МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ РАДІОЕЛЕКТРОНІКИ

МЕТОДИЧНІ ВКАЗІВКИ
до лабораторних робіт з дисципліни «ЕМПІРИЧНІ МЕТОДИ ПРОГРАМНОЇ ІНЖЕНЕРІЇ»

ХАРКІВ 2024

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ
УНІВЕРСИТЕТ РАДІОЕЛЕКТРОНІКИ

МЕТОДИЧНІ ВКАЗІВКИ

до лабораторних робіт з дисципліни
«Емпіричні методи програмної інженерії»
для студентів усіх форм навчання
першого (бакалаврського) рівня вищої освіти
спеціальності 121 – Інженерія програмного забезпечення,
освітня програма «Програмна інженерія»

Електронне видання

ЗАТВЕРДЖЕНО
кафедрою Програмної інженерії.
Протокол № 5 від 13.11.2023.

ХАРКІВ 2024

Методичні вказівки до лабораторних робіт з дисципліни «Емпіричні методи програмної інженерії» для студентів усіх форм навчання першого (бакалаврського) рівня вищої освіти спеціальності 121 – Інженерія програмного забезпечення, освітня програма Програмна інженерія [Електронний ресурс] / Упоряд.: І.В. Груздо, О.С. Назаров. – Електронне видання. – Харків: ХНУРЕ, 2024. – 127 с.

Упорядники: І.В. Груздо,
О.С. Назаров

Рецензенти: О.В. Золотухін, доцент кафедри штучного інтелекту, заступник декана факультету КН, заступник відповідального секретаря приймальної комісії, кандидат технічних наук, доцент, ХНУРЕ.

ЗМІСТ

ЗАГАЛЬНІ ПОЛОЖЕННЯ

Під час вивчення дисципліни «Емпіричні методи програмної інженерії» передбачено проведення лабораторних робіт (для усіх форм навчання) у такому обсязі: лабораторні роботи – 4, практичні заняття – 3.

Студенти денної форми виконують 4 лабораторні роботи у комп'ютерному класі з викладачем, а студенти заочної форми – 2 лабораторні роботи в комп'ютерному класі з викладачем та 2 – самостійно під час виконання контрольної роботи.

Метою проведення практичних занять і лабораторних робіт з дисципліни «Емпіричні методи програмної інженерії» є закріплення теоретичного матеріалу та набуття практичних навичок, що спрямовані на засвоєння принципів застосування емпіричних методів у галузі програмної інженерії, знання яких необхідні сучасному програмісту при розробці алгоритмів для розв'язання задач у різних галузях виробництва, економіки, науки і техніки мовами програмування.

1 СТАТИСТИЧНА ОБРОБКА ЕМПІРИЧНИХ ДАНИХ. ОБЧИСЛЕННЯ ТОЧКОВИХ ХАРАКТЕРИСТИК ВИБІРКИ

1.1 Мета роботи

Метою лабораторної роботи є формулювання навичків підбору оптимального опису вибірок/популяцій та визначенні сукупності необхідних критеріїв. Обчислення, формулювання та обґрунтування необхідних висновків на основі вибіркових даних в галузі дослідження даних, а саме перевірка статистичних гіпотез, яка як раз і привносить науковість в дослідження даних.

1.2 Опис роботи

Сучасні технології проектування та розробки програмного забезпечення надають потужні інструменти для їх програмної реалізації. Проте можуть виникати ситуації, коли необхідно обрати з великої кількості існуючих рішень необхідне. При цьому, інколи постає завдання виявити оптимальне рішення з безлічі можливих варіантів.

Набір даних зазвичай є вибіркою з якоїсь більшої популяції, або генеральної сукупності. Іноді ця вибірка занадто велика, щоб бути виміряною повністю. Іноді вона незмірна за своєю природою, тому що вона дуже велика за розміром (Терабайти інформації) або тому що до неї в тому вигляді що вона нам потрібна, не можна отримати одразу безпосередній доступ. У будь-якому випадку треба (вынужденны) робити висновок, виходячи з даних, які є зараз.

Зараз все збільшується обсяг даних про події, що відбуваються навколо. Найчастіше є потреба швидко відповісти на питання за домовою наявної інформації, для цього якнайкраще підходить процес, пов’язаний з перевіркою статистичних гіпотез.

Існує ряд питань і бізнес-завдань, відповіді на які можна знайти, застосувавши статистичні методи, наприклад.

Класифікувати користувачів, щоб ефективніше працювати з рекламними кампаніями.

Оцінити зміни в дизайні сайту. Наприклад, дізнатися за допомогою A / B тестування, як зменшення кількості полів в формі замовлення вплинуло на конверсію. При проведенні A / B тестів слід враховувати такі поняття, як статистична потужність, довжина вибірки, довірчий інтервал і статистична значущість.

Зрозуміти, наскільки критичною є просадка або зростання тієї чи іншої метрики. Для цього необхідно визначити інтервал допустимих значень основних метрик на сайті.

Спрогнозувати поведінку користувача на сайті за тими чи іншими показниками. Визначити потенційних покупців і запустити для них рекламні кампанії.

Перевірка статистичних гіпотез є найважливішим класом задач математичної статистики. За допомогою даного інструменту можна підтвердити або відкинути припущення про властивості випадкової величини шляхом застосування методів статистичного аналізу для елементів вибірки.

Гіпотеза (від грец. hypothesis, означає передбачення, основа) — це науково обґрунтоване передбачення про існування деякого предмета мислення або пояснення причин чи закономірних зв’язків між ними.

Статистична гіпотеза – це припущення про вид розподілу або про величини невідомих параметрів генеральної сукупності, яка може бути перевірена на основі вибіркових показників.

Загальна гіпотеза – це припущення, яке пояснює причину явища або групи явищ у цілому.

Часткова гіпотеза – припущення, яке пояснює якусь окрему частину чи окрему властивість явища або події.

Статистична гіпотеза про вид розподілу має назву «критерій згоди».

Статистичною гіпотезою називається кожна несуперечлива множина тверджень $\{H_0, H_1, \dots, H_{k-1}\}$ щодо властивостей розподілу випадкової величини.

Будь-яке з тверджень H_i називається альтернативою гіпотези. Найпростішою гіпотезою є двоальтернативна: $\{H_0, H_1\}$. В цьому випадку альтернативу H_0 називають нульовою гіпотезою, а H_1 – конкуруючою гіпотезою.

Нульовою (H_0) називають *висунуту гіпотезу*.

Альтернативною (H_1) – гіпотезу, що суперечить нульовій.

Якщо висунута гіпотеза буде відкинута, то вживають альтернативну до неї.

Під час перевірки будь-якої статистичної гіпотези **можливі варіанти:**

– гіпотеза H_0 правильна і її приймають (правильне рішення);

– гіпотеза H_0 неправильна і її відкидають, приймаючи гіпотезу H_1 (правильне рішення);

– гіпотеза H_0 правильна, але її відкидають згідно з правилом перевірки (неправильне рішення) – це *помилка першого типу*;

– гіпотеза H_0 неправильна, але її слід приймати згідно з правилом перевірки (неправильне рішення) – це *помилка другого типу*.

Альтернативні гіпотези приймаються тоді і тільки тоді, коли спростовується нульова гіпотеза. Це буває у випадках, коли відмінності, скажімо, в середніх арифметичних експериментальної і контрольної груп настільки значущі (статистично достовірні), що ризик помилки відкинути нульову гіпотезу і прийняти альтернативну не перевищує одного з трьох прийнятих рівнів значущості статистичного виведення:

Перший рівень – 5% (p = 0,05); де допускається ризик помилки у висновку в п’яти випадках зі ста теоретично можливих таких же експериментів при строго випадковому відборі випробовуваних для кожного експерименту;

Другий рівень – 1%, т. е. відповідно допускається ризик помилитися тільки в одному випадку зі ста;

Третій рівень – 0,1%, т. е. допускається ризик помилки тільки в одному випадку з тисячі. Останній рівень значущості пред'являє дуже високі вимоги до обґрунтування достовірності результатів експерименту і тому рідко використовується. У дослідженнях, котрі мають потребу в дуже високому рівні достовірності, приймають 5% рівень значущості.

Щоб побудувати «вдалу» наукову гіпотезу яку можна перевірити емпірично, необхідно пам'ятати, що гіпотеза:

– не повинна містити понять, які емпірично не можуть бути конкретизовані;

– не повинна містити ціннісних суджень;

– не повинна містити занадто багато обмежень і припущень;

– повинна перевірятися;

– може мати кілька частин.

Приклади простої гіпотези :

– H_0: Зв'язок між ознаками вибірки відсутній. H_1: Існує зв'язок між ознаками вибірки.

– H_0: є різниця в числі звернень в порівнянні з попереднім 30-денним інтервалом збору статистики з сайту. H_1: не має різниці в числі звернень на сайті в порівнянні з попереднім періодом.

– H_0: Розподіл даної вибірки є нормальним; H_1: Розподіл даної вибірки відрізняється від нормального.

– H_0: Темпи росту рослин не залежить від присутності кадмію в грунті; H_1: На темпи росту рослини впливають різних концентрацій кадмію в грунті.

– H_0: Відсутність активної інтелектуальної діяльності є фактором ризику розвитку нервовопсихічних захворювань в літньому віці; H_1: Відсутність активної інтелектуальної діяльності ніяк не впливає на винекнення нервовопсихічних захворювань в літньому віці.

Приклад. Треба перевірити гіпотезу, що чоловікам більше подобається зелений колір, ніж червоний. Якщо буде продемонстровано різні варіанти

кнопки двом чоловікам і один натисне червону кнопку, а інший - зелену, чи можна говорити, що гіпотеза спростована? Звичайно ні, тому що один з двох чоловіків міг виявитися любителем яскравих кольорів або дальтоніком. Якщо ж ви покажете кнопки, наприклад, тисячі відвідувачів сайту чоловічої статі, то вже зможете визначити, кнопка якого кольору подобається чоловікам. Тобто, чим більше довжина (розмір) вибірки для тесту, тим вище його статистична потужність. Не варто довіряти тестам, статистична потужність яких нижче 80%.

Проста та складна гіпотези.

– проста гіпотеза містить тільки одне припущення:

H_0: математичне очікування нормального розподілу дорівнює 3 ($\mu = 3$, σ – відомо);

– складна гіпотеза складається з кінцевого або нескінченного числа простих гіпотез:

H_0: математичне очікування нормального розподілу менше 3 ($\mu < 3$, σ – відомо)

($\mu < 3$ складається з нескінченної кількості простих виду H_i: $\mu = b_i$, де b_i – будь-яке число, менше 3).

Приклади складної гіпотези:

–Я очікую, що втрата ваги може тривати довше, ніж через шість тижнів H_1: $\mu > 6$. Нульова гіпотеза складається в тому, що ви очікуєте, що ця гіпотеза ніяк не відбудеться. У цьому випадку, якщо втрата ваги не відбудеться за шість тижнів, то це повинно відбутися протягом часу, дорівнює або менше ніж за шість тижнів. Це можна записати математично як H_0: $\mu \leq 6$

Загальна схема перевірки гіпотез.

– формулюють нульову й альтернативну гіпотези;

– задають величину рівня значущості критерія α;

– вибирають деяку функцію – статистику від результатів спостережень – і при обох гіпотезах знаходять закони її розподілу;

– за допомогою закону розподілу на основі обраного рівня значущості область можливих значень статистики розбивають на дві або три частини;

– роблять вибірку і за її результатами обчислюють статистику. З'ясовують, в яку зі сфер потрапляє її значення. Якщо величина знаходиться у полі, де правдоподібна основна гіпотеза, то вважають що експеримент не суперечить основній гіпотезі.

Алгоритм перевірки статистичних гіпотез.

1. За вибірковими даними формулюють основну H_0 та альтернативну H_1 гіпотези.

2. Задають рівень значущості α (0,05 або 0,01).

3. Залежно від H_0 визначають статистичний критерій K, що має відомий розподіл.

4. За вибіркою і формулою критерію K розраховують спостережуване значення критерія $K_{набл}$.

5. Залежно від виду H_1 визначають вид критичної області W й критичні точки у таблиці додатка для розподілу критерію K.

6. За результатами перевірки: $K_{набл} W?$ – роблять висновок про прийняття або відхилення гіпотези H_0.

7. Формулюють загальний висновок згідно поставленого завдання.

Існує дві групи статистичних критеріїв: непараметричні та параметричні. Непараметричні використовуються для аналізу даних, виміряних в шкалах найменувань та порядку. Параметричні критерії застосовуються до даних, виміряних в шкалі інтервалів та шкалі відношень.

Якщо в дизайні експерименту використовуються залежні виміри змінних, це свідчить про те, що між рядами отриманих даних існує певний зв'язок.

Таблиця 1.1 – Застосування різних типів критеріїв в залежності від шкали вимірювання залежної змінної.

Тип критерію		Шкала вимірювання
Непараметричні	-	Найменувань
Непараметричні	-	Порядку
-	Параметричні	Інтервалів
-	Параметричні	Відношень

Можна виділити принаймні три типи таких зв’язків:

1. Повторні виміри залежної змінної.

2. Природні зв’язки між досліджуваними.

3. Використання прийому співвіднесення досліджуваних чи груп при побудові експериментальної та контрольної груп.

Дизайн з використанням повторних вимірів залежної змінної свідчить про те, що залежна змінна вимірюється до та після експериментального випробування, або виміри залежної змінної здійснюються в одних і тих самих досліджуваних в різних експериментальних умовах. У такому випадку отримані ряди даних будуть залежними.

Коли для участі в експерименті залучаються природно зв’язані між собою досліджувані (близнюки, діти та їх батьки, пари закоханих, тощо), маємо аналізувати отримані дані теж як «зв’язані».

Співвіднесення досліджуваних (використання корельованих пар чи корельованих груп) теж дає залежні ряди виміряних даних. Уявіть, що дослідник намагається проконтролювати таку побічну змінну, як розвиток інтелекту. Він може виміряти IQ всіх досліджуваних і співвіднести їх за даним параметром, створивши корельовані групи.

Якою ж має бути вибірка, щоб результат був достовірним? Це залежить від того, який *статистичної потужності і значимості очікується від тесту*.

Є декілька видів статистичного аналізу: дискриптивний аналіз, висновковий аналіз, аналіз відмінностей, аналіз зв’язків, прогнозний аналіз.

Для проведення дискриптивного аналізу використовують дві групи вимірювань: вимірювання «центральної тенденції» (мода, медіана, середнє арифметичне) та вимірювання варіації – опис відсотка відмінностей результатів від найбільш типових значень. Для цього застосовують аналіз розподілу частот, розмаху варіації, середнього квадратичного відхилення.

Висновковий аналіз визначає можливість поширення висновків дослідження на всю сукупність і розмір супутньої похибки.

Аналіз відмінностей ґрунтується на гіпотезі, що дві сукупності не мають відмінностей між собою, а відмінності мають випадковий характер.

Аналіз зв’язків присвячений визначенню ступеня впливу одного чинника на інший. Виділяють два типи зв’язків: лінійний та нелінійний.

Метою прогнозного аналізу є прогнозування розрахунку події на основі неявної інформації. Особливо важливою є історична інформація, представлена у форматі часових рядів.

Приклад 1. Опис основних джерел і канали даних та які саме дані необхідно аналізувати.

Основним результатом аналізу даних є побудова звітів, що відображають взаємозв’язок двох типів даних: метрик і параметрів (dimensions). Метриками називають кількісні показники використання інтернет-ресурсу. Метрикою може бути лічильник, середнє значення або ставлення двох метрик. Найбільш часто використовуються наступні метрики. Запити - число http-запитів за певний період часу. Перегляди сторінок - число запитів від відвідувачів сайту на передачу гіпертекстових документів за певний період часу. Візити (сесії) – число візитів користувачів інтернет-ресурсу за певний період часу. Під візитом розуміється період активного звернення одного користувача до інтернет-ресурсу; всі запити до файлів ресурсу, що надійшли від користувача протягом цього періоду, вважаються які належать до одного візиту користувача.

Унікальні відвідувачі - число користувачів, що визначаються своїми унікальними характеристиками і відвідали інтернет-ресурс хоча б один раз за розглянутий період часу. Відзначимо дві основні особливості метрик, які використовуються в аналітиці. По-перше, вони орієнтовані на аналіз активності реальних, «морського» відвідувачів сайту. По-друге, існують окремі метрики, значення яких неможливо визначити з абсолютною достовірністю. Так, ми ніколи не можемо точно визначити моменти початку і закінчення візиту користувача. Причина цього криється в тому, що http протокол є протоколом передачі даних без встановлення з'єднання і збереження стану і передбачає, що запити обробляються незалежно один від одного. Іншим прикладом такої неоднозначності обчислення метрик може служити проблема визначення унікальних користувачів, які відвідали сайт за певний відрізок часу. Для визначення унікальних відвідувачів передбачено цілий ряд ознак: IP адреса машини користувача, який він використовує браузер, реєстраційні дані користувача. Але жоден з цих ознак не може бути використаний з 100% вірогідністю для однозначної ідентифікації користувача: IP-адреса користувача найчастіше не є постійним і змінюється від одного візиту користувача до іншого; сам користувач може використовувати для доступу до інтернет-ресурсу різні програми-браузери. Для обчислення подібних проблемних метрик не існує єдиних загальновизнаних методик, кожен з розробників засобів аналітики використовує власні алгоритми їх визначення, засновані на різних вихідних даних і різних початкових припущеннях. Але метрики є тільки кількісними показниками використання інтернет-ресурсу і не уявляють самостійної цінності. Тому в web-аналітиці вони розглядаються щодо набору атрибутів, що характеризують користувача сайту і його активність під час відвідування web-ресурсу. Такі атрибути отримали назву параметрів (dimensions). Прикладами параметрів є такі характеристики користувачів, як їх IP-адреса і географічне положення, які вживали браузери і операційні системи, адреси відвіданих користувачами сторінок, введені пошукові запити. Значення параметрів, на

відміну від метрик, є текстовими величинами. Для кожного з параметрів може бути визначена одна або кілька метрик. Так, можна для кожної адреси сторінки інтернет-ресурсу обчислити число звернень користувачів до цієї сторінки за певний відрізок часу. Типовий звіт web-аналітики є двовимірною таблицею, де рядках відповідають можливі значення параметра, а колонки заповнені величинами метрик, обчисленими для цього значення. При цьому, звичайно, треба розуміти, що не всі види метрик можуть бути визначені для довільного параметра. Більшість існуючих засобів аналітики допускає створення звітів, що включають кілька параметрів одночасно, хоча звіти з числом параметрів більше двох зустрічаються досить рідко.

Щоб полегшити аналіз представлених в звітах даних, можна розбити безліч обчислених значень метрик на групи, використовуючи як критерій розбиття значення параметра (параметрів). Незважаючи на очевидність ідеї розбиття даних на основі значень параметрів, інтерпретація отриманих при цьому результатів вимагає деякої обережності. Наприклад, підсумкове число «унікальних відвідувачів» за місяць, як правило, менше суми всіх «унікальних відвідувачів» за кожен день місяця. Пояснення цього факту досить просте - один і той же відвідувач міг відвідувати сайт в різні дні місяця. При цьому він враховується як «унікальний відвідувач» для кожного дня місяця, коли користувач звертався до сайту. Аналогічна ж статистика за весь місяць розглядає його як одного «унікального користувача». Точно так же сума нових і повторних відвідувачів інтернет-ресурсу за певний відрізок часу може не збігтися з загальною кількістю відвідувачів, зареєстрованих за цей же період. Це неминуче станеться, якщо один з «нових» відвідувачів звернеться повторно до сайту протягом аналізованого відрізка часу. Таким чином, підсумовування значень метрик дає правильну оцінку тільки для таких розбиття даних, які виключають подвійний облік подій, що розглядаються даної метрикою.

Аналіз отриманих даних зазвичай включає в себе і аналіз підвибірок, обсяги яких менше основний вибірки. Тому помилка для висновків по

підвибірках більше, ніж помилка по вибірці в цілому. Якщо планується аналіз підгруп / сегментів, обсяг вибірки повинен бути збільшений (в розумних межах).

Генеральна сукупність – сумарна чисельність об'єктів спостереження (люди, домогосподарства, підприємства, населені пункти і т.д.), що володіють певним набором ознак (стать, вік, дохід, чисельність, оборот і т.д.), обмежена в просторі і часі. Приклади генеральних сукупностей: Всі жителі Харкова (1 443 886 осіб в 2021 році). Роздрібні торгові точки, які здійснюють продаж продуктів харчування (20 тисяч на початок 2021 року) і т.д.

Вивчити всі елементи генеральної сукупності не є можливим, тому для її опису використовують вибірку.

Вибірка (Вибіркова сукупність) – частина об'єктів з генеральної сукупності, відібраних для вивчення, з тим щоб зробити висновок про всю генеральної сукупності. Для того щоб висновок, отримане шляхом вивчення вибірки, можна було поширити на всю генеральну сукупність, вибірка повинна мати властивість репрезентативності.

Помилка вибірки – відхилення середніх характеристик вибіркової сукупності від середніх характеристик генеральної сукупності. Іншими словами, завжди присутній ймовірність виходу середніх значень досліджуваної ознаки за межі встановленого довірчого інтервалу.

Помилка вибірки є випадковою і завжди пов'язана з її *об'ємом* – числом спостережень n, які утворюють вибірку. *Як правило, обсяг вибірки n значно менше обсягу всієї генеральної сукупності.* При цьому чим більший об'єм вибірки, тим нижче випадкова помилка вибірки.

Мінімізувати випадкову похибку вибірки можливо шляхом *розрахунку мінімального допустимого обсягу вибірки*.

Одноступінчата вибірка. Її особливість полягає в тому, що після визначення кластера або страти вивчення піддається кожна одиниця виділеної групи.

Багатоступенева вибірка. У ній, на відміну від одноступінчастої вибірки, вивчення піддаються в повному обсязі одиниці виділених груп, а відбувається наступний відбір окремих одиниць. Багатоступінчастий відбір зазвичай використовується в великомасштабних дослідженнях, де вибірка формується послідовно на двох і більше ієрархічних рівнях.

Один з важливих питань, на які потрібно відповісти при плануванні дослідження, – це оптимальний обсяг вибірки. Занадто маленька вибірка не зможе забезпечити прийнятну точність результатів опитування, а надто велика призведе до зайвих витрат. В математичній статистиці використовуються різні підходи в залежності від обсягу вибірки.

Умовно *за обсягом елементів вибірки* ділять на три типи: малі, середні та великі. До *малих* відносять вибірки, обсяг яких не перевищує 30 од. ($N \leq 30$). *Середня* вибірка задовольняє умові $30 < n \leq 200$ одиниць. Поняття великої вибірки до кінця не визначено, але вважається, що *вибірка є великою за обсягом*, якщо кількість її елементів перевищує 200 ($n > 200$)

При великому обсязі вибірки (поняття «великий обсяг» залежить від цілей і методів обробки.

Згідно висновків К.А. Отдельнової – дослідження можна класифікувати по 3 рівням точності: орієнтоване знайомство, дослідження середньої точності, дослідження підвищеної точності. Такі три рівня точності вельми умовно з практичної точки зору можна поділити наступним чином. Рівень точності «орієнтоване знайомство» відповідає пілотному дослідженню, «дослідження середньої точності» – підійде для дослідження, результати якого можна буде опублікувати в якості наукової статті з подальшим більш глибоким вивченням, ну а «дослідження підвищеної точності» – для дисертаційного дослідження і формування остаточних висновків.

Формула для розрахунку обсягу вибірки наступна:

$$n = \frac{t^2 p(100 - p)}{\Delta^2}, \qquad (1.1)$$

де: t – довірчий рівень, статистична величина, значення якої для досліджень в соціальній сфері прийнято 1,96 (при 95% точності статистичного висновку). Довірчий рівень встановлює сам дослідник відповідно до своїх вимог до надійності отриманих результатів. Найчастіше застосовуються довірчі рівні, рівні 0,95 або 0,99.

p – % об'єктів, у яких імовірно проявляється ознака, важлива для проведеного дослідження;

Δ – допустима помилка в %, задається довільно при плануванні дослідження.

Помилка вибірки буває двох видів – статистична і систематична. Статистична похибка залежить від розміру вибірки. Чим більше розмір вибірки, тим вона нижче.

Для простої випадкової вибірки розміром 400 одиниць максимальна статистична помилка (з 95% довірчою ймовірністю) становить 5%, для вибірки в 600 одиниць – 4%, для вибірки в 1100 одиниць – 3% Зазвичай, коли говорять про помилку вибірки, мають на увазі саме статистичну помилку.

Систематична помилка залежить від різних факторів, що постійний вплив на дослідження і зміщують результати дослідження в певний бік. У деяких випадках, коли відомі істинні розподілу, систематичну помилку можна нівелювати введенням квот або перезважуванням даних, але в більшості реальних досліджень навіть оцінити її буває досить проблематично.

Іноді, якщо обсяг генеральної сукупності точно відомий, наприклад – це всі працівники певної організації або всі автомобілі певної марки і року випуску, то можливо ще сильніше знизити необхідний обсяг вибірки, користуючись для його розрахунку наступною формулою:

$$n = \frac{t^2 pN(100-p)}{\Delta^2 N + t^2 pN(100-p)}, \qquad (1.2)$$

де: N – обсяг генеральної сукупності.

Зазвичай, якщо доступна для дослідження вибірка становить менше 5% від генеральної сукупності, то ця сукупність вважається великою і розрахунки проводяться за вищенаведеними правилами. Але якщо розрахунковий обсяг вибірки становить досить велику частку від генеральної сукупності, тобто якщо обсяг доступної вибірки перевищує 5% від генеральної сукупності, то в обсяг вибірки, розрахований за формулами (1.1) або (1.2), вводиться понижувальний коефіцієнт

$$n_0 = n * \sqrt{\frac{N-n}{N-1}} \tag{1.3}$$

Часто для прийняття рішень або з метою формування оціночних показників необхідно згрупувати об’єкти за певною ознакою.

Розраховувати довжину вибірки вручну зовсім не обов’язково – є величезна кількість зручних онлайн-калькуляторів.

Розбиття на групи (квотні вибірки) також необхідно для візуального відображення статистичних даних, яке дозволяє наочно продемонструвати об’єкти, з яким рівнем певного показника вони зустрічаються і з якою частотою. Спочатку виділяється певна кількість груп об’єктів (наприклад, чоловіки у віці 20 – 30 років, 31 – 45 років та 46 – 60 років; люди із доходом до 30 тисяч, з доходом від 30 до 60 тисяч і з доходом понад 60 тисяч). Для кожної групи задається кількість об’єктів, які повинні бути обстежені. Кількість об’єктів, які повинні потрапити в кожну з груп, задається, найчастіше, або пропорційно заздалегідь відомої частці групи в генеральній сукупності, або однаковим для кожної групи. Усередині груп об’єкти відбираються довільно.

Виділення типів в результаті класифікації або групування даних забезпечує їх однорідність. Однорідність узагальнює даних визначає стійкість всіх статистичних показників.

Розбиття на групи також необхідно для візуального відображення статистичних даних, яке дозволяє наочно продемонструвати об’єкти, з яким рівнем певного показу теля вони зустрічаються і з якою частотою.

Угруповання досліджуваного явища може бути проведена за одним або кількома ознаками. Якщо угруповання утворена за однією ознакою, то є простий. Якщо для виділення груп береться одночасно два або більше ознаки, при цьому групи, утворені за однією ознакою, поділяються на підгрупи за іншою ознакою, то таке угрупування називається складною.

При побудові складного угруповання виникають два питання: яка послідовність розбиття одиниць об'єкта за видами ознак і яку кількість ознак слід використовувати. Як правило, рекомендується спочатку проводити угруповання по описовим (атрибутивною) ознаками, значення яких мають яскраво виражені якісні відмінності, а потім – за кількісними. Зі збільшенням кількості групованих ознак в складних угрупованнях дуже швидко росте число груп. Тому рекомендується утворювати групи не більше, ніж за трьома ознаками. При більшій деталізації дуже важко аналіз статистичних даних.

При виконанні угруповання велике значення надається ознакам групування. *Ознака групування* – це ознака, відповідно до якої проводиться розбиття одиниць сукупності на окремі групи. Від її вибору залежать результати статистичного дослідження і правильність отриманих висновків. В якості підстави угруповання слід використовувати істотні ознаки. Вибір ознаки групування повинен передувати глибокий теоретичний аналіз досліджуваного явища. Тільки після того, як визначені соціальна сутність і економічний характер досліджуваного явища і чітко сформульована мета вивчення, можна приступити до вибору ознаки групування.

Число груп залежить від:

1. завдання дослідження;

2. виду ознаки, покладеної в основу угруповання;

3. чисельності сукупності;

4. ступеня варіації ознаки;

Одиниці аналізованої сукупності можуть бути розбиті за одною і тою ж ознакою на різну кількість груп. Наприклад, групуючи населення за віком з

метою визначення трудових ресурсів країни все населення ділиться на три групи: населення молодше працездатного віку, працездатне населення і населення старше працездатного віку.

При *угрупованню за атрибутивною (описовому) ознакою* питання про кількість груп вирішується порівняно просто - за кількістю градацій, видів стану цієї ознаки.

Наприклад, угруповання населення за статтю утворює дві групи, за формами власності – на п’ять груп: державна, муніципальна, приватна, змішана, власність іноземних юридичних осіб. Характеристика типів підприємств за їх величиною, обмежується трьома групами: дрібні, середні і великі.

Якщо атрибутивний (описовий) ознака має безліч найменувань (наприклад, професія – в галузі зв’язку налічується кілька десятків їх найменувань), то для статистичної характеристики складу працівників утворюють укрупнені групи (керівники, фахівці, робітники, інші). Таке об’єднання засноване на вивченні сутності виробничих процесів.

Угруповання за кількісною ознакою дуже різноманітні. При виборі числа груп в сукупності з кількісною ознакою необхідно, щоб в кожну групу потрапила достатня кількість одиниць сукупності. Тільки в цьому випадку узагальнюючі характеристики кожної групи (середні, відносні показники) будуть стійкими, невипадковими, характерними.

Порівняно легше утворюються групи за кількісними ознаками, які мають дискретну (переривчастість) варіацію і приймають цілі значення.

Якщо кількісна ознака змінюється в широких межах і має безліч різних значень, то кожна група утворюється у вигляді *інтервалів*.

Угруповання може бути виконана *з рівними і нерівними інтервалами*.

Рівні інтервали вживаються в тих випадках, коли ознака змінюється більш-менш рівномірно в обмежених межах, наприклад маса листа, посилки, заробітна плата певної категорії працівників.

Величина інтервалу залежить від розмаху варіювання ознаки і чисельності досліджуваної сукупності і в разі рівних інтервалів може визначаться по формулі Стерджесса.

Формула Стерджесса використовується для визначення величини інтервалу:

$$i = \frac{x_{\max} - x_{\min}}{k} = \frac{x_{\max} - x_{\min}}{1 + 3{,}322 * \lg N} \qquad (1.4)$$

де і – інтервал, тобто різниця між максимальним $x_{\max}$ і мінімальним $x_{\min}$ значеннями ознаки в кожній групі; N – чисельність одиниць сукупності; k – число груп, яке оптимально при величині 1 + 3,322 * lg N.

Недолік формули Стерджесса полягає в тому, що її застосування дає хороші результати для великої сукупності одиниць і коли розподіл одиниць за ознакою, покладеною в основу угруповання, близько до нормального.

Число груп можна визначити також за закономірністю наведеній в табл. 1.2.

Таблиця 1.2 – Визначення числа груп за закономірністю

Чисельність одиниць сукупності	15..24	25..44	45..89	90..179	180..359	360..719	720..1439
Кількість груп	5	6	7	8	9	10	11

Іншим способом виконання угруповання є використання середнього квадратичного відхилення σ. Якщо величина інтервалу дорівнює 0,5 σ, то сукупність розбивається на 12 груп, якщо 2 / 3 σ або σ, то сукупність поділяється на 9 або 6 груп.

При $i = \sigma$, k = 6 інтервали груп виглядають наступним чином:

від $\bar{x} - 3\sigma$ до $\bar{x} - 2\sigma$

від $\bar{x}-2\sigma$ до $\bar{x}-\sigma$

від $\bar{x}-\sigma$ до $\bar{x}$

від $\bar{x}$ до $\bar{x}+\sigma$

від $\bar{x}+\sigma$ до $\bar{x}+2\sigma$

від $\bar{x}+2\sigma$ до $\bar{x}+3\sigma$

Однак при визначенні груп даними методами можливо отримання порожніх або нечисленних груп. Якщо розмах варіації сукупності ознаки великий і його значення варіюється нерівномірно, то використовують угруповання з нерівними інтервалами.

Нерівні інтервали вживаються в тих випадках, коли ознака змінюється нерівномірно. З нерівних інтервалах найчастіше вживаються прогресивно зростаючі або спадаючі інтервали.

Величина інтервалів, змінюються в:

– арифметичній прогресії, визначається за формулою:

$$i_{j+1} = i_j + a, \tag{1.5}$$

– в геометричній прогресії , визначається за формулою:

$$i_{j+1} = i_j + q \tag{1.6}$$

де а – постійна величина (позитивна для прогресивно-зростаючих інтервалів, негативна – для прогресивно-спадаючих);

q – константа – позитивне число (для прогресивно-зростаючих інтервалів q> 1, для прогресивно-спадаючих – q <1).

Після визначення ознаки групування і границь груп будується ряд розподілу.

Статистичний ряд розподілу – це впорядкований розподіл одиниць сукупності на групи за певною варіаційною ознакою. Залежно від ознаки, покладеної в основу утворення ряду розподілу, розрізняють атрибутивні і варіаційні ряди розподілу.

Атрибутивними називають ряди розподілу, побудовані за описовим (якісним) ознаками. Варіаційними називають ряди розподілу, побудовані за кількісною ознакою.

Щоб пояснити протікаючи в будь-якії типової групі процеси, необхідно проаналізувати її структуру. Це досягається за допомогою *структурних угруповань*. Структурна угруповання характеризує структуру сукупності за будь-якою однією ознакою.

Структурні угруповання здійснюються на основі диференціації соціально-економічних явищ за їх складовими (структурі). За допомогою таких угруповань можуть вивчатися: склад населення за статтю, віком, розміром середньодушового доходу. У зміні структури виявляються найважливіші закономірності розвитку соціально-економічних явищ.

Якщо для типологічного угруповання частіше використовуються відкриті і нерівні інтервали, то для структурного угруповання більш характерні закриті рівні інтервали. *Структурна угруповання* – це ряд розподілу. Воно дозволяє вивчати інтенсивність варіації ознаки групування. На основі структурного угруповання можна вивчати динаміку структури сукупності.

Для характеристики зміни структури сукупності використовують узагальнюючі показники структурних зрушень:

– лінійний коефіцієнт абсолютних структурних зрушень (середній абсолютний показник зміни структури)

$$S_a = \frac{\sum_{j=1}^{m} \left| d_{j1} - d_{j0} \right|}{m} \qquad (1.7)$$

– квадратичний коефіцієнт абсолютних структурних зрушень (середній квадратичний показник зміни структури)

$$S_\sigma = \sqrt{\frac{\sum_{j=1}^{m} \left(d_{j1} - d_{j0} \right)^2}{m}} \qquad (1.8)$$

де m – число виділених груп в сукупності,

d_{j1} – питома вага j-ї групи в загальній чисельності сукупності в звітному (поточному) періоді,

d_{j0} – питома вага j-ї групи в загальній чисельності сукупності в минулому (базисному) періоді,

– індекс відмінностей

$$I_{разг} = \frac{1}{2}\sum \left| d_{j1} - d_{j0} \right| \qquad (1.9)$$

Узагальнюючі показники структурних зрушень відображають, на скільки процентних пунктів в середньому змінився питома вага окремих структурних груп у загальній чисельності сукупності в поточному періоді в порівнянні з базисним. При незначних змінах структури сукупності ці показники близькі до нуля, величина цих показників тим більше, чим значніше абсолютні зміни питомих ваг груп. Верхньої межі зміни коефіцієнти не мають. Квадратичний коефіцієнт більш чутливо реагує на структурні зміни (ніж середньоарифметичний) – використання квадратичного коефіцієнта краще.

Індекс відмінностей є найбільш простим зведеним коефіцієнтом, який на відміну від попередніх коефіцієнтів має не тільки нижню, а й верхню межу: $0 \le I_{разг} \le 1$.

Розподіл угруповань на типологічні і структурні досить умовно. Якщо задати, наприклад, межі сукупного доходу, що відповідають певним типам добробуту (доходи нижче прожиткового мінімуму, від прожиткового мінімуму до середніх, середній клас, і т.д.), то можна з повним правом назвати отриману угруповання типологічним.

Аналітичні угруповання є основою для характеристики взаємозв'язку між *явищами*. Зміна будь-якого явища обумовлюється впливом на нього інших явищ, з якими воно пов'язане. При дослідженні взаємозв'язків прийнято явища і ознаки поділяти на факторні і результативні.

Факторною називається ознака, що викликає зміну іншої ознаки, і залежної від неї ознаки. Остання носить назву *результативної ознаки*.

Аналітичне угруповання ставить собі за мету встановити кількісне вираження ступеня зв'язку між факторною та результативною ознаками в конкретних умовах місця і часу.

Особливості аналітичної угруповання:

– основа аналітичного угруповання є факторна ознака,

– кожна група характеризується середнім значенням результативної ознаки.

Аналітичні угруповання будуються за факторною ознакою. Аналітичні угруповання дають можливість проявитися взаємозв'язку наступним чином: зі зростанням значення факторної ознаки систематично зростає чи спадає середнє результативне.

За аналітичним угрупованням можна визначити показник тісноти зв'язку між факторною та результативною ознакою. Його обчислення ґрунтується на правилі розкладання (складання) дисперсії, згідно з яким загальна дисперсія дорівнює сумі міжгрупової дисперсії і середньої з внутрішньогрупових дисперсій:

$$\sigma^2 = \delta^2 + \overline{\sigma_j^2} \tag{1.10}$$

Міжгрупова дисперсія характеризує варіацію результативної ознаки, викликану ознакою-фактором, тому її також називають факторної дисперсією і визначають за формулою

$$\delta^2 = \frac{\sum_j \left(\bar{y}_j - \bar{y}\right)^2 n_j}{\sum_j n_j} \tag{1.11}$$

де $\bar{y}$ – середнє значення результативної ознаки в сукупності.

$\bar{y}_j$ – середнє значення результативної ознаки в j -ї групі,

n_j – число одиниць в j -ї групі.

Внутрішньогрупова дисперсія виникає за рахунок інших чинників (не пов'язаних з досліджуваним). Внутрішньогрупова дисперсія називається також залишковою і для кожної групи розраховується за формулою

$$\sigma_j^2 = \frac{\sum_{i=1}^{n_j} \left(y_{ij} - \bar{y}_j \right)^2}{n_j} \tag{1.12}$$

де y_{ij} – значення результативної ознаки у i -й одиниці сукупності j -ї групи.

Потім на основі внутрішньогрупових дисперсій для окремих груп визначається середня величина внутрішньогрупової дисперсії для всієї сукупності

$$\overline{\sigma_j^2} = \frac{\sum_{1}^{m} \sigma_j^2 n_j}{\sum_{1}^{m} n_j} \tag{1.13}$$

Ставлення між груповий дисперсії до загальної називається коефіцієнтом детермінації

$$\eta^2 = \frac{\delta^2}{\sigma^2} \tag{1.14}$$

Коефіцієнт детермінації показує, яка частина варіації ознаки-результату викликана варіацією ознаки-фактора. Квадратний корінь з коефіцієнта детермінації називається емпіричним кореляційним відношенням

$$\eta = \sqrt{\frac{\delta^2}{\sigma^2}} \tag{1.15}$$

За значенням емпіричного кореляційного відношення можна судити про тісноті зв'язку між ознаками. Зазвичай дотримуються наступної шкали:

$\eta \le 0{,}3$ – зв'язок слабкий,

$0{,}3 < \eta \le 0{,}5$ – зв'язок помітний,

$0{,}5 < \eta \le 0{,}7$ – зв'язок помірно тісний,

$0{,}7 < \eta \le 0{,}9$ – зв'язок тісний,

$\eta > 0{,}9$ – зв'язок дуже тісний.

Одним з варіантів вибіркового методу для вивчення використання робочого часу і часу роботи устаткування є *метод моментних спостережень*. Цей метод полягає в реєстрації даних в певні, заздалегідь обрані моменти. Попередньо складається список всіх можливих станів або видів даних в часі.

Так як статистична звітність такої інформації не містить, то моментні спостереження є єдиним її джерелом. Метод полягає в тому, що в заздалегідь обрані моменти фіксується стан процесу (наприклад, робота - перерва). Його особливість полягає в тому, що за охопленням об'єкта сукупності метод є суцільним, а за часом – вибірковим. Інтервали часу між моментами спостереження можуть бути рівними і нерівними. Тривалість інтервалу і число моментів спостереження визначаються відповідно до теорії вибіркового методу. Моменти спостереження можуть відбиратися за таблицями випадкових чисел або періодично через певні проміжки часу за способом механічного відбору.

Необхідна чисельність моментів спостереження визначається як

$$n_{повт} = \frac{0{,}25\, t^2}{\Delta^2} \qquad (1.16)$$

В силу незворотності часу відбір моментів в вибіркову сукупність завжди є без повторним. Але оскільки загальна кількість існуючих моментів часу (генеральна сукупність) дуже велике, в моментному спостереженні використовують формули помилок повторного відбору.

Середня помилка моментного спостереження розраховується за формулою

$$\mu = \sqrt{\frac{w(1-w)}{n}} \tag{1.17}$$

де w – коефіцієнт втрат часу, частка простоїв у всьому часу.

Інколи є потреба провести детальне дослідження по кожному клієнту, при цьому клієнтів мільйони і треба продивитися всю базу даних та прорахувати (просчитать) необхідні дані, рішень цієї проблеми є семплірування [4].

Аналіз даних включає побудову таблиць і графіків, дослідження числових характеристик змінних: відсотків, заходів центральної тенденції (середнє, мода, медіана), заходів положення (квантилі), розсіювання (дисперсія, стандартне відхилення, коефіцієнт варіації) і форми (асиметрія, ексцес) розподілів.

Найбільш поширеною формою статистичних показників є *середні величини*.

Головне значення середніх величин полягає в їх узагальнюючої функції, тобто заміні безлічі різних індивідуальних значень ознаки середньою величиною, що характеризує всю сукупність явищ.

Якщо середня величина узагальнює якісно однорідні значення ознаки, то вона є типовою характеристикою ознаки в даній сукупності. Так, для осіб з досить однорідним рівнем доходу, наприклад, пенсіонерів, можна визначити типові частки витрат на купівлю предметів харчування в їхньому бюджеті.

Однак не завжди середні величини характеризують типові значення ознак в однорідних за цією ознакою сукупностях. На практиці сучасна статистика значно частіше використовує середні величини, узагальнюючі явно неоднорідні явища. Наприклад, вироблений національний дохід на душу населення (пенсіонери, бюджетники, безробітні, банківські працівники).

Середня величина національного доходу на душу населення, середня врожайність по всій країні, середнє споживання різних продуктів харчування – це характеристики держави як єдиної системи, так звані системні середні.

Але своє основне властивість – бути типової характеристикою – середня виконає в тому випадку, якщо вона буде отримана з якісно однорідної

сукупності. Якщо сукупність неоднорідна, то загальні середні (системні середні) повинні бути замінені і доповнені груповими середніми, тобто середніми, розрахованими по якісно однорідними групами.

Розрізняють такі види середніх:

– середня арифметична;

– середня гармонійна;

– середня геометрична;

– середня квадратична.

У теорії статистики здебільшого вказується, яка середня необхідна в тому чи іншому випадку і як правильно її розраховувати.

Названі середні відносяться до класу степенних. Загальні формули степенних середніх мають такий вигляд:

проста – $\overline{x} = \sqrt[z]{\Sigma x^z / n}$

зважена – $\overline{x} = \sqrt[z]{\Sigma x^z f / \Sigma f}$

де x – варіант, тобто варіюється, величина ознаки що змінюється; n – число одиниць сукупності (число варіантів); f – частота ознаки (вага і-того варіанту); z – показник ступеня середньої.

Проста і зважена с степеневі середні – це середні одного і того ж виду, але їх обчислення залежить від вихідних даних. Якщо вихідні дані не систематизовані (не згруповані), то застосовується формула простої статечної середньої, якщо вони згруповані і представлені варіаційним рядом, то використовується формула зваженої статечної середньої. Варіаційний ряд розподілу – впорядкування одиниць досліджуваного явища за групами в зростаючому або спадаючому порядку.

Критерій вибору виду середньої: середня тільки тоді буде вірною узагальнюючої характеристикою сукупності, коли при заміні всіх варіантів середньої загальний обсяг варьоючьої ознаки залишиться незмінним. Цей критерій був запропонований А.Я. Боярським.

Таблиця 1.3 – Зміна значення показника ступеня середньої визначає її вид:

Найменування середньої	Умовне позначення	Значення z	Формула середньої	
			проста	зважена
середня гармонічна	$\bar{x}_h$	–1	$n / \Sigma \frac{1}{x}$	$\Sigma f / \Sigma \frac{f}{x}$
середня геометрична	$\bar{x}_q$	0	$\sqrt[n]{\Pi x}$	$\sqrt[\Sigma f]{\Pi(x^f)}$
середня арифметична	$\bar{x}_a$	+1	$\frac{\Sigma x}{n}$	$\frac{\Sigma xf}{\Sigma f}$
середня квадратична	$\bar{x}_k$	+2	$\sqrt{\Sigma x^2 / n}$	$\sqrt{\Sigma x^2 f / \Sigma f}$

Таким чином, в залежності від того, як утворюється загальний обсяг варьоючьої ознаки, вибирають ту чи іншу середню: якщо обсяг ознаки утворюється як сума варіантів, то використовується середня арифметична, якщо як сума зворотних значень – то середня гармонійна, якщо як твір варіантів – то середня геометрична, якщо як сума квадратів значень ознаки – то середня квадратична.

Різні види середніх при одних і тих самих вихідних даних приймають неоднакове значення.

Наприклад, є класи кваліфікації 1, 2, 3-й.

Обчислимо середні:

гармонічна $\bar{x}_h = 3/\left(1+\frac{1}{2}+\frac{1}{3}\right) =$ *1,64 ;*

геометрична $\bar{x}_q = \sqrt[3]{1 \cdot 2 \cdot 3} =$ *= 1,82 ;*

арифметична $\bar{x}_a = (1+2+3)/3 =$ *2,0 ;*

квадратична $\bar{x}_k = \sqrt[2]{(1+2^2+3^2)/3} =$ *2,16* .

Таке співвідношення значень середніх виглядає наступним чином: $x_h<x_q<x_a<x_k$ і називається правилом мажорантності середніх. Воно

представляється показником ступеня; чим більше показник ступеня у формулі середньої, тим більше її величина.

У статистиці найбільш часто в розрахунках вдаються до середньої арифметичної, оскільки вона відповідає природі економічних явищ.

Середня геометрична використовується при розрахунках показників динаміки, середня квадратична - показників варіації.

1. Добуток середньої на суму частот завжди дорівнює сумі добутків окремих варіантів на відповідні їм частоти, тобто обсягом варьоючьої ознаки:

$$\bar{x} \cdot \Sigma f = \Sigma xf$$

2. Сума відхилень варіант від середньої арифметичної дорівнює нулю: $\Sigma(x - \bar{x}) f = 0$.

3. Якщо всі варіанти зменшити (збільшити) на якесь довільне число А, то нова середня арифметична зменшиться (збільшиться) на цю ж величину:

$$\frac{\Sigma(\mathrm{x} \pm \mathrm{A}) \cdot \mathrm{f}}{\Sigma \mathrm{f}} = \bar{\mathrm{x}} \pm \mathrm{A}.$$

4. Якщо всі варіанти зменшити (збільшити) в k раз, то нова середня відповідно зменшиться (збільшиться) в k раз:

$$\frac{\sum \frac{\mathrm{x}}{\mathrm{k}} \cdot \mathrm{f}}{\Sigma \mathrm{f}} = \frac{\bar{\mathrm{x}}}{\mathrm{k}}; \quad \frac{\Sigma \mathrm{x} \cdot \mathrm{k} \cdot \mathrm{f}}{\Sigma \mathrm{f}} = \bar{\mathrm{x}} \cdot \mathrm{k}.$$

5. Якщо всі частоти (ваги) зменшити (збільшити) в k раз, то середня арифметична від цього не зміниться:

$$\frac{\Sigma \mathrm{x}(\mathrm{f} / \mathrm{k})}{\Sigma(\mathrm{f} / \mathrm{k})} = \frac{1/\mathrm{k} \sum \mathrm{xf}}{1/\mathrm{k} \sum \mathrm{f}} = \bar{\mathrm{x}}$$

6. Сума квадратів відхилень варіант від середньої арифметичної менше, ніж сума квадратів їх відхилень від будь-якої іншої довільної величини С:

$$\sum (x - c)^2 f = \sum (x - \bar{x})^2 f + \sum (\bar{x} - c)^2 f$$

тобто $\sum (x - \bar{x})^2 f < \sum (x - c)^2 f$ на величину $\sum (\bar{x} - c)^2 f$

Властивості середньої арифметичної дозволяють спростити її розрахунок: для цього можна відняти з усіх варіантів значень ознаки постійне число, наприклад, серединний варіант, розділити варіанти на будь-яку величину (наприклад, в рядах з рівними інтервалами на значення інтервалу), висловити частоти у відсотках.

Обчислення середньої арифметичної із застосуванням перших двох властивостей середньої арифметичної називається способом відліку від умовного нуля:

$$\bar{x} = \frac{\sum\left(\frac{x - A}{k}\right)f}{\Sigma f} k + A \qquad (1.18)$$

де A – постійне число (серединний варіант); k – величина інтервалу.

Залежно від виду вихідних даних можна визначати середні величини за формулою як середньої арифметичної, так і середньої гармонійної.

Насправді, якщо відомий обсяг варьоючьої ознаки $\Sigma w = \Sigma xf$ і значення варіант x, то підставивши в формулу замість Σw значення Σxf , отримуємо:

$$\bar{x} = \frac{\Sigma w}{\Sigma w / x} = \frac{\Sigma xf}{\Sigma xf / x} = \frac{\Sigma xf}{\Sigma f} \qquad (1.19)$$

Це формула середньої арифметичної зваженої.

Перевірка гіпотез має різноманітне застосування, як у бізнесі та економіці, так і в багатьох інших сферах. Перевірка гіпотез, це основна процедура підбиття підсумків про поведінку генеральної сукупності.

Наприклад, на підприємстві про якість продукції роблять висновки за результатами вибіркового контролю. Якщо вибіркова частка браку не перевищує заздалегідь установленої (нормативної) величини p_0, то партія продукції приймається. Оскільки висновки про відповідність якості продукції встановленим вимогам робляться на підставі вибіркової перевірки, судження про якість продукції не можна розглядати як категоричне. Тут ідеться лише про

припущення (гіпотезу), що частка браку в усій партії (генеральній сукупності) менша або дорівнює p_0.

Існує велика кількість різноманітних методів перевірки статистичних гіпотез. При виборі методу для вирішення певного конкретного завдання необхідно виходити з відповідей на такі питання [5]:

- якою є мета перевірки гіпотези;
- у яких шкалах виміряні аналізовані дані;
- чи є аналізовані вибірки незалежними або спряженими;
- скільки вибірок необхідно порівняти.

Розглянуті в далі методи застосовують при порівнянні двох вибірок [5]. При більшій кількості вибірок використовують методи дисперсійного аналізу.

Зазвичай при перевірці нульової гіпотези використовують певні модельні розподіли, що приблизно відповідають розподілу досліджуваного параметра – це **статистичні критерії.** На практиці як критерії найчастіше використовують – нормальний розподіл, χ2-розподіл, розподіли Стьюдента і Фішера. **Значенням критерію, що спостерігається**, називають його величину, яку розраховують за досліджуваними вибірками.

Для перевірки гіпотези весь вибірковий простір поділяють на дві області, що не перетинаються: критичну (w) та область прийняття (W – w). **Критичною областю** називають сукупність значень критерію, за яких нульову гіпотезу слід відхилити. **Областю прийняття гіпотези (областю допустимих значень)** називають сукупність значень критерію, за яких нульову гіпотезу приймають. Перевірка гіпотези передбачає розрахунок значення критерію і перевірку його потрапляння до області прийняття гіпотези.

Інколи існують ситуації коли є необхідним знати закон розподілу досліджуваної ознаки генеральної сукупності. Якщо закон розподілу невідомий, але є міркування для припущення певного його вигляду (назвемо його А, де як А може виступати рівномірний, показниковий, нормальний розподіл тощо), тоді висувають ***гіпотезу Н: ознака генеральної сукупності***

розподілена за законом A. У цій гіпотезі йдеться про вигляд невідомого розподілу.

Іноді закон розподілу ознаки генеральної сукупності відомий, але його параметри (числові характеристики) невідомі. Якщо є міркування припустити, що невідомий параметр θ дорівнює певному значенню θ_0, то висувають *гіпотезу H:* $\theta = \theta_0$. Ця гіпотеза вказує на припущену величину параметра відомого розподілу.

Можливі також інші гіпотези: про рівність параметрів ознак двох різних розподілів, про незалежність вибірок тощо.

Перевірка нульової гіпотези H_0 проводиться статистичними методами, тому її називають **статистичною перевіркою гіпотези**. Основна нульова гіпотеза, яка на початковому етапі перевірки завжди вважається правильною, насправді може бути як правильною, так і помилковою. Тому за результатами статистичної перевірки нульової гіпотези може бути прийнято як правильне, так і помилкове рішення. Правильне рішення може бути прийнято у двох випадках: коли, за результатами перевірки не відхиляється правильна нульова гіпотеза та відхиляється хибна нульова гіпотеза. Помилкове рішення може бути прийнято теж у двох випадках: коли відхиляється правильна нульова гіпотеза та не відхиляється хибна.

Інакше кажучи, у результаті прийняття помилкового рішення можуть бути допущені помилки двох типів:

1. буде відхилено правильну нульову гіпотезу (помилка першого типу);

2. не буде відхилено хибну нульову гіпотезу (помилка другого типу).

Існує два можливих типи помилок.

Помилка першого типу має місце за умови відхилення істинної нульової гіпотези. Наприклад, зазначена вище помилка матиме місце за умови засудження невинної особи.

Помилка другого типу має місце за умови невідхилення помилкової нульової гіпотези, або виправдання винного відповідача.

Імовірність помилки першого типу позначається α та називається **рівнем значущості**. Імовірність помилки другого типу позначається β. Імовірності помилок α та β є взаємопов'язаними, спроба знизити одну з них призведе до збільшення іншої.

У табл. 3.4 наведено узагальнену термінологію та поняття перевірки гіпотез.

Таблиця 1.4 – Термінологія та поняття перевірки гіпотез

ВИСНОВОК	H_0 істинна (відповідач невинен)	H_0 помилкова (відповідач винен)
ВІДХИЛИТИ H_0 (Засудити відповідача)	Помилка першого типу P (імовірність помилки першого типу) $= \alpha$	Правильне рішення (Справедливий вирок)
НЕ ВІДХИЛЯТИ H_0 (Виправдати відповідача)	Правильне рішення (Справедливий вирок)	Помилка другого типу P (імовірність помилки другого типу) $= \beta$

Імовірність припущення помилки першого типу – це ймовірність невідхилення альтернативної гіпотези за умови, що нульова гіпотеза справедлива:

$$\alpha = P(H_1/H_0), \qquad (1.20)$$

$$\updownarrow$$

α = P(невідхилення H1/H0 правильна.

Найчастіше рівень значущості α задається наперед. У статистичних дослідженнях найчастіше використовують такі його значення: 0,001; 0,005; 0,01; 0,05. Наприклад, прийняття рівня значущості 0,05 означає, що в п'яти випадках із ста є ризик, що буде отримано помилку першого типу.

Імовірність не припуститися помилки першого типу називається ***рівнем надійності*** та позначається γ.

$$\gamma = 1 - \alpha = 1 - P(H_1/H_0) = P(H_0/H_0), \qquad (1.21)$$

$$\updownarrow$$

γ = Р(невідхилення H_0/H_0 правильна).

Зауваження Під час контролю якості продукції ймовірність визнати нестандартними стандартні вироби називають ризиком виробника, а ймовірність визнати придатними браковані вироби називають ризиком споживача.

Імовірність припущення помилки другого типу – це ймовірність невідхилення хибної гіпотези H_0 :

$$\beta = P(H_0 / H_1), \qquad (1.22)$$

$$\updownarrow$$

β = Р(невідхилення H_0/H_1 правильна).

Єдиним способом одночасного зменшення ймовірностей похибок першого та другого типу є збільшення обсягу вибірки.

Перевірку статистичної гіпотези можна здійснити лише з використанням даних вибірки. Для цього слід обрати деяку випадкову статистичну характеристику (вибіркову функцію), точний або наближений розподіл якої відомий, і за допомогою цієї характеристики здійснити перевірку основної нульової гіпотези. Позначимо обрану випадкову статистичну характеристику (вибіркову функцію) через *U*.

Статистичним критерієм перевірки гіпотези (або просто критерієм) називають випадкову величину *U*, розподіл якої (точний або наближений) є відомим, та застосовується для перевірки справедливості основної гіпотези. Тобто, це правило, яке дозволяє відхилити чи не відхилити гіпотезу H_0 на основі вибірки.

Іншими словами, визначення статистичного критерію – це визначення правила перевірки основної статистичної гіпотези.

Завдання математичної статистики полягає в тому, щоб вибрати статистику, яка б якомога краще відображала властивості нульової H_0 та

альтернативної гіпотези H_1, і відшукати значення критерію, яке в найбільш повній мірі дозволяло б розділити вірну та невірну гіпотези. Це завдання має два аспекти, а саме:

–Вибір найбільш придатної статистики.

–Вибір критичного значення статистики.

Потужність критерію – це ймовірність відхилення хибної гіпотези H_0 , або ймовірність запобігання помилки другого типу:

$$1-\beta = P(H_1 / H_1), \qquad (1.23)$$

$$\updownarrow$$

$$1 - \beta = P(\text{невідхилення } H_1/H_1 \text{ правильна}).$$

Якщо випадкова величина U розподілена за нормальним законом, то критерій позначають не літерою U, а літерою Z. Якщо статистична характеристика розподілена за законом Фішера-Снедекора, то її позначають F. У разі розподілу статистичної характеристики за законом Стьюдента її позначають T, а у разі розподілу за законом «*xi* - квадрат» - χ^2.

За великих обсягів вибірки ($n > 30$) закони розподілу статистичних критеріїв

T, F, χ^2 наближаються до нормального розподілу.

Після визначення статистичного критерію перевірки гіпотези, обчислюється спостережуване значення u^* статистичного критерію U за даними вибірки**.**

Спостережуваним значенням u* статистичного критерію U називають значення відповідного критерію, обчислене за даними вибірки.

Множину всіх можливих значень статистичного критерію U поділяють на дві підмножини A та $\overline{A}$, для яких виконується умова:

$$A \cap \overline{A} = \emptyset, \; A \cup \overline{A} = = \Omega.$$

Областю допустимих значень A називають множину значень критерію, за яких основна гіпотеза H_0 не відхиляється.

Критичною областю $\overline{A}$ називають сукупність значень критерію, за яких основна гіпотеза H_0 відхиляється. **Критичними точками** (межами) критерію U називають точки $u_{кр}$, які відокремлюють критичну область $\overline{A}$ від області допустимих значень А.

Якщо спостережуване за даними вибірки значення u* потрапляє в критичну область $\overline{A}$, то гіпотеза H_0 відхиляється.

Якщо спостережуване за даними вибірки значення u* потрапляє в область допустимих значень А, то гіпотеза H_0 не відхиляється.

Критичні точки – межі критичних областей – знаходять із таблиць розподілу ймовірностей випадкової величини *U*, відповідно до заданого рівня значущості.

Розрізняють три види критичних областей: ***правостороння, лівостороння і двостороння.***

Правосторонньою критичною областю є критична область, яка задається нерівністю: $U > u$кр (рис. 1.1).

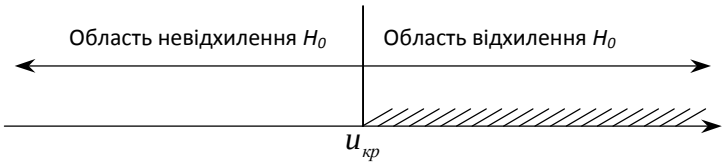


Рисунок 1.1 – Правостороння критична область

Критичну точку $u_{кр}$ цієї області за обраного рівня значущості α визначають зі співвідношення:

$$P(U > u_{кр}) = \alpha. \qquad (1.24)$$

Лівосторонньою критичною областю є критична область, яка задається нерівністю: $U < u_{кр}$ (рис. 1.2).

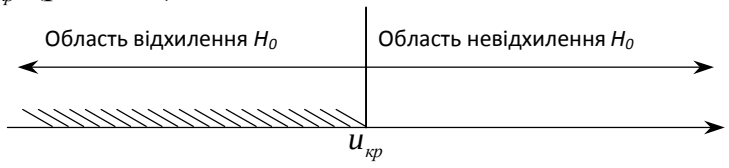


Рисунок 1.2 – Лівостороння критична область

Значення $u_{кр}$ знаходять за умовою:

$$P(U < u_{кр}) = \alpha. \qquad (1.25)$$

Двосторонньою критичною областю є критична область, яка задається двома нерівностями : $U < u_{кр}^{лів} u_{кр}$, $U > u_{кр}^{пр}$ (рис. 1.3).

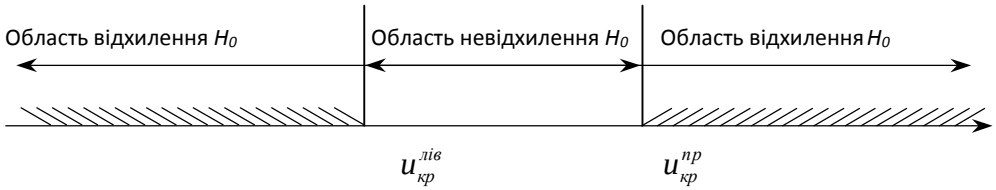


Рисунок 1.3 – Двостороння критична область

Критичні точки $u_{кр}^{лів}$, $u_{кр}^{пр}$ знаходять за умови:

$$P(U < u_{кр}^{лів}) + P(U > u_{кр}^{пр}) = \alpha$$

У разі, коли двостороння критична область симетрична відносно нуля:

$$P(U < u_{кр}^{лів}) + P(U > u_{кр}^{пр}) = \frac{\alpha}{2} \; i \; u_{кр}^{пр} = -u_{кр}^{лів} = u_{кр},$$

Статистичні висновки завжди є результатом перевірки статистичних гіпотез. Критерії перевірки статистичних гіпотез передбачають обчислення за даними вимірювань певних статистик та порівняння одержаних значень з теоретично розрахованими критичними величинами. Залежно від того, які статистики потрібно розрахувати, критерії перевірки статистичних гіпотез поділяють на:

– параметричні критерії;

– непараметричні критерії.

Параметричні критерії основані на обчисленні та порівнянні певних параметрів розподілу випадкових величин як то: математичного очікування, дисперсії, моментів вищих порядків тощо. Ці критерії вимагають апріорного знання (апріорі – наперед, до проведення досліду) закону розподілу випадкових величин або допущення при тип закону розподілу та перевірки його за результатами експерименту. Кожен параметричний критерій справедливий при тому чи іншому законі розподілу і може давати помилкові результати, якщо закон розподілу інший. Більшість параметричних критеріїв передбачає нормальний закон розподілу ймовірностей. Часто цей закон розподілу не

перевіряють, а вважають що він виконується, наприклад, на основі міркувань, що на випадкову величину впливає багато чинників, кожен з яких не сильно змінює саму величину. Допущення про нормальний закон розподілу дуже часто відповідає дійсності і параметричні критерії, основані на нормальному законі розподілу дають вірні результати.

Непараметричні критерії перевірки статистичних гіпотез використовують коли інформація про розподіл сама мінімальна, а вигляд закону розподілу не відомий. Застосовують їх також у випадках, коли вимірювання здійснюється у ранговій шкалі.

Непараметричні критерії є найбільш універсальними. Проте вони не такі потужні як параметричні критерії. Параметричні критерії більш потужні і дозволяють робити статистичні висновки при меншій кількості даних, тоді, коли непараметричні критерії не дають чіткої відповіді.

Історично так склалось, що параметричні критерії були розроблені значно раніше. Тому вони найбільш поширені. Кількість параметричних критеріїв досить значна і теоретично обґрунтовані всі особливості застосування кожного критерію. Непараметричні критерії вивчати і розробляти почали дещо пізніше і продовжують в наш час.

Важливою характеристикою критерію є його потужність і асимптотична потужність. Для багатьох непараметричних критеріїв доведено, що їх асимптотична потужність не менша ніж для параметричних критеріїв при нормальному законі розподілу випадкової величини і може перевищувати її при відхиленні закону розподілу від нормального. Тому використання непараметричних критеріїв у багатьох випадках корисне і теоретично обґрунтоване. Фактично кожний параметричний критерій має свій непараметричний аналог.

Серед великої кількості непараметричних критеріїв найбільш використовуємі такі:

– Критерій знаків.

– Знаковий ранговий критерій Вілконсона.

– Ранговий критерій Вілконсона.

– Коефіцієнт рангової кореляції.

Критерій знаків самий простий та математично обґрунтований з непараметричних критеріїв хоча і рідко вживаний критерій.

Критерій знаків служить для порівняння двох вибірок. Найчастіше його застосовують у випадках, коли умовами вимірювань допускається тільки попарне порівняння випадкових величин x_1 і x_2, та в деяких інших випадках. Нульова гіпотеза H0 критерію знаків така: вважають, що випадкові величини x_1 і x_2 мають одинаків розподіл а їх математичні очікування μ_1 і μ_2 рівні. Альтернативна гіпотеза H_1 полягає в тому, що розподіл величин x_1 і x_2, різний.

Цей критерій, як і більшість інших має два варіанти, що відрізняються формулюванням нульової гіпотези а саме: односторонній критерій знаків та двохсторонній критерій. Математично формулювання гіпотез для цих критеріїв має вигляд:

Для одностороннього критерію:

$$\begin{aligned} H_0 &: \mu_{X_1} = \mu_{X_2} \\ H_1 &: \mu_{X_1} > \mu_{X_2} \end{aligned} \tag{1.26}$$

Для двох стороннього критерію:

$$\begin{aligned} H_0 &: \mu_{X_1} = \mu_{X_2} \\ H_1 &: \mu_{X_1} \neq \mu_{X_2} \end{aligned} \tag{1.27}$$

Зауваження. У принципі односторонніх критеріїв два. Крім записаного вище може бути інший. Величина x_1 і x_2 можна вибрати довільно, тобто одній виборці присвоювати індекс 1 а другій 2, проте такий вибір повинен бути зроблений до експерименту і **до експерименту** повинна бути вибрана конкретна гіпотеза тому вважають, що є тільки один односторонній критерій. В ході експерименту, особливо під час аналізу його результатів, **зміна гіпотези не допускається.** Це зумовлене тим, що така зміна впливає на ймовірність події на основі якої розраховано критерій.

Обґрунтування критичних величин даного критерію основане на ймовірності результату. Якщо величини x_1 і x_2 мають однаковий розподіл (вірна гіпотеза H_0), то події коли в досліді буде знак + чи знак – (при відкиданні випадків $x_1 = x_2$) є рівно імовірні, тобто імовірність їх дорівнює ½

$$P\{x_1 - x_2 > 0\} = P\{x_1 - x_2 < 0\} = 1/2$$

Імовірність того що в n дослідах різниця x_1 - x_2 буде додатна в m випадках обраховується згідно формули біноміального розподілу.

$$P\{m\} = \frac{1}{2^2}\left(\left(\frac{n}{m+1}\right) + \left(\frac{n}{m+2}\right) + \ldots + \left(\frac{n}{n}\right)\right). \qquad (1.28)$$

де n – об'єм вибірки,

m – кількість дослідів зі знаком +.

На основі неї розраховане критичне значення. Статистичні таблиці подають значення при заданому рівні значимості α. Односторонній критерій вимагає відкинути гіпотезу H_0 на рівні значимості α, якщо число випадків з додатною різницею більше ніж критична величина $m_{\alpha n}$. Двохсторонній критерій вимагає відхилити нульову гіпотезу на рівні значимості 2α, якщо число випадків з додатною або від'ємною різницею більше від $m_{\alpha n}$. Величина $m_{\alpha n}$ табульована.

Використовують критерій знаків у такому порядку:

1. Задають вибіркові величини x_1 і x_2 та рівень значимості на якому потрібно зробити статистичний висновок про статистичну відмінність чи однорідність величин x_1 і x_2.

2. $(x_{1,1}x_{1,2}...x_{1,k})$ і $(x_{2,1}x_{2,2}...x_{2,k})$ в довільному порядку та визначають знаки різниці кожної пари значень $x_{1i} - x_{2i}$. Результати в яких $x_i = y_i$ відкидають.

3. Підраховують кількість результатів m зі знаком + та кількість всіх пар n за винятком відкинутих.

4. Знаходять за математичними таблицями критичну величину $m_{\alpha n}$ для вибраного рівня значимості α та кількості дослідів.

5. Порівнюють величину m з критичною. Якщо $m \geq m_{\alpha n}$ то гіпотезу H_0 відкидають на рівні значимості α. Якщо $m < m_{\alpha n}$ то гіпотезу не відкидають.

6. Роблять висновок про підтвердження чи не підтвердження нульової гіпотези на рівні значимості α. Тобто чи вибірки однорідні чи вони різні. Використовують цей висновок в подальшому як встановлений факт.

7. Якщо виявляється, що $m < m_{\alpha n}$, то додатково можна скористатись двохстороннім критерієм. Критичне значення вибирають на рівні значимості α/2 і перевіряють умову

$$n - m < m_{\alpha/2n}.$$

Часто на практиці обробки результатів вимірювань використовують спочатку більш прості критерії і якщо потужність їх недостатня, то застосовують більш потужні критерії. Критерій знаків використовують як до сукупності безперервних ознак, так і для оцінки відмінності рангових ознак (бали і т.п.) при достатньому числі їх градацій.

Ранговий критерій Вілкоксона (Wilcoxon). Критерій знаків є досить простим і зручним у використанні. Проте, як було відмічено, він недостатньо потужний. Підвищити потужність можна, якщо крім знаку можна врахувати величину різниці в ранговій шкалі. Якщо дані допускають не тільки визначити знак (факт краще-гірше) а і оцінити та впорядкувати дані за величиною різниці, присвоївши певний ранг, то для аналізу можна використати так званий знаковий ранговий критерій Вілкоксона.

Згідно знакового рангового критерію Вілкоксона для перевірки однорідності чи відмінності двох вибірок $(x_{1,1}x_{1,2}...x_{1,n})$ і $(x_{2,1}x_{2,2}...x_{2,n})$ визначають величину різниці $Z_i = |x_{1i} - x_{2i}|$, за модулем (без врахування знаку), виконують ранжування величин Z_i, присвоївши їм ранг R_i по мірі зростання та додають до рангу R_i знак який відповідає знаку різниці $x_{1i} - x_{2i}$. Таким чином одержують знаковий ранг різниці показників. За величинами знакових рангів розраховують статистику Вілкоксона, рівну

$$W = \sum_{i=1}^{S} T_i \qquad (1.29)$$

де T_i – впорядковані додатні ранги;

s – кількість додатних рангів.

Використовують одержане значення для перевірки справедливості нульової гіпотези. Перевірку гіпотези виконують шляхом порівняння значення статистики W з критичною величиною на рівні значимості α. У випадку коли кількість пар даних вибірки більша 26 критичне значення для статистики Вілкоксона можна розрахувати згідно з формулою:

$$W_{\alpha б т} = \frac{n(n+1)}{4} + \sqrt{\frac{n(n+1(2n+1)}{24} * Z(1-\alpha)} \qquad (1.30)$$

де n – об’єм вибірки;

α – рівень значимості критерію;

$Z(1-\alpha)$ – квантіль нормального розподілу.

У випадку, коли кількість пар даних у вибірках менше 26, слід скористуватись таблицею знакового розподілу Вілкоксона ***(Wilcoxon)***. За допомогою таблиці визначають імовірність одержаного результату та роблять висновок враховуючи цю імовірність.

Критерій Вілкоксона для зв’язаних сукупностей – це непараметричний метод, який використовується для оцінки значущості відмінностей двох зв’язаних сукупностей кількісних ознак.

Практичний розрахунок критерію включає наступні етапи:

–Знайти різниці парних варіантів.

–Визначити ранги одержаних різниць (без урахування знаків, пари спостережень, різниці яких виявилися рівними нулю, з подальшої оцінки виключаються).

–Визначити суму рангів одержаних різниць, що мають однакові знаки і узяти меншу з них (Т).

–Встановити достовірність відмінностей. При кількості спостережень менше 26 порівнюють знайдену суму з критичними значеннями з таблиці, інакше розраховують по спеціальній формулі випадкову змінну (u).

Критерій Вілкоксона є більш потужній ніж критерій знаків Його слід використовувати за наявності у порівнюваних сукупностей кількісних ознак значного числа різниць з протилежними знаками.

Параметричні критерії перевірки статистичних гіпотез. У випадках коли вимірювання випадкових величин здійснене у сильних шкалах (шкалі різниці чи шкалі відношень), для перевірки статистичних гіпотез використовують параметричні критерії.

Параметричні критерії дозволяють отримати числові значення і є більш потужними, тобто працюють там де непараметричні критерії не дозволяють відкинути нульову гіпотезу. Обмеження використання параметричних критеріїв зумовлені законом розподілу випадкових величин. Більшість параметричних критеріїв передбачає, що випадкові величини розподілені згідно нормального закону. Вони вимагають знання апріорі (до досліду) характеру закону розподілу випадкових величин, або перевірки його за даними експерименту.

Для практичного використання теорія рекомендує ряд критеріїв, потужність яких при заданих умовах є максимальною. Значна кількість критеріїв використовує розподіл Стьюдента, це так звані t-критерії. Практично всі критерії мають два варіанти, а саме двохсторонній критерій та односторонній критерій.

Використовують критерії перевірки статистичних гіпотез у такому порядку:

–формулюють нульову й альтернативну гіпотези;

–вибирають рівень значимості α;

–проводять експеримент і вимірюють значення випадкових величин;

–розраховують статистику потрібну для перевірки гіпотези;

–вибирають згідно із статистичними таблицями критичну величину;

–порівнюють одержане значення з критичною величиною;

–роблять висновок чи дані експерименту суперечать нульовій гіпотезі в користь альтернативної чи ні.

Слід мати на увазі, як це вже відмічалося, що статистичні висновки завжди мають імовірнісний характер. Вони можуть підтвердити, що дані експерименту не суперечать нульовій гіпотезі, або суперечать їй, і тоді нульову гіпотезу потрібно відкинути як невірну. Але довести, що нульова гіпотеза є вірна статистичні дані не можуть.

Статистичні гіпотези найбільш часто використовують для вирішення таких завдань:

–Перевірка параметрів розподілу а саме, чи дані досліду підтверджують певні значення математичного очікування, дисперсії, а чи параметрів вищих моментів.

–Перевірка гіпотези зсуву, тобто гіпотези чи математичні очікування вибірок відрізняються між собою.

–Перевірка гіпотези відносно однорідності дисперсій, тобто чи дисперсії вибірок є однаковими.

–Перевірка гіпотези узгодженості, тобто гіпотези чи відповідає вибірка певному закону розподілу ймовірностей випадкової величини.

–Перевірка значимості коефіцієнту кореляції, чи відрізняється його експериментальне значення від нуля.

–Перевірка значимості коефіцієнтів регресії, наскільки суттєвими є коефіцієнти певного рівняння регресії.

Найбільш вживані параметричні критерії наведено у табл. 1.5. При потребі більш детальний перелік та опис критеріїв можна знайти в спеціальній літературі.

Таблиця 1.5 – Найбільш вживані параметричні критерії перевірки статистичних гіпотез

№ п/п	Нульова H_0 та альтернативна H_1 гіпотези	Статистика	Розподіл статистики
1	2	4	5
1	$H_0: m_x = m_0$ $H_1: m_x \neq m_0$	$t = \frac{\bar{x} - m_0}{\sigma / \sqrt{n}}$ $df = n-1$	Z - статистика, σ^2 відомо Нормальний розподіл
2	$H_0: m_x = m_0$ $H_1: m_x > m_0$		
3	$H_0: m_x = m_0$ $H_1: m_x < m_0$		
4	$H_0: m_x = m_0$ $H_1: m_x \neq m_0$	$t = \frac{\bar{x} - m_0}{S / \sqrt{n}}$ $df = n-1$	t- статистика Стьюдента, σ_x^2 невідомо Нормальний розподіл
5	$H_0: m_x = m_0$ $H_1: m_x > m_0$		
6	$H_0: m_x = m_0$ $H_1: m_x < m_0$		
7	$H_0: \sigma_x = \sigma_0$ $H_1: \sigma_x \neq \sigma_0$	$X^2 = (n-1)\frac{S^2}{\sigma_0^2}$ $df = n-1$	X^2-розподіл з σ_0^2 відома степенями свободи
8	$H_0: \sigma_x = \sigma_0$ $H_1: \sigma_x > \sigma_0$		
9	$H_0: \sigma_x = \sigma_0$ $H_1: \sigma_x < \sigma_0$		
10	$H_0: m_x = m_y$ $H_1: m_x \neq m_y$	$U = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}}$	Нормальний розподіл σ_x^2, σ_y^2 відомі, n, m - об'єми виборок
11	$H_0: m_x = m_y$ $H_1: m_x > m_y$		
12	$H_0: m_x = m_y$ $H_1: m_x < m_y$		
13	$H_0: m_x = m_y$ $H_1: m_x \neq m_y$	$t = \frac{\bar{x} - \bar{y}}{\sqrt{(n-1)S_x^2 * (n-1)S_y^2} * \sqrt{\frac{nm(n+m-2)}{n+m}}}$ $df = n+m-2$	t-розподіл σ_x^2, σ_y^2 невідомі, n, m – об'єми вибірок

Продовження табл. 1.5

№ п/п	Нульова H_0 та альтернативна H_1 гіпотези	Статистика	Розподіл статистики
1	2	4	5
14	$H_0: m_x = m_y$ $H_1: m_x > m_y$		
15	$H_0: m_x = m_y$ $H_1: m_x < m_y$		
16	$H_0: \sigma_x^2 = \sigma_y^2$ $H_1: \sigma_x^2 \neq \sigma_y^2$	$F = \frac{S^2_{max}}{S^2_{min}}$	F – розподіл Фішера σ_x^2, σ_y^2 невідомі, n, m –об'єми вибірок
17	$H_0: \sigma_x^2 = \sigma_y^2$ $H_1: \sigma_x^2 > \sigma_y^2$	$F = \frac{S_x^2}{S_y^2}$	
18	$H_0: \sigma_x^2 = \sigma_y^2$ $H_1: \sigma_x^2 < \sigma_y^2$	$F = \frac{S_x^2}{S_y^2}$	
19	$H_0: R = 0$ $H_1: R \neq 0$	$t = \frac{R}{\sqrt{\frac{1-R}{n-2}}}$ $df = n-2$	t-розподіл Стьюдента
20	$H_0: m_x = m_y$ $H_1: m_x \neq m_y$	$F = \frac{S^2}{S^2_{\sum}}$	F – розподіл Фішера S_2 $S_{\sum}$ невідомі, вибіркові дисперсії усієї сукупності даних і даних в середині вибірок; n, m – об'єми вибірок

У табл. 1.5 використані такі позначення:

H_0:, H_1: – нульова й альтернативна гіпотези;

m, σ_x^2 – математичне очікування й дисперсія розподілу випадкових величин;

$\overline{X}$, S^2 – вибіркові значення, оцінки математичного очікування (середнє значення величини) й вибіркова дисперсія:

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

У колонці 2 таблиці вказані нульова й альтернативна гіпотеза та попередня інформація, що має бути відома для використання даного критерію.

У колонці 3 таблиці наведено статистику, яку потрібно розрахувати за вибірковими даними для перевірки гіпотези й число ступенів свободи – df.

В колонці 4 приведено характеристики розподілу статистики. Літерами позначено тип закону розподілу, якому підлягають значення статистики використаної для перевірки гіпотези.

U – нормальний розподіл (Гауса);

t – розподіл Стьюдента (t- розподіл);

$\chi 2$ – розподіл хі- квадрат;

F – розподіл Фішера (F-розподіл).

Рядом з типом розподілу вказано та рівень значимості α.

Критерій 10 дозволяє і його варіанти 11, 12 порівняти дві вибірки і визначити чи вони однорідні, є вибірками з однієї генеральної сукупності, чи відрізняються за математичним очікуванням. Тобто цей критерій перевіряє ефект зсуву вибірок одна відносно іншої. Досить часто ефект зсуву називають «ефектом обробки». Термін «ефект обробки» виник з того що досить часто зустрічається практичне завдання визначити як той чи інший вплив, та чи інша обробка впливає на характеристику виробів, приводить вона до покращення характеристик чи ні.

Критерій 13 і його різновиди 14, 15 аналогічно до попереднього критерію тільки застосовуються у випадку відсутності інформації щодо дисперсії величин у вибірках. Якщо об'єм вибірки є великим n > 200 то допустимо використовувати критерій 10 де $\sigma_x^2 \approx S_x^2$ $\sigma_y^2 \approx S_x^2$, але у випадку невеликих вибірок n <100 при невідомій дисперсії потрібно використовувати критерій 13.

Критерій 16 і його варіанти 17 та 18 використовують для порівняння дисперсій двох вибірок, чи можна їх вважати однаковими чи вони різні.

Критерій 19 призначений для перевірки значення коефіцієнту кореляції. Він використовується в кореляційному аналізі. Якщо нульова гіпотеза відносно

коефіцієнту кореляції відхилена, то можна твердити про наявність кореляційного зв'язку між двома випадковими величинами.

Критерій 20 – найбільш вживаний критерій перевірки чи однаковим є значення математичного очікування двох вибірок. Його часто застосовують замість використання t критерію для визначення ефекту зсуву. Особливістю його використання є те, що в даному випадку не розраховують середніх значень для кожної вибірки а розраховують тільки дисперсію значень у вибірках. Якщо дві вибірки відрізняються математичним очікуванням величини, то очевидно, що дисперсія об'єднаної вибірки створеної із всіх елементів двох вибірок буде більша ніж дисперсія кожної окремої вибірки. За відношенням дисперсій можна визначити чи однаковими є математичні очікування.

Критерій χ^2 (Пірсона) для простої гіпотези. Нехай $\{X_1, X_2, \ldots, X_n\}$ вибірка з генеральної сукупності F. Перевіряється гіпотеза $H_0 : F = F_1$ проти альтернативи $H_1 : F \neq F_1$.

Уявімо вибірку у вигляді асоційованого ряду, розбивши передбачувану область значень випадкової величини на m інтервалів. Нехай n_i – число елементів вибірки потрапивших в i-ий інтервал, а p_i – теоретична ймовірність попадання в цей інтервал за умови істинності H_0. Складемо статистику $\rho(\vec{X}) = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i}$, яка характеризує суму квадратів відхилення спостережуваних значень n_i від очікуваних np_i по всіх інтервалах групування.

Теорема Пірсона. Якщо H_0 вірна, то при фіксованому m та $n \to \infty$

$$\rho(\vec{X}) = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i} \Rightarrow \chi^2_{m-1} \tag{1.31}$$

(закон розподілу статистики $\rho(\vec{X})$ прагне до розподілу хі-квадрат з $m-1$ ступенем свободи).

Таким чином, статистику $\rho(\vec{X})$ можна використовувати в якості статистики критерію згоди для перевірки гіпотези про вид закону розподілу, який буде мати вигляд:

$$\delta(\vec{X}) = \begin{cases} H_0, & \rho(\vec{X}) < \tau_{1-\alpha} \\ H_1, & \rho(\vec{X}) \geq \tau_{1-\alpha} \end{cases}, \quad \rho(\vec{X}) = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i}, \tag{1.32}$$

де $\tau_{1-\alpha}$ – квантиль розподілу χ^2_{m-1}.

Цей критерій називається критерієм χ^2 або критерієм згоди Пірсона.

Досягнутий (спостережуваний) рівень значимості критерію:

$$\alpha_{спост} = P(\rho(\vec{X}) > \rho_{спост}) = 1 - F_{\chi^2_{m-1}}(\rho_{спост}), \tag{1.33}$$

$\rho_{спост}$ – обчислене за вибіркою значення статистики критерію.

Зауваження. Критерій не є спроможним для альтернатив, для яких $\tilde{p}_i = p_i$ для всіх $i \in \{1,2,\ldots,m\}$. Тому, слід прагнути до якомога більшій кількості інтервалів групування. Однак, з іншого боку, збіжність до розподілу хі-квадрат величини $\frac{(n_i - np_i)^2}{np_i}$ забезпечується ЦПТ, причому похибка наближення істотно зростає, якщо очікуване значення np_i для комірки занадто мале ($np_i < 5$). Тому зазвичай число інтервалів m вибирають таким чином, щоб для кожного осередку величина $np_i \geq 5$.

Критерій χ^2 (Пірсона) для складної гіпотези. Нехай $\{X_1, X_2, \ldots, X_n\}$ вибірка з генеральної сукупності F. Перевіряється складна гіпотеза $H_0 : F = F_\theta$, де θ – невідомий параметр розподілу F (або вектор параметрів), проти альтернативи $H_1 : F \neq F_\theta$.

Нехай вибірка по колишньому представлена у вигляді асоційованого ряду і n_i – число елементів вибірки потрапили в i-ий інтервал, $i \in \{1,2,\ldots,m\}$. Статистику (1.31) не можемо в цьому випадку використовувати для побудови критерію Пірсона, тому що не можемо обчислити теоретичні значення ймовірностей p_i, які залежать від невідомого параметра θ. Нехай θ^* – оцінка

параметра θ, а $p_i^*(\theta^*)$ – відповідні їй оцінки ймовірностей p_i. складемо статистику $\rho(\vec{X}) = \sum_{i=1}^{m} \frac{\left(n_i - np_i^*\right)^2}{np_i^*}$.

Теорема Пірсона. Якщо H_0 вірна, та l – число компонент вектора θ (Число невідомих параметрів розподілу), то при фіксованому m и $n \to \infty$

$$\rho(\vec{X}) = \sum_{i=1}^{m} \frac{\left(n_i - np_i^*\right)^2}{np_i^*} \Rightarrow \chi^2_{m-l-1} \tag{1.34}$$

(Закон розподілу статистики $\rho(\vec{X})$ прагне до розподілу хі-квадрат з $m-l-1$ ступенем свободи).

Таким чином, критерій Пірсона для параметричної гіпотези матиме вигляд:

$$\delta(\vec{X}) = \begin{cases} H_0, & \rho(\vec{X}) < \tau_{1-\alpha} \\ H_1, & \rho(\vec{X}) \geq \tau_{1-\alpha} \end{cases}, \qquad \rho(\vec{X}) = \sum_{i=1}^{m} \frac{\left(n_i - np_i^*\right)^2}{np_i^*}, \tag{1.35}$$

де $\tau_{1-\alpha}$ – квантиль розподілу χ^2_{m-l-1}.

Досягнутий (спостережуваний) рівень значимості критерію:

$$\alpha_{набл} = P(\rho(\vec{X}) > \rho_{набл}) = 1 - F_{\chi^2_{m-l-1}}(\rho_{набл}). \tag{1.36}$$

Зауваження. Взагалі кажучи, оцінки, які використовуються для побудови статистики критерію хі-квадрат, повинні бути визначені з умови мінімуму статистики $\rho(\vec{X})$. Тому бажано уточнити оцінки, знайдені в інший спосіб (методом максимальної правдоподібності або методом моментів) шляхом мінімізації $\rho(\vec{X})$.

Критерій згоди λ (Колмогорова)

Перевіряється гіпотеза $H_0 : F\xi(x) = F_0(x)$

проти альтернативи $H_1 : F\xi(x) \neq F_0(x)$,

де $F\xi(x)$ – функція розподілу генеральної сукупності,

$F_0(x)$ – гіпотетична функція розподілу (повністю відома функція). Вона передбачається безперервною.

Для перевірки гіпотези використовується статистика

$$\lambda = \Delta\sqrt{n} \tag{1.37}$$

$$\Delta = \max_x | F_0(x) - F_\xi^*(x)$$

максимальний модуль відхилення гіпотетичної функції розподілу $F_0(x)$ від емпіричної функції розподілу $F_\xi^*(x)$.

Якщо гіпотеза H_0 вірна, то статистика λ (1.37) має розподіл, що наближається при $n \to \infty$ до розподілу Колмогорова. Критерій для перевірки гіпотези має наступний вигляд:

$$P(\lambda > \lambda_\alpha) = \alpha,$$

де $\lambda_\alpha - 100\alpha$ – відсоткове відхилення розподілу Колмогорова (табл.1.6).

Таблиця 1.6 – Процентні відхилення розподілу Колмогорова

α	0.01	0.02	0.03	0.04	0.05
λ_α	1.627	1.520	1.45	1.40	1.358

Коефіцієнт вірогідності (t-критерій Стьюдента).

Критерій був розроблений Вільямом Госсетом, який працював gsl псевдонім Стьюдент. Це розподіл ймовірностей, пов'язаний з нормальним розподілом. Виникає, коли потрібно оцінити середню статистичної вибірки, коли розмір вибірки, що використовується для оцінки, малий і дисперсії невідомі.

Безпосередньо критерій Стьюдента спрямований на оцінку відмінностей величин середніх значень двох вибірок, які розподілені по нормальному закону. Головною перевагою критерію є широта його застосування. Він може бути використаний для зіставлення середніх у пов'язаних та непов'язаних вибірках, причому вибірки можуть бути не рівні за величиною.

Умовою застосування даного критерію є:

–Вимірювання може бути проведено в шкалі інтервалів і відносин.

–Порівнянні вибірки повинні бути розподілені за нормальним законом.

Парний t-критерій Стьюдента, є модифікацією критерію та використовується для порівняння двох залежних (парних) вибірок. Залежними є вимірювання, виконані у одних і тих же людей, але в різний час. Нульова гіпотеза свідчить про відсутність відмінностей між порівнюваними вибірками, альтернативна – про наявність статистично значущих відмінностей.

Основною умовою застосування є залежність вибірок, тобто порівнювані значення повинні бути отримані при повторних вимірах одного параметра.

Розглядають дві незалежні нормальні вибірки з генеральних сукупностей, що мають рівні або нерівні, але відомі чи рівні невідомі дисперсії.

Значення критерію Стьюдента розраховують за формулою:

$$t = \frac{\left|\bar{x}_A - \bar{x}_B\right|}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}, \qquad (1.38)$$

де σ_A^2, σ_B^2 – відомі внутрішньогрупові дисперсії;

n_A та n_B – чисельності груп. Для m груп рівної чисельності статистика має *t*-розподіл з кількістю степенів вільності $m(n-1)$.

У випадку, коли обсяги вибірок є малими або істотно розрізняються, а їх дисперсії є рівними, останні замінюють вибірковим середнім квадратичним відхиленням, яке розраховують за формулою:

$$S^2 = \frac{S_1^2(n_1-1) + S_2^2(n_2-1)}{n_1+n_2-2}; \qquad (1.39)$$

якщо стандартні відхилення вибірок оцінюють за самими вибірками, або:

$$S^2 = \frac{S_1^2 n_1 + S_2^2 n_2}{n_1+n_2-2}, \qquad (1.40)$$

якщо їх оцінюють незалежно. Формула для визначення розрахункового значення критерію у цьому разі набуває вигляду:

$$t = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}. \qquad (1.40)$$

Відповідна статистика має розподіл Стьюдента з $k = n_1 + n_2 - 2$ пенями вільності.

При застосуванні критерію Стьюдента у вигляді (1.40) необхідно спочатку перевірити гіпотезу про рівність дисперсій.

Критичні точки симетричні стосовно нуля. Нульову гіпотезу відхиляють: якщо $|t| = t_{оо}(\alpha/2, k)$ при конкуруючій гіпотезі $\bar{x}_1 \neq \bar{x}_2$; якщо $t > t_{оо}(\alpha, k)$ при конкуруючій гіпотезі $\bar{x}_1 > \bar{x}_2$ якщо $t < t_{оо}(\alpha, k)$ при конкуруючій гіпотезі $\bar{x}_1 < \bar{x}_2$.

При аналізі спряжених вибірок їх порівняння здійснюють з метою визначення наявності ефекту від певного фактора, наприклад, впливу змін у технології на якість виробленої продукції. Вимога щодо рівності дисперсій при цьому не висувається. Нульова гіпотеза полягає у відсутності різниці між середніми. Значення критерію розраховують за формулою:

$$t = \frac{\sum_{i=1}^{n} \delta_i}{\sqrt{\frac{n\sum_{i-1}^{n} \delta_i^2 - \left(\sum_{i=1}^{n} \delta_i\right)^2}{n-1}}}, \qquad (1.41)$$

де n – кількість елементів у кожній із вибірок;

$\delta_i = x_i - y_i$, x_i та y_i – відповідні значення елементів першої та другої вибірок.

Іноді цей критерій називають одновибірковим критерієм Стьюдента. Відповідна статистика має розподіл Стьюдента з кількістю степенів вільності n −1.

Таблиця 1.7 – Таблиця критичних значень критерію Стьюдента t при числі спостережень n<30

Число ступенів свободи = n - 1	Рівень ймовірності безпомилкового прогнозу (у відсотках)		
	95 (0.05)	**99** (0.01)	**99.9** (0.001)
1	12,7	63,6	636,6
2	4,3	9,9	31,6
3	3,1	5,8	12,9
4	2,7	4,6	8,6
5	2,5	4,0	6,8
6	2,4	3,7	5,9
7	2,3	3,5	5,4
8	2,3	3,3	5.1
9	2,2	3,2	4,7
10	2.2	3,1	4,6
11	2,2	3,1	4,4
12	2,2	3,0	4,3
13	2,1	3,0	4,2
14	2,1	2,9	4,1
15	2,1	2.9	4,0
16	2,1	2,9	4,0
17	2,1	2,8	3,9
18	2,1	2,8	3,9
19	2,0	2,8	3,8
20	2,0	2,8	3,8
21	2,0	2,8	3,8
22	2,0	2,8	3,7
23	2,0	2,8	3.7
24	2,0	2,7	3,7

Продовження таблиці 1.7

Число ступенів свободи = n - 1	Рівень ймовірності безпомилкового прогнозу (у відсотках)		
	95 (0.05)	**99** (0.01)	**99.9** (0.001)
25	2.0	2.7	3,7
26	2,0	2,7	3,7
27	2,0	2.7	3,6
28	2,0	2,7	3,6
29	2,0	2.7	3,6
30	2,0	2,7	3,6

Таблиця 1.8 – Значення критерію Стьюдента t при числі спостереження n>30

Величина критерію Стьюдента t	Ймовірність безпомилкового прогнозу	
	в одиницях	у відсотках
1,0	0,6827	68,3
1,5	0,8664	86,6
2,0	0,9545	95,5
2,5	0,9876	98,8
3,0	0,9973	99,7
3,5	0,9995	99,95
4,0	0,9999	99,99

Парний t-критерій може використовуватися тільки при порівнянні двох вибірок. Якщо необхідно порівняти три і більше повторних вимірів, слід використовувати однофакторний дисперсійний аналіз для повторних вимірів.

Точний критерій Фішера.

Завдяки йому можна дослідити взаємозв'язок певних факторів і результату, порівнювати частоту патологічних станів між двома групами досліджуваних і т.д.

Точний критерій Фішера – це критерій, який використовується для порівняння двох відносних показників, що характеризують частоту певної ознаки, що має два значення. Вихідні дані для розрахунку точного критерію Фішера зазвичай групуються у вигляді чотирьохпільної таблиці.

Його значення розраховують за формулою:

$$F = s_1^2 / s_2^2, \qquad (1.42)$$

де s_1^2, s_2^2 – значення оцінок більшої та меншої дисперсій, відповідно. Кількості степенів вільності для пошуку критичного значення обирають рівними $n_1 - 1$ та $n_2 - 1$. Гіпотезу про рівність дисперсій порівнюваних сукупностей відхиляють, якщо обчислене значення перевищує табличне при заданому довірчому рівні. При цьому, якщо конкуруючою є однобічна гіпотеза $s_1^2 > s_2^2$, то як критичну точку беруть значення оберненого розподілу Фішера, що відповідає рівню значущості α при заданій кількості степенів вільності. Якщо ж конкуруючою є двобічна гіпотеза $s_1^2 = s_2^2$, то критичною точкою буде значення оберненого розподілу Фішера, що відповідає рівню значущості α / 2.

Точний критерій Фішера можна застосовувати коли порівнянні змінні виміряні в номінальній шкалі і мають тільки два значення, наприклад, артеріальний тиск в нормі або підвищений, вихід сприятливий або несприятливий, післяопераційні ускладнення є чи ні.

1.3 Методичні вказівки щодо організації самостійної роботи студентів

Перед лабораторною роботою слід повторити матеріал за курсом лекцій і за рекомендованою літературою за темою лабораторної [1, лекц.1–3].

Особливо слід звернути увагу на такі питання:

– Види статистичних даних.

– Критерії розподілу ознак.

– Формулювання статистичної гіпотези.

– Шкали вимірювання.

– Перевірка статистичних гіпотез.

– Алгоритм перевірки статистичних гіпотез.

– Обсяг вибірки.

– Розбиття вибірки на групи.

– Показники ступеня середньої та їх вибір.

1.4 Завдання до лабораторної роботи

1. Вивчити теоретичну частину заняття.

2. Обрати відповідно до варіанту предметну область, визначити (зробити припущення та сформулювати нульову та альтернативну статистичні гіпотези), яке завдання необхідно вирішити відповідно до специфіки діяльності (з якими даними необхідно працювати, звідки дані отримуються та які саме данні).

3. Відповідно до п.2:

– Описати основні джерела і канали даних та які саме дані необхідно аналізувати (БД фірми, зовнішні дані з відкритих джерел / соціальних мереж / сайтів конкурентів, метадані і т.д.). Дані структуровані або не структуровані. Чи необхідна попередня обробка даних.

– Описати які саме дані потрібні для аналізу (наприклад: повинно мати загальну статистику звернень до сайту за поточну добу і включає наступні метрики: загальне число запитів і число переглядів сторінок сайту, число унікальних відвідувачів і число сесій, загальний обсяг переданого трафіку. Ця статистика наводиться окремо для запитів, що виходять від реальних відвідувачів сайту, і запитів, згенерованих іншими джерелами (програмами-

роботами, вірусами). У звітах, враховуються тільки запити, які виходять від реальних відвідувачів.)

– Як відбувається збільшення обсягів даних (скільки в день нових даних, скільки в місяць).

4. Обґрунтувати який розмір вибірки необхідне для проведення якісного дослідження за темою з п.2. (наведіть поетапний розрахунок).

5. З відкритих джерел взяти необхідні дані в необхідному розмірі за відповідною темою (наприклад з https://www.kaggle.com/fernandol/countries-of-the-world). Вказати звідки були взяті дані.

6. Визначити мінімальне та максимальне значення вибірки. Розбити значення вибірку на групи відповідно до необхідного методу, вибір обґрунтуйте.

7. Виконати перевірку висунутих гіпотез, аналіз та інтерпретацію отриманих результатів. Зробити пояснення чому саме було обрано відповідний критерії перевірки для вашої гіпотези (на оцінку добре та відмінно). На оцінку задовільно вирішити 3 задачі з запропонованих нижче, рішення навести покроково.

8. Сформулювати висновки і пояснити чому вийшов такий результат.

9. Оформити звіт що демонструє роботу і здати викладачеві в електронному вигляді.

10. Відповісти на контрольні питання (парний варіант дає відповіді на парні питання, непарний варіант дає відповіді непарні питання).

Запропоновані теми для виконання завдання

1. Рекламне агентство.
2. Готель.
3. Культурно-спортивний центр.
4. Лікарня.

5. Фірма з продажу автомобілів.
6. Банк.
7. Архітектурна організація.
8. Видавництво.
9. Центру служби зайнятості.
10. Фірма з продажу та оренди житла.
11. Фірма по влаштуванню свят.
12. Супермаркет.
13. Фірма по ремонту квартир.
14. Фірма з оренди автомобілів.
15. Торгове підприємство з продажу меблів.
16. Дизайнерське бюро.
17. Магазин зоотоварів.
18. Авіапредпріятіе, що займається перевезенням вантажів.
19. Фірма з продажу різних груп товарів (по типу Амазон, Розетка, Каста).
20. Фірма з продажу косметичних засобів.
21. Велика IT фірма (Epam, Nix т.і.).
22. М’ясо комбінат.
23. Фірма з продажу електроніки.
24. Туристична фірма.

Запропоновані задачі для виконання завдання

Задача 1. Вимірювалась швидкість руху автомобілів Результати вимірів наведено в табл. 1.9.

Таблиця 1.9 – Дані для задачі 1

xi ,	56	60	64	68	72	70	80
ni	2	4	6	8	3	1	1

Вважаючи, що X – швидкість автомобіля – є випадковою величиною, яка має нормальний закон розподілу, перевірити з надійністю 99% правильність нульової гіпотези: $H_0: M(X) = 70$, $H_a: M(X) \neq 70$

Задача 2. Два студенти деякого факультету бажають поїхати на стажування за кордон. Результати попередніх тестів визначили рівнозначність їх кандидатур. Єдиний критерій, який залишився – показники успішності за навчальний рік (табл.1.10). Визначити, хто із студентів поїде на стажування за кордон з надійністю 95%.

Таблиця 1.10 – Дані для задачі 2

	x_1	x_2
1.	91,5	91,4
2.	90,3	96,9
3.	87,4	99,8
4.	85,5	84,7
5.	97,0	91,3
6.	100	86,5
7.	93,2	94,4
8.	87,9	93,4
9.	100	87,5
10.	93,8	100,

Задача 3. Для перевірки впливу методики навчання на якість підготовки фахівців вибирають чотири групи студентів, які після закінчення навчання за різними методиками тестувалися. Результати тестування наведено в табл.1.11.

Таблиця 1.11 – Дані для задачі 3

Ступінь впливу фактора A (методики)	Оцінка
A_1	60, 80, 75, 80, 85, 70
A_2	75, 66, 85, 80, 70, 80, 90
A_3	60, 80, 65, 60, 86, 75
A_4	95, 85, 100, 80

При рівні значущості $\alpha = 0.05$ з’ясувати вплив методики навчання на якість підготовки учнів.

Задача 4. Дана вибірка обсягом $n = 100$ з невідомого розподілу F див табл. 1.12.

Таблиця 1.12 – Дані для задачі 4

4,81	7,03	4,95	0,25	13,00	26,52	1,40	3,19	0,07	1,99
11,48	15,45	5,17	14,65	8,09	0,38	2,34	1,14	0,39	1,56
2,58	17,15	0,47	1,75	13,74	11,50	8,75	1,08	0,51	2,68
0,53	9,04	3,82	1,01	5,13	6,80	4,52	6,69	3,04	9,41
0,61	7,58	4,26	0,14	3,60	1,27	2,97	8,63	3,46	0,57
0,21	20,35	5,96	3,81	3,35	1,93	1,70	0,71	1,97	4,87
21,17	6,28	0,12	6,02	4,92	1,06	2,94	10,82	3,57	8,04
4,49	5,35	1,07	1,44	0,07	1,61	8,54	14,11	9,63	7,90
0,74	2,96	0,04	5,23	16,01	12,32	0,15	1,36	16,36	5,48
9,88	5,14	6,81	1,27	7,33	10,11	1,88	1,52	1,14	5,62

Потрібно, використовуючи критерій хі-квадрат, на рівні значущості $\alpha = 0{,}05$ перевірити гіпотезу про розподіл вибіркових даних по показовому закону з щільністю $f(x) = \lambda e^{-\lambda x}$, $x > 0$, де параметр $\lambda > 0$ – невідомий. Вказати досягнутий рівень значущості.

1.5 Опис програмного забезпечення

Під час виконання лабораторної роботи використовується таке програмне забезпечення: будь-який текстовий редактор Microsoft Office Word або OpenOffice; браузери Chrome, Edge, Firefox, Opera.

1.6 Зміст звіту

1. Тема і мета роботи.

2. Варіант для якого проводилися розрахунки за відповідною темою(вказати тему).

3. Послідовність виконуваних у процесі роботи дій.

4. Опис процесу роботи.

5. Висновки з роботи.

6. Відповіді на контрольні питання (парний варіант дає відповіді на парні питання, непарний варіант дає відповіді непарні питання).

1.7 Контрольні запитання та завдання

1. Що таке генеральна сукупність і вибірка?
2. Які властивості повинна мати вибірка?
3. Що таке якість даних?
4. Які цілі підготовки даних до аналізу? Які завдання в неї входять?
5. Повторна і безповторна вибірки.
6. Способи формування вибірок.
7. Визначення обсягів вибірок — розрахунок обсягів повторної і безповторної вибірок за різних умов.
8. Групування емпіричних даних. Дискретний розподіл вибірки.
9. Інтервальний розподіл вибірки.
10. Формулювання статистичних гіпотез.
11. Загальна схема перевірки гіпотез.
12. Помилки першого і другого роду при перевірці гіпотез.
13. Прийняття рішень на основі перевірки статистичних гіпотез.
14. Непараметричні й параметричні критерії перевірки статистичних гіпотез.
15. Статистичні гіпотези в програмному забезпеченні.
16. Що таке вибірка і генеральна сукупність?
17. Назвіть характеристики варіативності випадкової величини.

18. Що таке частота (частотність) появи випадкової величини.

1.8 Додаткова література та електронні ресурси

1. Мир статистических гипотез. Статистика в IT [Електроний ресурс]. https://habr.com/en/articles/558836/ (дата звернення: 1.09.2023).

2. Онлайн калькулятор для розрахунку розміру вибірки. [Електроний ресурс]. https://socio-lab.vntu.edu.ua/download/Calculator.html (дата звернення: 1.09.2023).

3. Розрахунок довірчого інтервалу Он-лайн калькулятор [Електроний ресурс]. https://socio-lab.vntu.edu.ua/download/Calculator.html (дата звернення: 1.09.2023).

4. Семплювання та точність обчислень [Електроний ресурс]. https://habr.com/en/articles/458890/ (дата звернення: 1.09.2023).

5. Методи аналізу даних : навчальний посібник для студентів / В.Є. Бахрушин. – Запоріжжя : КПУ, 2011. – 268 с.

6. Реєстр засобів статичного аналізу коду. [Електронний ресурс]. – Режим доступу: http://en.wikipedia.org/wiki/List_of_tools_for_static_code_analysis. (дата звернення: 1.09.2023)

7. A. Karpov How to complement TDD with static analysis [Електронний ресурс]. – Режим доступу: https://pvs-studio.com/en/blog/posts/cpp/a0080/ (дата звернення: 1.09.2023)

8. . Реєстр каркасів тестування для різних мов програмування. [Електронний ресурс]. – Режим доступу: http://en.wikipedia.org/wiki/List_of_unit_testing_frameworks. (дата звернення: 1.09.2023)

2 МЕТОДИ КІЛЬКІСНОГО І ЯКІСНОГО ДОСЛІДЖЕННЯ, ТА ПІДБІР КРИТЕРІЇВ ДЛЯ ДОСЛІДЖЕННЯ. ВИЯВЛЕННЯ І ВИКОРИСТАННЯ ФОРМАЛІЗОВАНИХ ЗАКОНОМІРНОСТЕЙ

2.1 Мета та завдання роботи

Метою лабораторної роботи є ознайомитися з виявлення і використання формалізованих закономірностей для аналізу в процесі ЖЦ ПЗ. Набуття знань та практичних навичок відносно обчислення, формулювання та обґрунтування необхідних висновків на основі отриманих та відібраних даних в галузі дослідження закономірностей, що виникають в процесі розробки та функціонування програмного забезпечення.

2.2 Опис роботи

Використання кількісних методів в дослідженнях дає можливість, по-перше, виділити та формально описати найбільш важливі й суттєві закономірності функціонування в системі та об’єктів у вигляді моделей. Також на основі сформульованих за певними правилами логіки вхідних даних і співвідношень, методами дедукції зробити висновки, які адекватні об’єкту дослідження в міру зроблених припущень.

Виявлення закономірностей дозволяє розробити описові моделі прийняття рішень – концепції, які допомагають зрозуміти та прогнозувати дії в ситуації вибору. Досягнення теорії прийняття рішень дозволяють уникнути помилок при здійсненні вибору на множині альтернатив, а також розуміти, прогнозувати та навіть змінювати рішення.

Факторний і дисперсійний аналізи.

Факторний аналіз – статистичний метод, який використовується при обробці великих масивів експериментальних даних. Мета факторного аналізу: скоротити число змінних (редукція даних) і визначити структуру взаємозв'язків між ними. Можна також сказати, що в завдання факторного аналізу входить структурна класифікація змінних.

Важливою відмінністю факторного аналізу від інших статистичних методів в тому, що його не можна застосовувати для обробки первинних, або як кажуть «сирих», експериментальних даних, тобто отриманих безпосередньо при/від випробуваного об'єкта.

Матеріалами для факторного аналізу служать кореляційні зв'язки, а точніше, коефіцієнти кореляції Пірсона, які обчислюються між змінними показниками (параметрами), включеними в обстеження.

Таким чином, факторному аналізу піддаються кореляційні матриці, або, як їх називають інакше, *матриці інтеркореляцій*. Найменування стовпців і рядків в цих матрицях однакові, так як вони представляють собою перелік змінних, включених в аналіз. Матриці інтеркореляцій завжди квадратні, тобто число рядків в них дорівнює числу стовпців, і симетричні, тобто, на головній діагоналі матриці стоять одні й ті ж коефіцієнти кореляції. У табл.2.1 наведено приклад такої матриці.

Таблиця 2.1

\	А	Б	В	Г	Д
А	1,0	0,2	0,7	0	0,9
Б	0,2	1,0	0,1	0,9	0
В	0,7	0,1	1,0	0,6	0,4
Г	0	0,9	0,6	1,0	0,8
Д	0,9	0	0,4	0,8	1,0

Очевидно, що якщо *коефіцієнт кореляції* (rk) між якимись показниками дорівнює нулю, то ці показники незалежні один від одного, при коефіцієнтах кореляції від 0,3 до 0,4 – слабка кореляція (залежність), при rk = 0,5 – 0,75– хороша кореляція, при 0,8-0,95 – дуже хороша кореляція, при rk = 1– залежність детермінована.

Слід зазначити, що вихідна таблиця даних може складатися з будь-якого числа рядків і стовпців, але матриця інтеркореляцій повинна бути квадратної, так як і в стовпчиках, і в рядках записуються одні й ті ж показники.

Головне поняття факторного аналізу – *фактор*. Це штучний статистичний показник, що виникає в результаті спеціальних перетворень таблиці коефіцієнтів кореляцій. Процедура вилучення факторів з матриці інтеркореляцій називається *факторизацією матриці*. В результаті факторизації з кореляційної матриці може бути вилучено різну кількість факторів, але не перевищує числа показників (рядків або стовпців) матриці. Однак фактори, які виявляються в результаті факторизації, як правило, нерівноцінні за своїм значенням. Елементи факторной матриці – коефіцієнти кореляції – часто називаються «факторними навантаженнями», або «факторними вагами».

Для того щоб краще засвоїти сутність факторного аналізу, розберемо більш детально такий приклад.

При розробці нового автомобіля необхідно виробити споживчі вимоги до конструкції його дверей. Припустимо, що при колективній виробці споживчих вимог до конструкції двері передбачуваного до випуску автомобіля покупцями висловлені наступні вимоги:

- двері повинні легко відкриватися (Т1),
- двері не повинні пропускати пилу (Т2),
- двері повинні бути чітко зафіксована при її повному відкритті (Т3),
- двері не повинні пропускати дорожнього шуму (Т4),
- двері повинні легко закриватися, без сильного хлопку (Т5),
- двері повинні бути чітко пригнана до кузова (Т6),

– двері не повинні іржавіти (Т7).

В реальній ситуації було висловлено значно більше число вимог, але для прикладу наведеного кількості споживчих вимог досить. Намалюємо таблицю попарних кореляцій rk (матрицю інтеркореляцій) між споживчими вимогами до дверей автомобіля (табл.2.2):

Таблиця 2.2 – Матриця інтеркореляцій

\	Т1	Т2	Т3	Т4	Т5	Т6	Т7
Т1	1,0	0,2	0,8	0,3	0,7	0,4	0
Т2	0,2	1,0	0	0,9	0,4	0,8	0,1
Т3	0,8	0	1,0	0	0,7	0,3	0
Т4	0,3	0,9	0	1,0	0,3	0,8	0
Т5	0,7	0,4	0,7	0,3	1,0	0,4	0,1
Т6	0,4	0,8	0,3	0,8	0,4	1,0	0,1
Т7	0	0,1	0	0,1	0,1	0,1	1,0

Коефіцієнти кореляції відображають спорідненість між собою споживчих вимог.

При аналізі величин коефіцієнтів кореляції rk легко виділити групи вимог, добре взаємопов’язаних, тобто, що мають загальне призначення, крім самого поняття «двері». Назвемо ці групи:

А – двері повинні бути зручні в експлуатації (вимоги Т1, Т3, Т5),

Б – двері повинні бути герметична (вимоги Т2, Т4, Т6).

Очевидно, що вимога Т7 (нержавіючий матеріал обшивки дверей) – дуже важлива, але вона відноситься до матеріалу двері і має слабке відношення до конструкції двері. Швидше за все, ця вимога потрапить в загальні вимоги по автомобілю в наступному вигляді: металева обшивка автомобіля повинна бути виконана з нержавіючих матеріалів.

Таким чином, змістовний аналіз всіх вимог показав, що шість з них характеризують дві узагальнених вимоги: зручність в експлуатації і

герметичність. Назвемо ці узагальнені вимоги факторами і застосуємо до них факторний аналіз.

Представимо в табл. 2.3 ці два фактори А і Б у вигляді стовпців, а змінні (споживчі вимоги) – у вигляді рядків. При цьому кожному фактору в рядку буде відповідати середнє значення коефіцієнта кореляції відповідних змінних за цим фактором. Як було зазначено вище, коефіцієнти кореляції в факторной матриці (табл. 2.3) називаються факторними навантаженнями (вагами).

Таблиця 2.3 – Змінна Фактор А Фактор Б

Т1	0,83	0,30
Т2	0,30	0,90
Т3	0,83	0,10
Т4	0,40	0,90
Т5	0,80	0,40
Т6	0,35	0,87
Т7	0	0,1

Як видно з табл.2.3, факторні навантаження (або ваги) А і Б для різних споживчих вимог значно відрізняються. Факторне навантаження А для вимоги Т1 відповідає тісноті зв’язку, яка характеризується коефіцієнтом кореляції, що дорівнює 0,83, тобто хороша (тісна) залежність.

Факторне навантаження Б для того ж вимоги дає $rk = 0,3$, що відповідає слабкій тісноті зв’язку. Як і передбачалося, фактор Б дуже добре корелюється з споживчими вимогами Т2, Т4 і Т6.

З огляду на те, що факторне навантаження А, так само як і факторне навантаження Б, впливають на споживчі вимоги, що не належать до їхньої групи, з тіснотою зв’язку не більше 0,4 (тобто слабо), то можна вважати, що представлена вище матриця інтеркореляцій (табл. 2.2) визначається двома незалежними факторами, які в свою чергу визначають шість споживчих вимог (за винятком Т7).

Змінну Т7 можна було виділити в самостійний фактор, так як ні з однією споживчою вимогою вона не має значущого кореляційного навантаження (понад 0,4). Але, це не слід робити, так як фактор «двері не повинні іржавіти» не має безпосереднього відношення до споживчим вимогам по конструкції двері.

Таким чином, при затвердженні технічного завдання на проектування конструкції дверей автомобіля саме назви отриманих факторів будуть вписані як споживчі вимоги, за якими необхідно знайти конструктивне рішення у вигляді інженерних характеристик.

Аналіз подій і пошук закономірностей за допомогою методу асоціативних правил.

Алгоритми асоціативних правил дозволяють знаходити закономірності між залежними подіями.

Метою аналізу встановлення залежності вигляду “якщо в транзакції зустрівся набір елементів *X*, то можна зробити висновок, що інший набір елементів *Y* також повинен з’явитися в цій транзакції”.

Транзакція – це множина подій, що відбулися одночасно. Нехай є база даних, що складається з транзакцій щодо придбання. Кожна транзакція – це набір товарів, які придбав покупець за один візит. Таку транзакцію ще називають ринковим кошиком.

Нехай є список транзакцій. Необхідно знайти закономірності між цими подіями. Як в умові, так і в наслідку правила повинні знаходитися елементи транзакцій.

Нехай $I = \{i_1, i_2, ..., i_n\}$ – множина елементів, що входять у транзакції.

Асоціативним правилом є пара <*X*, *Y*>, в якій *X* -> *Y* – з *X* випливає *Y*; *D* – загальна кількість транзакцій;

S – частина транзакцій у % від загальної кількості транзакцій *D*.

Основні характеристики правил, що отримуються на основі таких залежностей, – це *підтримка* і *вірогідність*.

Правило “з *X* випливає *Y*” має підтримку *s*, якщо *s* % транзакцій з усього набору *D* містять набори елементів *X* і *Y*.

Вірогідність правила показує, яка імовірність того, що з *X* випливає *Y*. Правило “з *X* випливає *Y*” справедливе з вірогідністю *C*, якщо *C* % транзакцій із усієї множини *D*, що містять набір елементів *X*, також містять набір елементів *Y*.

Також асоціативні правила характеризуються значеннями мінімальної підтримки та мінімальної вірогідності.

Ці значення є граничними і враховуються при побудові правил. Мінімальні значення цих показників дозволяють обмежити кількість знайдених правил (низьке значення підтримки зумовлює генерування великої кількості правил, що вимагає великої кількості обчислювальних ресурсів і великих часових витрат на подальше доопрацювання).

Але якщо для показника підтримки будуть обрані великі значення, алгоритми пошуку будуть знаходити правила, що вже відомі або очевидні.

Виявлення нетривіальних правил, що будуть становити найбільший інтерес для прийняття рішення, – це одна з головних задач при пошуку асоціативних залежностей. Існує набір нескладних правил, що дозволяють знаходити найбільш цікаві залежності.

Зменшення мінімальної підтримки призводить до збільшення кількості потенційно цікавих правил, при цьому потрібні значні обчислювальні ресурси. Занадто мале значення підтримки робить правило статистично необґрунтованим.

Зменшення порога вірогідності також призводить до збільшення кількості правил. Також не рекомендується задавати занадто мале значення для вірогідності, наприклад, правило з вірогідністю 5 % має малу цінність і фактично не є правилом.

Правило з великим значенням підтримки має велику статистичну цінність, але на практиці може свідчити про те, що воно добре відоме і не

викликає інтересу, або про те, що товари, що входять до правила, є очевидними лідерами.

Занадто велика вірогідність правила знижує його практичну цінність, тому що може свідчити про те, що незалежно від розташування товарів й інших зусиль працівників супермаркету, покупці все одно будуть купувати ці товари в одній транзакції.

Дерево правил – це дворівневе дерево, що може бути побудоване як за умовою, так і за наслідком. При побудові дерева правил за умовою на першому (верхньому) рівні знаходяться вузли з умовами, а на другому (нижньому) рівні – вузли з наслідком. Отже може вирішувати задачу, наприклад, «що буде, якщо виконається задана умова». Побудова дерева за наслідками здійснюється в зворотному порядку. У даному випадку зможе вирішувати задачі, наприклад – «що значною мірою впливає на виконання даного наслідку».

Аналіз подій і пошук закономірностей можна зробити також за допомогою матриці трасування.

Збір вимог є початковим і невід'ємним етапом процесу розробки програмних систем. Він полягає у визначенні набору функцій, які необхідно реалізувати в продукті. Процес збору вимог реалізується частково в спілкуванні з замовником, частково за допомогою мозкових штурмів розробників. Результатом є формування набору вимог до системи, іменованої технічним завданням.

Фіксація вимог (Requirement Capturing), з одного боку, визначається бажаннями замовника в реалізації тієї чи іншої властивості. З іншого боку в процесі збору вимог може виявитися помилка, яка призведе до певних наслідків, усунення яких забере непередбачені ресурси – додаткове кодування, перепланування.

Помилка може бути тим серйозніше, ніж пізніше вона буде виявлена, особливо якщо це пов'язано з великою кількістю специфікацій. Тому однією зі

складових етапу фіксації вимог, поряд зі збором є верифікація вимог, а саме перевірка їх на несуперечливість і повноту.

Автоматизована верифікація вимог може здійснюватися лише після специфікації або формалізації вимог.

Специфікація вимог до ПЗ – процес формалізованого опису функціональних і не функціональних вимог, вимог до характеристик якості відповідно до стандарту якості ISO / IEC 12119-94, які будуть враховуватися при створенні програмного продукту на етапах ЖЦ ПЗ. У специфікації вимог відбивається структура ПЗ, вимоги до функцій, показниками якості, яких необхідно досягти, і до документації. У специфікації задається в загальних рисах архітектура системи і ПЗ, алгоритми, логіка управління і структура даних. Визначаються також системні вимоги, нефункціональні вимоги і вимоги до взаємодії з іншими компонентами (БД, СУБД, протоколи мережі та ін.).

Формалізація включає в себе визначення компонентів системи і їх станів; правил взаємодії компонентів і визначення умов в формальному вигляді, які повинні виконуватися при зміні станів компонентів. Для формального опису поведінки системи використовуються мови інженерних специфікацій, наприклад, UML. Як формальної моделі для опису вимог використовуються базові протоколи, які дозволяють використовувати дедуктивні засоби верифікації в поєднанні з моделюванням поведінки систем шляхом трасування.

Вважаючи, що всі вимоги чітко ідентифіковані і пронумеровані, можна сконструювати матрицю **залежностей вимог** (requirements dependency matrix) (або **матрицю взаємодії** (interaction matrix вимог)). У стовпці і рядку заголовка перераховані впорядковані ідентифікатори вимог, як показано в табл. 2.4.

Комірки нижче першої верхньої та першої лівої заповнюються вимогами (наприклад функціональними та не функціональними), далі зазначають, чи перекриваються дві будь-які вимоги, суперечать один одному або незалежні один від одного (порожні клітинки) - ставляться відповідні позначки (наприклад: +; -; X; V). Суперечливі вимоги необхідно обговорити з

замовниками і при можливості переформулювати для пом'якшення протиріч (фіксацію протиріччя, видиму для подальшої розробки, необхідно зберегти). Вимоги, що перекриваються, також повинні бути сформульовані заново, щоб виключити збіги.

Таблиця 2.4 – Матриця залежностей вимог

Вимога	**Т2.1**	**Т2.2**	**Т2.3**	**Т2.4**
Т1.1	Х		Х	Х
Т1.2	Конфлікт	Х		
Т1.4			Х	Х
Т1.4		Перекриття	Перекриття	

Матриця залежності вимог являє собою простий, але ефективний метод виявлення протиріч перекриттів, коли кількість вимог щодо невелика. В іншому випадку описаний метод все ж можна застосувати, якщо вдається згрупувати вимоги за категоріями, а потім порівняти їх окремо в межах кожної категорії.

Матриця відповідності вимог – це двовимірна таблиця, яка містить відповідність функціональних вимог (functional requirements) продукту і підготовлених тестових сценаріїв (test cases). У заголовках колонок таблиці розташовані вимоги, а в заголовках рядків – тестові сценарії. На перетині – відмітка, що означає, що вимога поточної колонки покрито тестовим сценарієм поточного рядка.

Матриця відповідності вимог використовується QA-інженерами для валідації покриття продукту тестами (валідація – відповідно до стандарту ДСТУ ISO 9000-2008 (відповідає ISO 9000: 2005), валідація визначена наступним чином: «Підтвердження на основі наданням об'єктивних доказів того, що вимоги, призначені для конкретного використання або застосування, виконані»). Матриця відповідності вимог є невід'ємною частиною тест-плану.

Матриця трасування (матриці трасуванню) (traceability matrix) – якщо розбити слово трасування на складові стане легше зрозуміти і запам’ятати, трасованість (з англ. Traceability, Trace – хвіст, ability – здатність). Простежуємо залежності (хвости) між вимогами і тестами.

Матриця трасування (матриці трасуванню) – спосіб візуалізації зв’язків між елементами системи в формі таблиці.

У процесах збору вимог і проектування програмно-технічних систем матриці трасування використовуються для швидкої оцінки зв’язків між артефактами проектування, такими як:

- вимоги і тести,
- замовник і релізи (спринти) *,
- вимоги і підсистеми,
- вимоги і функціональні специфікації,
- функціональні і нефункціональні вимоги,
- вимоги і моделі системи,
- варіанти використання (Use Cases) і підсистеми,
- помилки і тести,
- помилки і модулі системи.

Конкретний набір матриць трасування визначається складом проектних даних - типами використовуваних артефактів, які в свою чергу визначаються прийнятою в організації методологією збору вимог і проектування.

Для побудови матриці трасування використовується наступний підхід:

1. Вибираються елементи даної системи для рядків і стовпців.

2. При наявності зв’язку необхідного типу між елементом рядка і елементом стовпця у відповідній клітинці ставиться будь-який зручний символ.

Більш складні матриці трасування можуть відображати не тільки факт наявності зв’язку, але і її атрибути.

Розглянемо простий і наглядний приклад.

Цей приклад знайомий всім з часу навчання в школі, технікумі, університеті. В даному прикладі система – курс навчання. А цікавить нас матриця трасування – табель відвідуваності занять.

Стовпцями даної матриці є елементи системи – заняття, рядками – елементи системи – студенти. Якщо студент відвідував заняття, то ставиться відмітка про відвідування. Таким чином, ми відображаємо зв'язок студента і заняття.

Таблиця 2.5 – Ілюстрація матриці трасування – табель відвідуваності

	Заняття 1	Заняття 2	Заняття 3	Заняття 4	Заняття 5	Заняття 6
БОДРОВ В.	□					
ХАЛІЛОВ Т.	□					□
НІКІТІН С.	□					
МОСКВИН О.	□		□			
ВЛАСОВ П.						
АКІМОВ А.						
МАСЛОВ О.	□					□
ФІНАГІН О.	□	□	□		□	

З цієї матриці з практичної точки зору можна оцінити загальну ситуацію по групі. Наскільки група в цілому відвідує заняття. Якщо матриця розріджена, то, як правило, загальний обсяг знань малий, якість конспектів погана і ділитися один з одним їм буде нічим, а значить група погано здасть залік і іспит і, швидше за все, доведеться призначати додаткові заняття і буде багато роботи на дод. сесії. З точки зору дисципліни це означає збільшення ресурсів і тривалості проекту.

Якщо з 10-12 чоловік немає трьох, які відвідували всі заняття, це означає, що у групи немає повного і достовірного конспекту. Це також призведе до складнощів на іспиті. А значить викладачеві необхідно планувати або попереджувальні заходи, або потім боротися з наслідками збільшення навантаження в сесію.

Якщо матриця сильно розріджена, необхідно терміново вживати заходів і перш за все необхідно зрозуміти з яких причин група не ходить - не цікаво? Не встигають? Чи не здали попередні роботи? Немає мотивації? Змушені працювати і пропускати заняття?

Також можемо оцінити приватну ситуацію по кожному студенту і адекватно реагувати на дії студентів, розуміючи з яких позицій вони здійснюються.

Трасування забезпечує повноту тестування і готує основу для планування тестів. Матриця трасування може бути самостійним документом або може бути включена як частина документації за вимогами або частина плану тестування.

Таблиця 2.6 – Приклад матриці відповідності вимог

Requirements Traceability Matrix		Root Folder: Contract Processing	**Requirement**	Agree on	Check	Create contract	Determine	See customer off	Send contact	Sign contact	Determine net price	Inform customer	Send original	Contract processing	Check if	Develop proposal	Explain contact	Quotation	Sales order
			#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
		Total	Req	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Root Folder: Modeling			Covered																
Test	**#**	Test	Relate	4	4	4	4	4	4	4	2	2	2	1	0	0	0	0	0
Contact processing – path 2	1	1	X	10	X	X	X					X	X						
Contact processing – path 1	2	1	X	8				X	X	X	X								
Agree on	3	1	X	2	X														
Check	4	1	X	2		X													
Create contract	5	1	X	2			X												
Determine	6	1	X	2				X											
See customer off	7	1	X	2					X										
Send contact	8	1	X	2						X									
Send original	9	1	X	2									X						
Sign contract	10	1	X	2							X								
Contact processing	11	1	X	1										X					

Матриця відповідності вимог – це двовимірна таблиця, яка містить відповідність функціональних вимог (functional requirements) продукту і підготовлених тестових сценаріїв (test cases). У заголовках колонок таблиці (див. табл. 2.6) розташовані вимоги, а в заголовках рядків – тестові сценарії. На перетині – відмітка, що означає, що вимога поточної колонки покрито тестовим сценарієм поточного рядка. Матриця відповідності вимог використовується QA-інженерами для валідації покриття продукту тестами. Та є невід'ємною частиною тест-плану.

Прив'язка вимоги і тест-кейса може бути:

– 1 до 1 (атомарний вимога, яке покривається одним тест-кейсом, даний тест-кейс покриває тільки ця вимога);

– 1 до n (вимога, яке покривається декількома тест-кейсами, дані тест-кейси покривають тільки ця вимога);

– n до n (вимога, яке покривається декількома тест-кейсами, дані тест-кейси покривають це та інші вимоги).

Таблиця 2.7 – Приклад матриці відповідності вимог

Requirement Identifiers	**Reqs Tested**	**REQ1 UC 1.1**	**REQ1 UC 1.2**	**REQ1 UC 1.3**	**REQ1 UC 2.1**	**REQ1 UC 2.2**	**REQ1 UC 2.3.1**	**REQ1 UC 2.3.2**	**REQ1 UC 2.3.3**	**REQ1 UC 2.4**	**REQ1 UC 3.1**	**REQ1 UC 3.2**
Test Cases	321	3	2	3	1	1	1	1	1	1	2	3
Tested Implicitly	77											
1.1.1	1	✕										
1.1.2	2		✕	✕								
1.1.3	2	✕										
1.1.4	1			✕								
1.1.5	2	✕										
1.1.6	1		✕									
1.1.7	1			✕								
1.2.1	2				✕		✕					
1.2.2	2					✕		✕				
1.2.3	2								✕	✕		
1.3.1	1										✕	
1.3.2	1										✕	
etc…												
5.6.2	1											

З приводу останнього пункту хочеться відзначити, що

– коли одна вимога в матриці трасуванню покривається декількома тестами, це може говорити про надмірність тестування. В такому випадку треба проаналізувати, наскільки вимога атомарна.

– якщо виконанням всіх тест-кейсів ми забезпечуємо повноту покриття, а самі тест-кейси НЕ дублюють один одного - це не буде надмірним тестуванням.

Дисперсійний аналіз – метод математичної статистики, який застосовують для аналізу результатів спостережень, які залежать від різних одночасно діючих факторів. Задачі дисперсійного аналізу – вибір найбільш важливих факторів, оцінка їхнього впливу і подібне.

Метод найменших квадратів дозволяє одержати точкові оцінки коефіцієнтів прийнятої залежності Y = φ (X). Але тому, що коефіцієнти рівняння регресії – величини випадкові, вимагають перевірки й сама залежність і її коефіцієнти.

Модулі програми, які більше схильні до помилок, ніж інші, так само є хорошими кандидатами для регресійного тестування.

Регресія повинна включати в себе тестові випадки, які перевіряють функціональність з підвищеним ризиком і найбільш часті шляху виконання програми. Після проходження цієї стадії потрібно більш детально розглядати інші області. Нижче представлені деякі принципи, які зроблять набір тестів для регресії краще і більш оптимальним:

– включати в регресію тести, які ідентифікували помилки в попередніх версіях програмного забезпечення;

– використання кількості і типу помилок, виявлених за допомогою тестів, щоб постійно покращувати процес регресійного тестування;

– можна використовувати матрицю регресії, щоб позначати більш зачеплені помилками області, для подальшої передачі цієї інформації розробникам;

– тести для регресії повинні оновлюватися в кожній ітерації для збільшення охоплення в ширину і глибину;

– будь-які зміни у вимогах повинні супроводжуватися зміною набору тестів для регресії, щоб відповідати новому функціоналу додатка;

– можна і потрібно автоматизувати регресивні тести, які виконуються в стабільному тестовому середовищі;

– аналіз результатів тестування виявляє компоненти програми, більш схильні до помилок, щоб надалі сфокусувати там свої зусилля.

В [1] запропоновані наступні принципи для відстеження статусу регресійного тестування:

– число знову використаних тестів;

– число доданих Тест-кейсів в Репозиторій інструменту або базу даних тестів;

– число знову пройдених тестів після внесення змін до програми;

– заплановану кількість тестів для запуску;

– число запланованих і пройдених тестів.

Існує кілька видів дисперсійного аналізу. Необхідний варіант вибирається з урахуванням числа факторів і наявних вибірок з генеральної сукупності.

1. Однофакторний дисперсійний аналіз використовується для перевірки гіпотези про подібність середніх значень двох чи більше вибірок, що належать одній генеральної сукупності.

2. Двофакторний дисперсійний аналіз з повтореннями представляє собою більш складний варіант дисперсійного аналізу з декількома вибірками для кожної групи даних, його називають також дисперсійним аналізом при класифікації з групуванням.

3. Двофакторний дисперсійний аналіз без повторення представляє собою двофакторний аналіз дисперсії, що не включає більше однієї вибірки на групу, його називають також дисперсійним аналізом при класифікація з пересічними факторами.

Дисперсійний аналіз факторів. Звернемо увагу на одну важливу властивість коефіцієнта кореляції між змінними: зведений в квадрат він показує, яка частина дисперсії (розкиду) ознаки є спільною для двох змінних.

Або, кажучи простіше, наскільки сильно ці змінні перекриваються. Так наприклад, якщо дві змінні Т1 і Т3 з кореляцією 0,8 перекриваються зі ступенем 0,64 (0,8 в квадраті), то це означає, що 64% дисперсії тієї та іншої змінної є загальними, тобто збігаються. Можна також сказати, що спільність цих змінних дорівнює 64%.

Нагадаємо, що факторні навантаження в факторній матриці (табл. 2.3) є теж коефіцієнтами кореляції, але між факторами і змінними (споживчими вимогами). Тому зведене в квадрат факторне навантаження (дисперсія) характеризує ступінь спільності (або перекриття) даної змінної і даного чинника. Визначимо ступінь перекриття (дисперсію D) обох факторів зі змінною (споживчим вимогою) Т1. Для цього необхідно обчислити суму квадратів ваг факторів з першої змінної, тобто $0{,}83 \bullet 0{,}83 + 0{,}3 \bullet 0{,}3 = 0{,}70$. Таким чином спільність змінної Т1 з обома факторами становить 70%. Це досить вагоме перекриття.

У той же час, низька спільність може свідчити про те, що змінна вимірює або *відображає щось, що якісно відрізняється від інших змінних, включених в аналіз*. Це має на увазі, що дана змінна не поєднується з факторами по одній з причин: або змінна вимірює інше поняття (як, наприклад, змінна Т7), або змінна має велику помилку вимірювання, або існують ознаки, що спотворюють дисперсію.

Слід зазначити, що значимість кожного фактора також визначається величиною дисперсії між змінними і факторним навантаженням (вагою).

Для того щоб обчислити власне значення фактора, потрібно знайти в кожному стовпці факторної матриці (табл. 2.3) суму квадратів факторного навантаження для кожної змінної. Таким чином, наприклад, дисперсія фактора А (DA) складе

$$2{,}42 = 0{,}83 \bullet 0{,}83 + 0{,}3 \bullet 0{,}3 + 0{,}83 \bullet 0{,}83 + 0{,}4 \bullet 0{,}4 + 0{,}8 \bullet 0{,}8 + 0{,}35 \bullet 0{,}35.$$

Розрахунок значущості фактора Б показав, що $DБ = 2{,}64$, тобто значимість фактора Б вище, ніж фактора А.

Якщо власне значення фактора розділити на число змінних (у нашому прикладі їх 7), то отримана величина покаже, яку частку дисперсії (або обсяг інформації) γ вихідної кореляційної матриці складе цей фактор. Для фактора А $\gamma = 0{,}34$ (34%), а для фактора Б – $\gamma = 0{,}38$ (38%). Підсумувавши результати, отримаємо 72%. Таким чином, два фактора, будучи об'єднані, заповнюють тільки 72% дисперсії показників вихідної матриці. Це означає, що в результаті факторизації частина інформації у вихідній матриці була принесена в жертву побудови двофакторної моделі. В результаті – втрачено 28% інформації, яка могла б відновитися, якби була прийнята шестифакторна модель.

Де ж допущена помилка, враховуючи, що всі розглянуті змінні, що мають відношення до вимог по конструкції двері, враховані? Найбільш ймовірно, що значення коефіцієнтів кореляції змінних, що відносяться до одного фактору, дещо занижені. З урахуванням проведеного аналізу можна було б повернутися до формування інших значень коефіцієнтів кореляції в матриці інтеркореляцій (таблиця 2.2).

На практиці часто стикаються з ситуацією, що число незалежних факторів досить велике, щоб їх усіх врахувати в рішенні проблеми або з технічної або економічної точки зору. Існує ряд способів з обмеження числа факторів. Найбільш відомий з них – аналіз Парето.

При цьому відбираються ті чинники (у міру зменшення значущості), які потрапляють в (80-85)% межу їх сумарної значимості.

Факторний аналіз можна використовувати при реалізації методу структурування функції якості (QFD), що широко застосовується за кордоном при формуванні технічного завдання на новий виріб.

1. Перевірка адекватності рівняння регресії експериментальним даним виконується за критерієм Фішера

$$F = \frac{D_{ya}}{D_{yo}} \tag{2.1}$$

де D_{ya} – дисперсія адекватності. Вона визначається за формулою:

$$D_{ya} = \frac{1}{n - s} \sum_{i=1}^{n} (y_{ip} - m_{yi})^2 , \tag{2.2}$$

де n – число дослідів;

s – кількість шуканих параметрів апроксимуючої залежності;

y_{ip} – розрахункове значення функції в і-й точці при апроксимації залежністю $Y = \varphi(X)$;

m_{yi} – середнє значення у в і-м досліді;

D_{yo} – дисперсія досліду. Вона визначається на підставі даних паралельних дослідів:

$$D_{yo} = \frac{1}{m * n} \sum_{i=1}^{n} D_{yi} , \tag{2.3}$$

де m – число паралельних дослідів в і-й точці;

n – число дослідів;

m*n – загальне число вимірів;

D_{yi} – дисперсія і-го досліду, обумовлена за формулою

$$D_{yi} = \frac{\sum_{j=1}^{m} (y_{ij} - m_{yi})^2}{m - 1} \tag{2.4}$$

де m_{yi} – середнє значення У у і-м досліді.

Отримане значення F порівнюють із табличним $F_т$. Якщо $\boldsymbol{F < F_m}$, то гіпотеза про адекватність *не відкидається*.

На практиці дисперсійний аналіз застосовують у задачах, де потрібно оцінити вплив деякого фактору F на кількісну ознаку X. Суть дисперсійного аналізу зводиться до порівняння дисперсії, обумовленої впливом фактору F (факторної дисперсії) з дисперсією, обумовленою випадковими причинами

(залишковою дисперсією). Очевидно, коли вплив фактору F є значимим, то й відмінність факторної дисперсії від залишкової дисперсії повинна бути значимою. І навпаки, якщо вплив фактору незначимий, то факторна й залишкова дисперсія відрізняються незначимо.

Нехай значення ознаки X отримані в результаті спостереження p різних груп досліду із числом спостережень в j-й групі, рівним q. Середнє значення ознаки X у кожній j-й групі (групова середня) визначиться за формулою

$$\bar{x}_{грj} = \frac{\sum_{i=1}^{q} x_{ij}}{q} \tag{2.5}$$

Результати спостережень зведені в табл. 2.8.

Загальне число спостережуваних значень ознаки X дорівнює pq. Загальна середня визначається за формулою

$$\bar{x} = \frac{\sum_{j=1}^{p}\sum_{i=1}^{q} x_{ij}}{pq} \tag{2.6}$$

Таблиця 2.8 – Результати спостережень

Номер досліду	Номер групи					
	1	2	…	j	…	p
1	x_{11}	x_{12}	…	x_{1j}	…	x_{1p}
2	x_{21}	x_{22}	…	x_{2j}	…	x_{2p}
…	…	…	…	…	…	…
i	x_{i1}	x_{i2}	…	x_{ij}	…	x_{ip}
…	…	…	…		…	
q	x_{q1}	x_{q2}	…	x_{qj}	…	x_{qp}
групова середня	$\bar{x}_{гр1}$	$\bar{x}_{гр2}$	…	$\bar{x}_{грj}$	…	$\bar{x}_{грp}$

Можна оцінити повне розсіювання ознаки X, викликане як випадковими причинами, так і впливом фактору F, визначивши суму квадратів відхилень всіх

спостережуваних значень x_{ij} від загальної середньої. Вона називається **загальною** сумою квадратів відхилень і визначається за формулою

$$S_{общ} = \sum_{j=1}^{p} \sum_{i=1}^{q} (x_{ij} - \bar{x})^2 \tag{2.7}$$

Вважають, що фактор F впливає на різні групи значень ознаки. Розсіювання за фактором або розсіювання між групами можна оцінити, визначивши суму квадратів відхилень групових середніх від загальної середньої. Її називають **факторною** сумою квадратів відхилень і визначають за формулою

$$S_{факт} = q * \sum_{j=1}^{p} (\bar{x}_{грj} - \bar{x})^2 \tag{2.8}$$

Уважають, що на значення ознаки в j-й групі фактор F впливає однаково, а їхнє розсіювання обумовлене впливом випадкових причин. Суму квадратів відхилень спостережуваних значень ознаки X від своєї групової середньої $\bar{x}_{грj}$ називають **залишковою** сумою квадратів відхилень. Залишкова сума квадратів відхилень характеризує розсіювання всередині групи й визначається формулою:

$$S_{ост} = \sum_{j=1}^{p} \sum_{i=1}^{q} (x_{ij} - \bar{x}_{грj})^2 \tag{2.9}$$

Можна показати, що справедливе співвідношення

$$S_{общ} = S_{факт} + S_{ост}, \tag{2.10}$$

яке найчастіше використовують для визначення залишкової суми квадратів відхилень

$$S_{ост} = S_{общ} - S_{факт} . \tag{2.11}$$

Оскільки дисперсійний аналіз припускає порівняння дисперсій, то використовуючи загальні, факторну й залишкову суми квадратів відхилень, визначають відповідні дисперсії.

Загальна дисперсія:

$$s^2_{общ} = \frac{S_{общ}}{pq - 1} \tag{2.12}$$

де pq-1 = n-1– число ступенів свободи загальної дисперсії.

Факторна дисперсія:

$$s^2_{факт} = \frac{S_{факт}}{p-1} \tag{2.13}$$

де p-1– число ступенів свободи факторної дисперсії;

p– число груп впливу фактору F.

Залишкова дисперсія:

$$s^2_{ост} = \frac{S_{ост}}{p(q-1)} \tag{2.14}$$

де p(q-1) – число ступенів волі залишкової дисперсії, обумовлене як різниця між числами ступенів свободи загальної й факторної дисперсій:

$$(pq - 1) - (p - 1) = p(q-1).$$

Припустимо, що вплив фактора F відсутній. У цьому випадку групові середні $\bar{x}_{грj}$ приймають різні значення в результаті впливу тільки випадкових причин, а виходить, розрізняються незначимо. Відповідно факторна й залишкова дисперсії є незміщеними оцінками невідомої генеральної дисперсії й також розрізняються незначимо. У такій задачі формулюють нульову гіпотезу про рівність факторної й залишкової дисперсій. Якщо зрівняти оцінки цих дисперсій за критерієм F, то критерій укаже, що гіпотезу можна прийняти.

Якщо нульова гіпотеза про рівність групових середніх (а отже факторної й залишкової дисперсій) помилкова, то зі зростанням розбіжності між груповими середніми буде збільшуватися факторна дисперсія й спостережуване значення критерію F. При $F_{набл} > F_{кр}$ нульова гіпотеза про рівність факторної й залишкової дисперсій буде відкинута.

2.3 Методичні вказівки щодо організації самостійної роботи студентів

Перед лабораторною роботою слід повторити матеріал за курсом лекцій та рекомендованою літературою за темою лабораторної роботи [1, лекц. 2–5].

Особливо слід звернути увагу на такі питання.

– Факторний аналіз.
– Дисперсійний аналіз.
– Дисперсійний аналіз факторів.
– Коефіцієнти кореляції Пірсона.
– Матриці інтеркореляцій.
– Аналіз подій і пошук закономірностей.
– Метод асоціативних правил.
– Матриці трасування та їх використання.

2.4 Завдання до лабораторної роботи

1. Вивчити теоретичний матеріал.
2. Вибрати тип програмного забезпечення, в рамках якого буде виконуватися розрахунок прогнозних показників якості ПЗ (з тих що розроблялося раніше починаючі з 2 по 4 курс). У проекту повинно бути ТЗ та Специфікація.
3. Присвисти опис ПЗ в рамках якого буде виконуватись розрахунки. ТЗ та Специфікацію включити до звіту по лабораторній роботі, як додатки 1 та 2.
4. Побудувати матриці залежностей вимог (вимоги і функціональні специфікації; функціональні і нефункціональні вимоги).
5. З відкритих джерел взяти необхідні дані в необхідному розмірі за відповідною темою (наприклад з https://www.kaggle.com/fernandol/countries-of-the-world). Вказати звідки були взяті дані.
6. Сформулювати ціль дослідження (приклад дослідження – необхідно визначити, від чого залежить рішення клієнта щодо покупки комп’ютера; залежність виконання покупки від рівня доходу, сімейного стану, типу роботи, регіону проживання).

7. Визначити які саме дані/фактори є головними відносно дослідження (з п.6.). Побудувати нову таблицю з п.5 вибравши стовпчики для аналізу, прибравши фактори, які не впливають на залежну змінну.

8. Розрахувати коефіцієнти лінійних рівнянь регресії, що зв’язують вихідні параметри з вхідними факторами. Провести регресійний аналізи для встановлення наявності зв’язку між вхідними і вихідними змінними досліджуваного об’єкта (процесу).

9. Виконати факторний аналіз аналіз для ПЗ з п.2 (скласти таблицю попарних кореляцій, виконати розрахунки і зробити висновки на підставі отриманих значень). Опишіть результати та запропонуйте їх інтерпретацію.

10. Сформулювати висновки і пояснити чому вийшов такий результат.

11. Оформити звіт що демонструє роботу і здати викладачеві в електронному вигляді.

12. Відповісти на контрольні питання (парний варіант дає відповіді на парні питання, непарний варіант дає відповіді непарні питання).

2.5 Опис програмного забезпечення

Під час виконання лабораторної роботи використовується таке програмне забезпечення: будь-який текстовий редактор Microsoft Office Word або OpenOffice; браузери Chrome, Edge, Firefox, Opera. Також використовується безкоштовна версія обраного ПЗ для підрахунку статистичних даних (ПЗ студент обирає самостійно відповідно до особистих вподобань та апаратних можливостей комп’ютера).

2.6 Зміст звіту

1. Тема і мета роботи.
2. Послідовність виконуваних у процесі роботи дій.

3. Опис процесу роботи з скріншотами роботи.

4. Висновки з роботи.

5. Відповіді на контрольні питання (парний варіант дає відповіді на парні питання, непарний варіант дає відповіді непарні питання).

2.7 Контрольні запитання та завдання

1. Емпірична і теоретична функції розподілу випадкових величин.
2. Що таке «Матриця інтеркореляцій», та для чого вона використовується?
3. Що таке змістовний аналіз вимог, та для чого він використовується?
4. Перевірка однорідності дисперсій двох вибірок.
5. Перевірка однорідності дисперсій декількох вибірок.
6. Перевірка однорідності середніх двох вибірок.
7. Перевірка приналежності двох вибірок однієї сукупності (при невідомому законі розподілу).
8. Задачі й основні положення регресійного аналізу.
9. Загальна схема регресійного аналізу, рівняння регресії.
10. Лінійна і поліноміальна регресія.
11. Оцінка коефіцієнтів рівняння регресії методом найменших квадратів.
12. Множинна лінійна регресія.
13. Аналіз рівняння регресії.
14. Назвіть вимоги, яким повинні задовольняти фактори, що підлягають включенню в рівняння регресії.
15. Як проводиться статистична оцінка значимості коефіцієнтів рівняння регресії?
16. Назвіть можливі причини незначимості коефіцієнтів рівняння регресії.
17. В чому полягає основна ідея методу дисперсійного аналізу?

18. Як формуються оцінки дисперсії: загальна, між серіями і залишкова (по середині серії)?

19. Яким чином здійснюється кількісне оцінювання впливу факторів?

20. Якого типу практичні задачі, зазвичай, вирішуються методом дисперсійного аналізу?

21. Як математично формулюється задача однофакторного аналізу?

22. Які основні передумови застосування дисперсійного аналізу?

2.8 Додаткова література та електронні ресурси

1. I. Burnstein, T. Suwanassart, R. Carlson. Developing a Testing Maturity Model for Software Test Process Evaluation and Improvement.

2. . Лежнюк П.Д., Рубаненко О.Є., Лук'яненко Ю. В. Основи теорії планування експерименту. Лабораторний практикум. - Вінниця: ВНТУ, 2006. - 167 с.

3. Системи підтримки прийняття рішень [Текст] : навчальний посіб- ник для самостійного вивчення дисципліни / [уклад.: С. М. Братушка, С. М. Новак, С. О. Хайлук] ; Державний вищий навчальний заклад "Українська академія банківської справи Національного банку України". – Суми : ДВНЗ "УАБС НБУ", 2010. – 265 с.

3 ЗАВДАННЯ ПРОВЕДЕННЯ КВАНТОВИХ ОБЧИСЛЕНЬ. ВИЯВЛЕННЯ ШАБЛОНІВ ДЛЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

3.1 Мета та завдання роботи

Метою лабораторної роботи є отримання навичок з основ квантової теорії інформації, квантових обчислень і зв'зку. Розуміти опис стану та динаміки квантово-механічних систем; визначати стан системи після вимірювання; пояснити принципи надщільного кодування та квантової телепортації. Пошук, спрощення структури, обробка даних та пошук причинно-наслідкових зв'язків в аналізуємих даних для підбору моделей.

3.2 Опис роботи

В 1994 році П. Шор приголомшив науковий світ, запропонувавши квантовий алгоритм, що дозволяє проводити швидку факторизацію великих чисел. У порівнянні з кращим з відомих на сьогодні класичних методів квантовий алгоритм Шора дає багаторазове прискорення обчислень.

У 1996 році колега Шора по роботі в Lucent Technologies Л. Гровер запропонував квантовий алгоритм швидкого пошуку в неупорядкованій базі даних. (Приклад такої бази даних – телефонна книга, у якій прізвища абонентів розташовані не за алфавітом, а довільним образом.) Задача пошуку, вибору оптимального елемента серед численних варіантів дуже часто зустрічається в економічних, військових, інженерних задачах, у комп'ютерних іграх. Алгоритм Гровера дозволяє не тільки прискорити процес пошуку, але і збільшити приблизно в два рази число параметрів, що враховуються при виборі оптимуму.

Основним елементом квантового комп'ютера являється регістр із L кубітів. Перед початком обчислень усі кубіти переводяться в деякий

початковий стан, наприклад, "0". Потім кожен кубіт індивідуально переводиться у змішаний стан, що відповідає умові розв'язуваної задачі. Після цього над регістром, як над єдиним цілим, проводяться послідовні операції. Результат обчислення зчитується наприкінці роботи. Таким чином, квантовий комп'ютер має три основні етапи роботи: ініціалізацію, виконання операцій над кубітами та зчитування результату обчислень.

Слід пам'ятати, що з одного боку, зменшення базового елементу в обчислювальному процесі призводить до того, що ми вже не можемо нехтувати квантовими ефектами. З іншого боку, для багатьох корисних завдань час їх рішення зростає експоненціально від розміру задачі.

Квантовий комп'ютер використовує звичну обчислювальним машинам двійкову систему числення, «всередині» у нього тільки нулі і одиниці. Однак термін «кубіт» (q-bit, «біт» квантового комп'ютера) позначає принципова відмінність від біта: про стан кубіта в кожен момент часу можна сказати, що у нього всередині - нуль або одиниця. Щоб з'ясувати це, треба «зняти» дані - відкрити коробку з котом Шредінгера і зрозуміти, чи живий кубіт («1») або мертвий («0»).

Квантові обчислення забезпечуються можливістю зафіксувати взаємозв'язок сукупності (регістра) кубітів, що знаходяться в суперпозиції. Кубіти можна ввести в так зване заплутане (загальне, єдине) стан, коли вимір одного кубіта фіксує не тільки його стан, а й стан всіх N-кубітів в регістрі. Якщо N-кубіти в регістрі заплутані, тоді однією операцією квантовий комп'ютер може відразу, одночасно, обробити 2N біт даних.

Це дає, по-перше, грандіозне зростання розмірності оброблюваних даних: при N = 50 регістр заплутаних кубітів еквівалентний за обсягом даних, що зберігаються 10 в 18-го ступеня біт. По-друге, дозволяє вирішувати згадані вище завдання, недосяжні для класичних комп'ютерів.

На сьогоднішній день, є безкоштовний сервіс IBM Q Experience, який швидко зібрав понад 150 тис. активних користувачів по всьому світу.

Відповідно до результатів його використання видно, що квантові комп'ютери відкривають безкраї можливості для пошуку і застосування креативних рішень. Людство скоро зможе по-новому поглянути на проблеми, які раніше здавалися нам неприступними.

Імовірно квантові обчислення і алгоритми дозволять вирішити складні питання, а саме [7]:

– пошук в масивах неструктурованих даних (радикальне прискорення обробки великих даних);

– розкладання чисел на прості множники (алгоритм Шора, важливий для подолання криптографічного захисту даних – квантовий комп'ютер за секунди здатний зробити те, на що у суперкомп'ютера підуть мільярди років);

– швидке генерування послідовності справді випадкових чисел (практичне застосування – одноразові ключі для гарантовано захищеної передачі даних по відкритому каналу зв'язку);

– моделювання квантових систем – молекул і матеріалів (практичне застосування - фармакологія, засоби захисту від біологічної зброї), причому для вирішення таких завдань достатній «малопотужний» квантовий комп'ютер з регістром до 100 кубіт.

Фізична реалізація квантових комп'ютерів знаходиться в стадії досліджень і експериментів, а розвиток алгоритмів квантових обчислень забезпечується імітацією квантових комп'ютерів за допомогою пристроїв, позбавлених квантової природи.

В останні роки аналітична обробка великих даних привертає все більшу увагу як в світі, так і в Україні, найбільш наближеним практичним рішенням для обробки та аналізу великих даних є «Інтелектуальний аналіз даних».

Інтелектуальний аналіз даних (ІАД, Data Mining), або розвідка даних - термін, що застосовується для опису здобуття знань у базах даних, дослідження даних, обробки зразків даних, очищення і збору даних. Це процес виявлення кореляції, тенденцій, шаблонів, зв'язків і категорій.

При розвідці даних багаторазово виконуються операції і перетворення над "сирими" даними (відбір ознак, стратифікація, кластеризація, візуалізація і регресія), що призначені для знаходження:

– структур, які інтуїтивно зрозумілі для людей і краще розкривають суть бізнес-процесів, що лежать в основі їх протікання;

– моделей, які можуть передбачити результат або значення певних ситуацій, використовуючи історичні або суб’єктивні дані.

Інтелектуальний аналіз даних – процес автоматичного пошуку прихованих закономірностей або взаємозв’язків між змінними у великих масивах необроблених даних, що поділяється на задачі класифікації, моделювання і прогнозування. Класичне визначення цього терміна дав у 1996 р. один із засновників цього напряму Г. П’ятецький-Шапіро.

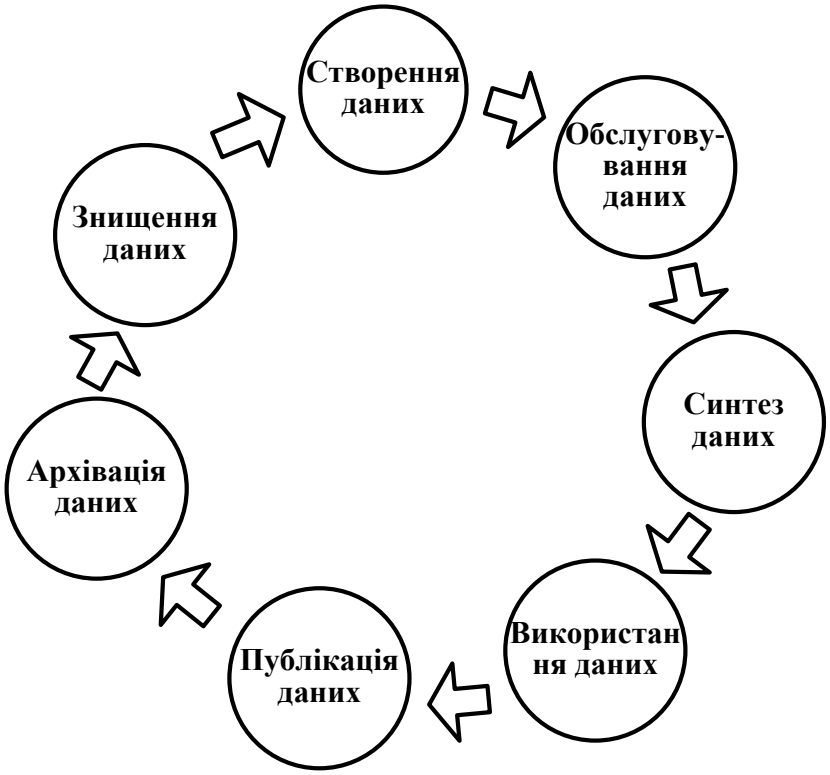


Рисунок 3.1 – Життєвий цикл даних

Метою інтелектуального аналізу даних є побудова аналітичних моделей, оптимізованих для вирішення конкретних класів прикладних задач. Аналітичні моделі будуються для навчальних екземплярів даних з певної проблемної

області. Як тільки модель побудована, вона перевіряється на тестових екземплярах даних. Після підтвердження адекватності та точності побудованої аналітичної моделі, її можна застосовувати до нових некласифікованих даних.

У відповідності зі стилем пошуку закономірностей в навчальних даних методи інтелектуального аналізу даних ділять на наступні категорії:

− навчання з вчителем: модель будується по заздалегідь класифікованим навчальним даними і описує закономірності між значеннями атрибутів ознак і значенням атрибуту класу;

− навчання без вчителя: в навчальних даних відсутній атрибут класу, закономірності шукаються між значеннями усіх атрибутів, які вважаються атрибутами ознак;

− часткове навчання з вчителем: кількість класифікованих навчальних даних набагато менша ніж кількість некласифікованих даних, тому спочатку для класифікованих даних вирішується задача навчання з вчителем і будується аналітична модель, після чого вирішується задача навчання без вчителя з підкріпленням за побудованою аналітичною моделлю;

− навчання з підкріпленням: при появі нових навчальних даних, попередня модель не будується заново, а успадковує знайдені раніше закономірності, котрі показали коректні результати класифікації на тестових наборах даних.

На етапі постановки завдання потрібно визначити, що є метою аналізу. Зокрема, потрібно відповісти на ряд питань, головне з яких - що саме необхідно визначити в результаті аналізу. Також в цьому списку:

Чи потрібно буде робити прогнози на підставі моделі інтелектуального аналізу даних або просто знайти змістовні закономірності і взаємозв’язку?

Якщо потрібно прогноз, який атрибут набору даних необхідно спрогнозувати? Як пов’язані стовпці? Якщо існує кілька таблиць, як вони пов’язані? Яким чином розподіляються дані? Чи є дані сезонними? Чи дають

дані точне уявлення про предметну область? Як правило, в процесі постановки завдання аналітик працює спільно з фахівцями в предметної області.

Етап підготовки даних включає визначення джерел даних для аналізу, об'єднання даних і їх очищення. Дані, що використовуються можуть перебувати в різних базах і на різних серверах. Більш того, якісь то дані можуть бути представлені у вигляді текстових файлів, електронних таблиць, перебувати в інших форматах.

У процесі об'єднання і перетворення даних часто використовуються можливості різних BI систем. Це дозволяє істотно автоматизувати процес підготовки. Зібрані таким чином дані, як правило, потребують додаткової обробки, званої очищенням. В процесі очищення при необхідності може проводитися видалення «викидів» (нехарактерних і помилкових значень), обробка відсутніх значень параметрів, чисельне перетворення (наприклад, нормалізація) і т.д.

Наступним етапом є вивчення даних, що дозволить зрозуміти, наскільки адекватно підготовлений набір являє досліджувану предметну область. Тут може проводитися пошук мінімальних і максимальних значень параметрів, аналіз розподілів значень і інших статистичних характеристик, порівняння отриманих результатів з уявленнями про предметну область.

Четвертий етап – побудова моделей. Спочатку створюється структура даних, а потім для структури створюється одна або декілька моделей. Модель включає вказівку на алгоритм інтелектуального аналізу даних і його параметри, а також аналізовані дані. При визначенні моделі можуть використовуватися різні фільтри. Таким чином, не всі наявні в описі структури дані будуть використовуватися кожної створеної для неї моделлю.

Модель може проходити навчання, що полягає в застосуванні обраного алгоритму до навчального набору даних. Після цього в ній зберігаються виявлені закономірності.

П’ятий етап – перевірка моделі. Тут метою є оцінка якості роботи створеної моделі перед початком її використання в «виробничому середовищі». Якщо створювалося кілька моделей, то на цьому етапі робиться вибір на користь тієї, що дасть найкращий результат. При вирішенні прогнозних (Предсказательни) завдань інтелектуального аналізу якість видається моделлю прогнозу можна оцінити на перевірочному наборі даних, для якого відомо значення прогнозованого параметра.

Слід пам’ятати, що згодом характеристики предметної області можуть змінюватися, що потребують і зміни шаблонів інтелектуального аналізу даних. Може знадобитися перенавчання існуючих моделей або створення нових.

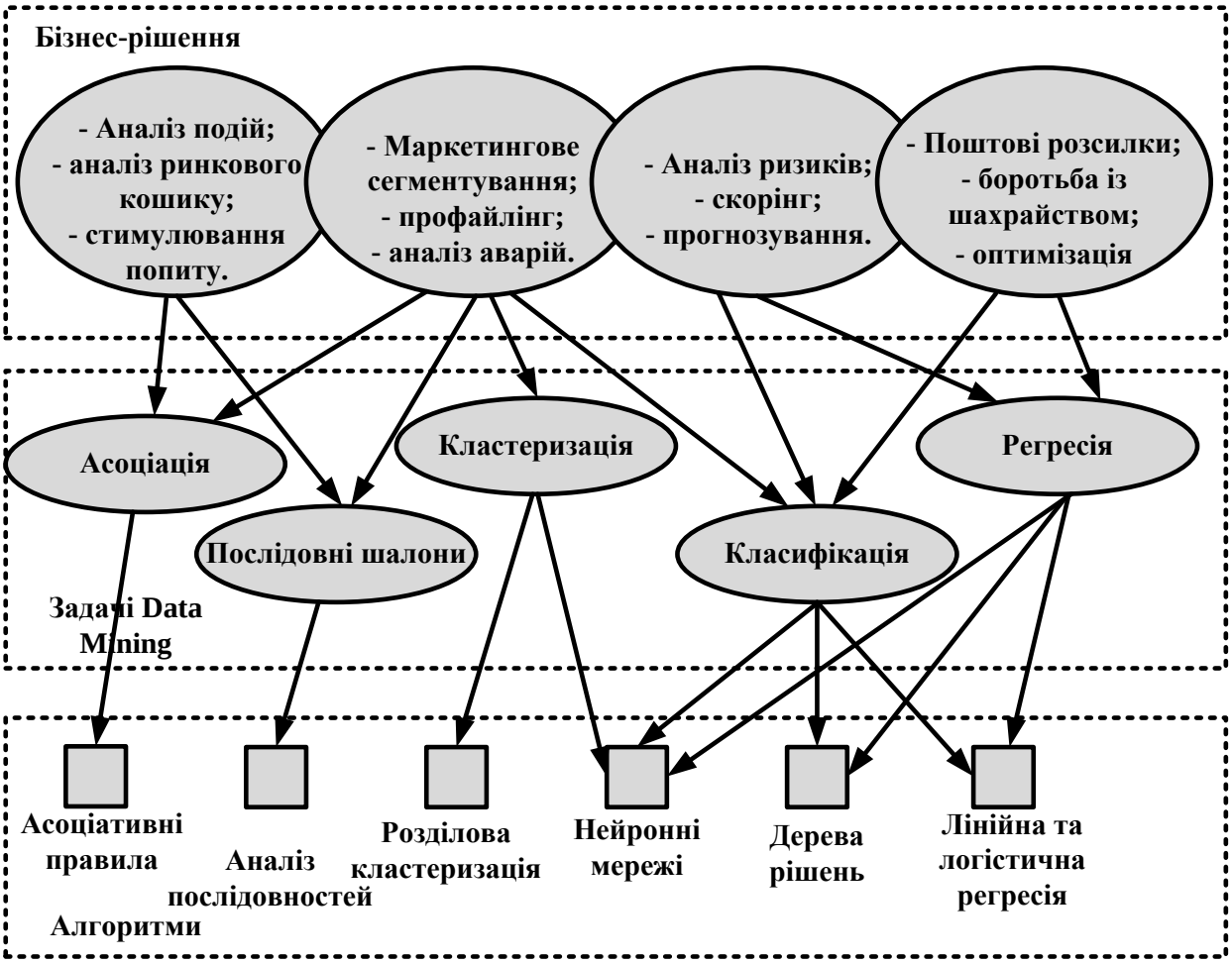


Рисунок 3.2 – Методи Data Mining в системах BI

За визначенням SAS Institute, Data Mining – це процес виділення, дослідження і моделювання великих обсягів даних для виявлення невідомих до цього структур (patterns) з метою досягнення переваг у бізнесі.

За визначенням Gartner Group, Data Mining – це процес, мета якого - виявляти нові кореляції, зразки і тенденції у результаті просіювання великого обсягу даних з використанням методик розпізнавання зразків і статистичних та математичних методів.

В основу технології Data Mining покладено концепцію шаблонів (patterns), що є закономірностями, які властиві вибіркам даних і можуть бути подані у формі, зрозумілій людині.

Організація процесу інтелектуального аналізу даних визначається стандартами, які містять покрокові рекомендації, задачі та цілі для всіх етапів процесу аналізу даних. Консорціумом компаній NCR, SPSS та DaimlerChrysler розроблено поширений в світі стандарт ***The Cross Industry Standard Process for Data Mining (CRISP-DM)***, який є продовженням підходів Knowledge Discovery in Databases (KDD) та SEMMA (Sample – збір даних, Explore – дослідження зв'язків, Modify – модифікування, Model – моделювання, Assess – оцінювання).

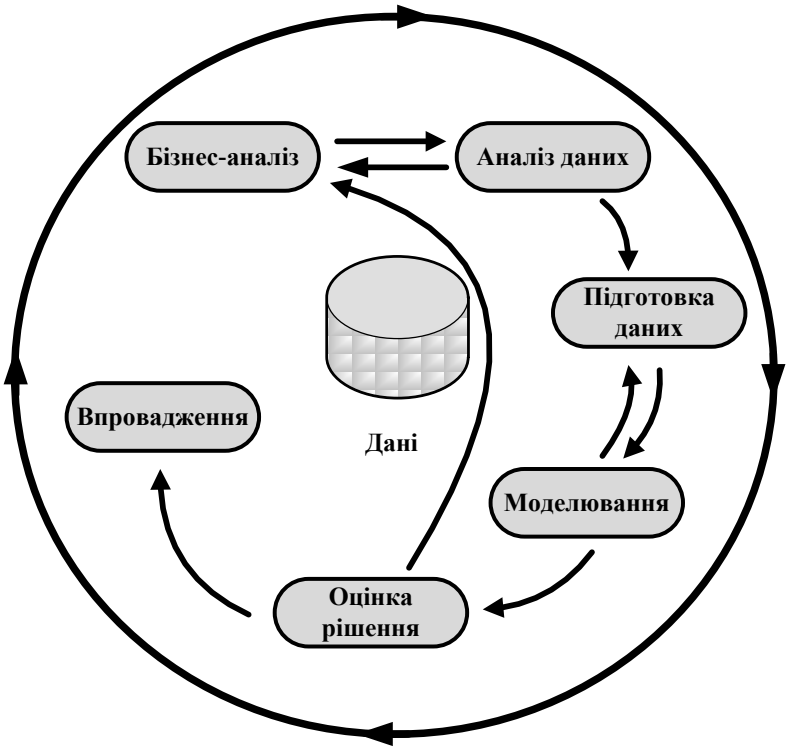


Рисунок 3.4 – Структурна модель CRISP-DM

На рисунку 3.4 показано життєвий цикл проекту інтелектуального аналізу даних у вигляді структурної моделі CRISP-DM. На початковому етапі

"Розуміння бізнес-контексту" виконується дослідження бізнес-цілей і вимог, приймається рішення про те, чи може застосування інтелектуального аналізу задовольнити бізнес-цілі, і визначається те, які дані необхідно зібрати для побудови адекватної аналітичної моделі. На наступному етапі «Підготовки даних» створюється і вивчається початковий набір даних, щоб визначити, чи придатний він для подальшої обробки.

Якщо кількість даних невелика, може знадобитися збір нових даних на базі більш строгих критеріїв. Аналіз даних, отриманих на даному етапі, може також призвести до перегляду бізнес-контексту – можливо, знадобиться перегляд мети застосування інтелектуального аналізу даних.

Кожен з цих етапів в свою чергу ділиться на завдання. На виході кожного завдання повинен виходити певний результат. Завдання зображені на рис. 3.5.

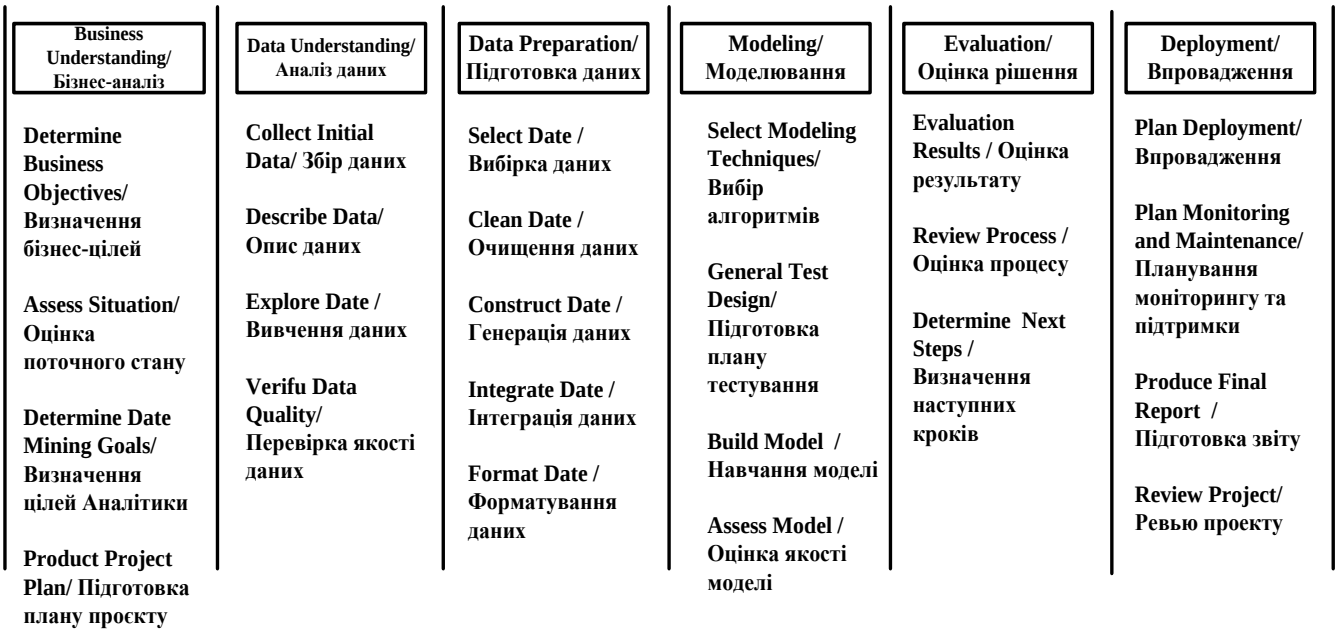


Рисунок 3.5 – Структурна модель CRISP-DM

Дані, в яких координати вимірюються в різних одиницях виміру, числа іноді записані словами, іноді латинськими цифрами, а іноді у вигляді відсканованого зображення почерку лаборанта, є неструктурованими даними.

Зазвичай Великі дані описуються за допомогою наступних характеристик [2].

1. Обсяг (Volume) – кількість згенерованих і збережених даних. Розмір даних визначає значимість і потенціал даних, а також те, чи можуть вони бути розглянуті як Великі дані.

2. Різноманітність (Variety) – тип даних. Великі дані можуть складатися з тексту, зображень, аудіо, відео. Великі дані при зіставленні один з одним можуть доповнювати відсутні дані.

3. Швидкість (Velocity) – швидкість. Тут мається на увазі швидкість, з якою дані генеруються і обробляються. Дуже часто Великі дані використовуються в режимі реального часу.

4. Мінливість (Variability) – суперечливість наборів даних може перешкоджати їх обробці й керуванню ними.

5. Достовірність (Veracity) – якість даних безпосередньо впливає на точність проведення аналізу даних.

Архітектура системи обробки Великих даних. Для роботи з Великими даними використовуються складні системи, в яких можна виділити кілька компонентів або шарів (Layers). Зазвичай виділяють чотири рівні компонентів таких систем: прийом, збір, аналіз даних і представлення результатів (рис. 3.6) [3].

Цей поділ є значною мірою умовною так як, з одного боку, кожен компонент в свою чергу може бути розділений на подкомпоненти, а з іншого деякі функції компонентів можуть перерозподілятися в залежності від розв’язуваної задачі і використовуваного програмного забезпечення, наприклад, виділяють зберігання даних в окремий шар.

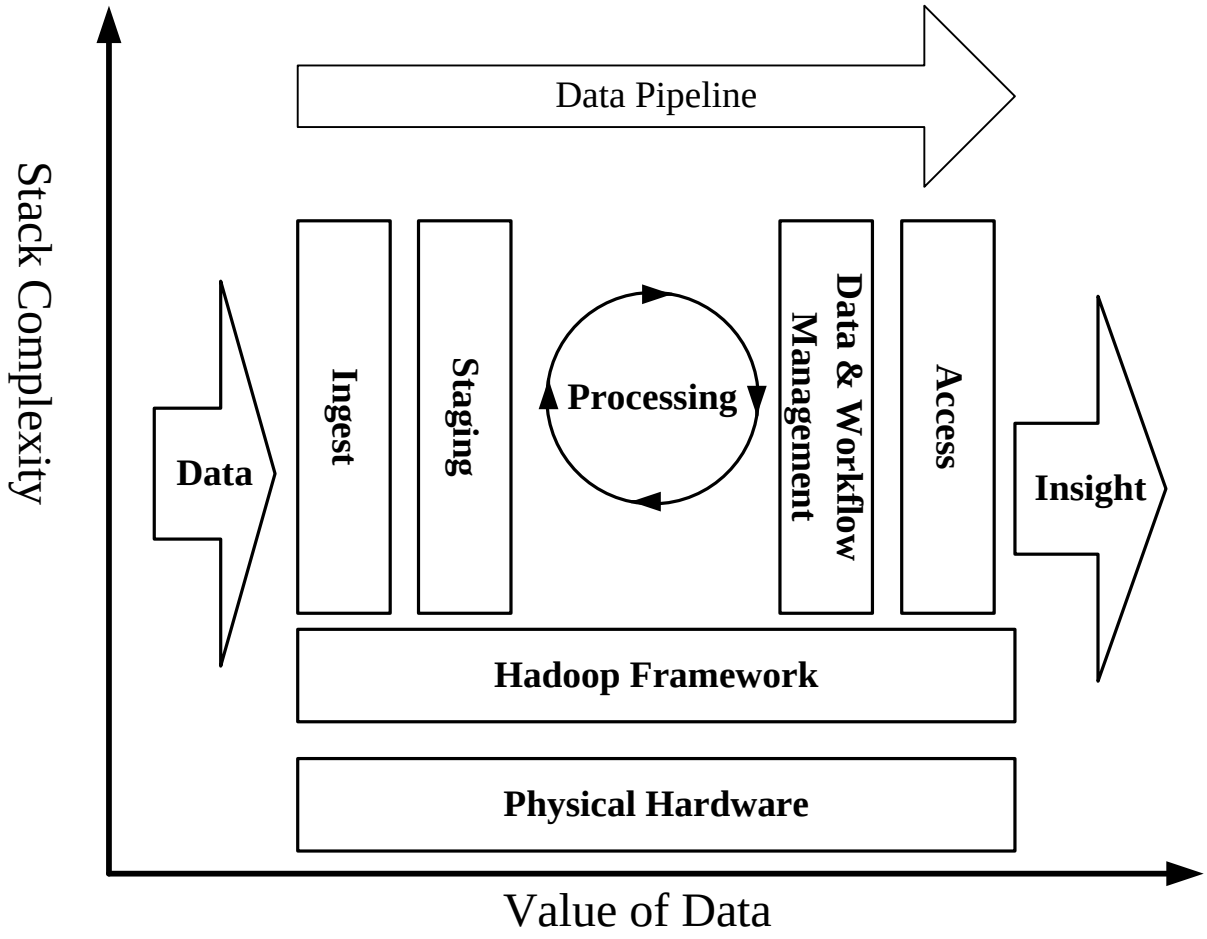


Рисунок 3.6 – Стек роботи з Великими даними.

Для роботи з Великими даними розробниками систем створюються моделі даних, змістовно пов’язані з реальним світом. Розробка адекватних моделей даних є складною аналітичну задачу, виконувану системними архітекторами і аналітиками. Модель даних дозволяє створити математичну модель взаємодій об’єктів реального світу і включає в себе опис структури даних, методи маніпуляції даними і аспекти збереження цілісності даних.

3.3 Методичні вказівки щодо організації самостійної роботи студентів

Перед лабораторною роботою слід повторити матеріал за курсом лекцій та за рекомендованою літературою за темою лабораторної роботи [1, лекц. 4–6].

Особливо слід звернути увагу на такі питання:

– Особливості технологій глибинного аналізу даних (Data Mining) у додатках BI

– Інтелектуальний аналіз даних та вилучення знань з даними.

– Основи технологій моніторингу, реєстрації та обробки великих даних.

– Визначення концепції Big Data при управлінні проектом.

– Концепція Web Mining.

3.4 Завдання до лабораторної роботи

1. Вивчити теоретичну частину заняття.

2. Вибрати тип програмного забезпечення, в рамках якого буде виконуватися розрахунок (з тих що розроблялося раніше починаючі з 2 по 4 курс, або з обраних на ПЗ №1-2).

3. Привести опис ПЗ в рамках якого буде виконуватись розрахунки.

4. Вивчити в інтернеті стан питання «проведення квантових обчислень», які є рішення, алгоритми. Описати у вигляді словесного алгоритму, або у вигляді блок схемі один з алгоритмів аналізу даних – Гровера, Шора. Відповідно до наданого опису спробувати вирішити задачу в рамках обраного ПЗ, пояснення відобразити в звіті.

5. З відкритих джерел взяти необхідні дані в необхідному розмірі за відповідною темою (наприклад з https://www.kaggle.com/fernandol/countries-of-the-world). Вказати звідки були взяті дані.

6. Табличка, які саме дані використовувалися для проведення експериментів (джерела даних посиланням, кількість і характеристики, приклади).

7. Обрати програмне забезпечення для проведення інтелектуального *аналізу великих даних, надати обґрунтування, чому саме воно було обрано*.

8. Опис контрольного прикладу: що використовували, скільки об’єктів на вході, які ознаки вибрали, скільки і які кластери отримали. Скріншоти, що

показують кінцеві результати. Доводи або порівняльне дослідження за деяким критерієм між вашим рішенням і обраним аналогом.

9. Побудувати одну (мінімум) модель інтелектуального аналізу даних на основі обраного алгоритму машинного навчання.

10. Отримати деякий передбачення (прогнозування), пояснити.

11. Сформулювати висновки і пояснити чому вийшов такий результат.

12. Оформити звіт що демонструє роботу і здати викладачеві в електронному вигляді.

13. Відповісти на контрольні питання (парний варіант дає відповіді на парні питання, непарний варіант дає відповіді непарні питання).

3.5 Опис програмного забезпечення

Під час виконання лабораторної роботи використовується таке програмне забезпечення: будь-який текстовий редактор Microsoft Office Word або OpenOffice; браузери Chrome, Edge, Firefox, Opera. Безкоштовна версія/тріал версія системи інтелектуального аналізу даних (Elasticsearch/ Fluentd/Kibana/Deductor/ Power BI Desktop або інша за вибором студенту).

3.6 Зміст звіту

1. Тема і мета роботи.
2. Послідовність виконуваних у процесі роботи дій.
3. Опис процесу роботи з скріншотами роботи.
4. Висновки з роботи.
5. Відповіді на контрольні питання (парний варіант дає відповіді на парні питання, непарний варіант дає відповіді непарні питання).

3.7 Контрольні запитання та завдання

1. Які є принципи побудови квантового комп’ютера?
2. Універсальні набори квантових вентилів.
3. Квантові вимірювання.
4. Квантовий паралелізм.
5. Алгоритм Дойча.
6. Алгоритм Гровера.
7. Квантове перетворення Фур’є.
8. Квантовий алгоритм знаходження періоду функції.
9. Життєвий цикл проекту інтелектуального аналізу даних.
10. За допомогою яких характеристик описуються великі дані?
11. Дайте визначення інтелектуального аналізу даних.
12. Навіщо використовувати BI-системи для аналізу логів?
13. Що таке генеральна сукупність і вибірка?
14. Які властивості повинна мати вибірка?
15. Навіщо потрібно виконувати очищення даних перед їх аналізом.
16. Від чого залежить успішність процесу аналіз даних?
17. Які фази життєвого циклу дослідження даних по CRISP-DM?
18. Які п’ять характеристик притаманні Великим даними?
19. Які існують базові принципи обробки Великих даних?
20. В чому різниця між структурованими та неструктурованими даними?

3.8 Додаткова література та електронні ресурси

1. CRISP-DM: перевірена методологія для Data Scientist-ів [Електронний ресурс]. https://habr.com/en/companies/lanit/articles/328858/ (дата звернення: 1.09.2023).

2. Data Analytics: An Essential Beginner's Guide To Data Mining, Data Collection, Big Data Analytics For Business, And Business Intelligence Concepts Paperback – February 4, 2018 [Електронний ресурс]. https://www.amazon.com/Data-Analytics-Essential-CollectionIntelligence/dp/ 1985097974/ (дата звернення: 1.09.2023).

3. 10 Best Big Data Analytics Tools for 2023 – With Uses & Limitations [Електронний ресурс]. https://data-flair.training/blogs/best-big-data-analytics-tools/ (дата звернення: 1.09.2023).

4. Practical Statistics for Data Scientists Andrew Bruce and Peter Bruce., Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, 2020. — 363 с. [Електронний ресурс]. https://www.researchgate.net/profile/Janine-Zitianellis/post/Can_anyone_please_suggest_a_books_on_machine_learning_using_R_Programming/attachment/613a5b83647f3906fc975a71/AS%3A1066204907204608%401631214467436/download/Practical+Statistics+for+Data+Scientists+50%2B+Essential+Concepts+Using+R+and+Python+by+Peter+Bruce%2C+Andrew+Bruce%2C+Peter+Gedeck.pdf (дата звернення: 1.09.2023).

5. Т. Крохмальский Вступ до квантових обчислень: Навчальний посібник. - Львів : ЛНУ імені Івана Франка, 2018. - 204 с. [Електронний ресурс]. http://ktf.lnu.edu.ua/books/Krokhmalskii-VKO.pdf (дата звернення: 1.09.2023).

6. Song_Y._Yan Quantum Computational Number Theory PublisherSpringer ISBN-103319798464 LanguageEnglish Publication Year2018. 252 p.

7. Про квантові комп'ютери простими словами [Електронний ресурс] https://root-nation.com/ua/articles-ua/tech-ua/ua-pro-kvantovi-kompyuteri/ . (дата звернення: 1.09.2023).

8. Енциклопедія з квантової інформатики. [Електронний ресурс] http://www.quantiki.org/ (дата звернення: 1.09.2023).

9. Лекції BigData from Zinoviev Alexey [Электронный ресурс]. https://www.youtube.com/playlist?list=PL972if8tX2vpdaa4OcZ76sZ1fBYqJbPcB (дата звернення: 1.09.2021).

4 МЕТОДИ ПРОГНОЗУВАННЯ С УРАХУВАННЯМ СУЧАСНИХ ОБЧИСЛЮВАЛЬНИХ АПАРАТІВ

4.1 Мета та завдання роботи

Метою лабораторної роботи є ознайомлення з емпіричними методами прогнозування. Розуміти розміри явища та вміти знаходити невідомі проміжних рівнів ряду динаміки; вміти продовжувати кількісні характеристики сукупностей за межі досліджуваного явища в майбутнє на базі встановлених закономірностей за попередній термін.

4.2 Опис роботи

Виявлені закономірності серед різних даних є дуже важливим в різних дослідження, на їх основі можна розробити описові моделі прийняття рішень, які допоможуть зрозуміти та зпрогнозувати різні ситуації (поведінка людини, як зміняться популярність товару і т.д.) вибору.

Характер внутрішніх зв’язків в об’єктах дослідження відображається у динаміці структурних змін. Фіксація стану процесу визначається інтегрованим динамічним рядом, якій досліджується на основі аналітичних показників, моделювання рядів динаміки, а прогнозування розвитку подій здійснюється за допомогою екстраполяції.

Вивчення рядів динаміки різних суспільних явищ дає базу для прогнозування і для знаходження невідомих рівнів ряду.

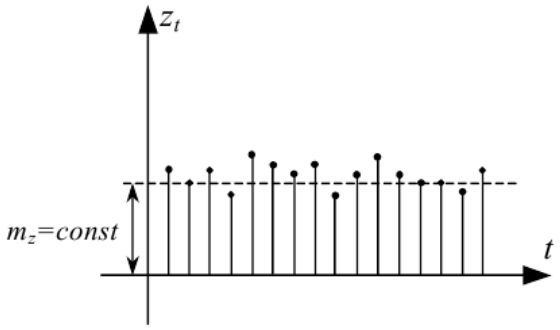
Інтерполяція – це спосіб побудови рядів динаміки за попередній період, коли з якихось причин були відсутні відомості про розміри явища, або для знаходження невідомих проміжних рівнів ряду динаміки. Відсутність цих даних може бути обумовлена різними причинами: був відсутній облік цих явищ

в попередній час, змінилася методика обчислення показника тощо. Для того щоб обчислити невідомі рівні ряду динаміки, провадять математичні розрахунки різної складності.

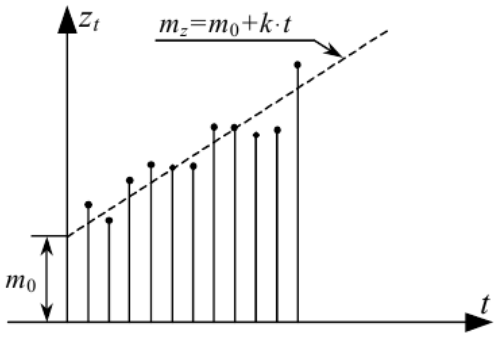
Невідомі рівні ряду динаміки знаходять або на базі сусідніх відомих значень ряду динаміки шляхом обчислення їх середньої арифметичної простої, або на базі взаємозв`язку цього явища з іншими, кількісний вираз яких відомий. При застосуванні методу інтерполяції робиться припущення, що загальна тенденція, яку ми маємо зараз, мала місце і в попередній інтервал часу. Завжди при застосуванні цього методу проводяться математичні розрахунки різної складності.

Динамічним рядом (рядом динаміки) називається послідовність показників, які характеризують зміну явища (процесу, об’єкта) у часі. Окремі спостереження динамічного ряду називаються рівнями.

Як і будь-який інший випадковий процес, часовий ряд може бути стаціонарним (рис. 4.1, а) або нестаціонарним (рис. 4.1, б).



а)



б)

Рисунок 4.1 – Графіки реалізації стаціонарного (а) та нестаціонарного (б) часових рядів

На основі рядів абсолютних величин утворюються ряди динаміки середніх і відносних величин, тому ряди абсолютних величин розглядаються як вихідні, а ряди відносних і середніх величин – як похідні.

Різновидами рядів динаміки відносних величин є ряди темпів зростання (або зниження) певного показника, зміни структури сукупності (наприклад, питомої ваги міського або сільського населення), зміни показники інтенсивності (рівень народжуваності і смертності на 1000 осіб).

Вид ряду динаміки обумовлюється сутністю, внутрішнім змістом досліджуваного явища. Розміри одних явищ фіксуються на будь-яку дату, момент часу, наприклад: чисельність працівників, вартість активів, основних і оборотних коштів і ін. Розміри інших явищ встановлюються при розгляді їх протягом будь-якого проміжку часу як підсумок діяльності за певний період. Наприклад, обсяг послуг, виручка від реалізації послуг, видатки виробництво послуг, прибуток завжди відносяться до якого-небудь проміжку (інтервалу) часу: дня, місяця і т.д.

Залежно від способу вираження періоду часу, за який характеризується розвиток явища, ряди динаміки поділяються на два вида: моментні і інтервальні.

Моментні ряди – рівні рядів характеризують значення показника (явища) станом на певні моменти часу (дату).

Інтервальні ряди – рівні рядів характеризують значення показник, досягнуте за певний період (інтервал) часу.

Показники інтервальних рядів можна складати, показники моментних рядів не можна, вони не володіють властивістю підсумовування. Можна, наприклад, скласти дані про щоденний обсяг послуг (трафіку) і отримати місячний підсумок, потім, служив місячні підсумки, отримати обсяг послуг

(трафіку) за рік. Якщо ж скласти дані про кількість працівників початку кожного місяця року, то така сума буде позбавлена економічного сенсу.

Стаціонарний ряд – часовий ряд даних, основні статистичні характеристики якого (середнє значення і дисперсія) залишаються постійними

Стаціонарні часові ряди передбачають, що процес породження наявних даних є лінійним. Вони не мають тренду або періодичної зміни середнього та дисперсії.

Перш ніж аналізувати ряд динаміки, необхідно переконатися, що рівень явища за один період можна порівнювати з рівнем явища за інший період, тобто бути впевненим в порівнянності рівнів ряду. Проблема порівнянності рівнів виступає в рядах динаміки особливо гостро, так як аналіз охоплює, як правило, великий період часу, протягом якого могли статися всілякі зміни, що призвели до непорівнянності статистичних даних.

Несумісність статистичних даних пояснюється різними причинами, найважливішими з яких є зміни цін, територіальних меж, одиниць вимірювання або рахунки, в методології обліку і розрахунку показників, кола охоплених статистичними спостереженням об'єктів.

Щоб вирішити питання про порівнянності статистичних даних, потрібно знати про всі зміни, що відбулися за аналізований період в досліджуваному явищі.

Несумісність статистичних даних найбільш часто виникає внаслідок зміни цін, тарифів.

Задача інтерполювання функції розв’язується шляхом побудови деякого аналітичного виразу, який співпадає зі значеннями таблично заданої функції в скінченній кількості табличних значень аргументу. Тому, задача інтерполювання функції в деякому розумінні обернена до задачі табулювання функції: при табулюванні від аналітичного способу задання функції переходять до табличного, а при інтерполюванні – за табличними значеннями функції

будується деякий аналітичний вираз, тобто формула, що задає шукану функцію наближено.

Екстраполяція – це спосіб продовження кількісних характеристик сукупностей за межі досліджуваного явища в майбутнє на базі встановлених закономірностей за попередній термін. За допомогою способу екстраполяції можуть бути зроблені висновки, одержані внаслідок вивчення однієї частини сукупності та поширені на його іншу аналогічну частину.

В основі використання способу екстраполяції лежить припущення, що фактори, які обумовили розвиток даного явища, залишаються незмінними і протягом наступного періоду. Цей спосіб в останні роки найчастіше застосо-вується для прогнозування явищ лише на короткий проміжок часу.

Використовуючи спосіб екстраполяції, можна прогнозувати чисельність населення, його міграцію, а також зміни в захворюваності.

Операція екстраполювання, взагалі кажучи, менш точна, ніж операція інтерполювання, і її слід застосовувати тоді, коли:

– функція біля кінців таблиці змінюється плавно;

– відстань від кінців таблиці, на якій екстраполюють, невелика (менша ніж відстань між сусідніми вузлами).

Досягнення дескриптивної теорії прийняття рішень дозволяють нам уникнути помилок при здійсненні вибору на множині альтернатив, а також розуміти, прогнозувати та навіть змінювати рішення інших людей.

Однією з основних концепцій, розроблених у рамках дескриптивної теорії прийняття рішень, є *концепція обмеженої раціональності*.

Концепція обмеженої раціональності є протиставленням концепції очікуваної корисності. Відповідно до концепції обмеженої раціональності, в ситуації вибору люди інтуїтивно використовують «стратегії спрощення», які дозволяють їм уникнути переробки величезних масивів інформації:

– розглядають лише невелику кількість альтернатив та їхніх можливих наслідків;

– спрощують проблему оцінки альтернатив за критеріями – вони встановлюють рівень прийнятних результатів за всіма можливими наслідками реалізації альтернатив;

– вибирають першу альтернативу, яка відповідає всім установленим рівням прийнятних результатів.

За твердженням О. Кулагіна, «відповідно до концепції обмеженої раціо нальності Г. Саймона, люди не прагнуть «оптимізувати», а хочуть відчути себе «задоволеними».

Тобто люди вибирають не найкращий варіант із всіх можливих, а лише той, який відповідає базовому набору їхніх вимог. Прийняті в такий спосіб рішення не вважаються раціональними або нераціональними – вони трактуються як «обмежено раціональні».

Перевірку гіпотез стосовно сталості середнього значення та дисперсії часового ряду можна здійснити кількома способами. Найпростішими з них є перевірка значущої відмінності двох середніх значень для деяких підмножин вибірки (наприклад, для першої та останньої третин усього обсягу даних) за z – критерієм (критерій перевірки гіпотези про рівність середніх двох нормально розподілених вибірок) і для дисперсії, якщо справедливе припущення про нормальний розподіл, можна використати F-критерій. Розглянемо два поширені методи: метод перевірки різниць середніх рівнів і метод Форстера-Стьюарта.

Метод перевірки різниць середніх рівнів. Реалізація цього методу передбачає такі чотири кроки.

Крок перший. Вхідний часовий ряд $y_1, y_2, y_3, \ldots, y_n$ розподіляють на дві приблизно однакові за кількістю спостережень частини: в першій частині n_1 першої половини рівнів вхідного ряду, у другій – решта рівнів $n_2 (n_1 + n_2 = n)$.

Крок другий. Для кожної з цих частин розраховують середні значення й дисперсії:

$$\bar{y}_1 = \frac{\sum_{t=1}^{n_1} y_t}{n_1};\ \hat{\sigma}^2{}_1 = \sum_{t=1}^{n_1} (y_t - \bar{y}_1)^2 /(n_1 - 1);\ y_2 = \frac{\sum_{t=1}^{n_2} y_t}{n_2};\ \hat{\sigma}^2{}_2 = \sum_{t=1}^{n_2} (y_t - \bar{y}_2)^2 /(n_2 - 1).$$

Крок третій. Перевірка рівності (однорідності) дисперсій обох частин ряду за допомогою *F*-критерію**,** що порівнює розрахункове значення цього критерію:

$$F = \begin{cases} \hat{\sigma}^2{}_1 \big/ \hat{\sigma}^2{}_2, \text{ якщо } \hat{\sigma}^2{}_1 > \hat{\sigma}^2{}_2 \\ \hat{\sigma}^2{}_2 \big/ \hat{\sigma}^2{}_1, \text{ якщо } \hat{\sigma}^2{}_1 < \hat{\sigma}^2{}_2 \end{cases} \tag{4.1}$$

із табличним (критичним) значенням критерію Фішера *F*α із заданим рівнем значущості α. Якщо розрахункове значення *F* менше за табличне *F*α, то гіпотезу про рівність дисперсій приймають, і можна переходити до четвертого кроку. Якщо *F* більше або дорівнює *F*α, гіпотезу про рівність дисперсій відхиляють і доходять висновку, що цей метод не дає відповіді щодо наявності тренду.

На *четвертому кроці* перевіряють гіпотезу про відсутність тренду за допомогою *t*-критерію Стьюдента. Для цього визначають розрахункове значення критерію Стьюдента за формулою:

$$t = \frac{\left|\bar{y}_1 - \bar{y}_2\right|}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{4.2}$$

де σ – оцінка середньоквадратичного відхилення різниць середніх:

$$\hat{\sigma} = \sqrt{\frac{(n_1 - 1)\hat{\sigma}^2{}_1 + (n_2 - 1)\hat{\sigma}^2{}_2}{n_1 + n_2 - 2}}$$

Якщо розрахункове значення *t* менше за табличне *t*α, то нульову гіпотезу не відхиляють, тобто тренд відсутній, інакше – тренд є. Зазначимо, що в цьому разі табличне значення *t*α приймають для числа ступенів вільності, яке дорівнює $n_1 + n_2 - 2$, до того ж цей метод застосовують суто для рядів із монотонною тенденцією. Недолік методу полягає у неможливості правильно

визначити існування тренду в тому разі, коли часовий ряд містить точку зміни тенденції у середині ряду.

Приклад 4.1. Застосуємо метод перевірки різниць середніх рівнів для двох часових рядів: доходів консолідованого бюджету (млн. грн.) і доходів консолідованого бюджету (% ВВП). Для цього початкові часові ряди поділяють на дві однакові частини: перша охоплює 1999−2000 рр., друга – 2001-2002 рр. Кількість кварталів-спостережень в обох частинах однакова: $n_1 = n_2 = 8$. Результати розрахунків наведено в табл. 4.1. На рівні значущості $\alpha = 0{,}05$, тобто з імовірністю 0,95, із числом ступенів вільності $k_1 = n_1 - 1 = 8 - 1 = 7$ і $k_2 = n_2 - 1 = 8 - 1 = 7$ табличне значення критерію Фішера дорівнює F_α = 3,79.

Таблиця 4.1 – Результати розрахунків

Доходи	Роки	Середнє значення	Дисперсія	F	$\hat{\sigma}$	*t*
млн. грн.	1999-2000 2001-2002	9860,2 13695,8	8349206 4459451	1,87	2733,44	2,8
% до ВВП	1999-2000 2001-2002	26,0 24,5	6,41 1,92	3,34	2,2	1,37

Для обох часових рядів *F* розрахункові менші за табличне значення F_α, тобто приймається гіпотеза про рівність дисперсій.

На рівні значущості із числом ступенів свободи $n_1 + n_2 - 2 = 16 - 2 = 14$ табличне значення *t*-розподілу дорівнює $t_a = 2{,}145$.

Для часового ряду доходів, виражених у млн. грн., *t*-розрахункове перевищує табличне значення t_a, тобто нульова гіпотеза не приймається, тренд існує.

Для часового ряду доходів, виражених у відсотках до ВВП, *t*-розрахункове менше за табличне значення t_a, тобто приймається гіпотеза про відсутність тренду.

Метод Форстера-Стьюарта. Реалізація методу передбачає чотири кроки.

Крок перший. Порівнюють кожен рівень вхідного часового ряду, починаючи з другого рівня, з усіма попередніми, при цьому визначають дві числові послідовності:

$$k_a = \begin{cases} 1, \textit{ якщо } y_t \textit{ більше всіх попередніх рівнів} \\ 0, \textit{ в іншому разі} \end{cases} \tag{4.3}$$

$$l_t = \begin{cases} 1, \textit{ якщо } y_t \textit{ меньше всіх попередніх рівнів} \\ 0, \textit{ в іншому разі} \end{cases} \tag{4.4}$$

$t = 2,3,..., n.$

Крок другий. Розраховують величини *c* і *d:*

$$c = \sum_{t=2}^{n} (k_t + l_t) \tag{4.5}$$

$$d = \sum_{t=2}^{n} (k_t - l_t) \tag{4.6}$$

Величина *c*, яка характеризує зміну рівнів часового ряду, набуває значення від 0 (усі рівні ряду однакові) до $n - 1$ (ряд монотонний). Величина *d* характеризує зміну дисперсії часового ряду та змінюється від $[-(n - 1)]$ — ряд поступово згасає, до $(n - 1)$ — ряд поступово розхитується.

Крок третій Перевіряється гіпотеза стосовно того, чи можна вважати випадковими:

1. відхилення величини *c* від математичного сподівання ряду, в якому рівні розташовані випадково,

2. відхилення величини *d* від нуля. Цю перевірку здійснюють на підставі обчислення *t*-відношення відповідно для середньої та для дисперсії:

$$t_c = \frac{|c - \hat{\mu}|}{\hat{\sigma}_1}; \ \hat{\sigma}_1 = \sqrt{2 \ln n - 3{,}4253}; \tag{4.7}$$

$$t_d = \frac{|d - 0|}{\hat{\sigma}_2}; \ \hat{\sigma}_2 = \sqrt{2 \ln n - 0{,}8456}; \tag{4.8}$$

де $\hat{\mu}$ – оцінка математичного сподівання ряду; $\hat{\sigma}_1$ – оцінка середньоквадратичного відхилення для величини *c*; $\hat{\sigma}_2$ – оцінка середньоквадратичного відхилення для величини *d*.

Фрагмент розрахованих значень величин $\hat{\mu}$, $\hat{\sigma}_1$ і $\hat{\sigma}_2$ для різних *n* наведено в табл. 4.2.

Таблиця 4.2 – Фрагмент розрахованих значень величин $\hat{\mu}$, $\hat{\sigma}_1$ і $\hat{\sigma}_2$

n	10	20	30	40
$\hat{\mu}$	3,858	5,195	5,990	6,557
$\hat{\sigma}_1$	1,288	1,677	1,882	2,019
$\hat{\sigma}_2$	1,964	2,279	2,447	2,561

Крок четвертий. Розрахункові значення t_c і t_d порівнюють із табличним значенням *t*-критерію із заданим рівнем значущості t_a. Якщо розрахункове значення *t* менше за табличне t_a, то гіпотезу про відсутність відповідного тренду приймають, в іншому разі тренд існує. Наприклад, якщо t_c більше табличного значення t_a, а t_d менше t_a, то для заданого часового ряду існує тренд у середньому, а тренду дисперсії рівнів ряду немає.

Приклад 4.2.

Застосування методу Форстера-Стьюарта для двох часових рядів: доходів консолідованого бюджету (млн. грн.) та доходів консолідованого бюджету (% до ВВП) дає розрахунки, наведені в табл. 4.3.

Таблиця 4.3 – Приклад застосування методу Форстера-Стьюарта для двох часових ряді

Доходи	$\sum k_t$	$\sum l_t$	*c*	*d*	t_c	t_d
млн. грн.	8	0	8	8	3,28	4,07
% до ВВП	4	1	5	3	0,9	1,53

На рівні значущості $a = 0{,}05$, тобто з імовірністю 0,95 та з числом ступенів волі n – 2 = 16 – 2 = 14 табличне значення критерію Стьюдента дорівнює t_a = 2,145.

Для часового ряду доходів, виражених у млн. грн., розрахункові значення t_c і t_d перевищують табличне значення t_a, тобто нульова гіпотеза не приймається, існує тренд як середнього, так і дисперсії ряду.

Для часового ряду доходів, виражених у відсотках до ВВП, розрахункові значення t_c і t_d менші за табличне значення t_a, тобто приймається гіпотеза про відсутність тренду в тенденції й дисперсії ряду.

Позначимо:

y_1 – початкове значення рівня динамічного ряду;

y_n – кінцеве значення рівня динамічного ряду;

y_i – умовно прийнятий (i-й) рівень динамічного ряду;

n – кількість елементів динамічного ряду.

Основні аналітичні показники динамічного ряду, які використовуються у прогнозуванні:

а. абсолютний приріст:

1. ланцюговий

$$\Delta' y_i = y_i - y_{i-1} \tag{4.9}$$

2. базисний

$$\Delta y_i = y_i - y_1 \tag{4.10}$$

б. середній абсолютний приріст

$$\Delta y = \frac{y_n - y_1}{n-1} = \frac{\sum_{i=1}^{n-1} \Delta' y_i}{n-1} \tag{4.11}$$

в. коефіцієнт росту:

1. ланцюговий

$$K_{Pi} = \frac{y_i}{y_{i-1}} \tag{4.12}$$

2. базисний

$$K_{Pi} = \frac{y_i}{y_1} \tag{4.13}$$

3. за весь період

$$K_{Pn} = \frac{y_n}{y_1} \tag{4.14}$$

г. коефіцієнт приросту

$$K_{np} = k_p - 1 \tag{4.15}$$

д. середній коефіцієнт росту

$$\bar{k}_p = \sqrt[n-1]{\frac{y_n}{y_1}} \tag{4.16}$$

е. середній коефіцієнт приросту

$$\bar{k}_{np} = \bar{k}_p - 1 \tag{4.17}$$

ж. абсолютний розмір 1% приросту:

1. ланцюговий

$$\Delta y_{1\%} = \frac{y_i - y_{i-1}}{\frac{y_i - y_{i-1}}{y_{i-1}} 100} = \frac{y_{i-1}}{100} \tag{4.18}$$

2. за весь період

$$\bar{\Delta} y_{1\%} = \frac{\bar{\Delta} y}{\bar{k}_{np}} \tag{4.19}$$

з. коефіцієнт випередження (відставання)

$$k = \frac{y_i}{y_{i-1}} : \frac{x_i}{x_{i-1}} \tag{4.20}$$

Добуток ланцюгових коефіцієнтів росту дорівнює базисному коефіцієнту росту за весь період, тобто

$$k'_{p1} \bullet k'_{p2} \bullet k'_{p3} .... \bullet k'_{pn} = k_{pn} \tag{4.21}$$

що може бути доведено таким чином.

На основі наведених аналітичних показників, які широко застосовуються для оцінки динамічних рядів, можна вивести залежності, що можуть бути використані для побудови прогнозів:

$$\hat{y}_{n+1} = y_n + \Delta' y_n = y_n - y_{n-1} \tag{4.22}$$

$$\hat{y}_{n+T} = y_n + \bar{\Delta} y \cdot T \tag{4.23}$$

$$\hat{y}_{n+1} = y_n + k_{Pn}; \quad k_{Pn} = \frac{y_n}{y_{n-1}} \tag{4.24}$$

$$\hat{y}_{n+T} = y_n \cdot \bar{k}_p^T \tag{4.25}$$

де $\hat{y}$ – прогнозні значення показника.

T – величина горизонту прогнозу *(*T = 1; 2; 3 ...)

Прогноз по формулі 4.23 називається ***екстраполяцію за середнім абсолютним приростом***. Екстраполяцію за середнім абсолютним приростом можна бути виконати в тому разі, коли загальна тенденція розвитку вважається лінійною.

Прогноз по формулі 4.25 називається ***екстраполяцію за середнім темпом зростання*.**

Екстраполяцію за середнім темпом зростання можна виконувати у разі, коли є підстави вважати, що загальна тенденція динамічного ряду характеризується експоненціальною кривою.

Приклад 4.3. В таблиці 4.4 наведені дані про середнє споживання кондитерських виробів на одну людину по області за рік. Використовуючи рівняння (4.23; 4.25), побудувати прогноз споживання кондитерських виробів на наступну п’ятирічку.

Використовуючи дані перших шести років – базисний рік та роки першої п’ятирічки, розрахуємо відповідно:

Середній абсолютний приріст

$$\bar{\Delta} y = \frac{y_k - y_0}{k-1} = \frac{15{,}0 - 10{,}7}{6-1} = \frac{4{,}3}{5} = 0{,}9 \text{ кг.}$$

Таблиця 4.4 − Середньорічне споживання кондитерських виробів по області

Номер року, t	Споживання кондитерських виробів на одну людину в рік, кг.	Номер року, t.	Споживання кондитерських виробів на одну людину в рік, кг.
1	2	3	4
1	10,7	7	15,9
2	11,5	8	17,2
3	12,2	9	18,1
4	13,4	10	19,8
5	15,0	11	21,2
6	15,0		

Середньорічний коефіцієнт росту

$$\bar{k}_p = \sqrt[k-1]{\frac{y_k}{y_1}} = \sqrt[6-1]{\frac{15,0}{10,7}} = \sqrt[5]{1,40} = 1,07$$

На основі залежності (4.23) складемо прогноз споживання кондитерських виробів на період $(\bar{e}+1) \div n$.

$\hat{y}_{k+1} = 15 + 0,9*1 = 15,9\ кг;$
$\hat{y}_{k+2} = 15 + 0,9*2 = 16,8\ кг;$
$\hat{y}_{k+3} = 15 + 0,9*3 = 17,7\ кг;$
$\hat{y}_{k+4} = 15 + 0,9*4 = 18,6\ кг;$
$\hat{y}_{k+1} = 15 + 0,9*5 = 19,5\ кг;$

Результати розрахунків зведені в таблицю та порівняні з фактичними даними.

Складемо прогноз споживання кондитерських виробів на основі формули (4.25):

$\hat{y}_{k+1} = 15*1,07^1 = 16\ кг;$
$\hat{y}_{k+2} = 5*1,07^2 = 17,2\ кг;$
$\hat{y}_{k+3} = 5*1,07^3 = 18,4\ кг;$
$\hat{y}_{k+4} = 5*1,07^4 = 19,7\ кг;$
$\hat{y}_{k+1} = 5*1,07^5 = 21\ кг;$

Результати прогнозу порівняні із фактичними даними та оцінена якість прогнозу (табл. 4.5)

Таблиця 4.5 − Оцінка якості прогнозу, складеного на основі середнього абсолютного приросту

№	Фактичне значення, кг	Прогнозоване значення, кг	Відхилення	
			Абсолютне (гр2-гр3), кг	Відносне (гр4:гр2)100%
1	2	3	4	5
1	15,9	15,9	0	0
2	17,2	16,8	0,4	2,3
3	18,1	17,7	0,4	2,2
4	19,8	18,6	1,2	6,1
5	21,2	19,5	1,7	8
	Середнє значення	—	0,7	3,7

Таблиця 4.6 − Оцінка якості прогнозу, складеного на основі середньорічного коефіцієнта росту

№	Фактичне значення, кг	Прогнозоване значення, кг	Відхилення	
			Абсолютне (гр2-гр3), кг	Відносне (гр4:гр2)100%
1	2	3	4	5
1	15,9	16	-0,1	-0,6
2	17,2	17,2	0	0
3	18,2	18,4	-0,3	-1,7
4	19,8	19,7	0,1	0,5
5	21,2	21	0,2	0,9
	Середнє значення	—	0,1	0,7

Порівнюючи результати прогнозів, поданих в табл. 4.5 та табл. 4.6, можна зробити висновок про те, що використання середньорічного коефіцієнта росту забезпечує більш високу точність прогнозу, про що свідчать відхилення за всі роки і в цілому за п'ятиріччя.

Для складання прогнозу за межі наявних даних, тобто на перспективу, розрахуємо середньорічний коефіцієнт росту на основі другої п'ятирічки з використанням базисного періоду

$$\bar{k}_p = \sqrt[5]{\frac{21,5}{15,0}} = \sqrt{1,413} = 1,071$$

Прогноз споживання кондитерських виробів на наступне п’ятиріччя складе:

$\hat{y}_{n+1} = 21{,}2 * 1{,}071^1 = 22{,}7\ кг;$

$\hat{y}_{n+2} = 21{,}2 * 1{,}071^2 = 24{,}3\ кг;$

$\hat{y}_{n+3} = 21{,}2 * 1{,}071^3 = 26\ кг;$

$\hat{y}_{n+4} = 21{,}2 * 1{,}071^4 = 28\ кг;$

$\hat{y}_{n+5} = 21{,}2 * 1{,}071^5 = 29{,}9\ кг;$

Прогноз споживання кондитерських виробів складено з урахуванням зберігання тенденцій, які склалися в «передісторії».

Суттєвим недоліком показників середнього абсолютного приросту та середнього коефіцієнта росту є те, що значення їх цілком залежить тільки від крайніх рівнів динамічного ряду. Проміжні значення, які багато в чому, а іноді і в вирішальній мірі визначають тенденцію змін показників, по суті в розрахунках не беруть участі. Зазначений недолік багато в чому усувається шляхом аналітичного вирівнювання рядів динаміки.

Одна з найважливіших завдань вивчення рядів динаміки - виявити основну тенденцію (закономірність) в зміні рівнів ряду, звану трендом. Закономірність в зміні рівнів ряду в одних випадках проявляється досить наочно, в інших - може затушовує коливаннями, викликаними випадковими і невипадковими причинами.

Можна сказати, що динаміка ряду включає три компоненти:

– довгострокове рух (тренд),

– короткочасне систематичне рух (наприклад, сезонні коливання),

– несистематичний випадкове рух, що викликає коливання рівнів щодо тренда.

Вивчаючи ряди динаміки, дослідники намагаються розділити ці компоненти і виявити основну закономірність розвитку явища в окремі періоди, звільнену від дії випадкових чинників. Для цього ряди динаміки піддають обробці.

Статистика застосовує ряд методів перетворення рядів динаміки, що дозволяють виявити і показати основну тенденцію в розвитку явища.

Внутрішньорічні коливання, що мають більш-менш регулярний характер, являють собою сезонну компоненту ряду динаміки.

Внутрішньорічні коливань схильні рівні багатьох показників. Наприклад, витрата електроенергії в літні місяці значно менше, ніж в зимові.

Закономірності в зміні рівнів ряду динаміки по місяцях року називаються сезонними коливаннями. Мірою сезонних коливань є індекси сезонності IS. Сукупність індексів сезонності відображає сезонну хвилю.

Індекси сезонності – відношення місячних рівнів ряду динаміки до середньомісячного рівня за рік:

$$I_S = \frac{y_i}{\bar{y}} \cdot 100\% \qquad (4.26)$$

Для обчислення індексів сезонності застосовують різні методи.

Якщо ряд динаміки не містить яскраво вираженої тенденції розвитку, то індекси сезонності обчислюють за емпіричними даними без їх попереднього вирівнювання.

У розглянутому методі розрахунку індексів сезонності використовувалися дані одного року. Цей метод досить простий, але в силу елемента випадковості місячні дані одного року недостатньо надійні для визначення міри сезонних коливань. Тому рекомендується використовувати місячні дані за ряд років (в основному за 3 роки).

За даними ряду років визначається середнє значення рівня для кожного місяця $\bar{y}_i$, а також середньомісячний рівень за весь період $\bar{y}$. Потім визначаються індекси сезонності за формулою:

$$I_{S_i} = \frac{\bar{y}_i}{\bar{y}} \cdot 100\% \qquad (4.27)$$

Для характеристики сили коливання рівнів динамічного ряду через сезонної нерівномірності використовується середньоквадратичне відхилення індексів сезонності (у відсотках) від 100%:

$$\sigma_S = \sqrt{\frac{\sum (I_S - 100\%)^2}{n}} \qquad (4.28)$$

Порівняння σ_S, обчислених за різні періоди, показує зрушення в сезонності. Так, зменшення σ_S свідчить про зменшення впливу сезонності на динаміку аналізованого показника.

Якщо ряд містить певну тенденцію розвитку, то перш ніж визначати сезонну хвилю, фактичні дані обробляються для виявлення основної тенденції розвитку. Т.ч. в даному випадку індекси сезонності можуть бути розраховані також як відношення фактичного рівня відповідного місяця до рівня,

1. розраховується за методом ковзної середньої або

2. визначеної за рівнянням тренду.

В останньому випадку для ряду динаміки попередньо визначається рівняння тренда, на підставі якого для кожного місяця розраховується теоретичне значення рівня ряду. В даному випадку індекс сезонності визначається:

$$I_S = \frac{y_i}{\hat{y}_i} \cdot 100\% \qquad (4.29)$$

де $\hat{y}_i$ - теоретичне значення рівня i -го місяця.

4.3 Методичні вказівки щодо організації самостійної роботи студентів

Перед лабораторною роботою слід повторити матеріал за курсом лекцій та за рекомендованою літературою за темою лабораторної роботи.

Особливо слід звернути увагу на такі питання:

1. Дисперсійний аналіз.

2. Дисперсійний аналіз факторів.

3. Аналіз подій і пошук закономірностей.

4. Інтерполяція та екстраполяції.

5. Метод перевірки різниць середніх рівнів.

6. Метод Форстера-Стьюарта.

4.4 Завдання до лабораторної роботи

1. Вивчити теоретичну частину заняття.

2. Вибрати тип програмного забезпечення, в рамках якого буде виконуватися розрахунок (з тих що розроблялося раніше починаючі з 2 по 4 курс, або з обраних на ПЗ №1-2, або ЛБ 1-2).

3. Привести опис ПЗ в рамках якого буде виконуватись розрахунки.

4. З відкритих джерел взяти необхідні дані в необхідному розмірі за відповідною темою (наприклад з https://www.kaggle.com/fernandol/countries-of-the-world). Вказати звідки були взяті дані.

5. Табличка, які саме дані використовувалися для проведення експериментів (джерела даних посиланням, кількість і характеристики, приклади).

6. На основі статистичних даних (з попередніх пунктів) сформувати ряд динаміки для статистичного показника, обраного відповідно до порядкового номера в журналі

7. Методом перевірки різниць середніх рівнів та методом Форстера-Стьюарта перевірити часовий ряд на стаціонарність.

8. Простими методами прогнозування на основі екстраполяції тенденції отримати значення показника на наступні 3 місяці/роки (залежить від ситуації по темі роботи, час прогнозу пояснити, чому саме на такий період).

9. Обчислити індекси сезонності, та зробити пояснення відносно отриманих даних.

10. Зробити висновки щодо зміни значення показника у найближчій перспективі.

11. Оформити звіт що демонструє роботу і здати викладачеві в електронному вигляді.

12. Відповісти на контрольні питання (парний варіант дає відповіді на парні питання, непарний варіант дає відповіді непарні питання).

4.5 Опис програмного забезпечення

Під час виконання лабораторної роботи використовується таке програмне забезпечення [1-7, лекц. 3-5, 7-9]:

– офісні програми для ПК MS Office Word, PowerPoint або LibreOffice, або Google Документы, або OpenOffice;

– браузери Edge, Firefox, Chrome, Opera.

4.6 Зміст звіту

1. Тема і мета роботи.
2. Послідовність виконуваних у процесі роботи дій.
3. Опис процесу роботи з скріншотами роботи.
4. Висновки з роботи.
5. Відповіді на контрольні питання (парний варіант дає відповіді на парні питання, непарний варіант дає відповіді непарні питання).

4.7 Контрольні запитання та завдання

1. Застосування методів інтерполяція та екстраполяції?
2. Інтуїтивні (експертні) методи прогнозування.
3. Перевірка гіпотези про існування тренда. Метод Фостера – Стюарта.
4. Поняття прогноз, прогнозування.

5. Поняття тенденція, динамічний ряд, тренд

6. Основні аналітичні показники, що використовуються в екстраполяційних методах.

7. Суть методу прогнозування на основі індексу сезонності.

8. Екстраполяція трендів. Види трендових моделей.

9. Переваги та недоліки методів екстраполяці?

10. Основні стратегії екстраполяції:

11. В чому відмінність інтерполяції від екстраполяції?

12. Чому іноді використання екстраполяція це погана ідея?

4.8 Додаткова література та електронні ресурси

1. Державної служби статистики України (http://www.ukrstat.gov.ua/) [Електронний ресурс]. (дата звернення: 1.09.2023).

2. Лежнюк П.Д., Рубаненко О.Є., Лук'яненко Ю. В. Основи теорії планування експерименту. Лабораторний практикум. - Вінниця: ВНТУ, 2006. - 167 с.

3. Системи підтримки прийняття рішень [Текст] : навчальний посіб- ник для самостійного вивчення дисципліни / [уклад.: С. М. Братушка, С. М. Новак, С. О. Хайлук] ; Державний вищий навчальний заклад "Українська академія банківської справи Національного банку України". – Суми : ДВНЗ "УАБС НБУ", 2010. – 265 с.

РЕКОМЕНДОВАНА ЛІТЕРАТУРА

1. Vigers K. Development of software requirements – 3rd Edition, Microsoft Press, 2016. – 736 с.

2. Розділи BABOK v3 [Електронний ресурс] https://www.maxzosim.com/babok-v3/ (дата звернення: 1.09.2023).

3. Автоматизація бізнес-процесів : [навч. посіб.] / Н. В. Косенко, Ю. Ю. Гусєва, І. В. Чумаченко, Ш. А. Омаров ; М-во освіти і науки України, Харків. нац. ун-т радіоелектроніки. – Харків : ХНУРЕ, 2019. – 80 с. – ISBN 978-966-659-261-6. – 5,20

4. Garaedaghi, Jamshid. Systems thinking: how to manage chaos and complex processes https://www.amazon.com/Systems-Thinking-Complexity-Designing-Architecture/dp/0750671637 .

5. Варенко В.М. Інформаційно-аналітична діяльність: Навч. посіб. / В. М. Варенко. – К.: Університет «Україна», 2014. – 417 с.

6. Мигаль В. Д. Інтелектуальні системи в технічній експлуатації автомобілів: монографія / В. Д. Мигаль. Х.: Майдан, 2018. 262 с. [Електронний ресурс]. https://dspace.khadi.kharkov.ua/dspace/bitstream/123456789/2316/1/migal_1_2018.pdf (дата звернення: 1.09.2023).

7. Practical Statistics for Data Scientists Andrew Bruce and Peter Bruce., Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, 2020. — 363 с. [Електронний ресурс]. https://www.researchgate.net/profile/Janine-Zitianellis/post/Can_anyone_please_suggest_a_books_on_machine_learning_using_R_Programming/attachment/613a5b83647f3906fc975a71/AS%3A1066204907204608%401631214467436/download/Practical+Statistics+for+Data+Scientists+50%2B+Essential+Concepts+Using+R+and+Python+by+Peter+Bruce%2C+Andrew+Bruce%2C+Peter+Gedeck.pdf (дата звернення: 1.09.2023).

8. Data Analytics: An Essential Beginner's Guide To Data Mining, Data Collection, Big Data Analytics For Business, And Business Intelligence Concepts Paperback – February 4, 2018 [Электронный ресурс]. https://www.amazon.com/Data-Analytics-Essential-CollectionIntelligence/dp/1985097974/ (дата обращения: 11.05.2024).

Додаток А

ПРИКЛАД ТИТУЛЬНОЇ СТОРІНКИ ЗВІТУ ПРО ВИКОНАНУ СТУДЕНТОМ ЛАБОРАТОРНУ РОБОТУ

Міністерство освіти і науки України
Харківський національний університет радіоелектроніки
Кафедра програмної інженерії

Лабораторна робота № 1
з дисципліни «Емпіричні методи програмної інженерії»
з теми: «Статистична обробка емпіричних даних.
Обчислення точкових характеристик вибірки»

Виконав
студ. гр. ПЗПІ-23-9
Петренко Микола Васильович

01 жовтня 2024 р.

Перевірила
к.т.н., доцент кафедри ПІ
Груздо І. В.

Харків 2024

Електронне навчальне видання

Методичні вказівки
до лабораторних робіт з дисципліни
«Емпіричні методи програмної інженерії»
для студентів денної та заочної форм навчання
першого (бакалавського) рівня вищої освіти
спеціальності 121 – Інженерія програмного забезпечення,
освітня програма «Програмна інженерія»

Упорядники: ГРУЗДО Ірина Володимирівна
НАЗАРОВ Олексій Сергійович

Відповідальний випусковий З.В. Дудар
Редактор О.Г. Троценко
Комп'ютерна верстка Л.Ю. Свєтайло

План 2024 (друге півріччя), поз. 20

Підп. до друку 19.10.2024 Формат 60х 84 $^{1}/_{16}$ Спосіб друку – ризографія
Умов. друк. арк. 4,5 Облік. вид. арк. 4,0 Тираж 20 прим.
Зам. № 1-20 Ціна договірна.

ХНУРЕ. Україна. 61166 Харків, просп. Науки, 14

Віддруковано в редакційно-видавничому відділі ХНУРЕ
Харків, 61166, просп. Науки, 14