класифікації, який ґрунтується на лінійній комбінації двох отриманих студентами оцінок (приклад оцінок для однієї групи у табл. 11.1).

Таблиця 11.1

Результати усного опитування студентів та письмового іспиту

№ пор.	Оцінка за усне опитування, бали	Оцінка за письмовий екзамен, бали
1	740	680
2	670	600
3	560	550
4	540	520
5	590	540
6	590	700
7	470	600
8	560	540
9	540	630
10	500	600
Середнє	576	596

Крім того, необхідно задати деяке критичне (порогове) значення цього критерію. Воно має бути таким: якщо значення критерію для деякого студента виявиться нижчим критичного, то цього студента відносять до однієї підгрупи, а якщо вищим критичного, то він буде віднесеним до іншої підгрупи.

Якщо дві підгрупи студентів, сформовані за конкретними значеннями критеріїв, подібні з точки зору екзаменаційних оцінок (тобто їх середні значення та коваріаційні матриці є достатньо близькими), то задовільна класифікація студентів може виявитись неможливою внаслідок значного перекриття між групами.

Процедура лінійного дискримінантного аналізу полягає у знаходженні такої лінійної комбінації змінних X_1 і X_2, щоб перекриття розподілів для цих двох груп було незначним. Лінійну функцію

$$y_i(k) = \beta_1 x_{i1}(k) + \beta_2 x_{i2}(k),\ i = 1, 2;\ k = 1, 2, 3, ..., n_i \qquad (11.2.1)$$

називають лінійною дискримінантною функцією з невідомими параметрами β_1, β_2. Індекс i позначає групу спостережень, а k — номер спостереження у групі. На відміну від регресійного аналізу змінна

y — це результат комбінування змінних X_i, а не множина значень, які необхідно апроксимувати за допомогою змінних X_i.

Геометрично рівняння (12.2.1) визначає площину, а тому проекція $x_{i1}(k)$ і $x_{i2}(k)$ на цю площину перетворює двовимірні оцінки в одновимірну оцінку $y_i(k)$.

Класифікація відбувається краще (чіткіше) при більших значеннях варіацій всередині груп значень, тому що зменшення варіації свідчить про зменшення інформативності даних. (*Варіація може вимірюватись коваріацією або сумою квадратів відхилень від середнього.*) При більших варіаціях відбувається краще розділення середніх. З іншого боку, велика варіація всередині групи небажана, оскільки будь-яка відстань між середніми буде тим менше значущою у статистичному смислі, чим більшою є варіація розподілів, які відповідають цим середнім.

Існує деяка оптимальна дискримінантна площина і один із методів пошуку цього оптимуму полягає у максимізації відношення:

$$\lambda = \frac{\text{Міжгрупова варіація}}{\text{Внутрішньогрупова варіація}}. \qquad (11.2.2)$$

Розглянемо коротко можливості максимізації цього відношення.

Дискримінантну функцію (ДФ) легко узагальнити на випадок, коли у кожній групі є p змінних (у такому випадку ДФ визначає гіперплощину):

$$y_i(k) = \beta_1 x_{i1}(k) + \beta_2 x_{i2}(k) + ... + \beta_p x_{ip}(k), \; i = 1, 2, \qquad (11.2.3)$$

де $x_{ij}(k)$ значення j-ї змінної для k-го спостереження в i-й групі. Для двох груп спостережень вектори середніх (проекції центрів) задаються рівняннями:

$$\mu_{y_1} = \beta_1\mu_{x_{11}} + \beta_2\mu_{x_{12}} + \ldots + \beta_p\mu_{x_{1p}}; \qquad (11.2.4)$$

$$\mu_{y_2} = \beta_1\mu_{x_{21}} + \beta_2\mu_{x_{22}} + \ldots + \beta_p\mu_{x_{2p}}, \qquad (11.2.5)$$

де $\mu_{y_1}, \mu_{y_2}, \mu_{x_{ij}}$ — середні значення. Рівняння (11.2.4) і (11.2.5) мають місце за означенням, оскільки $y(k)$ — це лінійна комбінація змінних $x(k)$. Використовуючи (11.2.3), (11.2.4) і (11.2.5), запишемо рівняння для відхилень від середніх:

$$y_i(k) - \mu_{y_i} = \beta_1\left(x_{i1}(k) - \mu_{x_{i1}}\right) + \beta_2\left(x_{i2}(k) - \mu_{x_{i2}}\right) + \\ + \beta_p\left(x_{ip}(k) - \mu_{x_{ip}}\right). \qquad (11.2.6)$$

Сума квадратів відхилень $y_i(k)-\mu_{y_i}$ буде мірою *варіації всередині групи*.

У вибірках даних всі середні генеральної сукупності замінюють вибірковими середніми:

$$\bar{y}_i=\frac{1}{n_i}\sum_{k=1}^{n_i}y_i(k);\quad \bar{x}_{ij}=\frac{1}{n_i}\sum_{k=1}^{n_i}x_{ij}(k);\quad j=1,2,\ldots,p.$$

Таким чином, для варіації в середині групи можна записати вираз:

$$\sum_{i=1}^{2}\sum_{k=1}^{n_i}\left[y_i(k)-\bar{y}_i\right]^2=\sum_{i=1}^{2}\sum_{k=1}^{n_i}\left\{\beta_1\left[x_{i1}(k)-\bar{x}_{i1}\right]+\beta_2\left[x_{i2}(k)-\bar{x}_{i2}\right]+\right.$$
$$\left.+\beta_p\left[x_{ip}(k)-\bar{x}_{ip}\right]\right\}^2. \qquad (11.2.7)$$

Введемо матричні позначення для вимірів, що відносяться до обох груп (ми стараємось розділити всі значення на дві групи):

$$\mathbf{X}^T=\left[\begin{array}{cccc|cccc} x_{11}(1) & x_{11}(2) & \cdots & x_{11}(n_1) & x_{21}(1) & x_{21}(2) & \cdots & x_{21}(n_2)\\ x_{12}(1) & x_{12}(2) & \cdots & x_{12}(n_1) & x_{22}(1) & x_{22}(2) & \cdots & x_{22}(n_2)\\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots\\ x_{1p}(1) & x_{1p}(2) & \cdots & x_{1p}(n_1) & x_{2p}(1) & x_{2p}(2) & \cdots & x_{2p}(n_2)\end{array}\right],$$

або

$$\mathbf{X}^T=[\mathbf{X}_1^T\ \mid\ \mathbf{X}_2^T],$$

де $\mathbf{X}_1^T$ — містить всі спостереження для p змінних у групі I, а матриця $\mathbf{X}_2^T$ — містить спостереження для всіх p змінних у групі II. Цим двом матрицям відповідають вектори середніх:

$$\bar{\mathbf{x}}_1^T=\left[\bar{x}_{11}\ \ \bar{x}_{12}\ \ \cdots\ \ \bar{x}_{1p}\right];$$
$$\bar{\mathbf{x}}_2^T=\left[\bar{x}_{21}\ \ \bar{x}_{22}\ \ \cdots\ \ \bar{x}_{2p}\right].$$

Середні визначаються за виразом:

$$\bar{\mathbf{x}}^T=\frac{1}{n}\mathbf{A}^T\mathbf{X}=\frac{1}{n}[1\ \ 1\ \ \cdots\ \ 1]\left[\begin{array}{cccc} x_1(1) & x_2(1) & \cdots & x_p(1)\\ x_1(2) & x_2(2) & \cdots & x_p(2)\\ \vdots & \vdots & & \vdots\\ x_1(n) & x_2(n) & \cdots & x_p(n)\end{array}\right]=$$

$$=\left[\bar{x}_1\ \ \bar{x}_2\ \ \cdots\ \ \bar{x}_p\right],$$

де $n=n_1+n_2$.

Коваріаційні матриці $\mathbf{X}_1^T\mathbf{X}_1$ і $\mathbf{X}_2^T\mathbf{X}_2$ знаходяться згідно формули

$$\mathbf{S}=\frac{1}{n-1}\mathbf{X}^T\mathbf{X},$$

де

$$\mathbf{X}^T\mathbf{X}=\begin{bmatrix} x_1(1) & x_1(2) & \cdots & x_1(n)\\ x_2(1) & x_2(2) & \cdots & x_2(n)\\ \vdots & \vdots & & \vdots\\ x_p(1) & x_p(2) & \cdots & x_p(n)\end{bmatrix}\times\begin{bmatrix} x_1(1) & x_2(1) & \cdots & x_p(1)\\ x_1(2) & x_2(2) & \cdots & x_p(2)\\ \vdots & \vdots & & \vdots\\ x_1(n) & x_2(n) & \cdots & x_p(n)\end{bmatrix}=$$

$$=\begin{bmatrix} \sum x_1^2 & \sum x_1x_2 & \cdots & \sum x_1x_p\\ \sum x_2x_1 & \sum x_2^2 & \cdots & \sum x_2x_p\\ \vdots & \vdots & & \vdots\\ \sum x_px_1 & \sum x_px_2 & \cdots & \sum x_p^2\end{bmatrix}=\begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p}\\ s_{21} & s_{22} & \cdots & s_{2p}\\ \vdots & \vdots & & \vdots\\ s_{p1} & s_{p2} & \cdots & s_{pp}\end{bmatrix}.$$

Якщо ввести вектор параметрів $\beta^T=[\beta_1,\beta_2,...,\beta_p]$, то вираз (11.2.7) можна переписати у вигляді:

$$\sum_{i=1}^{2}\sum_{k=1}^{n_i}\left[y_i(k)-\overline{y}_i\right]^2=\beta^T\mathbf{X}_1^T\mathbf{X}_1\beta+\beta^T\mathbf{X}_2^T\mathbf{X}_2\beta=\beta^T\left(\mathbf{X}_1^T\mathbf{X}_1+\mathbf{X}_2^T\mathbf{X}_2\right)\beta.$$

Згадаємо, що об'єднана коваріаційна матриця визначається так:

$$\mathbf{S}^*=\frac{1}{n_1+n_2-2}\left(\mathbf{X}_1^T\mathbf{X}_1+\mathbf{X}_2^T\mathbf{X}_2\right).$$

Звідси можна записати вираз для варіації в середині групи:

$$\sum_{i=1}^{2}\sum_{k=1}^{n_i}\left[y_i(k)-\overline{y}_i\right]^2=\beta^T\left[(n_1+n_2-2)\mathbf{S}^*\right]\beta.$$

Чисельник відношення (11.2.2), тобто варіацію між групами, можна представити через групові середні: $\left(\overline{y}_1-\overline{y}_2\right)^2$, що, у свою чергу, можна записати через вибіркові середні:

$$\left(\overline{y}_1-\overline{y}_2\right)^2=\left(\beta^T\overline{\mathbf{x}}_1-\beta^T\overline{\mathbf{x}}_2\right)^2=\beta^T\left(\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_2\right)\left(\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_2\right)^T\beta.$$

Тепер можна записати вибірковий аналог відношення (11.2.2):

$$\lambda=\frac{\text{Міжгрупова варіація}}{\text{Внутрішньогрупова варіація}}=\frac{1}{n_1+n_2-2}\times$$
$$\times\frac{\beta^T\left(\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_2\right)\left(\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_2\right)^T\beta}{\beta^T\mathbf{S}^*\beta},\qquad(11.2.8)$$

яке представляє собою ту функцію, яку необхідно максимізувати. Очевидно, що можна максимізувати таке відношення:

$$l = \frac{\beta^T (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \beta}{\beta^T \mathbf{S}^* \beta} \qquad (11.2.9)$$

і отримати такий самий результат, оскільки l пропорціональне λ. Прирівнюючи нулю перші похідні $\partial l / \partial \beta$, після деяких перетворень отримаємо:

$$c(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - \mathbf{S}^* \beta = 0, \qquad (11.2.10)$$

де $c = \beta^T (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) / l$ — деяка константа, що не дорівнює нулю. Звідси отримаємо вираз для оцінок коефіцієнтів:

$$\hat{\beta} = c\left(\mathbf{S}^*\right)^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2),$$

тобто значення коефіцієнтів β залежать також від ненульової константи c. Цю константу вибирають так: $c = 1$, або такою, щоб $\hat{\beta} = 1$. У такому випадку вираз для знаходження коефіцієнтів спрощується до такого:

$$\beta = \left(\mathbf{S}^*\right)^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2). \qquad (11.2.11)$$

Якщо підставити (11.2.11) у (11.2.9), то отримаємо величину:

$$l = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \left(\mathbf{S}^*\right)^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = D^2, \qquad (11.2.12)$$

яку називають *узагальненою відстанню або відстанню D^2 Махаланобіса* [4].

Приклад 11.1. Знайдемо тепер коефіцієнти дискримінантної функції для прикладу з оцінками за екзамен. Оцінки студентів двох груп наведено у табл. 11.2; при цьому відомо, що друга група завжди була відстаючою.

Знайдемо вектори і матриці, які необхідні для подальшого застосування.

Вектори середніх для обох груп:

$$\bar{\mathbf{x}}_1^T = [598{,}46;\ 710{,}77]^T, \quad \bar{\mathbf{x}}_2^T = [576;\ 596]^T.$$

Значення відхилень від середніх для кожної оцінки, які необхідні для подальших обчислень, наведені у табл. 11.3.

Таблиця 11.2

Оцінки за усний і письмовий іспити для студентів двох груп

№ пор.	Успішна група (1)		Неуспішна група (2)	
	Оцінка за усний іспит	Оцінка за письмовий іспит	Оцінка за усний іспит	Оцінка за письмовий іспит
1	750	590	740	680
2	360	600	670	600
3	720	750	560	550
4	540	710	540	520
5	570	700	590	540
6	520	670	590	700
7	590	790	470	600
8	670	700	560	540
9	620	730	540	630
10	690	840	500	600
11	610	680	–	–
12	550	730	–	–
13	590	750	–	–
Середнє	598,46	710,77	576,0	596,0

Таблиця 11.3

Оцінки за усний і письмовий іспити для студентів двох груп

№ пор.	Успішна група (1)		Неуспішна група (2)	
	Відхилення від середнього оцінки за усний іспит	Відхилення від середнього оцінки за письмовий іспит	Відхилення від середнього оцінки за усний іспит	Відхилення від середнього оцінки за письмовий іспит
1	2	3	4	5
1	151,54	–120,77	164,0	84,0
2	–238,46	–110,77	94,0	4,0
3	121,54	39,33	–16,0	–46,0
4	–58,46	–0,77	–36,0	–76,0
5	–28,46	–10,77	14,0	–56,0
6	–78,46	–40,77	14,0	104,0

Закінчення табл. 11.3

1	2	3	4	5
7	–8,46	79,23	–106,0	4,0
8	71,54	–10,77	–16,0	–56,0
9	21,54	19,23	–36,0	34,0
10	91,54	129,23	–76,0	4,0
11	11,54	–30,77	–	–
12	–48,46	19,23	–	–
13	–8,46	39,23	–	–

Добутки $\mathbf{X}^T\mathbf{X}$ для обох груп:

$$\mathbf{X}_1^T\mathbf{X}_1 = \begin{bmatrix} 151{,}54 & -238{,}46 & \cdots & -8{,}46 \\ -120{,}77 & -110{,}77 & \cdots & 39{,}23 \end{bmatrix} \times \begin{bmatrix} 151{,}54 & -120{,}77 \\ -238{,}46 & -110{,}77 \\ \vdots & \vdots \\ -8{,}46 & 39{,}23 \end{bmatrix} =$$

$$= \begin{bmatrix} 121569 & 25615 \\ 25615 & 56492 \end{bmatrix}.$$

$$\mathbf{X}_2^T\mathbf{X}_2 = \begin{bmatrix} 164 & 94 & \cdots & -76 \\ 84 & 4 & \cdots & 4 \end{bmatrix} \times \begin{bmatrix} 164 & 84 \\ 94 & 4 \\ \vdots & \vdots \\ -76 & 4 \end{bmatrix} = \begin{bmatrix} 56240 & 17240 \\ 17240 & 33240 \end{bmatrix}.$$

Обчислимо об'єднану коваріаційну матрицю:

$$\mathbf{S}^* = \frac{1}{n_1 + n_2 - 2}\left(\mathbf{X}_1^T\mathbf{X}_1 + \mathbf{X}_2^T\mathbf{X}_2\right) = \begin{bmatrix} 8467 & 2041 \\ 2041 & 4273 \end{bmatrix},$$

і обернену до неї:

$$\left(\mathbf{S}^*\right)^{-1} = \begin{bmatrix} 0{,}0001335 & -0{,}0000637 \\ -0{,}0000637 & 0{,}0002645 \end{bmatrix}.$$

Значення різниць між векторами середніх:

$$\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 = \begin{bmatrix} 598{,}46 \\ 710{,}77 \end{bmatrix} - \begin{bmatrix} 576 \\ 596 \end{bmatrix} = \begin{bmatrix} 22{,}46 \\ 114{,}77 \end{bmatrix}.$$

Отже, тепер можна знайти вектор параметрів дискримінантної функції:

$$\hat{\beta} = \left(\mathbf{S}^*\right)^{-1}\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right) = \begin{bmatrix} -0{,}0043 \\ 0{,}0289 \end{bmatrix},$$

а оцінка дискримінантної функції має вигляд:

$$y_i(k) = -0{,}0043\, x_{i1}(k) + 0{,}0289\, x_{i2}(k).$$

Можна показати — якщо ціна помилкової класифікації однакова для обох груп і якщо ймовірності належності спостережень однакові для кожної групи, то y^* (знайдене за дискримінантною функцією) буде знаходитись посередині між μ_{y_1} і μ_{y_2}. Для вибіркових значень можна записати, що:

$$y^* = \frac{1}{2}\left(\bar{y}_1 + \bar{y}_2\right),$$

але за допомогою рівнянь (11.2.4) і (11.2.5) можна отримати співвідношення

$$y^* = \frac{1}{2}\hat{\beta}^T\left(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2\right). \qquad (11.2.13)$$

Для нашого прикладу з оцінками студентів двох груп маємо:

$$y^* = \frac{1}{2}\begin{bmatrix} -0{,}0043 & 0{,}0289 \end{bmatrix}\begin{bmatrix} 1174{,}46 \\ 1306{,}77 \end{bmatrix} = 16{,}36.$$

На основі зроблених нами припущень правило класифікації можна сформулювати таким чином:

- відносити дані до групи I, якщо $y_i(k) \geq y^*$;
- відносити дані до групи II, якщо $y_i(k) < y^*$.

На останньому кроці визначимо вектор $\mathbf{y}^T$:

$$\mathbf{y}^T = [y_{11};\ y_{12};\ \ldots;\ y_{1n_1};\ y_{21};\ y_{22};\ \ldots;\ y_{2n_2}],$$

а також вибіркову дискримінантну функцію

$$y = \hat{\beta}\mathbf{X}^T. \qquad (11.2.14)$$

Тепер за допомогою (11.2.13) і (11.2.14) сформулюємо правило класифікації в остаточному вигляді:

- відносити дані до групи I, якщо $\hat{\beta}\mathbf{X}^T - \frac{1}{2}\beta^T\left(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2\right) \geq 0$;
- відносити дані до групи II, якщо $\hat{\beta}\mathbf{X}^T - \frac{1}{2}\beta^T\left(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2\right) < 0$. (11.2.15)

Пояснимо правила на прикладі першої пари оцінок на іспиті студентів у першій групі. Згідно з табл. 11.2 $x_{11} = 750$, а $x_{12} = 590$. Таким чином, оцінка допоміжної змінної, необхідної для аналізу, становить:

$$-0{,}0043 \cdot 750 + 0{,}0289 \cdot 590 - 16{,}36 = -2{,}5.$$

Оскільки отримане число є меншим нуля, то цей студент невірно віднесений (нами) до групи II (згідно з правилами (11.2.15)). Такі самі підрахунки стосовно класифікації оцінок студентів виконані для всіх значень, наведених у табл. 11.2. Отриманий результат такий:

- два студенти з групи I неправильно віднесені до групи II;
- два студенти з групи II неправильно віднесені до групи I.

Таким чином, з 23-х студентів 19 (або приблизно 83 %) класифіковано правильно.

11.3. Перевірка гіпотез у дискримінантному аналізі

Загалом є три типи критеріїв, які використовують у прийнятті рішень при застосуванні дискримінантного аналізу:

- критерій придатності дискримінантної функції загалом;
- критерій для прийняття рішень стосовно того, чи узгоджується деяка гіпотетична дискримінантна функція з дискримінантною функцією, знайденою за наявними даними;
- критерій для включення або виключення деякої змінної в дискримінантну функцію.

Критерій 1. Якщо дві генеральні сукупності однорідні, то способу розділення цих сукупностей за допомогою дискримінантної функції не існує. Якщо дві генеральні сукупності неоднорідні, то дискримінантна функція за означенням придатна для використання, оскільки вона ґрунтується на деякому оптимальному методі розділення сукупностей. Але якщо обидві сукупності розділені нормально з однаковими коваріаційними матрицями, то неоднорідність може полягати тільки у нерівності векторів середніх значень цих сукупностей.

Таким чином, для перевірки можливої придатності дискримінантної функції можна скористатись статистикою T^2 Хотелінга [4]

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)^T \left(\mathbf{S}^*\right)^{-1} \left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)$$

і порівняти її з критичним значенням, яке розраховується за виразом:

$$T^2_{\alpha;p,n_1+n_2-p-1} = \frac{\left(n_1 + n_2 - 2\right)}{n_1 + n_2 - p - 1} F_{\alpha;p,n_1+n_2-p-1},$$

де $F_{\alpha;p,n_1+n_2-p-1}$ — F-статистика (Фішера), значення якої можна взяти з таблиць для F-розподілу. Зазначимо, що F-статистика пов'язана з t-розподілом Стьюдента таким чином:

$$\left(t_{\alpha/2;\nu_2}\right)^2 = F_{\alpha;\nu_1,\nu_2}.$$

Також необхідно зазначити, що статистика T^2 і статистика D^2 Махаланобіса тісно пов'язані між собою, тобто

$$T^2 = \left[n_1 n_2 / \left(n_1 + n_2\right)\right] D^2,$$

тобто T^2 також можна розглядати як деяку міру відстані.

Критерій 2. Для того щоб перевірити, чи узгоджується гіпотетична дискримінантна функція

$$y_i(k) = c_1 x_{i1}(k) + c_2 x_{i2}(k) + ... + c_p x_{ip}(k),\ i = 1, 2;\ k = 1, 2, ..., n_i,$$

з дискримінантною функцією, знайденою на основі наявних даних, на першому кроці необхідно знайти узагальнену відстань D_0, яка відповідає гіпотетичній дискримінантній функції (це має місце, якщо вірна нульова гіпотеза), тобто [4]:

$$D_0^2 = \left(\mathbf{c}^T \Delta\bar{\mathbf{x}}\right)^T \left(\mathbf{c}^T \mathbf{S}^* \mathbf{c}\right)^{-1} \left(\mathbf{c}^T \Delta\bar{\mathbf{x}}\right), \qquad (11.3.1)$$

де $\Delta\bar{\mathbf{x}} = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$ і $\mathbf{c}^T = [c_1, c_2, ..., c_p]$. Потім виконується порівняння D_0^2 з D^2, знайденим за наявними даними.

Гіпотетична дискримінантна функція відхиляється як несумісна на даному рівні значущості з дискримінантною функцією, знайденою за наявними даними, якщо [4]

$$F = \frac{n_1 + n_2 - p - 1}{p - 1} \times \frac{m\left(D^2 - D_0^2\right)}{1 + m D_0^2} > F_{\alpha;p-1,n_1+n_2-p-1},$$

де $m = n_1 n_2 / (n_1 + n_2)(n_1 + n_2 - 2)$.

Критерій 3. Дискримінантна функція може бути повторно знайдена з виключенням деяких (неінформативних або малоінформативних) змінних. Існує також інший спосіб визначення надлишкових змінних, який полягає у порівнянні коефіцієнтів функції з їх отриманими на великих вибірках стандартними оцінками. Цей спосіб дуже схожий на метод перевірки одновимірних гіпотез стосовно коефіцієнтів регресії.

Однак цей підхід пов'язаний з деякими труднощами, які полягають у тому, що вектор оцінок коефіцієнтів $\hat{\beta}$ залежить від ненульової константи c у (11.2.10). На практиці часто порівнюють узагальнену відстань D_p^2, яка знаходиться на основі p змінних, з узагальненою відстанню D_{p+q}^2, яка базується на $(p + q)$ змінних. Як правило, виконується така нерівність [4]:

$$D_{p+q}^2 > D_p^2$$

і чим меншою буде різниця $D_{p+q}^2 - D_p^2$, тим меншим буде вплив додаткових q змінних.

Таким чином, нульова гіпотеза про те, що q додаткових змінних не збільшують роздільної здатності дискримінантної функції відхиляється на заданому рівні значущості, якщо [4]

$$F = \frac{n_1 + n_2 - p - q - 1}{q} \times \frac{m\left(D_{p+q}^2 - D_p^2\right)}{1 + m\,D_p^2} > F_{\alpha;q,n_1+n_2-p-q-1},$$

де m — визначається так само, як і раніше.

Випадок аналізу більше двох груп. Дискримінантний аналіз можна узагальнити на довільну кількість груп. Наприклад, може виникнути необхідність у класифікації об'єктів на три категорії: "*неякісні*", "*середні*" і "*добрі*", або у загальному випадку на K категорій. Як правило, такі задачі розв'язуються іншими методами.

11.4. Контрольні питання і вправи

1. Яка мета дискримінантного аналізу?
2. Узагальненням якого методу є дискримінантний аналіз?
3. Назвіть галузі застосування дискримінантного аналізу.
4. Поясніть складові лінійної дискримінантної функції:

$$y_i(k) = \beta_1 x_{i1}(k) + \beta_2 x_{i2}(k) + \ldots + \beta_p x_{ip}(k),\ i = 1, 2, \ldots$$

5. Поясніть сутність і мету використання відношення

$$\lambda = \frac{\text{Міжгрупова варіація}}{\text{Внутрішньогрупова варіація}}.$$

6. Що використовують у дискримінантному аналізі для визначення варіації всередині груп?
7. Поясніть складові формули

$$\lambda = \frac{\text{Міжгрупова варіація}}{\text{Внутрішньогрупова варіація}} = \frac{1}{n_1 + n_2 - 2} \times$$

$$\times \frac{\beta^T (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \beta}{\beta^T \mathbf{S}^* \beta}.$$

8. Яку величину називають відстанню Махаланобіса? Де вона використовується?
9. Які типи критеріїв застосовують у прийнятті рішень при використанні функцій дискримінантного аналізу?
10. Поясніть сутність статистики Хотелінга.
11. Поясніть сутність третього критерію, який використовують при аналізі функцій дискримінантного аналізу.

Розділ 12

СТАТИСТИЧНА ОБРОБКА ДАНИХ ЗА ДОПОМОГОЮ ОПТИМАЛЬНОГО ФІЛЬТРА

12.1. Принцип рекурсивного оцінювання

Розглянемо принцип рекурсивного оцінювання даних, який у подальшому буде використовуватись при виведенні алгоритмів оптимального оцінювання (фільтрації). Принцип рекурсивного оцінювання можна легко пояснити на прикладі знаходження поточного середнього значення часового ряду:

$$\bar{x}(k)=\frac{1}{k}\sum_{i=1}^{k}x(i), \tag{12.1.1}$$

де $\bar{x}(k)$ — це оцінка середнього значення послідовності $x(k)$. Формулу для середнього значення представимо у вигляді:

$$\bar{x}(k)=\frac{k-1}{k-1}\cdot\frac{1}{k}\left[\sum_{i=1}^{k-1}x(i)+x(k)\right]=\frac{k-1}{k}\cdot\frac{1}{k-1}\left[\sum_{i=1}^{k-1}x(i)+x(k)\right]=$$

$$=\left(1-\frac{1}{k}\right)\cdot\left[\frac{1}{k-1}\sum_{i=1}^{k-1}x(i)+\frac{1}{k-1}x(k)\right]=\left(1-\frac{1}{k}\right)\cdot\left[\bar{x}(k-1)+\frac{1}{k-1}x(k)\right]=$$

$$=\bar{x}(k-1)-\frac{1}{k}\bar{x}(k-1)+\frac{1}{k}x(k).$$

Тобто

$$\bar{x}(k)=\bar{x}(k-1)+\frac{1}{k}\cdot\left[x(k)-\bar{x}(k-1)\right]. \tag{12.1.2}$$

Це рівняння має рекурсивну форму, тому що поточна оцінка $\bar{x}(k)$ базується на попередній оцінці $\bar{x}(k-1)$. Другий член у правій частині останньої формули коригує оцінку в момент k. Таким чином, отримано рівняння у формі, яка подібна до рівняння фільтра Калмана [8]. Таке рівняння можна вважати рівнянням оцінювання поточного середнього у формі фільтра Калмана з нестаціонарним коефіцієнтом, що дорівнює $1/k$.

Перевага рекурсивного рівняння у порівнянні з нерекурсивною формою (12.1.1) полягає в тому, що рекурсивне рівняння не потре-

бує запам'ятовування всієї вибірки значень $x(k)$ і середнє значення оцінки може знаходитися у будь-якому інтервалі. Воно базується на оцінці, отриманій у попередній момент часу, та на поточних вимірах.

12.2. Дискретний фільтр Калмана для вільної динамічної системи

Дискретний фільтр Калмана описується множиною рекурсивних матричних рівнянь, які в лінійному випадку відносно легко розв'язуються на комп'ютері. У даному розділі ми виведемо рівняння алгоритму оптимальної фільтрації.

Постановка задачі. Нестаціонарна лінійна система описується у дискретному часі рівняннями зі змінними з часом коефіцієнтами (це означає залежність від часу k):

$$\mathbf{x}(k) = \mathbf{A}(k, k-1)\mathbf{x}(k-1) + \mathbf{B}(k, k-1)\mathbf{u}(k-1) + \mathbf{w}(k-1), \quad (12.2.1)$$

де $\mathbf{x}(k)$ — n-вимірний вектор станів системи; $\mathbf{u}(k-1)$ — m-вимірний вектор детермінованих вхідних величин (сигналів керування); $\mathbf{w}(k-1)$ — n-вимірний вектор випадкових зовнішніх збурень; $\mathbf{A}(k, k-1)$ — $(n \times n)$ матриця динаміки системи (вона містить коефіцієнти, що характеризують динаміку, тобто швидкість зміни станів у часі); $\mathbf{B}(k, k-1)$ — це $(n \times m)$ матриця коефіцієнтів керування. Подвійний часовий аргумент у вигляді $(k, k-1)$ означає, що величини з цим аргументом використовуються в момент k, але їх значення базуються на попередніх даних, які відомі на момент $k-1$ включно. Далі будемо записувати для простоти матриці $\mathbf{A}$ і $\mathbf{B}$ з одним аргументом, тобто $\mathbf{A}(k)$ та $\mathbf{B}(k)$. Очевидно, що стаціонарна система описується матрицями з постійними коефіцієнтами, які записують просто $\mathbf{A}$ і $\mathbf{B}$.

Нехай послідовність зовнішніх збурень $\mathbf{w}(k)$ задовольняє властивостям білого гаусового шуму з нульовим середнім значенням і коваріаційною матрицею $\mathbf{Q}$, тобто статистики шуму мають вигляд:

$$\begin{aligned} E\left[\mathbf{w}(k)\right] &= 0, \ \forall k; \\ E\left[\mathbf{w}(k)\mathbf{w}^T(j)\right] &= \mathbf{Q}(k)\delta_{kj}, \end{aligned} \quad (12.2.2)$$

де δ_{kj} — дельта-функція Кронекера, що визначається як $\delta_{kj} = \begin{cases} 0 & \text{для } k \neq j \\ 1 & \text{для } k = j \end{cases}$; $\mathbf{Q}(k)$ — додатно визначена коваріаційна матриця

зовнішніх збурень розмірності $(n \times m)$. Діагональні елементи матриці є дисперсії компонент вектора $\mathbf{w}(k)$.

Початковим станом системи $\mathbf{x}_0$ будемо вважати випадкові змінні з відомими статистиками:

$$E[\mathbf{x}_0] = \bar{\mathbf{x}}_0;\ \ E\left[\mathbf{x}_0\mathbf{x}_0^T\right] = \mathbf{M};\ \ E\left[\mathbf{w}(k)\mathbf{x}_0^T\right] = 0,\ \ \forall k. \qquad (12.2.3)$$

Нехай вектор вимірів $\mathbf{z}(k)$ вихідних змінних доступний у будь-який момент часу k, а його компоненти лінійно пов'язані з вектором стану і на них впливає шум, тобто

$$\mathbf{z}(k) = \mathbf{H}(k)\mathbf{x}(k) + \mathbf{v}(k), \qquad (12.2.4)$$

де $\mathbf{H}(k)$ — матриця спостережень розмірності $(r \times n)$, $\mathbf{v}(k)$ — r-вимірний вектор випадкових величин шуму вимірювання з відомими статистиками:

$$E\left[\mathbf{v}(k)\right] = 0,\ \ E\left[\mathbf{v}(k)\mathbf{v}^T(j)\right] = \mathbf{R}(k)\delta_{kj},$$

де $\mathbf{R}(k)$ — додатно визначена коваріаційна матриця шуму розмірності $(r \times r)$, діагональні елементи якої є дисперсіями адитивного шуму в кожному каналі вимірювань. Шум вимірювань також задовольняє властивостям білого гаусового шуму. Він вважається некорельованим із зовнішнім збуренням $\mathbf{w}(k)$ і початковим станом системи, тобто

$$E\left[\mathbf{v}(k)\mathbf{w}^T(j)\right] = 0, \quad \forall k, j;$$

$$E\left[\mathbf{v}(k)\mathbf{x}_0^T\right] = 0, \quad \forall k.$$

Для системи з вектором стану $x(k)$ необхідно знайти оцінку стану $\hat{\mathbf{x}}(k)$ в момент k як лінійну комбінацію оцінки $\hat{x}\ (k-1)$ в момент $k-1$ і самого останнього виміру (або статистичних даних) $z(k)$.

Оцінка $\hat{x}\ (k)$ повинна знаходитися як найкраща за мінімумом середнього значення суми квадратів оцінок помилок. Інакше кажучи, оцінка повинна бути такою, щоб

$$E\left[\left(\hat{\mathbf{x}}(k) - \mathbf{x}(k)\right)^T\left(\hat{\mathbf{x}}(k) - \mathbf{x}(k)\right)\right] = \min_K, \qquad (12.2.5)$$

де $\mathbf{x}(k)$ — точне значення вектора стану, яке може бути обчислене за допомогою детермінованої частини математичної моделі процесу; $\mathbf{K}$ — оптимальний матричний коефіцієнт фільтра, який необхідно знайти в результаті розв'язку оптимізаційної задачі.

Концепція вільної динамічної системи

Вільна динамічна система простіша, ніж повна система (12.2.1), в якій відсутні зовнішні вхідні дані. Така система описується однорідним лінійним різницевим рівнянням [8]

$$\mathbf{x}(k) = \mathbf{A}(k)\,\mathbf{x}(k-1), \qquad (12.2.6)$$

Це рівняння має простішу структуру, ніж рівняння (12.2.1) і за його допомогою можна легше вивести рівняння фільтра Калмана. Для цього спочатку треба отримати рівняння фільтра для вільної динамічної системи (12.2.6).

Формування рівняння фільтра

Виберемо структуру рівняння для оцінювання лінійної системи. Прогноз оцінки стану в момент часу k можна знайти за допомогою матриці динаміки $\mathbf{A}(k)$ з рівняння (12.2.6).

Рівняння (12.2.6) показує, що прогноз оцінки в момент k визначається як результат множення оцінки вектора стану в момент $(k-1)$ на матрицю $\mathbf{A}(k)$, тобто прогноз оцінки визначається тільки динамікою системи.

Для коригування прогнозу оцінки необхідно використати вимір $\mathbf{z}(k)$. З рівняння (12.2.4) видно, що очікуване значення $\hat{\mathbf{z}}(k)$ вихідного вектора вимірів $\mathbf{z}(k)$ в момент k буде $\mathbf{H}(k)\hat{\mathbf{x}}(k)$ або

$$\hat{\mathbf{z}}(k) = \mathbf{H}(k)\mathbf{A}(k)\hat{\mathbf{x}}(k-1).$$

Похибка прогнозованого значення виміру може бути знайдена як різниця між вимірюванням та очікуваною величиною, тобто

$$\mathbf{e}_p(k) = \mathbf{z}(k) - \mathbf{H}(k)\mathbf{A}(k)\hat{\mathbf{x}}(k-1).$$

Таким чином, оцінка $\hat{\mathbf{x}}(k)$ є лінійною комбінацією оцінки $\hat{\mathbf{x}}(k-1)$ у попередній момент і значення вихідного вектора $\mathbf{z}(k)$, що вимірюється, в момент k:

$$\hat{\mathbf{x}}(k) = \mathbf{A}(k)\hat{\mathbf{x}}(k-1) + \mathbf{K}(k)\left[\mathbf{z}(k) - \mathbf{H}(k)\mathbf{A}(k)\hat{\mathbf{x}}(k-1)\right]. \quad (12.2.7)$$

У цьому рівнянні на вибір матриці $\mathbf{K}(k)$ (матричний коефіцієнт фільтра) впливають кілька факторів, які вибираються таким чином, щоб мінімізувати математичне сподівання квадрату оцінки похибки, тобто задовольняють критерію (12.2.5). Матрицю $\mathbf{K}(k)$ називають матричним оптимальним фільтром Калмана.

Похибка оцінки вектора стану за визначенням

$$\mathbf{e}(k)=\hat{\mathbf{x}}(k)-\mathbf{x}(k);$$

$$E\left[\left(\hat{\mathbf{x}}(k)-\mathbf{x}(k)\right)^T\left(\hat{\mathbf{x}}(k)-\mathbf{x}(k)\right)\right]=E\left[\mathbf{e}^T(k)\mathbf{e}(k)\right];$$

$$E\left[\mathbf{e}^T(k)\mathbf{e}(k)\right]=tr\,E\left[\mathbf{e}(k)\mathbf{e}^T(k)\right]=tr\left[\mathbf{P}(k)\right],$$

де $\mathbf{P}(k)$ — коваріаційна матриця похибок оцінок вектора стану, яка асоціюється з оцінкою стану, визначеною рівнянням (12.2.7), а $tr\,[\mathbf{P}(k)]$ — сума діагональних елементів матриці $\mathbf{P}(k)$. Ця матриця необхідна для того, щоб знайти у подальшому вираз для матричного оптимального коефіцієнта фільтра.

Для того щоб вивести вираз для матриці $\mathbf{P}(k)$, запишемо спочатку вираз для вектора похибок оцінок $\mathbf{e}(k)$:

$$\begin{aligned}\mathbf{e}(k)&=\left[\mathbf{A}(k)\hat{\mathbf{x}}(k-1)+\mathbf{K}(k)\left(\mathbf{z}(k)-\mathbf{HA}\,\hat{\mathbf{x}}(k-1)\right)\right]-\mathbf{A}(k)\mathbf{x}(k-1)=\\&=\mathbf{A}(k)\mathbf{e}(k-1)-\mathbf{K}(k)\mathbf{HA}\,\hat{\mathbf{x}}(k-1)+\mathbf{K}(k)\mathbf{z}(k)=\\&=\mathbf{A}(k)\mathbf{e}(k-1)-\mathbf{K}(k)\mathbf{HA}\left[\mathbf{e}(k-1)+\mathbf{x}(k-1)\right]+\\&+\mathbf{K}(k)\left[\mathbf{Hx}(k)+\mathbf{v}(k)\right]=\left[\mathbf{I}-\mathbf{K}(k)\mathbf{H}\right]\mathbf{Ae}(k-1)-\\&-\mathbf{K}(k)\mathbf{HA}\,\mathbf{x}(k-1)+\mathbf{K}(k)\mathbf{HA}\,\mathbf{x}(k-1)+\mathbf{K}(k)\mathbf{v}(k)=\\&=\left[\mathbf{I}-\mathbf{K}(k)\mathbf{H}\right]\mathbf{Ae}(k-1)+\mathbf{K}(k)\mathbf{v}(k).\end{aligned}\qquad(12.2.8)$$

Тут $\mathbf{H}=\mathbf{H}(k)$, а $\mathbf{I}$ — матриця, на головній діагоналі якої стоять одиниці, а решта елементів нулі.

Тепер підставимо (12.2.8) у вираз для коваріаційної матриці $\mathbf{P}(k)=E\left[\mathbf{e}(k)\mathbf{e}^T(k)\right]$ і отримаємо:

$$\begin{aligned}\mathbf{P}(k)=&\left[\mathbf{I}-\mathbf{K}(k)\mathbf{H}\right]\mathbf{A}\,E\left[\mathbf{e}(k-1)\mathbf{e}^T(k-1)\right]\mathbf{A}^T\left[\mathbf{I}-\mathbf{H}^T\mathbf{K}^T(k)\right]+\\&+\mathbf{K}(k)E\left[\mathbf{v}(k)\mathbf{e}^T(k-1)\right]\mathbf{A}^T\left[\mathbf{I}-\mathbf{H}^T(k)\mathbf{K}^T(k)\right]+\\&+\left[\mathbf{I}-\mathbf{K}(k)\mathbf{H}(k)\right]\mathbf{A}\,E\left[\mathbf{e}(k-1)\mathbf{v}^T(k)\right]\mathbf{K}^T(k)+\\&+\mathbf{K}(k)E\left[\mathbf{v}(k)\mathbf{v}^T(k)\right]\mathbf{K}^T(k).\end{aligned}$$

За визначенням,

$$\begin{aligned}&E\left[\mathbf{e}(k-1)\mathbf{e}^T(k-1)\right]=\mathbf{P}(k-1);\\&E\left[\mathbf{v}(k)\mathbf{v}^T(k)\right]=\mathbf{R}(k).\end{aligned}\qquad(12.2.9)$$

Беручи до уваги, що $E\left[\mathbf{v}(k)\mathbf{v}^T(k-1)\right]=0$ та $E\left[\mathbf{v}(k)\mathbf{x}_0^T\right]=0$, так як $\mathbf{v}(k)$ і $\mathbf{x}_0^T$ некорельовані, то отримаємо

$$E\left[\mathbf{v}(k)\mathbf{e}^T(k-1)\right]=0,$$

і

$$E\left[\mathbf{e}(k-1)\mathbf{v}^T(k)\right]=0.$$

Тобто значення шуму $\mathbf{v}(k)$ в момент k не корельоване з вектором стану $\mathbf{x}(k-1)$ в момент $k-1$. Вектор $\hat{\mathbf{x}}(k-1)$, який залежить від $\mathbf{v}(k-1)$, некорельований з $\mathbf{v}(k)$, тому що немає кореляції між $\mathbf{v}(k-1)$ та $\mathbf{v}(k)$. Звідси отримуємо, що вектор $\mathbf{e}(k-1)=\hat{\mathbf{x}}(k-1)-\mathbf{x}(k-1)$ некорельований з $\mathbf{v}(k)$.

Апріорна коваріаційна матриця похибок оцінок вектора стану

Беручи до уваги (12.2.9) і останній вираз, запишемо $\mathbf{P}(k)$ так

$$\mathbf{P}(k)=\left[\mathbf{I}-\mathbf{K}(k)\mathbf{H}\right]\mathbf{P}'(k)\left[\mathbf{I}-\mathbf{K}(k)\mathbf{H}\right]^T+\mathbf{K}(k)\mathbf{R}(k)\mathbf{K}^T(k), \quad (12.2.10)$$

де

$$\mathbf{P}'(k)=\mathbf{A}\mathbf{P}(k-1)\mathbf{A}^T.$$

Матрицю $\mathbf{P}'(k)$ називають *апріорною* коваріаційною матрицею похибок оцінок вектора стану, тому що вона характеризує оцінку вектора стану до надходження вимірів $\mathbf{z}(k)$. Тому матрицю $\mathbf{P}(k)$ називають *апостеріорною* коваріаційною матрицею похибок оцінок.

Після деяких алгебраїчних перетворень співвідношення (12.2.10) маємо

$$\begin{aligned}\mathbf{P}(k)&=\left[\mathbf{P}'(k)-\mathbf{K}(k)\mathbf{H}\mathbf{P}'(k)\right]\left[\mathbf{I}-\mathbf{H}^T\mathbf{K}^T(k)\right]+\mathbf{K}(k)\mathbf{R}(k)\mathbf{K}^T(k)=\\&=\mathbf{P}'(k)-\mathbf{K}(k)\mathbf{H}\mathbf{P}'(k)-\mathbf{P}'(k)\mathbf{H}^T\mathbf{K}^T(k)+\\&+\mathbf{K}(k)\left[\mathbf{H}\mathbf{P}'(k)\mathbf{H}^T+\mathbf{R}(k)\right]\mathbf{K}^T(k).\end{aligned} \quad (12.2.11)$$

Таке рівняння називають рівнянням Ріккаті [8].

Матриця $\left[\mathbf{H}\mathbf{P}'(k)\mathbf{H}^T+\mathbf{R}\right]$ — симетрична, оскільки

$$\left[\mathbf{H}\mathbf{P}'(k)\mathbf{H}^T+\mathbf{R}\right]=\left[\mathbf{H}\mathbf{P}'(k)\mathbf{H}^T+\mathbf{R}\right]^T,$$

та невід'ємно визначена. Таку матрицю можна представити у вигляді:

$$\mathbf{H}\mathbf{P}'(k)\mathbf{H}^T+\mathbf{R}=\mathbf{S}(k)\mathbf{S}^T(k), \quad (12.2.12)$$

де матриця $\mathbf{S}(k)$ — на даний момент невідома, але буде визначена нижче.

Таким чином, формула для $\mathbf{P}(k)$ приймає вигляд

$$\mathbf{P}(k)=\mathbf{P}'(k)-\mathbf{K}(k)\mathbf{H}\mathbf{P}'(k)-\mathbf{P}'(k)\mathbf{H}^T\mathbf{K}^T(k)+\mathbf{K}(k)\mathbf{S}(k)\mathbf{S}^T(k)\mathbf{K}^T(k),$$

або в спрощеному вигляді без індексів

$$\mathbf{P}=\mathbf{P}'-\mathbf{K}\mathbf{H}\mathbf{P}'-\mathbf{P}'\mathbf{H}^T\mathbf{K}^T+\mathbf{K}\mathbf{S}\mathbf{S}^T\mathbf{K}^T. \qquad (12.2.13)$$

Праву частину цього рівняння необхідно привести до такого вигляду, щоб можна було:

- визначити коефіцієнт фільтра $\mathbf{K}(k)$ із умови мінімуму суми діагональних елементів матриці $\mathbf{P}(k)$;
- забезпечити невід'ємність діагональних елементів матриці у правій частині рівняння.

З цією метою введемо до розгляду наступний добуток

$$(\mathbf{KS}-\mathbf{G})(\mathbf{KS}-\mathbf{G})^T=\mathbf{KSS}^T\mathbf{K}^T-\mathbf{KSG}^T-\mathbf{GS}^T\mathbf{K}^T+\mathbf{GG}^T,$$

де $\mathbf{G}$ — невідома матриця, що існує за припущенням. Діагональні елементи матриці $(\mathbf{KS}-\mathbf{G})(\mathbf{KS}-\mathbf{G})^T$ повинні бути невід'ємними, тобто $\{(\mathbf{KS}-\mathbf{G})(\mathbf{KS}-\mathbf{G})^T\}\geq 0$.

Знайдемо матрицю $\mathbf{G}$ із умови

$$\mathbf{KHP}'+\mathbf{P}'\mathbf{H}^T\mathbf{K}^T=\mathbf{KSG}^T+\mathbf{GS}^T\mathbf{K}^T, \qquad (12.2.14)$$

і запишемо тепер рівняння (12.2.13) у вигляді

$$\mathbf{P}=\mathbf{P}'+(\mathbf{KS}-\mathbf{G})(\mathbf{KS}-\mathbf{G})^T-\mathbf{GG}^T. \qquad (12.2.15)$$

Перехід від (12.2.13) до (12.2.15) дає можливість записати наступні рівності (з використанням (12.2.14)):

$$\mathbf{KSG}^T=\mathbf{KHP}';$$
$$\mathbf{GS}^T\mathbf{K}^T=\mathbf{P}'\mathbf{H}^T\mathbf{K}^T$$

або $$\mathbf{GS}^T=\mathbf{P}'\mathbf{H}^T.$$

Тепер знайдемо $\mathbf{G}$:

$$\mathbf{G}=\mathbf{P}'\mathbf{H}^T(\mathbf{S}^T)^{-1}=\mathbf{P}'\mathbf{H}^T(\mathbf{S}^{-1})^T. \qquad (12.2.16)$$

Знаходження оптимального матричного коефіцієнта фільтра

У рівнянні (12.2.15) від $\mathbf{K}(k)$ залежить тільки середній член, який є добутком матриці $(\mathbf{KS}-\mathbf{G})$ на її транспоновану. Це забезпечує невід'ємність елементів головної діагоналі середнього члена у (12.2.15).

Таким чином, сума діагональних елементів матриці $\mathbf{P}(k)$ буде мінімальною, якщо покласти середній член (12.2.15) рівним нулю, тобто

$$\mathbf{KS} - \mathbf{G} = 0,$$

або $\mathbf{KS} = \mathbf{G}$, а звідси $\mathbf{K}(k)$ можна визначити як

$$\mathbf{K}(k) = \mathbf{GS}^{-1}. \qquad (12.2.17)$$

Підставимо у (12.2.17) значення, що визначається рівнянням (12.2.16) і отримаємо:

$$\mathbf{K}(k) = \mathbf{P}'\mathbf{H}^T(\mathbf{S}^{-1})^T\mathbf{S}^{-1} = \mathbf{P}'\mathbf{H}^T(\mathbf{SS}^T)^{-1},$$

а із врахуванням (12.2.12) остаточно можна записати:

$$\mathbf{K}(k) = \mathbf{P}'\mathbf{H}^T[\mathbf{HP}'\mathbf{H}^T + \mathbf{R}]^{-1}.$$

Враховуючи часовий аргумент, отримаємо рівняння

$$\mathbf{K}(k) = \mathbf{P}'(k)\mathbf{H}^T(k)[\mathbf{H}(k)\mathbf{P}'(k)\mathbf{H}^T(k) + \mathbf{R}(k)]^{-1}. \qquad (12.2.18)$$

Таким чином, коефіцієнт фільтра Калмана (12.2.18) знайдено із умови мінімуму середньоквадратичної похибки оцінок вектора станів об'єкта.

Апостеріорна коваріаційна матриця похибок оцінок вектора станів

Знайдемо формулу для коваріаційної матриці, яка пов'язана з оптимальною оцінкою станів. Розглянемо знову коваріаційну матрицю, що визначається рівнянням Ріккаті (12.2.11), тобто

$$\mathbf{P}(k) = \mathbf{P}'(k) - \mathbf{KHP}'(k) - \mathbf{P}'(k)\mathbf{H}^T\mathbf{K}^T(k) + \\ + \mathbf{K}(k)[\mathbf{HP}'(k)\mathbf{H}^T + \mathbf{R}(k)]\,\mathbf{K}^T(k).$$

Підставимо в це рівняння значення $\mathbf{K}(k)$, що визначається (12.2.18), і отримаємо:

$$\mathbf{P}(k) = \mathbf{P}'(k) - \mathbf{P}'(k)\mathbf{H}^T[\mathbf{HP}'(k)\mathbf{H}^T + \mathbf{R}]^{-1}\,\mathbf{HP}' = \\ = \mathbf{P}'(k) - \mathbf{K}(k)\mathbf{HP}'(k) = [\mathbf{I} - \mathbf{K}(k)\mathbf{H}]\mathbf{P}'(k). \qquad (12.2.19)$$

Тобто апостеріорна коваріаційна матриця похибок оцінок є меншою порівняно з апріорною на величину $\mathbf{K}(k)\mathbf{HP}'(k)$. Це пояснюється тим, що отриманий вимір $\mathbf{z}(k)$ в момент часу k сприяє зменшенню невизначеності оцінок вектора станів.

Результати

Матричне рівняння (12.2.7) використовується для знаходження оптимальних оцінок вектора стану в момент k за допомогою оптимального коефіцієнта фільтра $\mathbf{K}(k)$, який можна знайти із системи матричних рівнянь. Матричні рівняння (12.2.7), (12.2.10) і (12.2.19) утворюють рекурсивний алгоритм фільтра Калмана. Зрозуміло, що процес оцінювання може продовжуватись до нескінченності. Послідовність обчислюваних кроків показано нижче у вигляді алгоритму. Нагадаємо, що рівняння фільтра Калмана побудовані для вільної динамічної системи.

Алгоритм фільтрації (оптимального оцінювання стану) вільної динамічної системи

1. Задати початкові умови $\mathbf{x}_0$ для вектора стану і коваріаційної матриці похибок оцінок $\mathbf{P}_0$. Присвоїти значення коваріаційним матрицям збурень стану $\mathbf{Q}$ та похибок вимірів $\mathbf{R}$.

2. Знайти матричний оптимальний коефіцієнт фільтра:

$$\mathbf{K}(k) = \mathbf{P}'(k)\mathbf{H}^T[\mathbf{H}\mathbf{P}'(k)\,\mathbf{H}^T + \mathbf{R}]^{-1}.$$

3. Скористатись новими вимірами для знаходження поточної оцінки вектора стану:

$$\hat{\mathbf{x}}(k) = \mathbf{A}\hat{\mathbf{x}}(k-1) + \mathbf{K}(k)\left[\mathbf{z}(k) - \mathbf{H}\mathbf{A}\hat{\mathbf{x}}(k-1)\right].$$

4. Знайти апостеріорну коваріаційну матрицю похибок для оновлених оцінок:

$$\mathbf{P}(k) = [\mathbf{I} - \mathbf{K}(k)\mathbf{H}]\,\mathbf{P}'(k).$$

5. Знайти апріорну (для наступної оцінки вектора стану) коваріаційну матрицю похибок оцінок:

$$\mathbf{P}'(k) = \mathbf{A}\mathbf{P}(k-1)\mathbf{A}^T.$$

і перейти на крок 2 (наступна реалізація рівнянь фільтра).

Розглянутий фільтр може продовжувати функціонувати нескінченно довго при надходженні нових вимірів. Після перехідного процесу коефіцієнти фільтра встановлюються, як правило, на деяких постійних значеннях, які залишаються незмінними на довгих проміжках часу, якщо процес стаціонарний. Якщо фільтр працює нормально, то діагональні елементи коваріаційної матриці похибок оцінок вектора стану залишаються додатно визначеними (оскільки це

дисперсії похибок оцінок) і також прямують до невеликих сталих значень. Якщо дисперсії похибок оцінок стану не зменшуються, а зростають з часом, то такий режим називають розбіжним. Із збільшенням часу розрядна сітка може переповнюватися і функціонування алгоритму повністю припиняється. Очевидно, що виникнення такого режиму функціонування фільтра необхідно попереджувати. Методи підвищення ефективності алгоритму фільтрації будуть розглянуті нижче у цьому розділі, а в наступному параграфі отримаємо рівняння фільтра для динамічної системи, яка функціонує в умовах впливу зовнішніх збурень.

12.3. Дискретний фільтр Калмана для лінійної системи з детермінованими і стохастичними входами

Будемо вважати розширеною математичною моделлю лінійну систему з детермінованими і стохастичними входами (збуреннями):

$$\begin{aligned}\mathbf{x}(k) &= \mathbf{A}(k)\mathbf{x}(k-1)+\mathbf{B}(k)\mathbf{u}(k-1)+\mathbf{w}(k-1);\\ \mathbf{z}(k) &= \mathbf{H}(k)\mathbf{x}(k)+\mathbf{v}(k).\end{aligned} \qquad (12.3.1)$$

Доданок $\mathbf{B}(k)\mathbf{u}(k-1)$ першого рівняння системи (12.3.1) характеризує вплив детермінованого вхідного сигналу $\mathbf{u}(k-1)$; $\mathbf{w}(k-1)$ — стохастичні збурення, які впливають на лінійну систему.

Для початку знайдемо прогнозовану оцінку вектора стану, використовуючи попередню оцінку $\hat{\mathbf{x}}(k-1)$ для моменту $k-1$, тобто не будемо враховувати вимірювання в момент k. Стохастичне збурення стану (або шум) $\mathbf{w}(k-1)$ не залежить від вектора стану в момент $k-1$ і має нульове середнє значення. Тому вважаємо, що цей шум не буде впливати на оцінку вектора стану в момент k. Але $\mathbf{B}(k)\mathbf{u}(k-1) \equiv \mathbf{f}(k-1)$ — це відома векторна функція на інтервалі $[k-1, k]$ і, згідно з рівнянням (12.3.1), прогноз оцінки вектора стану може бути представлений у наступному вигляді:

$$\hat{\mathbf{x}}'(k) = \mathbf{A}(k)\hat{\mathbf{x}}(k-1)+\mathbf{f}(k-1). \qquad (12.3.2)$$

Після появи виміру вихідного сигналу $\mathbf{z}(k)$ в момент k можна знайти нову оцінку вектора стану за формулою

$$\hat{\mathbf{x}}(k) = \hat{\mathbf{x}}'(k)+\mathbf{K}(k)\left[\mathbf{z}(k)-\mathbf{H}(k)\hat{\mathbf{x}}'(k)\right], \qquad (12.3.3)$$

де $\mathbf{K}(k)$ — невідома матриця коефіцієнтів фільтра. (У подальшому будемо записувати матриці $\mathbf{A}$, $\mathbf{H}$ без аргументів для спрощення запису)

$$\begin{aligned}\mathbf{e}(k)=\hat{\mathbf{x}}(k)-\mathbf{x}(k)&=\left\{\mathbf{A}\hat{\mathbf{x}}(k-1)+\mathbf{f}(k-1)+\mathbf{K}(k)\left[\mathbf{z}(k)-\mathbf{H}\hat{\mathbf{x}}'(k)\right]\right\}-\\&\quad-\left[\mathbf{A}\mathbf{x}(k-1)+\mathbf{f}(k-1)+\mathbf{w}(k-1)\right]=\\&=\left[\mathbf{A}\mathbf{e}(k-1)+\mathbf{K}(k)\left[\mathbf{H}\mathbf{x}(k)+\mathbf{v}(k)\right]-\mathbf{K}(k)\mathbf{H}\hat{\mathbf{x}}'(k)-\mathbf{w}(k-1)\right]=\\&=\mathbf{A}\mathbf{e}(k-1)+\mathbf{K}(k)\mathbf{H}\left[\mathbf{A}\mathbf{x}(k-1)+\mathbf{f}(k-1)+\mathbf{w}(k-1)\right]-\\&\quad-\mathbf{K}(k)\mathbf{H}\left[\mathbf{A}\hat{\mathbf{x}}(k-1)+\mathbf{f}(k-1)\right]-\mathbf{w}(k-1)+\mathbf{K}(k)\mathbf{v}(k)=\\&=\mathbf{A}\mathbf{e}(k-1)-\mathbf{K}(k)\mathbf{H}\mathbf{A}\mathbf{e}(k-1)+\mathbf{K}(k)\mathbf{H}\mathbf{w}(k-1)-\mathbf{w}(k-1)+\mathbf{K}(k)\mathbf{v}(k)=\\&=\left[\mathbf{I}-\mathbf{K}(k)\mathbf{H}\right]\left[\mathbf{A}\mathbf{e}(k-1)-\mathbf{w}(k-1)\right]+\mathbf{K}(k)\mathbf{v}(k)\,.\end{aligned}\tag{12.3.4}$$

Тепер знайдемо апостеріорну коваріаційну матрицю похибок оцінок

$$\begin{aligned}\mathbf{P}(k)&=E\left[\mathbf{e}(k)\mathbf{e}^T(k)\right]=\\&=E\left\{\begin{matrix}\left[\left[\mathbf{I}-\mathbf{K}(k)\mathbf{H}\right]\left[\mathbf{A}\mathbf{e}(k-1)-\mathbf{w}(k-1)\right]+\mathbf{K}(k)\mathbf{v}(k)\right]\cdot\\\left[\left[\mathbf{I}-\mathbf{K}(k)\mathbf{H}\right]\left[\mathbf{A}\mathbf{e}(k-1)-\mathbf{w}(k-1)\right]+\mathbf{K}(k)\mathbf{v}(k)\right]^T\end{matrix}\right\}.\end{aligned}\tag{12.3.5}$$

Введемо апріорну коваріаційну матрицю похибок оцінок вектора стану $\mathbf{P}'(k)$:

$$\begin{aligned}\mathbf{P}'(k)&=E\left\{\left[\mathbf{A}\mathbf{e}(k-1)-\mathbf{w}(k-1)\right]\left[\mathbf{A}\mathbf{e}(k-1)-\mathbf{w}(k-1)\right]^T\right\}=\\&=\mathbf{A}\mathbf{P}(k-1)\mathbf{A}^T+\mathbf{Q}(k-1)\,,\end{aligned}\tag{12.3.6}$$

де $\mathbf{Q}(k-1)$ — коваріаційна матриця для вектора $\mathbf{w}(k)$.

Відповідно до (12.3.5) апостеріорна коваріаційна матриця $\mathbf{P}(k)$ визначається рівнянням

$$\mathbf{P}(k)=[\mathbf{I}-\mathbf{K}(k)\mathbf{H}]\mathbf{P}'(k)\,[\mathbf{I}-\mathbf{K}(k)\mathbf{H}]^T+\mathbf{K}(k)\mathbf{R}(k)\mathbf{K}^T(k).\tag{12.3.7}$$

Структура цього рівняння ідентична структурі рівняння для апостеріорної матриці, отриманої в попередньому параграфі для вільної динамічної системи. Це означає, що оптимальний матричний коефіцієнт фільтра $\mathbf{K}(k)$ для системи (12.3.1) визначається таким самим рівнянням, що і для вільної динамічної системи. Звідси випливає, що оптимальний дискретний фільтр для лінійної системи (12.3.1) визначається рівняннями: (12.3.2), (12.3.3), (12.2.19), (12.2.20) і (12.3.6).

Очевидно, що поява випадкового збурення $\mathbf{w}(k)$ веде до погіршення якості оцінок, про що свідчить додатковий член $\mathbf{Q}(k-1)$ у

правій частині рівняння (12.3.6) для знаходження $\mathbf{P}'(k)$. Детермінований сигнал $\mathbf{f}(k-1)$ впливає на прогноз оцінки вектора стану відповідно до рівняння (12.3.2).

Якщо середнє значення $E[\mathbf{w}(k-1)]$ випадкового збурення $\mathbf{f}(k-1)$ не дорівнює нулю і може бути виміряне (оцінене) у будь-який момент часу, то його можна досить просто врахувати в рівняннях фільтрації. У такому випадку значення $\mu_w = E[\mathbf{w}(k-1)]$ додається до детермінованого впливу $\mathbf{f}(k-1)$.

Крім сигналу керування $\mathbf{Bu}(k-1)$, детерміноване збурення $\mathbf{f}(k-1)$ може включати в себе вплив інших вхідних детермінованих сигналів, що впливають на динамічну систему. Аналогічно можна врахувати ненульове середнє шуму вимірів $\mathbf{v}(k)$.

Система рівнянь оптимального фільтра Калмана для лінійної системи, яка функціонує в умовах впливу детермінованих та випадкових збурень, наведена нижче у вигляді алгоритму фільтрації.

Алгоритм фільтрації для лінійної системи з детермінованими та випадковими входами

1. Математична модель процесу (лінійної системи):

$$\mathbf{x}(k) = \mathbf{A}(k)\mathbf{x}(k-1) + \mathbf{B}(k)\mathbf{u}(k-1) + \mathbf{w}(k-1);$$

$$\mathbf{z}(k) = \mathbf{H}(k)\mathbf{x}(k) + \mathbf{v}(k);$$

$$E\left[\mathbf{w}(k)\right] = 0,\ E\left[\mathbf{v}(k)\right] = 0,\ E\left[\mathbf{x}_0\right] = \bar{\mathbf{x}}_0,\ E\left[\bar{\mathbf{x}}_0\ \bar{\mathbf{x}}_0^T\right] = \mathbf{P}_0 = \mathbf{P}_0';$$

$$E\left[\mathbf{w}(k)\mathbf{w}^T(l)\right] = \mathbf{Q}(k)\delta(k-l),\ E\left[\mathbf{v}(k)\mathbf{v}^T(l)\right] = \mathbf{R}(k)\delta(k-l);$$

$$E\left[\mathbf{w}(k)\mathbf{v}^T(l)\right] = E\left[\mathbf{w}(k)\mathbf{x}_0^T\right] = E\left[\mathbf{v}(k)\mathbf{x}_0^T\right] = 0.$$

2. Матричний коефіцієнт фільтра:

$$\mathbf{K}(k) = \mathbf{P}'(k)\mathbf{H}^T[\mathbf{H}\mathbf{P}'(k)\ \mathbf{H}^T + \mathbf{R}]^{-1}.$$

3. Рівняння фільтрації:

$$\hat{\mathbf{x}}(k) = \mathbf{A}\hat{\mathbf{x}}(k-1) + \mathbf{Bu}(k-1) + \mathbf{K}(k)\left[\mathbf{z}(k) - \mathbf{HA}\hat{\mathbf{x}}(k-1)\right].$$

4. Апріорна коваріаційна матриця похибок оцінок:

$$\mathbf{P}(k) = [\mathbf{I} - \mathbf{K}(k)\mathbf{H}]\mathbf{P}'(k).$$

5. Апостеріорна коваріаційна матриця похибок оцінок для наступного циклу:

$$\mathbf{P}'(k+1) = \mathbf{A}\mathbf{P}(k)\mathbf{A}^T + \mathbf{Q}(k).$$

6. Перейти на крок 2.

Якщо значення коваріаційних матриць **Q** і **R** неможливо оцінити на основі наявної інформації про процес, то необхідно застосувати алгоритм адаптивної фільтрації [8], який дає можливість одночасно із станом оцінювати невідомі статистичні параметри процесу.

Статистичне моделюванні алгоритму фільтрації

Перед практичним застосуванням оптимального фільтра для обробки реальних даних, необхідно виконати його статистичне моделювання, яке полягає у багатократному прогоні алгоритму з використанням множини реалізацій випадкових процесів, що описують вплив збурень стану та шуму вимірів. Суть статистичного моделювання полягає у виконанні наступних кроків:

1. Задати число реалізацій алгоритму фільтрації, якому відповідає число реалізацій випадкових процесів. Як правило, вибирають число реалізацій $M \geq 100$. Покласти число C реалізацій (циклів) алгоритму фільтрації рівним M: $C = M$.

2. Згенерувати вибірки випадкових процесів $\mathbf{w}(k)$ і $\mathbf{v}(k)$, обсяг яких дорівнює обсягу вибірки експериментальних (статистичних) даних, і врахувати їх у вимірюваннях (згенерованих чи реальних).

3. Застосувати алгоритм фільтрації до вибірки вимірів обсягом N і зберегти на кожному кроці ($k = 1, ..., N$) діагональні елементи коваріаційної матриці похибок фільтрації $\mathbf{P}(k)$ (тобто дисперсії похибок оцінок). При кожному наступному прогоні алгоритму фільтрації до цих дисперсій необхідно додати нові, отримані для наступного прогону. Тобто, якщо розмірність вектора стану дорівнює n, то в результаті моделювання отримаємо n векторів дисперсій похибок оцінок розмірністю $[N \times 1]$.

4. Зменшити число реалізацій на 1: C: $C - 1$. Якщо $C \neq 0$, то перейти на крок 2. Інакше, перейти на крок 5. Знайти середнє значення по сукупності реалізацій для кожного діагонального елемента похибок фільтрації $\mathbf{P}(k)$, тобто

$$\overline{p}_{ii}(k) = \frac{1}{M-1}\sum_{l=1}^{M} p_{ii}(l, k), \quad k = 1,\ldots,N,$$

де N — обсяг вибірки даних.

5. Побудувати графіки залежностей усереднених дисперсій похибок оцінок від часу k, тобто $p_{ii}(k)$, $k = 1, ..., N$. Якщо алгоритм реалізовано коректно, то усереднені дисперсії будуть експоненціально спадати до невеликих сталих значень.

Статистичне моделювання алгоритму фільтрації дає можливість дослідити характеристики його функціонування на множині реалізацій випадкових процесів і тим самим отримати гарантію нормального функціонування в системі обробки реальних даних.

12.4. Заходи щодо підвищення якості оптимального фільтра

Показники нормального функціонування оптимального фільтра

Реалізація оптимального фільтра вимагає неперервного стеження за його нормальним функціонуванням. Нормальне функціонування фільтра характеризується наступними показниками:

1. Фільтр Калмана відноситься до рекурсивних алгоритмів оцінювання, характерною ознакою яких є наявність перехідного процесу оцінювання. Протягом перехідного процесу значення оцінок стану можуть досить сильно коливатись, а похибки оцінок приймають великі значення. Для стаціонарного процесу дисперсії похибок оцінок вектора стану мають бути спадаючими функціями часу і прямувати до деяких невеликих постійних сталих значень після закінчення перехідного процесу. Таким чином, контролювати функціонування фільтра можна за допомогою діагональних елементів матриці $\mathbf{P}(k)$.

2. Для стаціонарного процесу коефіцієнти підсилення фільтра $\mathbf{K}(k)$ повинні приймати постійні значення.

3. Графіки змінних стану $\mathbf{x}(k)$ не повинні мати значних імпульсів чи викидів по закінченню перехідного процесу.

4. Послідовність $\{\nu(k) = \mathbf{z}(k) - \mathbf{AH}\,\hat{\mathbf{x}}(k-1)\}$ повинна задовольняти властивостям білого шуму: $E[\nu(k)] = 0$ і $E[\nu(k)\,\nu(k-l)] = 0,\ l \geq 1$.

Остання властивість досить часто і успішно використовується у системах автоматичної діагностики об'єктів автоматичного керування для виявлення аномальних режимів процесів.

Причини неефективності алгоритму оптимальної фільтрації

У практиці впровадження алгоритмів оптимальної фільтрації досить часто зустрічаються випадки, коли елементи коваріаційної

матриці похибок фільтрації не прямують до невеликих сталих значень, а з часом набувають таких великих значень, які можуть перевищувати максимально допустимі для вибраного способу зберігання чисел у комп'ютері. Неефективність алгоритму фільтрації може бути зумовлена наступними причинами:

1. *Математична модель процесу* (матриці **A**, **B**, **H**) має недостатню ступінь адекватності, тому в процесі її побудови необхідно застосувати кілька статистичних критеріїв для аналізу якості моделі. Особливу увагу треба звернути на якість даних, на основі яких оцінено коефіцієнти (параметри) моделі. Якщо дані недостатньо інформативні, то побудувати модель, адекватну досліджуваному режиму функціонування процесу, практично неможливо.

2. Некоректне визначення початкових умов — коваріаційної матриці похибок оцінок початкового стану системи $\mathbf{P}(0)$, тобто, використовується наближене значення $\mathbf{P}_c(0)$, де індекс "*c*" означає субоптимальність.

3. *Значення коваріаційних матриць* **Q** і **R** невідомі або відомі із значними похибками. У такому випадку необхідно збирати додаткову інформацію про процес та/або застосовувати алгоритм адаптивної фільтрації. В окремих випадках можна спробувати "підібрати" невідомі статистичні параметри, виходячи з конкретних значень вимірюваних сигналів (статистичних даних). Якість фільтрації можна оцінювати за допомогою діагональних елементів коваріаційної матриці похибок фільтрації $\mathbf{P}(k)$.

4. *Вимірювальні* (*статистичні*) дані не відповідають вимогам якості (наявні пропуски, великі викиди, значні періоди з постійними значеннями).

5. *Припущення щодо розподілу та/або статистичних характеристик* збурень стану та похибок (шуму) вимірів не відповідають дійсності.

6. Іноді для спрощення алгоритму при застосуванні його у реальному часі використовують *постійні коефіцієнти фільтра*, що також може привести до великих похибок оцінок стану досліджуваного процесу.

7. Комбінації вказаних причин та *помилки програмування* алгоритму.

Можливі заходи щодо підвищення якості алгоритму

1. *Підвищення адекватності* математичної моделі процесу шляхом ускладнення її структури.

2. Застосування спеціальних методів *попередньої обробки даних* з метою їх нормування, заповнення пропусків та зменшення великих викидів.

3. *Застосування алгоритмів адаптивної фільтрації* у випадку, коли невідомі параметри коваріаційних матриць **Q** і **R**.

4. Якщо шум вимірів відсутній, то його необхідно ввести штучно, що сприяє покращенню зумовленості матриці $[\mathbf{HP}(k)\,\mathbf{H}^T + \mathbf{R}]$, для якої знаходиться обернена при визначенні коефіцієнтів фільтра.

5. Застосування спеціальних алгоритмів *знаходження обернених матриць* при визначенні матричного коефіцієнта фільтра. Наприклад, алгоритму квадратного кореня, *LU*-факторизації та ін. При можливості можна використовувати *алгоритм послідовної фільтрації*, який не потребує операції обернення матриці [8].

6. Використання *статистичного* (*Монте-Карло*) *моделювання* [11] з метою підвищення достовірності визначення характеристик алгоритму фільтрації перед його використанням за призначенням.

7. Необхідно *контролювати процес оцінювання* шляхом контролю діагональних елементів матриці $\mathbf{P}(k)$. Допускається примусове зменшення значень цих елементів, якщо це не пов'язано з похибками самого алгоритму оцінювання. Крім того, *послідовність* $\nu(k) = \mathbf{z}(k) - \mathbf{HA}\hat{\mathbf{x}}(k-1)$ повинна задовольняти вимогам: $E[\nu(k)] = 0$ і $E[\nu(k)\nu^T(l)] = \mathbf{G}\,\delta(k-l)$, тобто повинна бути некорельованою послідовністю з нульовим середнім.

8. Якщо конкретне застосування фільтра не припускає існування перехідного процесу оцінювання (в якому спостерігаються максимальні похибки оцінок), то необхідно використовувати *фільтри з постійними коефіцієнтами* [8], які визначаються наперед під час статистичного моделювання.

9. Необхідно *уникати випадків одночасного оцінювання* невідомих параметрів моделі та стану процесу, тому що це викликає появу нелінійностей в моделі і розбіжності алгоритму.

10. У випадку мікропроцесорної реалізації необхідно уникати використання цілочислової арифметики.

12.5. Приклади побудови оптимального фільтра

Приклад 12.1. Розглянемо просту скалярну модель випадкового кроку

$$x(k) = x(k-1) + \mathrm{w}(k), \quad x(0) = x_0,$$
$$z(k) = x(k) + \nu(k),$$

де $\mathrm{w}(k)$ і $\nu(k)$ — процеси білого шуму з нульовим середнім та дисперсіями Q і R, відповідно. Оскільки процес скалярний, то модель і рівняння фільтрації суттєво спрощуються. Так, рівняння фільтрації має вигляд:

$$\hat{x}(k) = \hat{x}(k-1) + K(k)\left[z(k) - \hat{x}(k-1)\right].$$

Апріорна коваріація оцінки вектора стану і коефіцієнт фільтра будуть

$$P'(k) = P(k-1) + Q, \quad K(k) = \frac{P(k-1)+Q}{P(k-1)+Q+R}.$$

З останнього рівняння видно, що значення коефіцієнта $K(k)$ є обернено пропорціональним дисперсії шуму вимірів. Це означає, що чим вищий шум вимірів, тим нижчим буде значення коефіцієнта фільтра і оцінка стану $\hat{x}(k)$ значною мірою визначається попередньою оцінкою $\hat{x}(k-1)$. Внаслідок малого значення коефіцієнта $K(k)$, величина $\nu(k) = z(k) - \hat{x}(k-1)$ не буде суттєво впливати на оцінки стану. У випадку малого рівня шуму вимірів (або похибок статистичних даних) значення $K(k)$ буде зростати і на поточне значення оцінки стану $x(k)$ будуть суттєво впливати поточні виміри.

Вплив збурення стану $\mathrm{w}(k)$ на коефіцієнт фільтра також можна простежити за допомогою отриманих рівнянь. Однак його вплив не такий чіткий, як дисперсія шуму вимірів внаслідок того, що дисперсія збурення стану є в чисельнику і знаменнику рівняння для $K(k)$. Очевидно, що $\mathrm{w}(k)$ безпосередньо впливає на апріорну коваріацію похибок оцінок $P'(k)$. Чим більше значення має дисперсія Q, тим більшими будуть коваріації $P'(k)$ і $P(k)$. Це означає, що Q безпосередньо впливає на похибки оцінок і, таким чином, зменшує надійність оцінок $\hat{x}(k)$ загалом.

Приклад 12.2. Класичний технічний приклад застосування фільтра у навігаційних системах.

Розглянемо прямолінійний рух літаючого апарата (літака) у повітряному просторі. На рух апарата впливає випадкове прискорення, зумовлене неоднорідністю густини атмосфери. Положення літаючого апарата (ЛА) вимірюється в дискретні моменти часу з періодом дис-

кретизації T_s. При цьому кожний вимір містить шумову складову, поява якої зумовлена похибками вимірювального пристрою (радара) та впливом атмосфери на корисний радіосигнал.

Будемо вважати, що положення ЛА вимірюється у прямокутних координатах, а точність вимірів не залежить від часу та координат; похибки вимірів мають нульове середнє і некорельовані між собою. Таким чином, положення $z(k)$ можна описати рівнянням:

$$z(k) = x(k) + \nu(k), \tag{12.5.1}$$

де $z(k) = z(kT_s)$ — вимір положення ЛА; $x(k) = x(kT_s)$ — дійсне положення ЛА; $\nu(k)$ — випадкова похибка вимірів, яка має наступні статистичні характеристики:

$$\begin{aligned} &E\left[\nu(k)\right] = 0,\ \forall k;\ \ E\left[\nu^2(k)\right] = \sigma_x^2 = \text{const},\ \forall k; \\ &E\left[\nu(k)\nu(l)\right] = 0,\ k \neq l. \end{aligned} \tag{12.5.2}$$

Припустимо, що координати ЛА вимірюються незалежно. Це дає можливість описати динаміку руху апарата по кожній координаті окремо за допомогою наступних рівнянь:

$$x(k) = x(k-1) + \nu(k-1)T_s + \frac{1}{2}a(k-1)T_s^2; \tag{12.5.3}$$

$$\nu(k) = \nu(k-1) + a(k-1)T_s, \tag{12.5.4}$$

де $x(k)$ — положення ЛА в поточний момент kT_s; $\nu(k) = x(k)$ — швидкість руху ЛА; $a(k)$ — випадкове прискорення ЛА, величина якого покладається незмінною протягом одного періоду дискретизації вимірів. Будемо також вважати, що середнє значення випадкового прискорення дорівнює нулю, а кореляція між значеннями $a(k)$ в різних періодах дискретизації відсутня, тобто

$$E\left[a(k)\right] = 0;\ E\left[a^2(k)\right] = \sigma_a^2 = \text{const},\ \forall k;\ E\left[a(k)a(l)\right] = 0,\ k \neq l. \tag{12.5.5}$$

Для того щоб скористатись рівняннями оптимальної фільтрації, запишемо рівняння руху ЛА у формі простору станів [1; 8]:

$$\mathbf{x}(k) = \mathbf{F}\mathbf{x}(k-1) + \mathbf{G}\mathbf{a}(k-1), \tag{12.5.6}$$

$$\mathbf{z}(k) = \mathbf{H}\mathbf{x}(k) + \mathbf{v}(k). \tag{12.5.7}$$

де

$$\mathbf{F} = \begin{bmatrix} 1 & T_s \\ 0 & 1 \end{bmatrix},\quad \mathbf{G} = \begin{bmatrix} \frac{T_s^2}{2} \\ T_s \end{bmatrix},\quad \mathbf{H} = \begin{bmatrix} 1 & 0 \end{bmatrix}, \tag{12.5.8}$$

$\mathbf{z}(k) = [x(k), \quad \nu(k)]^T$ — вектор стану ЛА.

Рівняння дискретного фільтра Калмана приймає вигляд:

$$\hat{\mathbf{x}}(k) = \mathbf{F}\hat{\mathbf{z}}(k-1) + \mathbf{K}(k)\left[\mathbf{z}(k) - \mathbf{HF}\hat{\mathbf{x}}(k-1)\right]. \qquad (12.5.9)$$

Величина $\hat{\mathbf{x}}(k,k-1) = \mathbf{F}\hat{\mathbf{x}}(k-1)$ — є прогнозом значення вектора стану на основі попередньої оптимальної оцінки. Із врахуванням цього позначення рівняння фільтрації (12.5.9) можна переписати у вигляді:

$$\hat{\mathbf{x}}(k,k) = \hat{\mathbf{x}}(k,k-1) + \mathbf{K}(k)\left[\mathbf{z}(k) - \mathbf{H}\hat{\mathbf{x}}(k,k-1)\right], \qquad (12.5.10)$$

де $\hat{\mathbf{x}}(k,k)$ — оцінка вектора стану в момент k на основі інформації, що є в наявності на цей момент включно.

Матричний коефіцієнт підсилення фільтра $\mathbf{K}(k)$ визначається за формулою

$$\mathbf{K}(k) = \mathbf{P}'(k)\,\mathbf{H}^T[\mathbf{HP}'(k)\,\mathbf{H}^T + \mathbf{R}]^{-1}, \qquad (12.5.11)$$

де $\mathbf{R} = \sigma_y^2$ — дисперсія шуму вимірів; $\mathbf{P}'(k)$ — апріорна (до отримання останнього виміру $z(k)$) коваріаційна матриця похибок оцінювання, що знаходиться за наступною формулою [8]

$$\mathbf{P}'(k) = \mathbf{FP}(k-1, k-1)\mathbf{F}^T + \mathbf{GQG}^T, \qquad (12.5.12)$$

де

$$\mathbf{P}(k-1, k-1) = [\mathbf{I} - \mathbf{K}(k)\mathbf{H}]\,\mathbf{P}'(k-1) \qquad (12.5.13)$$

— апостеріорна коваріаційна матриця похибок оцінювання, отримана на попередньому кроці виконання алгоритму; $\mathbf{Q} = \sigma_a^2$ — дисперсія випадкового процесу (прискорення руху ЛА) $a(k)$.

Із врахуванням виразу (12.5.11) для знаходження коефіцієнта фільтра (12.5.12) приймає вигляд:

$$\mathbf{P}(k-1, k-1) =$$
$$= \{\mathbf{I} - \mathbf{P}'(k-1)\,\mathbf{H}^T[\mathbf{HP}'(k-1)\,\mathbf{H}^T + \mathbf{R}]^{-1}\,\mathbf{H}\}\,\mathbf{P}'(k-1). \qquad (12.5.14)$$

Матриці $\mathbf{P}'$ і $\mathbf{P}$ мають різні значення, оскільки після врахування останнього виміру похибки оцінок зменшуються. Тобто між цими матрицями існує наступне співвідношення:

$$\|P\| < \|P'\|.$$

Однак, відповідно до рівняння (12.5.12), випадкові прискорення ЛА між моментами дискретизації вимірів ведуть до збільшення по-

хибок оцінювання. Фільтр переходить у стаціонарний стан функціонування тоді, коли збільшення похибок протягом періоду дискретизації внаслідок випадкових прискорень буде компенсуватися їх зменшенням завдяки приходу чергового виміру.

Аналіз фільтра в стаціонарному стані

У стаціонарному стані мають місце співвідношення:

$$\mathbf{P}'(k,k) = \mathbf{P}'(k-1,k-1) = \mathbf{P}',\ \mathbf{P}(k,k) = \mathbf{P}(k-1,k-1) = \mathbf{P}, \quad (12.5.15)$$

а тому (12.5.12), (12.5.14) приймають вигляд:

$$\mathbf{P}' = \mathbf{F}\mathbf{P}\mathbf{F}^T + \mathbf{G}\mathbf{Q}\mathbf{G}^T; \quad (12.5.16)$$

$$\mathbf{P} = \{\mathbf{I} - \mathbf{P}'\mathbf{H}^T[\mathbf{H}\mathbf{P}'\mathbf{H}^T + \mathbf{R}]^{-1}\mathbf{H}\}\,\mathbf{P}'\mathbf{F}^T. \quad (12.5.17)$$

Перепишемо (12.5.16) як

$$\mathbf{P}' - \mathbf{G}\mathbf{Q}\mathbf{G}^T = \mathbf{F}\mathbf{P}\mathbf{F}^T,$$

і підставимо у праву частину останньої рівності замість $\mathbf{P}$ (12.5.17):

$$\mathbf{P}' - \mathbf{G}\mathbf{Q}\mathbf{G}^T = \mathbf{F}\{\mathbf{I} - \mathbf{P}'\mathbf{H}^T[\mathbf{H}\mathbf{P}'\mathbf{H}^T + \mathbf{R}]^{-1}\mathbf{H}\}\,\mathbf{P}'\mathbf{F}^T. \quad (12.5.18)$$

Якщо ввести позначення

$$\mathbf{P}' = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix},\ \text{де}\ \ p_{12} = p_{21} \quad (12.5.19)$$

і врахувати значення матриць, задані (12.5.8), то (12.5.18) можна переписати у вигляді [1]

$$\begin{bmatrix} p_{11} - \dfrac{\sigma_a^2 T_s^4}{4} & p_{12} - \dfrac{\sigma_a^2 T_s^3}{2} \\ p_{12} - \dfrac{\sigma_a^2 T_s^3}{4} & p_{22} - \sigma_a^2 T_s^2 \end{bmatrix} = \frac{1}{1 + \dfrac{p_{11}}{\sigma_x^2}} \times \quad (12.5.20)$$

$$\times \begin{bmatrix} p_{11} + 2p_{12}T_s + p_{22}T_s^2 + \dfrac{\left(p_{11}p_{22} - p_{12}^2\right)T_s^2}{\sigma_x^2} & p_{12} + p_{22}T_3 + \dfrac{\left(p_{11}p_{22} - p_{12}^2\right)T_s}{\sigma_x^2} \\ p_{12} + p_{22}T_s + \dfrac{\left(p_{11}p_{22} - p_{12}^2\right)T_s}{\sigma_x^2} & p_{22} + \dfrac{p_{11}p_{22} - p_{12}^2}{\sigma_x^2} \end{bmatrix}.$$

Якщо прирівняти елементи матриці зліва (12.5.20) елементам справа, то можна записати систему алгебраїчних рівнянь відносно стаціонарних коваріацій похибок оцінок стану ЛА:

$$\frac{p_{11}}{\sigma_x^2} = \frac{\sqrt{1+2r}\left(\sqrt{1+r}+1\right)^2}{r^2}; \qquad (12.5.21)$$

$$\frac{p_{12}}{\sigma_x \sigma_a T_s} = \frac{1}{2}\frac{\left(\sqrt{1+2r}+1\right)^2}{r}; \qquad (12.5.22)$$

$$\frac{p_{22}}{\sigma_a^2 T_s^2} = \frac{1}{2}\left(\sqrt{1+2r}+1\right), \qquad (12.5.23)$$

де $r = 4\sigma_x / (\sigma_a T_s^2)$ — параметр, який можна розглядати як відношення "шум/сигнал", оскільки σ_x — середня квадратична похибка датчика, що вимірює положення ЛА; $\sigma_a T_s^2/2$ — середнє квадратичне відхилення ЛА щодо прискорення, яке зумовлене його випадковим прискоренням.

За допомогою рівнянь (12.5.21)–(12.5.24) можна знайти безрозмірні відношення, які характеризують елементи апостеріорної матриці дисперсій похибок оцінювання:

$$\frac{\hat{p}_{11}}{\sigma_x^2} = \frac{\sqrt{1+2r}}{r^2}\left(\sqrt{1+2r}-1\right)^2; \qquad (12.5.25)$$

$$\frac{\hat{p}_{12}}{\sigma_x \sigma_a T_s} = \frac{1}{2}\frac{\left(\sqrt{1+2r}-1\right)^2}{r}; \qquad (12.5.26)$$

$$\frac{p_{22}}{\sigma_a^2 T_s^2} = \frac{1}{2}\left(\sqrt{1+2r}-1\right). \qquad (12.5.27)$$

Величини p_{11}/σ_x^2 і $\hat{p}_{11}/\sigma_x^2$ є відношеннями середнього значення квадрата похибки оцінки положення ЛА до дисперсії похибки σ_x^2 датчика положення до і після моменту вимірювання положення ЛА, відповідно.

Приклад 12.3. Припустимо, що лінійна система описується різницевим рівнянням другого порядку, тобто має дискретну функцію передачі [8]

$$G(z) = \frac{Y(z)}{U(z)} = \frac{b_1 z^{-1}}{z + a_1 z^{-1} + a_2 z^{-2}} = \frac{0{,}6 z^{-1}}{1 - 1{,}6 z^{-1} + z^{-2}}, \qquad (12.5.28)$$

де U і Y — вхід і вихід системи, відповідно; z^{-1} оператор затримки у часі на один крок (ми використовували його раніше при формуванні функцій прогнозування методом мінімізації дисперсії оцінки прогнозу).

Модель (12.5.28) необхідно представити у просторі станів. Методика представлення моделей такого типу у просторі станів буде розглянута в наступному параграфі, а зараз наведено тільки кінцевий результат — матриці **A** і **B**, побудовані на основі коефіцієнтів моделі:

$$\mathbf{A}=\begin{bmatrix}0 & 1\\ -a_2 & -a_1\end{bmatrix};\quad \mathbf{B}=\begin{bmatrix}1 & 0\\ a_1 & 1\end{bmatrix}^{-1}\begin{bmatrix}b_1\\ b_2\end{bmatrix};$$

$$\mathbf{A}=\begin{bmatrix}0 & 1\\ -1 & 1{,}6\end{bmatrix};\quad \mathbf{B}=\begin{bmatrix}1 & 0\\ -1{,}6 & 1\end{bmatrix}^{-1}\begin{bmatrix}0{,}6\\ 0\end{bmatrix}=\begin{bmatrix}0{,}6\\ 0{,}96\end{bmatrix}.$$

Модель (13.5.28) у просторі станів має вигляд:

$$\mathbf{x}(k)=\begin{bmatrix}0 & 1\\ -1 & 1{,}6\end{bmatrix}\begin{bmatrix}x_1(k-1)\\ x_2(k-1)\end{bmatrix}+\begin{bmatrix}0{,}6\\ 0{,}96\end{bmatrix}u(k-1), \qquad (12.5.29)$$

де $x_1(k) = y(k)$; $x_2(k) = y(k-1)$ — компоненти вектора стану. Рівняння вимірів:

$$\mathbf{z}(k)=\mathbf{H}\mathbf{x}(k)+\mathbf{v}(k)=\begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix}x_1(k)\\ x_2(k)\end{bmatrix}+\nu(k)=x_1(1)+\nu(k). \qquad (12.5.30)$$

Для того щоб врахувати випадкові збурення, що діють на процес, у праву частину рівняння (12.5.29) необхідно ввести змінну **w**(k). Такі збурення, як правило, впливають на протікання реальних процесів, наприклад, у вигляді похибок неідеальної моделі, похибок виконання арифметичних операцій, неврахованих ефектів при моделюванні та інших факторів. Статистичні параметри процесу **w**(k) необхідно визначати апріорно, до впровадження фільтра у функціонуючу систему обробки даних. Якщо неможливо визначити точні значення статистичних параметрів, то використовують наближені, визначені в процесі статистичного моделювання фільтра.

Нехай коваріації похибок оцінок початкового стану процесу, а також коваріації збурень стану і похибок вимірів мають наступні значення:

$$\mathbf{P}_0=\mathbf{P}(0,0)=\begin{bmatrix}0{,}5 & 0{,}2\\ 0{,}2 & 0{,}5\end{bmatrix};\quad \mathbf{Q}=\begin{bmatrix}0{,}3 & 0\\ 0 & 0{,}1\end{bmatrix};\quad \mathbf{R}=\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}.$$

Використовуючи ці дані, знаходимо коефіцієнт фільтра для одного кроку:

$$\mathbf{P}(1,0)=\mathbf{A}\mathbf{P}(0,0)\mathbf{A}^T+\mathbf{Q}=\begin{bmatrix}0{,}8 & 0{,}6\\ 0{,}6 & 1{,}24\end{bmatrix};$$

$$\mathbf{K}(1)=\mathbf{P}(1,0)\mathbf{H}^T\left[\mathbf{H}\mathbf{P}(1,0)\mathbf{H}^T+\mathbf{R}\right]^{-1}=\begin{bmatrix}0{,}55 & 0\\ 0 & 0{,}33\end{bmatrix};$$

$$\mathbf{P}(1,1)=\left[\mathbf{I}-\mathbf{K}(1)\mathbf{H}\right]\mathbf{P}(1,0)=\begin{bmatrix}0{,}36 & 0{,}27\\ 0{,}336 & 1{,}042\end{bmatrix}.$$

Далі можна знаходити оцінку вектора стану з використанням виміру, який надходить в момент часу 1, і переходити до наступного циклу реалізації рівнянь фільтра.

Приклад 12.4. Розглянемо неперервний стаціонарний процес Гауса-Маркова з автокореляційною функцією [10]

$$r_x(\tau)=\exp\left(-|\tau|\right).$$

Для одиничної дисперсії і нульового середнього спектральною функцією цього процесу є

$$S_x(x)=\frac{2}{1-s^2}=\frac{\sqrt{2}}{1+s}\cdot\frac{\sqrt{2}}{1-s},$$

а формуючий фільтр, який перетворює білий шум у процес Гауса-Маркова, має функцію передачі [10]

$$G(t)=\frac{x(t)}{w(t)}=\frac{\sqrt{2}}{1+s},$$

де $w(t)$ — білий гаусівський шум з одиничною дисперсією; $x(t)$ — процес Гауса-Маркова; s — змінна Лапласа. Рівняння стану для цього процесу має вигляд:

$$\dot{x}(t)=-x(t)+\sqrt{2}\,w(t).$$

Для того щоб застосувати фільтр Калмана, необхідно згенерувати послідовність вимірів даного процесу. Виберемо період дискретизації вимірів T_s секунди з нульовим початковим значенням при $t = 0$.

Нехай дисперсія похибок вимірів складає R = 0,65. Перехідна матриця станів у даному випадку є скаляром і може бути визначена за допомогою експоненти, як

$$A(k) = \exp(-0{,}045) \approx 0{,}956.$$

Оскільки між виміром вихідної величини процесу і станом $x(t)$ нема різниці, то матриця (скаляр) вимірів $H(k)$ = 1. Дисперсію збурення (шуму) стану запишемо у вигляді

$$Q(k) = E\left[w^2(k)\right] = E\left[\int_0^{0,045} \sqrt{2}\exp(-u)w(u)du \times \int_0^{0,045} \sqrt{2}\exp(-u)w(u)du\right] =$$

$$= \int_0^{0,045} \left(\sqrt{2}\exp(-\nu)\right)^2 d\nu \approx 0,165.$$

Для того щоб розпочати процес оцінювання, необхідно задати початкові умови x_0 і P_0. Оскільки процес починається в момент $t = 0$, має нульове середнє та одиничну дисперсію, то покладемо початкові умови рівними $x_0 = 0$, $P_0 = 1$ при $t = 0$. Тепер усі параметри фільтра відомі і можна розпочинати рекурсивну процедуру оцінювання.

У спеціальній літературі з проблем оцінювання параметрів і станів динамічних систем наведено багато прикладів успішного застосування оптимального фільтра [1; 8]. Спектр реальних задач, де він може бути використаний, є надзвичайно широким. Зокрема, це навігаційні системи, аерокосмічна техніка, обробка часових рядів в експериментальній фізиці, сейсмологічних дослідженнях, економетриці, біології та інших галузях.

12.6. Оцінювання невимірюваних компонент вектора стану за допомогою оптимального фільтра

При моделюванні динамічних систем виникають проблеми оцінювання невимірюваних компонент вектора стану. Наприклад, якщо до вектора стану входить третя похідна або вища, то вимірювати такі величини дуже складно або неможливо. Іншим прикладом може бути необхідність вимірювання високих температур при плавленні тугоплавких матеріалів. Температури можуть бути настільки високими, що неможливо створити датчик, який може витримати таку температуру. Досить часто при моделюванні економічних систем також виникає проблема оцінювання невимірюваних компонент вектора стану. Така задача виникає у випадках, коли відповідні статистичні дані не були зібрані або ці дані мають нерегулярний характер. Задачу оцінювання невимірюваних компонент можна розв'язати за допомогою оптимального фільтра.

Оцінювання невимірюваних компонент вектора стану можливе у випадку, коли коваріаційна матриця похибок оцінок має ненульові відповідні елементи, що дає змогу знаходити коефіцієнти фільтра,

пов'язані з невимірюваними компонентами. Розглянемо цей механізм оцінювання. У даному випадку в рівнянні фільтрації

$$\hat{\mathbf{x}}(k) = \mathbf{A}(k)\hat{\mathbf{x}}(k-1) + \mathbf{K}(k)\left[\mathbf{z}(k) - \mathbf{HA}(k)\hat{\mathbf{x}}(k-1)\right]$$

розмірність вектора вимірів $\mathbf{z}(k)$ є меншою розмірності вектора стану $\mathbf{x}(k)$, тобто $\dim[\mathbf{z}] < \dim[\mathbf{x}]$.

Оптимальний коефіцієнт фільтра знаходиться за формулою

$$\mathbf{K}(k) = \mathbf{P}'(k)\mathbf{H}^T[\mathbf{HP}'(k)\mathbf{H}^T + \mathbf{R}]^{-1},$$

де $\dim[\mathbf{P}'(k)] = [n \times n]$; $\dim[\mathbf{H}^T] = [n \times r]$ за визначенням; n — розмірність вектора стану і $\dim[\mathbf{HP}'(k)\mathbf{H}^T + \mathbf{R}]^{-1} = [r \times r]$. Таким чином, $\dim[\mathbf{K}(k)] = [n \times r]$. Вектор неув'язок $\nu(k)$ у рівнянні оцінювання

$$\nu(k) = \mathbf{z}(k) - \mathbf{HA}\hat{\mathbf{x}}(k)$$

має розмірність $[r \times 1]$, а розмірність добутку $\dim[\mathbf{K}(k)\nu(k)] = [n \times 1]$. Наприклад, якщо $\dim[\mathbf{x}] = [3 \times 1]$, а $\dim[\mathbf{z}] = [2 \times 1]$, то добуток $\mathbf{K}(k)\nu(k)$ має вигляд:

$$\mathbf{K}(k)\nu(k) = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \\ K_{31} & K_{32} \end{bmatrix} \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix}$$

і оцінка вектора стану визначається за рівняннями:

$$\hat{x}_1(k) = \hat{x}_1(k, k-1) + K_{11}\nu_1 + K_{12}\nu_2;$$

$$\hat{x}_2(k) = \hat{x}_2(k, k-1) + K_{21}\nu_1 + K_{22}\nu_2;$$

$$\hat{x}_3(k) = \hat{x}_3(k, k-1) + K_{31}\nu_1 + K_{32}\nu_2.$$

Отже, невимірювана третя компонента вектора стану може бути оцінена, якщо матриці $\mathbf{P}'(k)$, $\mathbf{P}(k)$ і $\mathbf{K}(k)$ мають ненульові відповідні елементи.

12.7. Функція прогнозування на основі оптимального фільтра

Застосування фільтра Калмана розпочалось із навігаційних систем, в яких він успішно використовується для фільтрації вимірів радіолокаційних станцій, оцінювання невимірюваних компонент сигналів, екстраполяції та деяких інших задач. Однак область застосування алгоритмів оптимальної фільтрації постійно розширюється. На сьогоднішній день фільтр застосовують, у тій чи іншій формі, в

усіх областях техніки, експериментальній фізиці, біофізиці, економіці та інших галузях досліджень. Одним із напрямів сучасного застосування фільтра є прогнозування динаміки розвитку процесів різної природи.

Структура рівняння у просторі станів є дуже зручною для використання з метою знаходження прогнозу. Розглянемо рівняння динаміки стохастичної системи у просторі станів:

$$\mathbf{x}(k+1) = \mathbf{F}\mathbf{x}(k) + \mathbf{w}(k), \tag{12.7.1}$$

де $\mathbf{w}(k)$ — процес білого шуму з нульовим середнім та скінченною постійною коваріацією $\mathbf{Q}(k)$. Функцію прогнозування на один крок можна знайти як умовне математичне сподівання

$$\hat{\mathbf{x}}\left(k+1|k\right) = E_k\left[\mathbf{x}\left(k+1\right)\right] = \mathbf{F}\mathbf{x}\left(k\right), \tag{12.7.2}$$

де $\hat{\mathbf{x}}(k+1|k)$ — прогноз на один крок на основі інформації на момент k включно. Функцією (12.7.2) можна скористатись для знаходження прогнозу на довільне число кроків. Так, прогноз на два кроки має вигляд:

$$\hat{\mathbf{x}}\left(k+2\right) = \mathbf{F}\hat{\mathbf{x}}\left(k+1\right) = \mathbf{F}\cdot\mathbf{F}\mathbf{x}\left(k\right) = \mathbf{F}^2\mathbf{x}\left(k\right),$$

і на довільне число кроків s:

$$\hat{\mathbf{x}}\left(k+s\right) = \mathbf{F}^s\mathbf{x}\left(k\right). \tag{12.7.3}$$

Очевидно, що дисперсія похибки прогнозу буде зростати пропорційно кількості кроків s. Похибка прогнозу на один і два кроки складає

$$\mathbf{e}_f\left(1\right) = \mathbf{x}\left(k+1\right) - \hat{\mathbf{x}}\left(k+1\right) = \mathbf{F}\mathbf{x}\left(k\right) + \mathbf{w}\left(k\right) - \mathbf{F}\mathbf{x}\left(k\right) = \mathbf{w}\left(k\right);$$

$$\mathbf{e}_f\left(2\right) = \mathbf{x}\left(k+2\right) - \hat{\mathbf{x}}\left(k+2\right) = \mathbf{F}\mathbf{x}\left(k+1\right) + \mathbf{w}\left(k+1\right) - \mathbf{F}\left[\mathbf{F}\mathbf{x}\left(k\right)\right] =$$

$$= \mathbf{F}\mathbf{x}\left(k+1\right) + \mathbf{w}\left(k+1\right) - \mathbf{F}\left[\mathbf{x}\left(k+1\right) - \mathbf{w}\left(k\right)\right] = \mathbf{w}\left(k+1\right) + \mathbf{w}\left(k\right).$$

Тобто дисперсія прогнозу на s кроків визначається як

$$\operatorname{var}\left[\mathbf{e}_f\left(s\right)\right] = E\left\{\left[\mathbf{w}\left(k+s-1\right) + \mathbf{w}\left(k+s-2\right) + \ldots + \mathbf{w}\left(k\right)\right]\times\right.$$

$$\left.\times\left[\mathbf{w}^T\left(k+s-1\right) + \mathbf{w}^T\left(k+s-2\right) + \ldots + \mathbf{w}^T\left(k\right)\right]\right\} = s\mathbf{Q}_{\mathrm{w}}.$$

Отриманий результат аналогічний тому, що був отриманий для прогнозування на основі різницевих рівнянь, оскільки в обох випадках ми користуємось лінійними моделями одного класу.

Прогнозування навантаження на телефонну лінію

Приклад 12.5. Задача побудови моделі та прогнозування навантаження на телефонний канал розглядається в роботі [8]. Навантаження на канал носить коливальний характер з незначним позитивним трендом. Максимальне навантаження припадає на зимові місяці, а мінімальне на середину літа, що пояснюється піком відпускного сезону. На лінійний тренд процесу накладається шумова компонента, зумовлена тим, що коливання не носять чисто гармонійний характер. Загалом процес описується сумою коливальної та лінійної компонент з адитивним білим шумом. Лінійна частина моделі має вигляд [8]:

$$\ddot{x}(t) = f_1(t), \tag{12.7.4}$$

де $f_1(t)$ — білий шум з нульовим середнім. Коливання описані рівнянням

$$\ddot{y}(t) + \omega^2 y(t) = f_2(t), \tag{12.7.5}$$

де $f_2(t)$ — білий шум, незалежний від $f_1(t)$. Рівняння (12.7.4) і (12.7.5) записані у векторній формі:

$$x_1(t) = x(t); \tag{12.7.6}$$

$$x_2(t) = \dot{x}(t); \tag{12.7.7}$$

$$x_3(t) = y(t); \tag{12.7.8}$$

$$x_4(t) = \dot{y}(t). \tag{12.7.9}$$

Модель у просторі станів:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -\omega^2 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} 0 \\ f_1(t) \\ 0 \\ f_2(t) \end{bmatrix}. \tag{12.7.10}$$

Оскільки виміри мають дискретний характер з періодом дискретизації T_s, то модель перетворена у дискретну форму має вигляд:

$$\mathbf{x}(k+1) = \begin{bmatrix} 1 & T_s & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cos(\omega T_s) & \sin(\omega T_s) \\ 0 & 0 & -\sin(\omega T_s) & \cos(\omega T_s) \end{bmatrix} \mathbf{x}(k) + \mathbf{w}(k). \tag{12.7.11}$$

Вид елементів правої нижньої субматриці розмірності [2 × 2] зумовлений видом розв'язку диференціального рівняння (12.7.5). Значення елементів коваріаційної матриці шуму стану $\mathbf{Q}(k)$ залежать від амплітуд випадкових величин $f_1(t)$ і $f_2(t)$. З визначенням матриці $\mathbf{Q}(k)$, як правило, існують проблеми, а тому найчастіше значення її елементів добирають у процесі обчислювальних експериментів.

Рівняння вимірів у даному випадку має досить просту форму завдяки тому, що загальний вимір є сумою лінійної та коливальної частин. Таким чином, $z(k)$ — це скаляр, який визначається рівнянням

$$z(k) = [1;\ \ 0;\ \ 1;\ \ 0]\mathbf{x}(k) + \nu(k). \qquad (12.7.12)$$

Оскільки виміри навантаження на телефонний канал є досить точними, то елементи коваріаційної матриці $\mathbf{R}(k)$ шумів вимірів $\mathbf{v}(k)$ у даному випадку мають відносно малі значення. Після визначення елементів матриць $\mathbf{F}(k)$, $\mathbf{Q}(k)$, $\mathbf{H}(k)$ і $\mathbf{R}(k)$ розв'язувалась задача прогнозування навантаження на телефонний канал на один крок. Як зазначалось, прогноз визначається добутком $\hat{\mathbf{x}}(k+1|k) = \mathbf{F}(k)\hat{\mathbf{x}}(k-1)$, тобто для прогнозування можна скористатись звичайним алгоритмом фільтрації, який складається з таких кроків:

1. Початкові умови (ініціалізація): задати $\hat{\mathbf{x}}(0)$, $\mathbf{P}(0)$, $\mathbf{Q}$ і $\mathbf{R}$.
2. Знайти поточне значення коефіцієнта фільтра:

$$\mathbf{K}(k) = \mathbf{P}'(k)\mathbf{H}^T[\mathbf{H}\mathbf{P}'(k)\mathbf{H}^T + \mathbf{R}]^{-1}.$$

3. Знайти нову оцінку вектора стану:

$$\hat{\mathbf{x}}(k) = \hat{\mathbf{x}}(k|k-1) + \mathbf{K}(k)\left[\mathbf{z}(k) + \mathbf{H}\hat{\mathbf{x}}(k|k-1)\right].$$

4. Визначити елементи апостеріорної коваріаційної матриці:

$$\mathbf{P}(k) = [\mathbf{I} - \mathbf{K}(k)\mathbf{H}]\ \mathbf{P}'(k).$$

5. Знайти апріорну коваріаційну матрицю для наступного кроку і прогнозовану оцінку вектора стану:

$$\mathbf{P}'(k+1) = \mathbf{F}(k)\ \mathbf{P}(k)\ \mathbf{F}^T(k) + \mathbf{Q};$$
$$\hat{\mathbf{x}}(k+1|k) = \mathbf{F}(k)\hat{\mathbf{x}}(k).$$

Перейти на крок 2.

Для того щоб стабілізувати функціонування фільтра у перехідному процесі, було використано перших 16 вимірів навантаження на канал для оцінювання елементів коваріаційної матриці похибок оцінок вектора стану $\mathbf{P}(0)$. Для отримання значень похибок оцінок викорис-

тано зважений метод найменших квадратів з одночасним використанням всіх 16 вимірів.

Фактично використання наведеного алгоритму фільтрації розпочато з 17-го кроку і продовжувалось до того моменту, на який були відомі виміри. Іншим варіантом ініціалізації фільтра є використання довільних великих значень діагональних елементів матриці $\mathbf{P}(0)$. У результаті встановлено, що фільтр Калмана дає можливість знайти досить точний (прийнятний) однокроковий прогноз для вибраної змінної — навантаження на телефонну лінію. Загалом, у спеціальній літературі можна знайти численні приклади застосування фільтра в різних галузях техніки [8].

Фільтр Калмана також успішно застосовується для прогнозування динаміки змінних у багатьох сферах діяльності, у тому числі для прогнозування фінансових та економічних змінних і їх дисперсії, що підтверджено численними дослідженнями [10].

Прогнозування фінансової змінної

Приклад 12.6. Розглянемо ряд, що описує дисперсію вартості акцій однієї з компаній, що входить до числа провідних на Нью-Йоркській фондовій біржі. На основі часового ряду, що характеризує динаміку ціни акцій, побудовано модель дисперсії ціни [1; 10]

$$\varepsilon_1^2(k) = 0{,}0714 + 0{,}1187\,\varepsilon_1^2(k-1) + 0{,}1123\,\varepsilon_1^2(k-2) - \\ - 0{,}096\,\varepsilon_1^2(k-4) + 0{,}1257\,\varepsilon_1^2(k-5) + \varepsilon^2(k),$$

яка представлена також у просторі станів.

Прогноз дисперсії вартості акцій на 5 кроків виконано за допомогою трьох методів: безпосередньо за отриманою моделлю, за допомогою фільтра Калмана та методом подібних траєкторій [10]. Графік, який показує якість прогнозу, наведено на рис. 12.1 (по осі абсцис — число кроків прогнозування, а по осі ординат — значення похибки прогнозу).

Статистичні параметри отриманого прогнозу наведені у табл. 12.1.

Прогноз на один крок отримано за допомогою фільтра. З табл. 12.1 видно (вона містить лише результати оцінювання якості прогнозу), що фільтр дає змогу також знайти прийнятні значення прогнозу на п'ять кроків для досить складного процесу. Він є кращим за сумою квадратів похибок прогнозу. Численні обчислювальні експерименти з фільтром свідчать про те, що він, як правило, дає можливість знайти високоякісну оцінку однокрокового прогнозу. Однак прогнозу-

Рис. 12.1. **Порівняльний графік прогнозу дисперсії вартості акцій компанії Нью-Йоркської фондової біржі**

Таблиця 12.1

Метод прогнозування	Максимальне відхилення		Мінімальне відхилення		Сума квадратів похибок
	абсолютне	%	абсолютне	%	
За моделлю	0,055	73	0,003	8	0,0338
Фільтр Калмана	0,059	78	0,007	12	0,0274
Метод подібних траєкторій	0,058	77	0,007	10	0,0335

вання на більше число кроків може супроводжуватись високими похибками оцінок. Це можна пояснити використанням відносно простої моделі у просторі станів, яка враховує тільки попередній стан системи. Основною проблемою застосування оптимального фільтра залишається побудова математичної моделі процесу високого ступеня адекватності. Досить непростою, з обчислювальної точки зору, є також задача оцінювання і прогнозування станів нелінійних систем, яка потребує коректної лінеаризації нелінійної моделі. Однак всі ці проблеми, як правило, вирішуються шляхом коректного застосування методів статистичної обробки даних і теорії оцінювання.

12.8. Контрольні питання і вправи

1. Запишіть вираз для поточного середнього і поясніть принцип рекурсивного оцінювання. Чи можна застосовувати рекурсивне оцінювання на довгих часових інтервалах? Дайте докладне пояснення цій задачі.

2. Яка форма математичної моделі використовується у методі оптимальної фільтрації Калмана і які її переваги? Наведіть приклад моделі процесу, яка може бути використана для побудови алгоритму оптимальної фільтрації.
3. Дайте означення вільної динамічної системи. Що є причиною руху динамічної системи у даному випадку?
4. Запишіть модель у просторі станів загального вигляду для нестаціонарного динамічного процесу. Поясніть сутність можливої нестаціонарності.
5. Сформулюйте мету застосування оптимального фільтра. У чому полягає різниця між оптимальним і цифровим фільтром? Що необхідно знати про процес (дані вимірів) для того, щоб застосувати цифровий фільтр?
6. Яким чином враховують статистичні параметри зовнішнього випадкового збурення $\mathbf{w}(k)$ і шумів вимірів $\mathbf{v}(k)$ в алгоритмі оптимальної фільтрації?
7. За яким методом можна отримати оцінки коваріаційних матриць $\mathbf{Q}$ і $\mathbf{R}$? Що є діагональні елементи цих матриць?
8. Поясніть сутність квадратичного функціоналу

$$J = E\left\{\left[\mathbf{x}(k) - \hat{\mathbf{x}}(k)\right]^T \left[\mathbf{x}(k) - \hat{\mathbf{x}}(k)\right]\right\},$$

на мінімізації якого критерію ґрунтується знаходження оптимального матричного коефіцієнта фільтра Калмана? Яким чином можна отримати точні значення вектора стану $\mathbf{x}(k)$?
9. Які початкові умови необхідно задати оптимальному фільтру? Як їх можна отримати? Що викликають неправильні значення початкових умов?
10. У чому полягає різниця між апріорною та апостеріорною коваріаційними матрицями похибок оцінок вектора стану динамічної системи?
11. Якого типу диференціальне рівняння необхідно розв'язувати для знаходження оптимального коефіцієнта фільтра?
12. Яка суть діагональних елементів коваріаційної матриці похибок оцінок вектора стану? Як повинні змінюватись їх значення у випадку нормального функціонування фільтра?
13. Сформулюйте своїми словами суть процесу статистичного моделювання оптимального фільтра. З якою метою виконується статистичне моделювання?

14. Запишіть повністю алгоритм статистичного моделювання процесу оптимальної фільтрації для моделі АР(1).
15. Які заходи необхідно застосовувати для підвищення якості алгоритму фільтрації? Чи можна побудувати алгоритм оптимальної фільтрації без операції обернення матриці? У яких випадках?
16. Чому окремі компоненти вектора стану динамічної системи можуть бути не спостережуваними (не вимірюватись)?
17. Поясніть, яким чином можна оцінити невимірювані компоненти вектора стану системи за допомогою оптимального фільтра. Наведіть приклади, коли може виникнути така необхідність.
18. Рівняння для знаходження коефіцієнта фільтра можна отримати за допомогою диференціального числення. Для цього необхідно мінімізувати елементи головної діагоналі коваріаційної матриці похибок оцінок вектора стану, яка визначається рівнянням Ріккаті:

$$\mathbf{P} = \mathbf{P}' - \mathbf{KHP}' - \mathbf{P}'\mathbf{H}^T\mathbf{K}^T + \mathbf{K}(\mathbf{HP}'\mathbf{H}^T + \mathrm{R})\,\mathbf{K}^T.$$

Елементи головної діагоналі матриці $\mathbf{P}$ необхідно мінімізувати за допомогою відповідного вибору значень коефіцієнта фільтра $\mathbf{K}$. Тобто можна мінімізувати суму діагональних елементів матриці $\mathbf{P} = p_{11} + p_{22} + p_{nn}$. При знаходженні похідних будуть корисними формули матричного числення.

СПИСОК ЛІТЕРАТУРИ

1. *Бідюк П. І.* Прикладна статистика: конспект лекцій / І. П. Бідюк. — К.: НТУУ "КПІ", 2012. — 220 с.
2. *Гаркавий В. К., Ярова В. В.* Математична статистика / В. К. Гаркавий, В. В. Ярова. — К.: Професіонал, 2004. — 384 с.
3. *Бобик О. І., Берегова Г. І., Копитко Б. І.* Теорія ймовірностей і математична статистика / О. І. Бобик, Г. І. Берегова, Б. І. Копитко. — К.: Професіонал, 2007. — 560 с.
4. *Кобзарь А. И.* Прикладная математическая статистика / А. И. Кобзарь. — М.: Физматлит, 2006. — 815 с.
5. *Гмурман В. Е.* Теория вероятностей и прикладная статистика / В. Е. Гмурман. — М.: Высш. шк., 2002. — 479 с.
6. *Хастингс Н., Пикок Дж.* Справочник по статистическим распределениям / Н. Хастингс, Дж. Пикок. — М.: Статистика, 1980. — 95 с.
7. *Эфрон Б.* Нетрадиционные методы многомерного статистического анализа / Б. Эфрон. — М.: Финансы и статистика, 1988. — 263 с.
8. *Harvey A. C.* Forecasting, structural time series models and the Kalman filter / A. C. Harvey. — Cambridge: Cambridge University Press, 1989. — 554 p.
9. *Franses Ph. H.* Time series models for business and economic forecasting / Ph. H. Franses. — Cambridge: Cambridge University Press, 1998. — 280 p.
10. *Tsay R. S.* Analysis of financial time series / R. S. Tsay. — Chicago: Wiley & Sons, Ltd., 2010. — 715 p.
11. *Gelman A., Carlin J. B., Stern H. S., Rubin D. B.* Bayesian Data Analysis / A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin. — New York, Chapman and Hall, CRC Press, 2000. — 670 p.
12. *Rossi P., Allenby G. M.* Bayesian Statistics and Marketing / P. Rossi, G. M. Allenby. — Hoboken (New Jersey): Wiley & Sons, Ltd., 2005. — 350 p.
13. *Wooldridge J. M.* Econometric analysis of cross section and panel date / J. M. Wooldridge. — London: MIT Press, 2002. — 735 p.
14. *Branson W. H.* Macroeconomic theory and policy / W. H. Branson. — New York: Harper & Row, 1990. — 656 p.
15. *Johnson R. A., Wichern D. W.* Applied multivariate statistical analysis / R. A. Johnson, D. W. Wichern. — New Jersey: Peason Prentice Hall, 2007. — 773 p.
16. *Ликеш И., Ляга Й.* Основные таблицы математической статистики / И. Ликеш, Й. Ляга. — М.: Финансы и статистика, 1985. — 356 с.

Додаток

Відносні частки площ під кривою нормального розподілу

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

3,1	0,49903	3,6	0,49984
3,2	0,49931	3,7	0,49989
3,3	0,49952	3,8	0,49993
3,4	0,49966	3,9	0,49995
3,5	0,49977	4,0	0,50000

Висвітлено сучасні методи статистичного аналізу даних, представлених часовими перерізами або часовими рядами. Розглянуто методику статистичного аналізу даних у вигляді чотирьох основних етапів, методи збору та попередньої обробки даних. Наведено основи побудови регресійних моделей процесів довільної природи. Запропоновано методику застосування розв'язків дискретних рівнянь типу авторегресії та авторегресії з ковзним середнім до аналізу поведінки випадкових процесів з детермінованими складовими та знаходження оцінок коротко- і середньострокових прогнозів динаміки їх розвитку. Розглянуто елементи теорії прийняття статистичних рішень, моделювання ризику та перевірки гіпотез. Наведено методики факторного і дискримінантного аналізу. Окремий розділ присвячено оптимальному статистичному оцінюванню станів динамічних систем за допомогою фільтра Калмана з використанням дискретних моделей у просторі станів, що враховують збурення станів та похибки вимірів. Подано процедуру статистичного моделювання фільтра, яка надає можливість формувати прогнозуючі розподіли, а також приклади застосування методів обробки статистичних даних до розв'язування реальних задач.

Для студентів, аспірантів та викладачів, а також для інженерів, які спеціалізуються у галузі розв'язання задач статистичного аналізу даних, математичного моделювання і прогнозування динаміки розвитку фінансово-економічних процесів та процесів іншої природи, представлених статистичними або експериментальними даними.

Навчальне видання

Бідюк Петро Іванович
Ткач Борис Петрович
Харрінгтон Том

МАТЕМАТИЧНА СТАТИСТИКА

Навчальний посібник

Редактор *Ю. А. Носанчук*
Коректор *А. А. Тютюнник*
Комп'ютерне верстання *Н. В. Коваленко*
Художнє оформлення *О. О. Стеценко*

Підп. до друку 26.11.15. Формат $60\times84/_{16}$. Папір офсетний.
Друк офсетний. Ум. друк. арк. 20,46. Обл.-вид. арк. 13,75. Наклад 1000 пр.

Міжрегіональна Академія управління персоналом (МАУП)
03039 Київ-39, вул. Фрометівська, 2, МАУП

Видавець і виготовлювач
ДП «Видавничий дім «Персонал»
03039 Київ-39, просп. Червонозоряний, 119, літ. ХХ

Свідоцтво про внесення до Державного реєстру суб'єктів видавничої справи ДК № 3262 від 26.08.2008 р.