МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ДЕРЖАВНИЙ УНІВЕРСИТЕТ ЕКОНОМІКИ І ТЕХНОЛОГІЙ
НАВЧАЛЬНО-НАУКОВИЙ ТЕХНОЛОГІЧНИЙ ІНСТИТУТ
Кафедра Інжинірингу з галузевого машинобудування

КОНСПЕКТ ЛЕКЦІЙ
З ОСВІТНЬОГО КОМПОНЕНТА
«Сучасні інформаційні технології з аналізу даних»

для здобувачів спеціальності 133 «Галузеве машинобудування»
освітньо-наукового ступеня доктора філософії (PhD)
очної (денної) та заочної форм навчання

РЕКОМЕНДОВАНО
на засіданні
кафедри Інжинірингу з галузевого
машинобудування
(протокол №7 від «13» лютого 2025р.)

ПОГОДЖЕНО
на засіданні
Науково-методичної ради
Державного університету економіки і технологій
(протокол №9 від «18 березня 2025р.)

м. Кривий Ріг
2025 р.

Конспект лекцій з освітнього компонента «Сучасні інформаційні технології з аналізу даних» для освітньо- наукового ступеня доктора філософії (PhD) спеціальності 133 «Галузеве машинобудування» очної (денної) та заочної форм навчання / укладачі: В.ЗАСЕЛЬСЬКИЙ, І. ЗАСЕЛЬСЬКИЙ;
рецензент: О. УЧИТЕЛЬ. Кривий Ріг: ДУЕТ, 2025, 74 с.

Укладачі: Володимир ЗАСЕЛЬСЬКИЙ, д-р тех. наук., проф.Інжинірингу з галузевого машинобудування навчально-наукового Технологічного інституту ДУЕТ

Ігор ЗАСЕЛЬСЬКИЙ, канд. техн. наук, доцент кафедри Металургійних технологій навчально-наукового Технологічного інституту ДУЕТ

Рецензент: Олександр УЧИТЕЛЬ, докт. техн. наук, професор кафедри Інжинірингу з галузевого машинобудування навчально-наукового Технологічного інституту ДУЕТ

Відповідальний за випуск: Володимир ЗАСЕЛЬСЬКИЙ, в.о. завідувача кафедри Інжинірингу з галузевого машинобудування, д-р тех. наук., проф.

Конспект лекцій з освітнього компонента «Сучасні інформаційні технології з аналізу даних» для освітньо-наукового ступеня доктора філософії (PhD) 133 «Галузеве машинобудування» очної (денної) та заочної форм навчання. Розроблено у відповідності до навчальнихпланів з метою надання здобувачам допомоги в засвоєнні матеріалів курсу.

ЗМІСТ

Вступ.

Інформаційні технології є невід'ємною частиною сучасного світу, вони значною мірою визначають подальший економічний та суспільний розвиток людства. У цих умовах революційних змін вимагає й система навчання. Звідси можна сказати, що актуальність даного питання має місце у сучасному освітньому середовищі, адже нині якісне викладання дисциплін не може здійснюватися без використання засобів і можливостей, які надають комп'ютерні технології та Інтернет.

Інформаційні технології, ІТ – сукупність методів, виробничих процесів і програмно-технічних засобів, інтегрованих з метою збирання, опрацювання, зберігання, розповсюдження, показу і використання інформації в інтересах її користувачів.

Технології, що забезпечують та підтримують інформаційні процеси, тобто процеси пошуку, збору, передачі, збереження, накопичення, тиражування інформації та процедури доступу до неї.

Незамінним інструментом сучасних дослідників є програми для наукових розрахунків. Розвиток обчислювальних методів зробив можливим розв'язання різноманітних по складності наукових завдань за допомогою обчислювальної техніки. Програмне забезпечення (ПЗ) розробляється багатьма організаціями, як невеликими компаніями, так і великими корпораціями світового рівня. Випускаються спеціалізовані програми для самих різних дисциплін - математики, астрономії, хімії, фізики, біології, лінгвістики, інженерії, розробки штучного інтелекту тощо. ПЗ необхідне як при елементарному відтворенню графіків, так і при обробці великих масивів інформації, зібраної науковими приладами. Більше 90% учених використовують у своїй практиці програми для наукових розрахунків. 50% розробляють власні програми, а майже 70% вважають, що сучасні наукові дослідження неможливі без застосування комп'ютерної техніки і наукового ПЗ [1]. Якісна і стабільна програма для точних розрахунків – це запорука чистоти наукового дослідження і надійності отриманих даних. Єдина помилка в коді програми унеможливлює увесь дослідницький проект. Якщо дані отримані за допомогою програми, результати якої невідтворні, то і результати дослідження будуть недостовірними [1]. Обробка великих масивів статистичної інформації, необхідної для аналізу діяльності, планування виробництва підприємства, підвищення прибутковості галузі господарства, держави може бути виконана лише з використанням сучасних засобів інформаційних технологій. Використання достовірної і науково-обгрунтованої інформації приводить до зменшення витрат, підвищення якості та ефективності виробництва. У зв'язку із зростанням потреби статистичного аналізу даних практично в усіх сферах діяльність, а особливо в науковій, ринок ПЗ для статистичної обробки даних нестримно розвивається.

Тема 1. Основні поняття обробки даних.

Поняття даних, основні завдання обробки даних. Особливості обробки даних. Класифікація та загальний огляд етапів та методів обробки даних.

Найважливішими, з практичної точки зору, властивостями інформації є цінність, достовірність, актуальність та обробка інформації.

Цінність інформації — визначається користю та здатністю її забезпечити суб'єкта необхідними умовами для досягнення ним поставленої мети.

Достовірність — здатність інформації об'єктивно відображати процеси та явища, що відбуваються в навколишньому світу. Як правило достовірною вважається насамперед інформація яка несе у собі безпомилкові та істинні дані. Під безпомилковістю слід розуміти дані які не мають, прихованих або випадкових помилок. Випадкові помилки в даних обумовлені, як правило, неумисними спотвореннями змісту людиною чи збоями технічних засобів при переробці даних в інформаційній системі. Тоді як під істинними слід розуміти дані зміст яких неможливо оскаржити або заперечити.

Актуальність — здатність інформації відповідати вимогам сьогодення (поточного часу або певного часового періоду). Часові властивості Часові властивості визначають здатність даних передавати динаміку зміни ситуації (динамічність). При цьому можна розглядати або час запізнення появи в даних відповідних ознак об'єктів, або розходження реальних ознак об'єкта і тих же ознак, що передаються даними. Відповідно можна виділити: *Актуальність* — властивість даних, що характеризує поточну ситуацію; *Оперативність* — властивість даних, яка полягає в тому, що час їхнього збору та переробки відповідає динаміці зміни ситуації;

Ідентичність — властивість даних відповідати стану об'єкта. Властивість недоступності При розгляді захищеності даних можна виділити технічні аспекти захисту даних від несанкціонованого доступу та соціально-

психологічні аспекти класифікації даних за мірою їхньої конфіденційності та секретності (властивість конфіденційності).

Інші властивості інформації

Суспільна природа — джерелом інформації є пізнавальна діяльність людей, суспільства.

Мовна природа — інформація виражається за допомогою мови — знакової системи будь якої природи, яка служить засобом спілкування, мислення, висловлювання думки. Мова може бути природною, що використовується у повсякденному житті та служить формою висловлення думок і засобом спілкування між людьми а також штучною, створеною людьми з певною метою (наприклад, мова математичної символіки, інформаційно-пошукова, алгоритмічна та ін. мови). Невідривність від мови носія.

Дискретність — одиницями інформації як засобами висловлювання є слова, речення, уривки тексту, а у плані змісту — поняття, висловлювання, описання фактів, гіпотези, теорії, закони тощо. Незалежність від творців Старіння — головною причиною старіння інформації є не сам час, а поява нової інформації, з надходженням якої попередня інформація виявляється невірною, перестає адекватно передавати явища та закономірності матеріального світу, людського спілкування та мислення.

Розсіювання — існування у багатьох джерелах.

Види інформації Інформацію можна поділити на види за кількома ознаками:

<u>За формою подання:</u> За формою подання інформація поділяється на такі види:

Текстова — що передається у вигляді символів, призначених позначати лексеми мови;

Числова — у вигляді цифр і знаків, що позначають математичні дії;

Графічна — у вигляді зображень, подій, предметів, графіків;

Звукова — усна або у вигляді запису передачі лексем мови аудіальним шляхом.

За призначенням:

Масова — містить тривіальні відомості і оперує набором понять, зрозумілим більшій частині соціуму;

Спеціальна — містить специфічний набір понять, при використанні відбувається передача відомостей, які можуть бути не зрозумілі основній масі соціуму, але необхідні і зрозумілі в рамках вузької соціальної групи, де використовується дана інформація;

Особиста — набір відомостей про яку-небудь особистість, що визначає соціальний стан і типи соціальних взаємодій всередині популяції.

Дані (від лат. *data*, множина від лат. *datum* від лат. *dare* – давати, щось дане):

1) відомості, показники, необхідні для ознайомлення з ким, чим- небудь, для характеристики когось, чогось або для прийняття певних висновків, рішень;

2) здібності, якості, необхідні для чого-небудь;

3) форма представлення знань. Тексти, таблиці, інструкції, відомості про факти, явища і таке інше, представлені у буквено-цифровій, числовій, текстовій, звуковій або графічній формі. Дані можуть зберігатися на різних носіях, в тому числі в ЕОМ та пересилатися і піддаватися обробці.

Носіями даних може бути папір, магнітний диск, компакт-диск тощо.

У ході інформаційного процесу дані перетворюються із одного виду в інший за допомогою різних методів. Обробка даних вимагає здійснення багатьох операцій. Серед них можна виділити основні операції:

- *збирання даних* – це накопичення з метою забезпечення достатньої повноти для прийняття рішення;
- *формалізація даних* – приведення даних, що надходять від різних джерел до однакової форми;
- *фільтрація даних* – відсіювання “зайвих” даних, у яких нема необхідності для прийняття рішення;

- *сортування даних* – упорядкування даних за заданою ознакою, що дозволяє підвищити доступність даних;

- *архівація даних* – організація зберігання даних, що дозволяє зменшити витрати для зберігання даних і підвищує надійність інформаційного процесу;

- *захист даних* – заходи, що спрямовані на запобігання втрат, відтворення та модифікацію даних;

- *перетворення даних* – переведення даних із однієї форми в іншу або із однієї структури в іншу, яке часто пов'язане із зміною типу носія [2].

База даних (БД) – це структурована сукупність відомостей про об'єкти певної предметної області та окремих стандартизованих засобів для автоматизації їх обробки. Існують не лише комп'ютерні БД. База даних обліку залізничних квитків – це комп'ютерна БД, журнал академічної групи коледжу – приклад некомп'ютерної БД.

За технологією обробки даних БД поділяють на централізовані та розподілені. У централізованих БД всі дані зберігаються на одному комп'ютері, найчастіше на сервері, але доступ до них можливий з інших комп'ютерів мережі. В розподілених БД різні частини бази зберігаються на різних комп'ютерах. Наприклад, бази даних автоматизованих систем бухгалтерського обліку 1С: Підприємство та Парус-Підприємство є централізованими, а база електронних листів є розподіленою, бо електронні листи зберігаються на серверах і в клієнтах поштових систем. За способом доступу до даних БД поділяють на локальні та мережеві. Локальні БД призначені для обробки даних бази лише на одному комп'ютері, мережеві – для роботи з базою в мережі. Наприклад, автоматизовані системи бухгалтерського обліку в середніх і великих підприємствах використовують мережеві БД. Електронний телефонний довідник, встановлений на комп'ютері, базується на локальній БД. За характером збереження даних розрізняють документальні, фактографічні та документально-фактографічні БД. Документальні БД створені сукупністю неструктурованих текстових

документів (статті, книги, реферати, тексти законів) та графічних об’єктів. Прикладом може бути сукупність файлів окремого диску комп’ютера. Основна ідея фактографічних БД полягає в тому, що усі відомості про об’єкти мають свій формат. Інформація, яка заноситься до бази даних, має чітку структуру. Прикладом роботи з фактографічною БД є реєстрація обліку відвідувань та оцінювання знань студентів в журналі академічної групи.

Дані у базі організовують відповідно до моделі організації даних. Розрізняють наступні моделі організації даних:

• *реляційна* – у цьому випадку БД являє собою одну чи декілька взаємозв’язаних двовимірних таблиць. Для кожного зв’язку можна виділити основну і підпорядковану таблиці. Прикладом БД з реляційною моделлю є БД, що містить таблиці з даними співробітників та здобутою ними освітою;

• *ієрархічна* – БД являє собою сукупність об’єктів різних рівнів, при цьому об’єкти нижчого рівня підпорядковуються об’єктам вищого рівня. Прикладом БД з ієрархічною моделлю є файлова структура дисків, оскільки каталоги нижчих рівнів вкладаються в каталоги вищих рівнів, аж до кореневого каталогу;

• *сіткова* – БД являє собою сукупність об’єктів різних рівнів, проте схема зв’язків між ними може бути довільною (кожен об'єкт може бути підпорядкований будь-якому іншому об'єкту). Надалі під БД будемо розуміти реляційну БД з ієрархічними індексами, основними об'єктами якої є таблиці та зв’язки між ними. Структуру двовимірної таблиці утворюють стовпці і рядки. Їх аналогами в таблиці БД є поля і записи.

Система управління базою даних (СУБД) – це комплекс програмних і мовних засобів загального і спеціального призначення, необхідних для створення БД, підтримування її в актуальному стані, маніпулювання даними і організації доступу до них різних користувачів. Іншими словами, якщо БД – це сховище даних, то СУБД – це прикладна програма для їх обробки.

Автоматизований банк даних (АБД) – це система спеціальним чином організованих даних (баз даних), програмних, технічних, мовних,

організаційно-методичних засобів, які необхідні для забезпечення централізованого нагромадження та колективного багатоцільового використання даних. Зрозуміло, що основними складовими АБД є БД і СУБД. Існує багато систем управління базами даних, наприклад *Oracle, Microsoft Visual FoxPro, dBase, Clipper, C++ Builder, Borland Delphi, Microsoft SQL Server, Microsoft Access та ін*. Інформаційна система (ІС) обробки баз даних – це база даних і комплекс апаратно-програмних засобів для збереження та маніпулювання ними, це прикладна програма, з якою працюють користувачі (наприклад, ІС бухгалтерського обліку, продажу залізничних квитків та ін.).

Зазвичай, з БД працюють три категорії персоналу: постановники та проектувальники, розробники і користувачі.

Перша категорія – постановники та проектувальники. Їх завдання полягає, насамперед, в розробці структури таблиць бази даних і узгодженні її з замовником.

Друга категорія – розробники. Вони реалізують розроблену структуру таблиць та схему БД за допомогою обраної СУБД. Також розробники створюють і вдосконалюють інші об'єкти ІС, призначені для автоматизації роботи з БД.

Третя категорія – користувачі. Вони одержують ІС та початкову базу даних від розробників і займаються її веденням для задоволення інформаційних потреб організації. Користувачі опрацьовують лише ті дані, робота з якими передбачена на конкретному робочому місці.

У загальному випадку класифікацією (розпізнаванням образів) називають поділ досліджуваної сукупності об’єктів на однорідні в певному розумінні групи (класи) або зарахування кожного із заданої множини об’єктів до деякого із заздалегідь відомих класів. При цьому вирізняють три групи завдань: дискримінацію, кластеризацію й групування. Останні дві групи є близькими за метою (поділ даних на класи або групи близьких у певному розумінні об’єктів), а також за алгоритмами. Але принципова різниця між ними полягає у тому, що у першому випадку межі класів є природними, а у другому – умовними й їх можна встановлювати суб’єктивно [2] .

Тема 2. Перевірка статистичних гіпотез

Основні поняття. Параметричні тести. Непараметричні тести. Визначення моделей розподілу емпіричних даних. Приклад ідентифікації функції розподілу однорідної вибірки.

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

– якою є мета перевірки гіпотези;

– у яких шкалах виміряні аналізовані дані;

– чи є аналізовані вибірки незалежними або спряженими;

– скільки вибірок необхідно порівняти.

Розглянуті в цьому розділі методи застосовують при порівнянні двох вибірок. При більшій кількості вибірок використовують методи дисперсійного аналізу [3].

Гіпотезу, що перевіряють, називають нульовою гіпотезою (H0). Прикладами нульових гіпотез можуть бути такі твердження:

I. “Середні значення двох вибірок суттєво не відрізняються одне від одного”;
II. “Дисперсія першої вибірки суттєво перевищує дисперсію другої”;
III. “Розподіл вибірки відповідає нормальному закону з певними параметрами”.

Гіпотезу, що суперечить нульовій, називають конкуруючою, або альтернативною гіпотезою (H1). Для вказаних вище нульових гіпотез конкуруючими можуть бути такі твердження:

I. “Середні значення двох вибірок суттєво розрізняються одне від одного”;
II. “Дисперсія першої вибірки не перевищує істотно дисперсію другої”;
III. “Розподіл вибірки не відповідає нормальному закону із вказаними параметрами”.

Для однієї нульової гіпотези у загальному випадку можна

сформулювати багато різних альтернативних гіпотез. Розрізняють прості та складні гіпотези. Простою називають гіпотезу, що містить тільки одне твердження. Складні гіпотези складаються з декількох простих (при цьому кількість простих гіпотез може бути нескінченно великою).

Зазвичай при перевірці нульової гіпотези використовують певні модельні розподіли, що приблизно відповідають розподілу досліджуваного параметра. Їх називають статистичними критеріями. На практиці як критерії найчастіше використовують нормальний розподіл, χ^2 -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. Вирізняють *двобічні* й *однобічні* (лівобічні, правобічні) критичні області (рис. 1, 2). Їх використання залежить від вибору конкуруючої гіпотези.

Якщо розподіл імовірності спостережень, що відповідає нульовій гіпотезі H0, є відомим, то критичну область визначають так, щоб при виконанні H0 імовірність її відхилення була рівною заздалегідь заданій малій величині (рівню значущості).

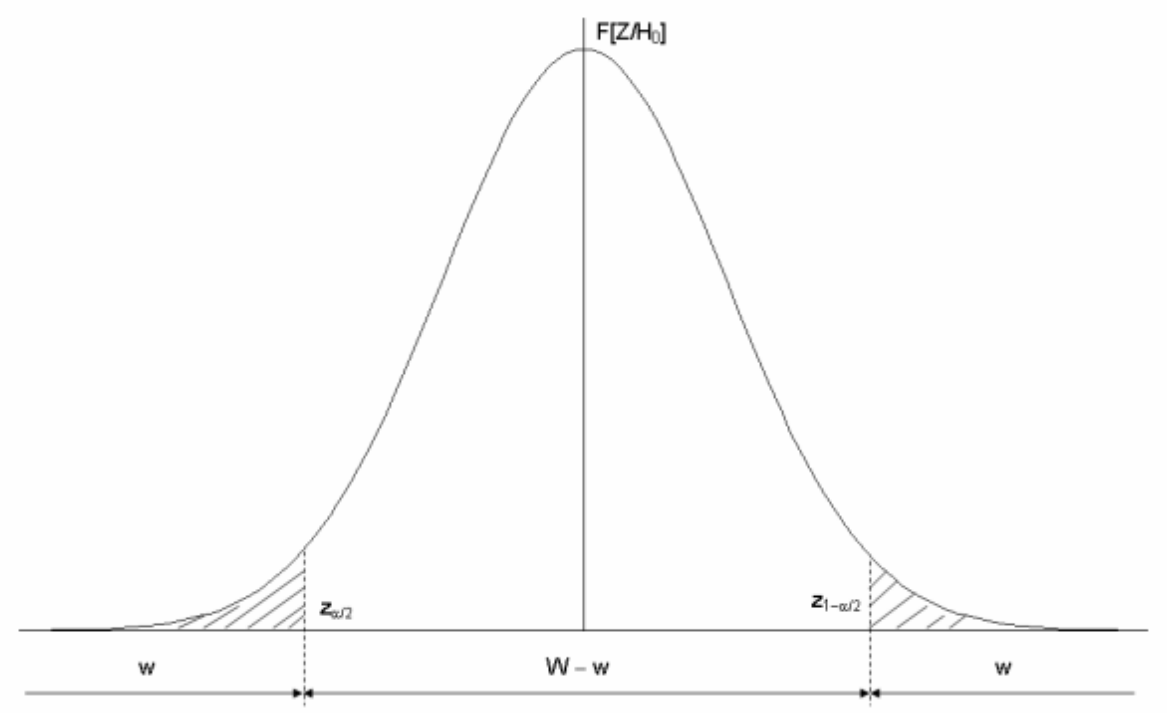


Рис. 1. Приклад двобічної критичної області

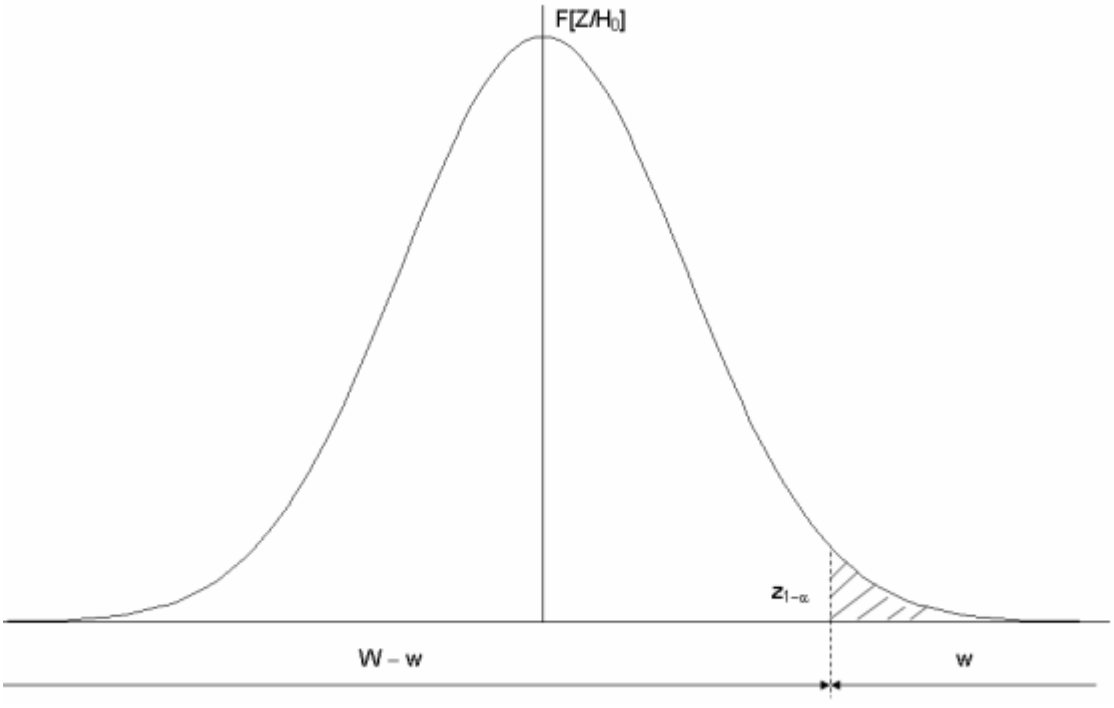


Рис. 2. Приклад правобічної критичної області

Критерії, що базуються на використанні заздалегідь заданого рівня значущості, називають *критеріями значущості*. Рівень значущості визначає розмір критичної області: що більшим є рівень значущості, то ширшою буде критична область.

Розглядають два типи помилок, що можуть виникати при перевірці статистичних гіпотез:

– помилкою першого роду є відхилення правильної нульової гіпотези, рівень значущості α є ймовірністю такої помилки;

– помилкою другого роду є прийняття помилкової нульової гіпотези. У деяких застосуваннях помилки першого та другого роду називають, відповідно, ризиком виробника та ризиком споживача.

Можливості сучасної комп'ютерної техніки та наявного програмного забезпечення дають змогу отримувати висновки іншим шляхом. Якщо за наявними емпіричними даними розрахувати значення критерію, то на наступному етапі можна визначити, для якого рівня значущості це значення буде критичним. Ураховуючи, що рівень значущості є ймовірністю відхилення правильної нульової гіпотези, ми можемо за його значенням зробити висновок про ймовірність правильності або помилковості нульової гіпотези. Залежно від того, задовольняє нас отримана ймовірність помилки чи ні, нульову гіпотезу приймають або відхиляють.

При перевірці гіпотез доцільно застосовувати різні методи, призначені для вирішення одних й тих самих завдань та однакових типів даних. Причинами розбіжності отримуваних при цьому результатів зазвичай є:

– помилки при введенні даних;

– непридатність окремих методик для типу даних, що розглядають;

– алгоритмічні помилки у програмах, що використовують для аналізу.

Залежно від наявності або відсутності можливості визначення напряму розбіжності порівнюваних вибірок, розрізняють однобічні та двобічні критерії. Перші застосовують, якщо наявні дані дають змогу вказати такий напрям, наприклад зробити висновок, що значення порівнюваної ознаки для

одної вибірки є вищим, ніж в іншої.

Двобічні критерії дають можливість зробити висновок лише про різницю вибірок за порівнюваною ознакою. Відповідно до цього говорять про *однобічні* й *двобічні* гіпотези. Для двобічних критеріїв рівень значущості є вдвічі більшим, ніж для відповідних однобічних. При використанні однобічних критеріїв рекомендується спочатку розраховувати двобічні. Якщо за двобічним критерієм різниці між вибірками немає, то наступне порівняння за однобічним є необґрунтованим [3] .

Параметричні тести

Критерії й тести, що застосовують для порівняння вибірок, поділяють на дві групи: параметричні й непараметричні. Особливістю параметричних критеріїв є припущення, що розподіл ознаки в генеральній сукупності підпорядковується певному відомому закону. Ця відповідність має бути доведена до застосування будь-якого з параметричних тестів. Переважна більшість параметричних тестів розроблена для нормально розподілених даних. Але для деяких типів гіпотез існують параметричні тести, призначені для вибірок, що підпорядковуються іншим законам розподілу. Як правило, параметричні критерії є потужнішими за непараметричні. Застосування непараметричних критеріїв у випадках, коли можна використовувати параметричні, призводить до збільшення ймовірності прийняття помилкової нульової гіпотези, тобто помилки другого роду [3].

Непараметричні тести

У багатьох випадках емпіричні дані не задовольняють нормальний розподіл. Тому для їх аналізу некоректно застосовувати параметричні тести. Серед непараметричних тестів важливе місце займають так звані робастні методи, що виявляють слабку чутливість до відхилень від стандартних умов і можуть використовуватися в широкому діапазоні реальних умов. При перевірці нульової гіпотези про однорідність вибірок числових даних рекомендується [3] використовувати омега-квадрат критерій або (за відсутності необхідних таблиць та програмного забезпечення) критерій

Смірнова. Критерій омега-квадрат (критерій Крамера – фон Мізеса) базується на розгляді відхилення між двома емпіричними функціями розподілу (або між емпіричною й теоретичною функціями розподілу при ідентифікації закону розподілу). Вперше його запропонували у 1928–1930 р. 53 шведський математик Карл Харальд Крамер та американський математик і механік Ріхард фон Мізес, який народився у Львові.

Визначення моделей розподілу емпіричних даних

На практиці часто виникає проблема перевірки відповідності емпіричного розподілу деякому заданому теоретичному. При цьому вирізняють прості та складні гіпотези. Якщо гіпотеза стверджує, що із A параметрів розподілу k мають задані значення, то гіпотезу вважають простою, коли k = A , і складною – якщо k < A . Різницю A − k називають кількістю степенів вільності гіпотези, а k – кількістю накладених обмежень. Особливу роль відіграє перевірка розподілу на нормальність, оскільки її прийняття дає змогу застосовувати більш досліджені параметричні критерії перевірки наступних гіпотез.

Для перевірки відповідності емпіричного розподілу теоретичному застосовують так звані критерії згоди: *ω^2 , Смірнова, χ^2, Ястремського, Бернштейна та інші. Критерій ω^2 (Крамера – фон Мізеса) запропонований в 1928–1930 р. К. Крамером та Р. фон Мізесом.*

Його використовують у випадках, коли необхідно перевірити нульову гіпотезу про відповідність ви- 62 бірки певному відомому закону розподілу.

Розрахункове значення обчислюють за формулою:

$$n\omega^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left[F(x_i) - \frac{2i-1}{2n}\right]^2,$$

де F (x) = теоретична функція розподілу,

n – обсяг вибірки.

При n > 40 критичні значення визначають згідно з табл. 1 [3]. Якщо значення параметрів розподілу визначають за вибіркою, то критичні значення суттєво зменшуються.

Таблиця 1

Критичні значення статистики ω^2

α	0,900	0,950	0,990	0,995	0,999
$n\omega^2(\alpha)$	0,3473	0,4614	0,7435	0,8694	1,1679

Ще одним способом перевірки типу розподілу є побудова емпіричної функції розподілу в певних координатах, що лінеаризують її графік. У статистичних пакетах *SPSS, Statistica* та інших реалізовано спеціальні засоби такої перевірки. Як приклад на рис. 3, 4 наведені лінеаризовані графіки емпіричних функцій розподілу вибірок з генеральних сукупностей, що підпорядковуються рівномірному розподілу на відрізку [−3, 3] і стандартному нормальному розподілу. На рис. 3 графіки побудовано в координатах, що мають лінеаризувати функцію нормального розподілу, а на рис. 4 – функцію рівномірного розподілу.

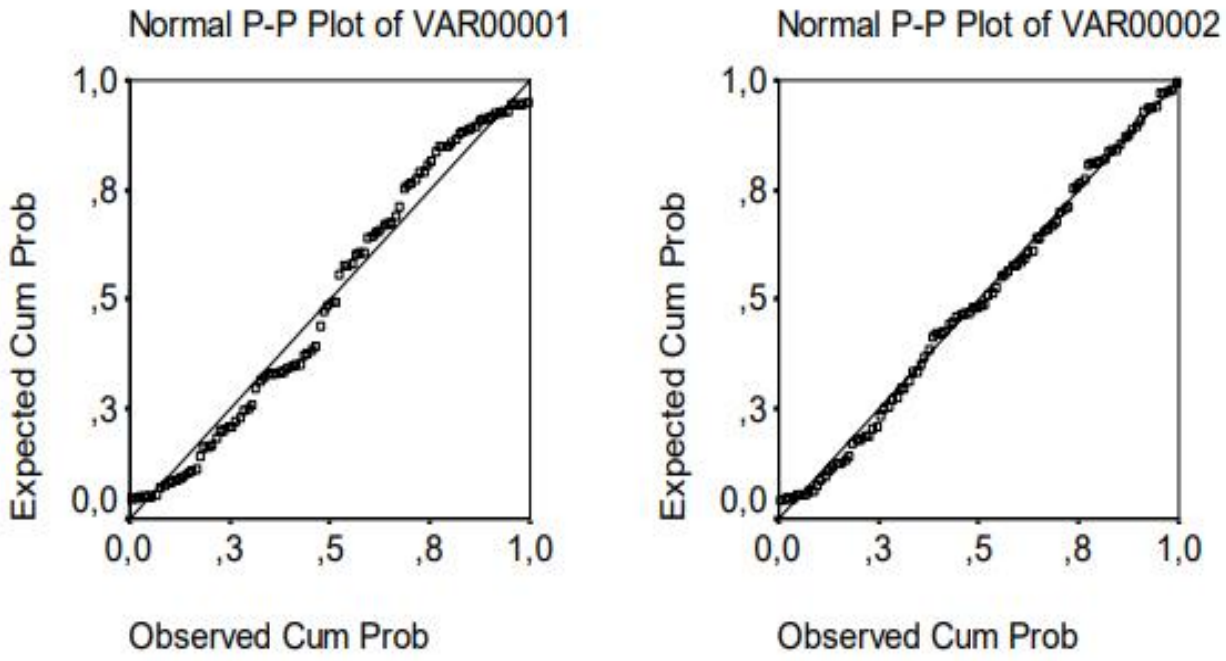


Рис. 3. P-P діаграми рівномірно й нормально розподілених вибірок у координатах, що лінеаризують функцію нормального розподілу

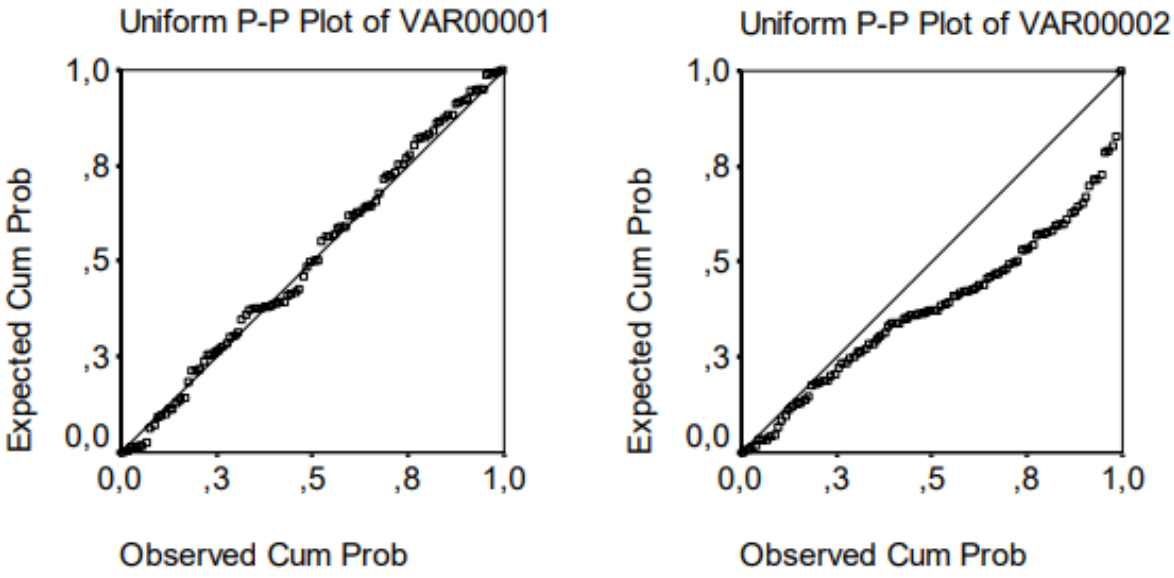


Рис. 4. P-P діаграми рівномірно й нормально розподілених вибірок у координатах, що лінеаризують функцію рівномірного розподілу

Приклад ідентифікації функції розподілу однорідної вибірки

Ідентифікація емпіричних функцій розподілу є відносно простою для однорідних вибірок. У цьому випадку її алгоритм може бути таким:

1. За допомогою P-P діаграм статистичних пакетів (*SPSS, Statistica* тощо) підбираємо найбільш придатний тип розподілу.

2. Використовуючи статистичні пакети або мінімізуючи суму квадратів залишків моделі за допомогою процедури “Пошук розв’язку” електронних таблиць *MS Excel*, уточнюємо параметри розподілу.

3.Перевіряємо адекватність підібраної моделі розподілу, використовуючи критерії Крамера – Уелча та Фішера (для нормального розподілу), ω^2 або Смирнова (для інших типів розподілу).

4. Розглянемо як приклад завдання ідентифікації моделі розподілу питомого електричного опору епітаксійних шарів кремнієвих композицій [1].

До робочого вікна пакету *SPSS* (рис.5) уводимо значення елементів досліджуваної вибірки, що є значеннями питомого електричного опору епітаксійного шару досліджуваної серії виробів. У головному меню обираємо пункти *Graphs / P-P Plots*. При цьому відчиняється діалогове вікно (рис. 6).

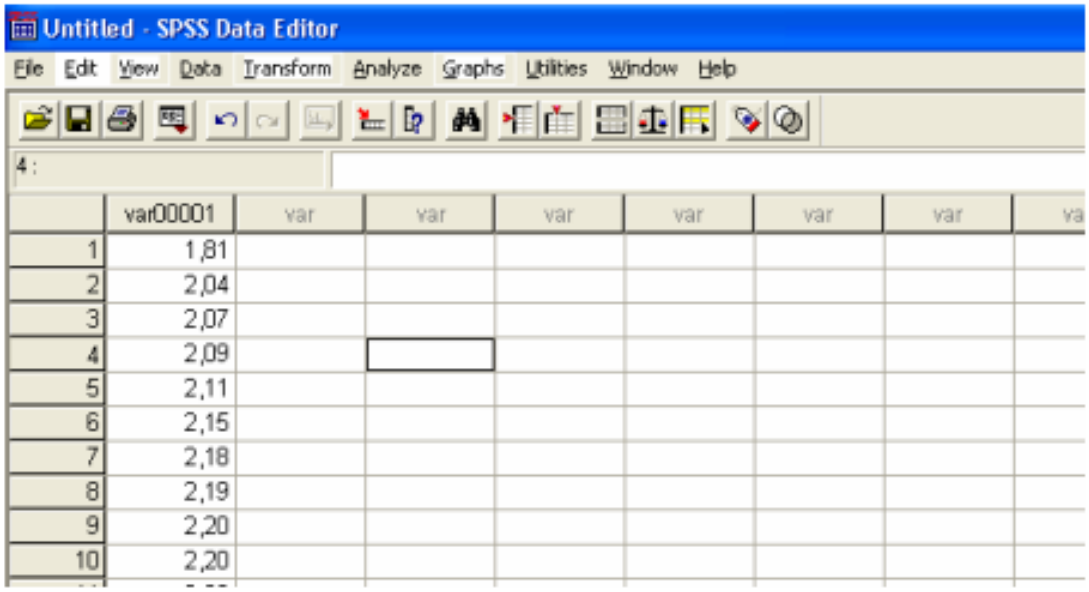


Рис. 5. Головне вікно для уведення даних пакету *SPSS*

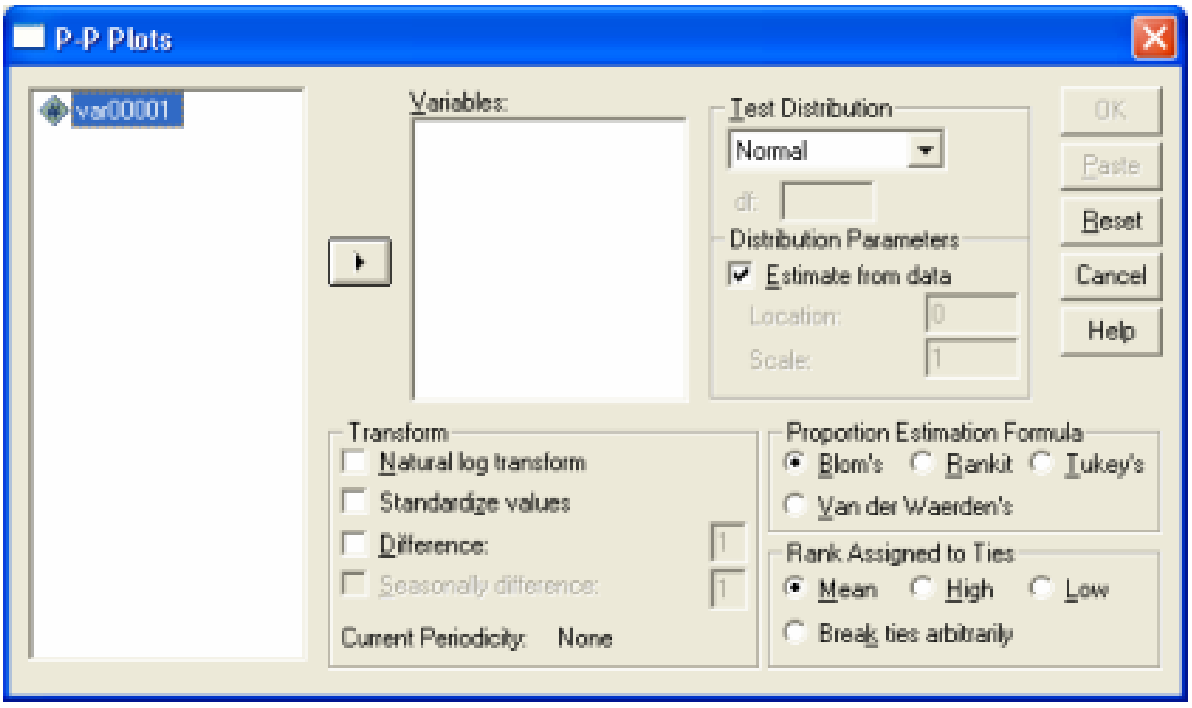
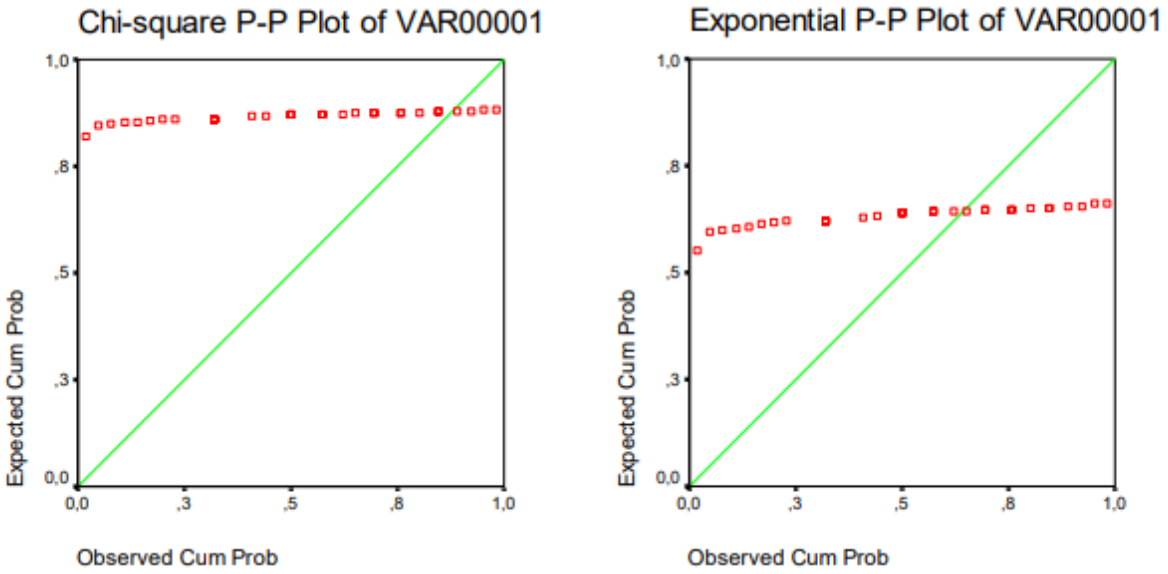


Рис.6. Діалогове вікно побудови P-P діаграм

У цьому вікні зазначаємо, для якої вибірки треба побудувати P-P діаграму, а також, який саме тип розподілу перевірятимемо. При цьому можливо вибрати такі типи розподілу: бета, χ^2 , експоненціальний, гама, напівнормальний, Лапласа, логістичний, логнормальний, нормальний, Парето, Стьюдента, Вейбула, рівномірний.

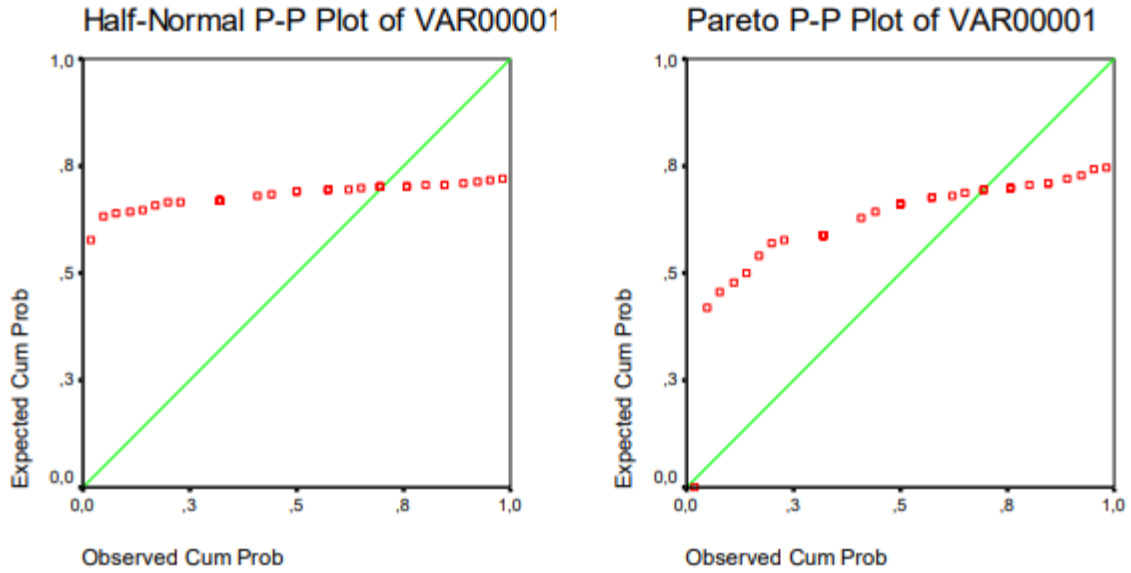
Для досліджуваної вибірки одержуємо такі результати.



а) **б)**

Рис.7. Р-Р діаграма досліджуваної вибірки

а) для χ^2 розподілу; б) для експоненціального розподілу



а) **б)**

Рис. 8. Р-Р діаграма досліджуваної вибірки

а) для напівнормального розподілу; б) для розподілу Парето

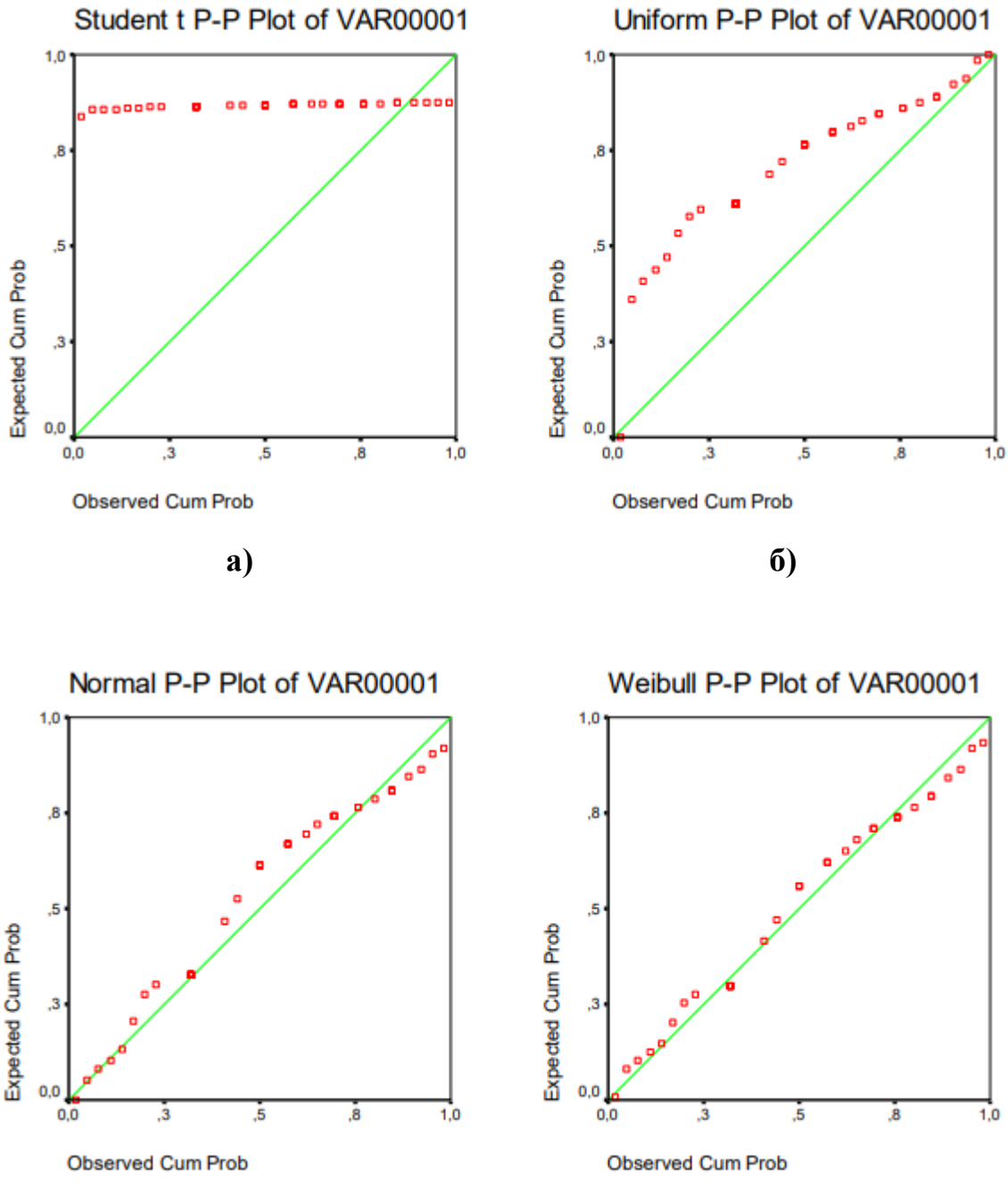


Рис. 9. P-P діаграма досліджуваної вибірки

а) для розподілу Стьюдента; б) для рівномірного розподілу

в) для нормального розподілу; г) для розподілу Вейбулла

З наведених P-P діаграм бачимо, що найбільш придатним для опису досліджуваної вибірки є розподіл Вейбулла.

На рис. 10 наведено графіки емпіричної функції розподілу, а також підібраних теоретичних моделей.

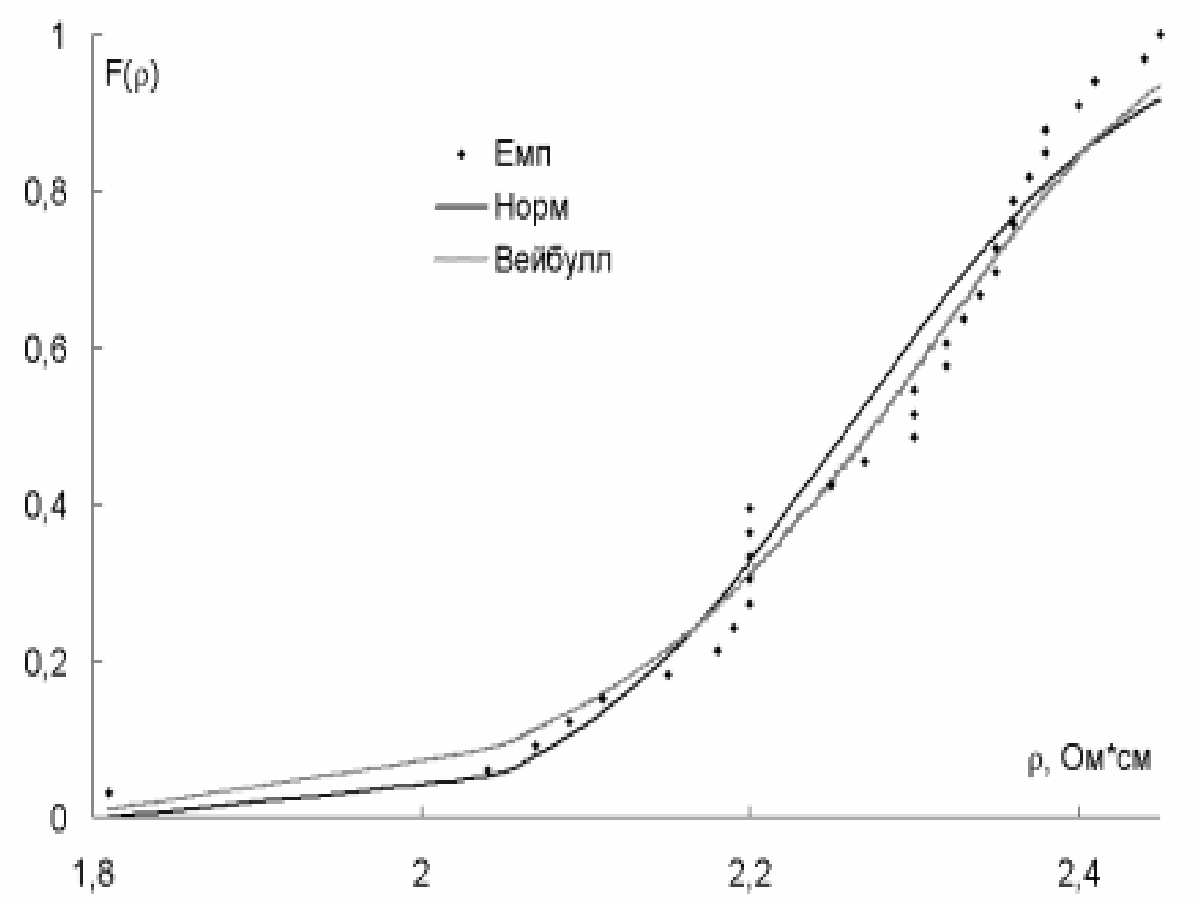


Рис. 10. Емпірична й теоретичні функції розподілу досліджуваної вибірки

Тема 3. Дисперсійний аналіз

Однофакторний аналіз. Двофакторний аналіз.

Дисперсійний аналіз є сукупністю статистичних методів, призначених для перевірки гіпотез про зв'язок між певною ознакою та досліджуваними факторами, які не мають кількісного опису, а також для встановлення ступеня впливу факторів та їх взаємодії.

У спеціальній літературі дисперсійний аналіз часто називають **ANOVA** (від англомовної назви *Analysis of Variations*). Вперше цей метод було розроблено **Р. Фішером в 1925 р.**

Факторами називають контрольовані чинники, що впливають на кінцевий результат. Рівнем фактора, або способом обробки, називають значення, що характеризують конкретний прояв цього фактора. Ці значення зазвичай подають у номінальній або порядковій шкалі вимірювань. Значення вимірюваної ознаки називають відгуком. Часто вихідні значення факторів вимірюють у кількісних або порядкових шкалах. Тоді постає проблема групування вихідних даних у ряди спостережень, що відповідають приблизно однаковим значенням фактора. Якщо кількість груп взяти надмірно великою, то кількість спостережень у них може виявитися недостатньою для отримання надійних результатів. Якщо її взяти надмірно малою, це може призвести до втрати суттєвих особливостей впливу досліджуваного фактора на систему. Загальну методологію групування описано в розділі 1. Вибір конкретного способу групування даних залежить від їх обсягу і характеру варіювання значень фактора. Кількість і розміри інтервалів при однофакторному аналізі найчастіше визначають за принципом рівних інтервалів або за принципом рівних частот. При багатофакторному аналізі застосовують три типи групування:

– групи з рівною кількістю спостережень;

– групи з різною кількістю спостережень;

– групи, кількості спостережень у яких відповідають певній пропорції. При цьому існують певні особливості обробки даних, залежно від типу групування, які не розглядаються у цьому посібнику.

Однофакторний аналіз

Основною метою однофакторного аналізу зазвичай є оцінка величини впливу конкретного фактора на досліджуваний відгук. Іншою метою може бути порівняння двох або декількох факторів один з одним з метою визначення різниці їх впливу на відгук, яку часто називають контрастом факторів. Попереднім етапом є перевірка нульової гіпотези про відсутність будь-якого впливу досліджуваного фактора (факторів), тобто гіпотези про те, що зміни значень ознаки в порівнюваних вибірках є випадковими, і всі дані належать до однієї генеральної сукупності.

Якщо нульову гіпотезу відкидають, то наступним етапом є кількісне оцінювання впливу досліджуваного фактора і побудова довірчих інтервалів для отриманих характеристик. У випадку, коли нульова гіпотеза не може бути відкинутою, зазвичай її приймають і роблять висновок про відсутність впливу. Але, якщо є підстави вважати, що такий вплив має бути присутнім (наприклад, це може випливати з теоретичних уявлень про об’єкт дослідження), то необхідно перевірити наявність інших факторів, що можуть його маскувати.

При однофакторному дисперсійному аналізі вихідні дані подають у вигляді таблиць, у яких кількість стовпчиків дорівнює кількості рівнів фактора, а кількість значень у кожному стовпчику – кількості спостережень при відповідному рівні фактора (табл.2). Для різних рівнів фактора кількість спостережень може бути різною. При цьому виходять з припущення, що результати спостережень для різних рівнів є вибірками з нормально розподілених сукупностей, середні значення та дисперсії яких є однаковими і не залежать від рівнів. Завданням аналізу є перевірка нульової гіпотези про рівність середніх значень сукупностей, що розглядаються [2] .

Таблиця 2

Форма таблиці спостережень при проведенні однофакторного дисперсійного аналізу

Результати вимірювань	Рівні фактора			
	1	2	…	k
1	x_{11}	x_{12}	…	x_{1k}
2	x_{21}	x_{22}	…	x_{2k}
…	…	…	…	…
n_i	$x_{n_1 1}$	$x_{n_1 2}$	…	$x_{n_1 k}$

Двофакторний аналіз

Двофакторний дисперсійний аналіз застосовують для пов'язаних нормально розподілених вибірок. Дані подають у вигляді таблиці 3, у стовпчиках якої наводять дані, що відповідають певному рівню першого фактора, а в рядках – дані, що відповідають рівням другого. Таблиця даних має розмірність n×k, де n і k – кількість рівнів першого та другого факторів, відповідно.

Таблиця 3

Таблиця даних двофакторного дисперсійного аналізу

Рівні фактора B	Рівні фактора A			
	1	2	…	k
1	x_{11}	x_{12}	…	x_{1k}
2	x_{21}	x_{22}	…	x_{2k}
…	…	…	…	…
n	x_{n1}	x_{n2}	…	x_{nk}

Основною відмінністю від таблиці однофакторного дисперсійного аналізу є можлива неоднорідність даних у стовпцях, якщо вплив другого фактора є суттєвим. На практиці часто використовують і складніші таблиці

двофакторного дисперсійного аналізу, зокрема такі, у яких кожна комірка містить набір даних (повторні вимірювання), що відповідають фіксованим значенням рівнів обох факторів.

У пакеті SPSS також передбачено можливість здійснення однофакторного дисперсійного аналізу. Для цього необхідно використовувати вікно *Analyze/Compare means/One way* **ANOVA.**

Вихідні вибірки розміщуємо по- 96 слідовно одна за одною в одному й тому самому стовпчику, як значення змінної VAR00001. Як значення змінної VAR00002 беремо номери вибірок, до яких належить відповідне значення VAR00001. Додатково ми можемо отримати основні показники описової статистики для кожної з вибірок та усієї сукупності в цілому; перевірити однорідність дисперсій (тест Левена); визначити, які саме вибірки істотно відрізняються від інших (тест Дункана), й отримати іншу корисну інформацію. Якщо для вибірок, що розглядаються у прикладі, замовити у пункті меню *“Options”* розрахунок параметрів описової статистики й перевірку однорідності дисперсій, а пункті *“Post Hoc”* – виконання тесту Дункана, одержимо результати, показані на рис. 11.

Descriptives

VAR00001

		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum	Between-Component Variance
						Lower Bound	Upper Bound			
1,00		200	50,3430	5,17707	,36607	49,6211	51,0649	36,15	64,18	
2,00		200	49,3706	4,88349	,34531	48,6896	50,0515	36,24	60,87	
3,00		200	49,8749	5,70403	,40334	49,0796	50,6703	34,38	62,45	
4,00		200	49,3866	5,28285	,37365	48,6500	50,1233	35,89	64,92	
Total		800	49,7438	5,27546	,18652	49,3777	50,1099	34,38	64,92	
Model	Fixed Effects			5,27008	,18633	49,3780	50,1095			
	Random Effects				,23150	49,0071	50,4805			,07549

Test of Homogeneity of Variances

VAR00001

Levene Statistic	df1	df2	Sig.
1,132	3	796	,335

ANOVA

VAR00001

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	128,617	3	42,872	1,544	,202
Within Groups	22107,895	796	27,774		
Total	22236,512	799			

VAR00001

	VAR00002	N	Subset for alpha = .05
			1
Duncan[a]	2,00	200	49,3706
	4,00	200	49,3866
	3,00	200	49,8749
	1,00	200	50,3430
	Sig.		,093

Рис. 11. Результати, отримані у пакеті SPSS

Тема 4. Кореляційний аналіз

Кореляційний аналіз кількісних ознак. Кореляційний аналіз порядкових ознак. Кореляційний аналіз номінальних ознак. Кореляційний аналіз змішаних ознак. Множинна кореляція

Кореляцією (кореляційним зв'язком) між випадковими величинами (ознаками) називають наявність статистичного або ймовірнісного зв'язку між ними. При цьому закономірна зміна певних ознак призводить до закономірної зміни середніх значень інших, пов'язаних з ними ознак. Кореляційним аналізом називають сукупність методів виявлення кореляційного зв'язку. Тому його можна застосовувати для формалізованого подання моделей зв'язків між окремими компонентами системи або між окремими процесами, що відбуваються в ній. Наявність кореляційного зв'язку не означає існування причинно-наслідкового зв'язку між досліджуваними ознаками. Вона може бути зумовлена тим, що обидві ознаки мають причинно-наслідковий зв'язок з певним іншим фактором. Наприклад, існує кореляція між цінами на нафту й на золото. Проте вона пояснюється тим, що обидві ціни виражаються у доларах США й залежать від динаміки його індексу. Кореляція також може бути випадковою. Сучасну класифікацію мір подібності запропонували австрійський та американський біостатистик та антрополог Роберт Сокал та британський таксономіст Пітер Сніс у 1963 р. Згідно з нею виокремлюють такі типи мір подібності [4,5]:

– міри асоціації, що відбивають різні співвідношення кількості ознак, що збігаються до загальної кількості ознак, а також близькі до них коефіцієнти спряженості (квантифіковані коефіцієнти зв'язку);

– вибіркові коефіцієнти зв'язку типу кореляції (нормовані косинусні міри);

– показники відстані у метричному просторі.

Перевірку зв'язку можна здійснювати лише для пов'язаних вибірок. Це означає, що між елементами обох досліджуваних вибірок існує взаємно однозначна відповідність, а кількість елементів у вибірках є однаковою.

Кореляційний аналіз здійснюють на початковому етапі вирішення всіх основних проблем статистичного аналізу даних [4]. У проблемі статистичного аналізу залежностей і побудови регресійних моделей він дає змогу встановити сам факт існування зв'язку між змінними та оцінити ступінь його прояву. У проблемі класифікації даних за допомогою кореляційного аналізу отримують вихідну інформацію у вигляді коваріаційних і кореляційних матриць та інших характеристик парних порівнянь. Це дає змогу визначити подібні один до одного або до певних еталонів об'єкти, сформувати класи подібних об'єктів і здійснити класифікацію. У

проблемі зменшення розмірності досліджуваного простору ознак також за допомогою коваріаційних і кореляційних матриць визначають ознаки, що можуть бути без втрати суттєвої інформації подані через інші наявні дані. Загальна методика перевірки гіпотези про існування зв'язку між ознаками передбачає три основних етапи: визначення типу даних; перевірку гіпотези про відсутність зв'язку і, в разі її відхилення, оцінювання сили зв'язку. Тип вихідних даних суттєво впливає на вибір методів і критеріїв, які можна застосовувати на наступних етапах аналізу. Для визначення сили зв'язку використовують різноманітні показники. Зазвичай їх прагнуть вибрати такими, щоб вони змінювалися від –1 до +1 або від 0 до 1. Значення, що є близькими за модулем до одиниці, свідчать про наявність сильного зв'язку. Близькі до нуля значення вказують або на відсутність будь-якого зв'язку, або на відсутність зв'язку того типу (найчастіше лінійного), для якого розроблено відповідний коефіцієнт. Знак коефіцієнта вказує на напрям зв'язку: прямий (для додатних значень) або зворотний (для від'ємних).

Кореляційний аналіз кількісних ознак

Для кількісних ознак найчастіше застосовують коефіцієнти кореляції Пірсона і Фехнера. Коефіцієнт кореляції ***Пірсона*** (коефіцієнт кореляційного відношення Пірсона, парний коефіцієнт кореляції, вибірковий коефіцієнт кореляції, коефіцієнт *Бравайса – Пірсона*) вимірює ступінь лінійного кореляційного зв'язку між кількісними скалярними ознаками. Він був запропонований К. Пірсоном у 1896 р. Часто, посилаючись на згадування К. Пірсона про ідеї математичного подання зв'язку, висловлені в 1846 р. відомим французьким фізиком та кристалографом Огюстом Браве, цей показник називають коефіцієнтом Бравайса – Пірсона (Бравайс – це викривлена транскрипція від французького ***Bravais***, що закріпилася в літературі з кореляційного аналізу).

Застосування коефіцієнта Пірсона як міри зв'язку є обґрунтованим лише за умови, що спільний розподіл пари ознак є нормальним. Тому перед його розрахунком слід перевірити виконання цієї гіпотези. Якщо вона справедлива, то квадрат коефіцієнта кореляції Пірсона дорівнює коефіцієнту детермінації. Значення коефіцієнта кореляції може змінюватися від −1 до +1. Значення −1 та +1 відповідають чіткій лінійній функціональній залежності, яка в першому випадку є спадною, а у другому – зростаючою.

Для функціональної залежності ***y const*** = коефіцієнт кореляції, як видно з наведеної формули, є невизначеним, оскільки в цьому випадку знаменник дорівнює нулю. Що ближчим є значення коефіцієнта кореляції до −1 або +1, то більш обґрунтованим є припущення про наявність лінійного зв'язку. Наближення його значення до нуля свідчить про відсутність лінійного зв'язку, але не є доказом

відсутності статистичного зв’язку взагалі. На рис.12 показано дві серії точок, координати яких відповідають двом парам спряжених вибірок.

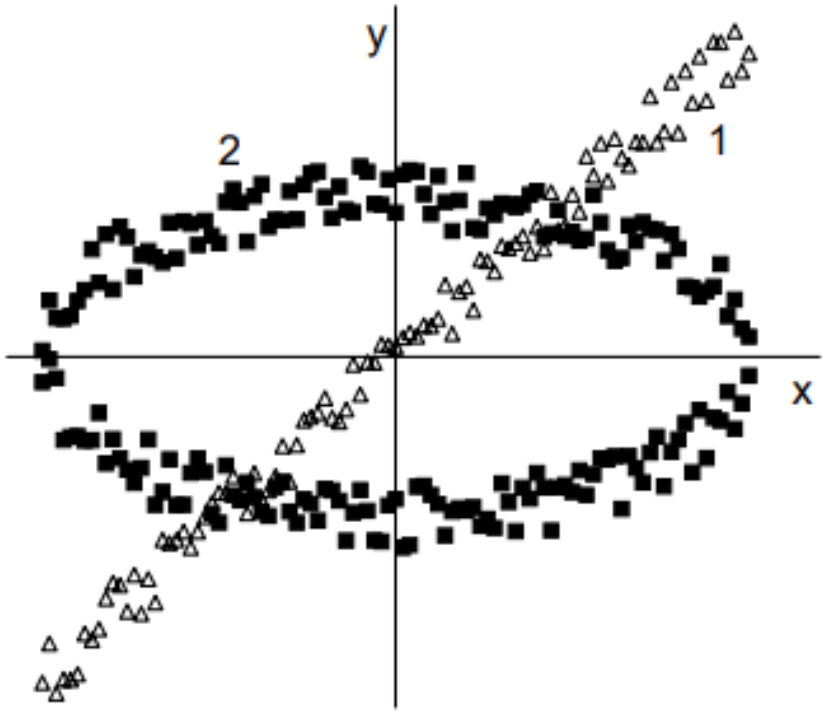


Рис. 12. Графічне зображення двох наборів тестових даних

Показаний приклад свідчить, що в багатьох випадках для попереднього аналізу припущення про наявність і тип зв’язку між певними ознаками доцільно нанести наявні дані на графік.

Кореляційний аналіз порядкових ознак

Під ранговою кореляцією розуміють статистичний зв’язок між порядковими ознаками. Вихідні дані зазвичай подають у вигляді табл. 4, де елемент x_{ik} є рангом *i*-го об’єкта за *k*-ю властивістю.

Таблиця 4.

Таблиця вихідних даних для рангового кореляційного аналізу

Порядковий номер об’єкта	Порядковий номер досліджуваної ознаки						
	0	1	2	…	*k*	…	*p*
1	x_{10}	x_{11}	x_{12}	…	x_{1k}	…	x_{1p}
2	x_{20}	x_{21}	x_{22}	…	x_{2k}	…	x_{2p}
…	…	…	…	…	…	…	…
i	x_{i0}	x_{i1}	x_{i2}	…	x_{ik}	…	x_{ip}
…	…	…	…	…	…	…	…
n	x_{n0}	x_{n1}	x_{n2}	…	x_{nk}	…	x_{np}

Завданнями аналізу в цьому випадку можуть бути: вивчення структури досліджуваних об’єктів; перевірка сукупної узгодженості ознак та умовне ранжирування об’єктів за ступенем тісноти зв’язку кожної з них з іншими ознаками; побудова єдиного групового впорядкування об’єктів (задача регресії на порядкових змінних).

Найхарактернішими типами структури є такі:

1. Аналізовані точки рівномірно розкидані по всій області їх можливих значень. Це означає відсутність будь-якого зв’язку між досліджуваними ознаками.

2. Частина точок утворює ядро (кластер) із точок, що розташовані близько одна до одної, а інші випадково розкидані навколо цього ядра. Це відповідає існуванню підмножини узгоджених ознак.

3. Аналізовані точки утворюють декілька кластерів, розташованих відносно далеко один від одного. Це відповідає наявності декількох таких підмножин ознак, що існує істотний статистичний зв’язок між ознаками, які належать до однієї і тієї самої підмножини, і не існує значущого зв’язку між ознаками, які належать до різних підмножин.

Прикладом завдання другого типу є визначення узгодженості думок групи експертів з наступним впорядкуванням їх за рівнем компетентності. Для цього розраховують коефіцієнти конкордації для різних сукупностей досліджуваних змінних. Вирішення завдань третього типу зводиться до побудови такого впорядкування, яке б у певному значенні було б найближчим до кожного з наданих впорядкувань досліджуваних ознак. Для цього часто застосовують середнє арифметичне або медіану наявних базових рангів. Це можна розглядати як задачу найкращого у певному розумінні відновлення невідомого ранжирування за наявними емпіричними даними, що зумовлює можливість її розгляду як задачі регресії.

Кореляційний аналіз змішаних ознак

Типовою ситуацією, коли необхідна перевірка зв’язку між номінальними ознаками, є обробка результатів соціологічних досліджень, що можуть містити такі комбінації ознак, як освіта, стать, професія, підтримка певної політичної партії, регіон проживання тощо. При дослідженні зв’язків між категоризованими ознаками вихідні дані подають у вигляді таблиці спряженості (табл. 5). До категоризованих зараховують номінальні ознаки, а також порядкові ознаки, для яких є відомим скінченний набір можливих градацій. Величини f_{ij} показують, скільки разів зустрічалася комбінація ознак, за якої рівень першої має значення i, а рівень другої – j; m_j є сумами стовпців, а n_i – сумами рядків. За даними табл. 5 можна оцінити значення ймовірностей.

Таблиця 5

Таблиця спряженості категоризованих ознак

Рівні ознаки 1	Рівні ознаки 2				Разом
	1	2	…	r	
1	f_{11}	f_{12}	…	f_{1r}	n_1
2	f_{21}	f_{22}	…	f_{2r}	n_2
…	…	…	…	…	…
c	f_{c1}	f_{c2}	…	f_{cr}	n_c
Разом	m_1	m_2	…	m_r	S

Існує велика кількість показників ступеня тісноти статистичного зв'язку, призначених для категоризованих змінних, які не є універсальними, а відображають окремі властивості такого зв'язку.

Хеммінгова відстань (метрика Хеммінга) $H = a+d$, також може застосовуватися для визначення кореляції. Проте, як і коваріація, вона не є безрозмірною величиною і може набувати будь-яких невід'ємних значень (верхньою межею є загальна кількість спостережень n). Цей показник був уведений відомим американським математиком Ричардом Веслі Хеммінгом у 1950 р.

Множинна кореляція

Про множинну кореляцію мова йде в тому випадку, коли певна ознака може бути пов'язана не з однією, а із сукупністю декількох інших ознак. У реальних дослідженнях можлива ситуація, коли на певну ознаку може впливати не одна, а декілька інших. В таких випадках парні показники кореляції будуть давати неправильну інформацію щодо наявності зв'язку між відповідними показниками, оскільки ці їх значення будуть викривлятися невраховуваними ознаками. Для уникнення помилок використовують частинні показники кореляції, що усувають такий вплив. Ідея введення таких показників вперше була висунута Г.У. Юлом у 1896 р., а пізніше розвинена ним та К. Пірсоном.

Множинний коефіцієнт кореляції мажорує будь-який парний або частинний коефіцієнт кореляції, що характеризує статистичні зв'язки досліджуваної ознаки. Додавання нових ознак не може зменшувати коефіцієнт множинної кореляції.

Подання математичних об'єктів називають *канонічним*, якщо кожному об'єкту однієї множини відповідає один і тільки один об'єкт іншої множини й ця відповідність є взаємно однозначною. *Канонічний кореляційний аналіз* здійснюють між двома сукупностями (групами) вибірок. Він призначений для визначення лінійної функції від перших p компонент і лінійної функції від q компонент, що

залишилися, таких, щоб коефіцієнт кореляції між цими лінійними функціями набув найбільшого можливого значення. Чисельності груп (кількість вибірок у першій та другій групах, p та q) можуть різнитися, але необхідною умовою є рівна кількість варіант у всіх вибірках, що становлять обидві групи.

Коефіцієнт конкордації призначений для дослідження зв'язків між порядковими ознаками, кількість яких є більшою, ніж два. Як міру узго- 122 дженості беруть суму квадратів відхилень сум рангів спостережень (об'єктів) від їх спільного середнього рангу.

Коефіцієнт конкордації Кендалла (W-коефіцієнт Кендалла) було запропоновано М. Кендаллом у 1939 р. Його значення може змінюватися в межах від нуля до одиниці, при цьому він дорівнює одиниці лише за умови, що всі досліджувані ранжирування збігаються. Коефіцієнт конкордації дорівнює нулю, якщо $k \geq 3$ і всі ранжирування є випадковими впорядкуваннями вихідної вибірки.

Для перевірки нульової гіпотези про наявність зв'язку скористаємося відповідною процедурою пакета *SPSS (Analyze/Correlate/Bivariate)*. У відповідному вікні задаємо: вибірки, зв'язок між якими необхідно перевірити; значення коефіцієнтів кореляції, які треба розрахувати; вказуємо характер гіпотези – однобічна чи двобічна; а також, за необхідністю, додаткові опції [4,5].

На рис. 14 наведено результати кореляційного аналізу, отримані у пакеті SPSS, а на рис. 13 – графік, з якого видно наявність близького до лінійного зв'язку між ознаками.

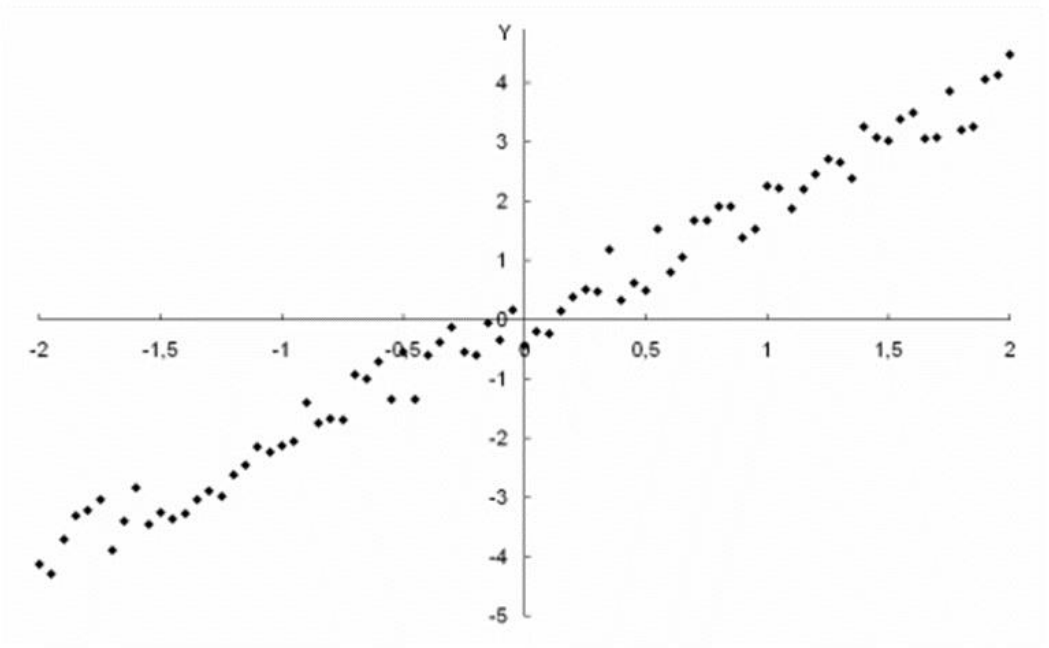


Рис. 13. Графік зв'язку між досліджуваними ознаками у випадку лінійного зв'язку

Correlations

		VAR00001	VAR00006
VAR00001	Pearson Correlation	1	,991**
	Sig. (2-tailed)		,000
	Sum of Squares and Cross-products	110,700	225,011
	Covariance	1,384	2,813
	N	81	81
VAR00006	Pearson Correlation	,991**	1
	Sig. (2-tailed)	,000	
	Sum of Squares and Cross-products	225,011	465,288
	Covariance	2,813	5,816
	N	81	81

**. Correlation is significant at the 0.01 level (2-tailed).

Correlations

			VAR00001	VAR00006
Kendall's tau_b	VAR00001	Correlation Coefficient	1,000	,927**
		Sig. (2-tailed)	.	,000
		N	81	81
	VAR00006	Correlation Coefficient	,927**	1,000
		Sig. (2-tailed)	,000	.
		N	81	81
Spearman's rho	VAR00001	Correlation Coefficient	1,000	,991**
		Sig. (2-tailed)	.	,000
		N	81	81
	VAR00006	Correlation Coefficient	,991**	1,000
		Sig. (2-tailed)	,000	.
		N	81	81

**. Correlation is significant at the 0.01 level (2-tailed).

Рис. 14. Результати кореляційного аналізу, отримані за допомогою пакета SPSS, у випадку лінійного зв'язку

Різниця між цими засобами полягає в тому, що за допомогою пакета аналізу ми отримуємо кореляційну матрицю для даних, що розташовані у декількох сусідніх стовпчиках або рядках робочого аркушу. При спробі розрахувати коефіцієнти кореляції для даних, між якими є розриви, буде отримано повідомлення про помилку.

Тема 5. Факторний аналіз
Метод головних компонент. Метод головних факторів. Інші методи факторного аналізу

При дослідженні складних систем часто немає можливості безпосередньо вимірювати величини, що визначають їх властивості (фактори). Більше того, нерідко є невідомими кількість та зміст цих факторів. Але можуть вимірюватися інші величини, що залежать від них. Якщо невідомий фактор впливає на декілька вимірюваних ознак, останні виявляють певний зв'язок, наприклад корельованість, між собою. Тому загальна кількість факторів може бути значно меншою, ніж кількість вимірюваних ознак. Для виявлення таких факторів використовують факторний аналіз. Зменшення кількості факторів може бути необхідним також для забезпечення збіжності алгоритмів подальшого аналізу даних, скорочення ресурсів пам'яті ЕОМ та часу, потрібних для їх обробки, бажанням візуалізувати отримані результати тощо. Основні ідеї факторного аналізу було сформульовано *Ф. Гальтоном* наприкінці XIX ст. Пізніше значний внесок у розвиток його методології зробили *Р. Кеттелл, К. Пірсон, Ч. Спірмен, Л. Терстоун, Г. Хотеллінг* та інші фахівці [4] .

Першим етапом факторного аналізу зазвичай є вибір нових ознак (факторів), які є лінійними комбінаціями старих і відображають переважну частку загальної мінливості вихідних даних. Тому вони зберігають основну частину інформації, що містили ці дані. Другим етапом є обертання факторів з метою спрощення їх інтерпретації. Об'єктом дослідження методами факторного аналізу, як правило, є кореляційна матриця, побудована із застосуванням коефіцієнта кореляції *Пірсона* для кількісних ознак. Основною вимогою до цієї матриці є її додатна напіввизначеність. Згідно з умовами Сильвестра для цього достатньо, щоб усі її головні мінори були невід'ємними. З додатної напіввизначеності кореляційної матриці випливає невід'ємність усіх її власних значень.

Методами факторного аналізу вирішують три основні групи проблем: – пошук передбачуваних неявних закономірностей, що визначаються впливом зовнішніх або внутрішніх чинників на досліджуваний процес; – виявлення та вивчення

статистичного зв'язку ознак з факторами або головними компонентами; – стискування інформації шляхом подання процесу за допомогою узагальнених факторів або головних компонент, кількість яких є меншою за кількість обраних спочатку ознак (параметрів), але достатньою для забезпечення відтворення кореляційної матриці з потрібною точністю. Розрізняють ***R-техніку*** та ***Q-техніку*** факторного аналізу. Перша з них розроблена британським психологом *Реймондом Б. Кеттеллом* і передбачає розрахунок коефіцієнтів кореляції між параметрами (ознаками), що утворюють матрицю вихідних даних. Її використовують для зменшення кількості параметрів. ***Q-техніку*** запропонував британський психолог *В. Стефенсон* в 1935–1936 р. й докладно описав *Р.Б. Кеттелл* у 1946 р. За її допомогою вивчають кореляцію між об'єктами або станами об'єктів. Її застосовують для зменшення кількості об'єктів. З формального погляду в першому випадку шукають кореляцію між стовпчиками таблиці спостережень (табл. 6), а у другому – між її рядками [5] .

Таблиця 6

Загальний вигляд таблиці спостережень для факторного аналізу

Номери об'єктів (станів)	Параметри об'єктів (станів)			
	1	2	...	p
1	x_{11}	x_{12}	…	x_{1p}
2	x_{21}	x_{22}	…	x_{2p}
...	…	…	…	…
n	x_{n1}	x_{n2}	…	x_{np}

Основними методами факторного аналізу є методи головних компонент, головних факторів, максимальної правдоподібності та центроїдний. Усі вони ґрунтуються на припущенні, що досліджувана залежність є лінійною. Вихідні дані мають підпорядковуватися багатовимірному нормальному розподілу, але центроїдний метод є досить стійким до відхилень від такого закону. Метою факторного аналізу є зменшення кількості змінних та визначення структури взаємозв'язків між змінними (класифікація даних).

З формального погляду його метою є одержання матриці факторного відображення. Її рядки є координатами кінців векторів, що відповідають n змінним у p'-вимірному факторному просторі. Близькість цих векторів свідчить про взаємну залежність змінних.

Якщо кількість факторів перевищує одиницю, зазвичай здійснюють обертання матриці факторного відображення для одержання більш простої її структури. Однією з проблем, що виникають при застосуванні факторного аналізу, є необхідність знаходження власних значень кореляційної матриці. Якщо вона є виродженою, ця задача може виявитися нерозв'язною. Для матриць високого порядку може відбуватися втрата значущості у процесі обчислень. У певних випадках проблему виродженості можна зняти виключенням лінійно залежних параметрів.

Обов'язковими умовами факторного аналізу є такі:

– всі досліджувані ознаки мають бути кількісними;

– кількість ознак має бути принаймні вдвічі більшою, ніж кількість змінних;

– вибірка має бути однорідною;

– вихідні змінні повинні мати симетричний розподіл.

Метод головних компонент

Метод головних компонент, або компонентний аналіз вперше був запропонований *К. Пірсоном* у 1901 р., який розглядав задачу найкращої (з погляду мінімізації суми квадратів відхилень) апроксимації сукупності точок прямими та площинами. Потім він був докладно розроблений американським статистиком й економістом *Гарольдом Хотеллінгом* у 1933 р. Його важливою перевагою є те, що він є єдиним математично обґрунтованим методом факторного аналізу.

За своєю сутністю метод полягає у виборі нової ортогональної системи координат у просторі спостережень. Як першу головну компоненту обирають напрям, вздовж якого масив спостережень має найбільшу дисперсію. Кожну наступну компоненту обирають також з умови максимізації частки дисперсії, що залишилася, вздовж неї, доповненої умовою ортогональності всім раніше обраним компонентам. При цьому із зростанням номера компоненти буде зменшуватися пов'язана з нею частка загальної дисперсії. Кількість компонент визначається значною мірою суб'єктивно, виходячи з розуміння того, яка величина загальної дисперсії відповідає випадковій мінливості, що відображає похибку вимірювань, вплив неконтрольованих випадкових чинників тощо.

Вибір критерію інформативності в методі головних компонент передбачає, що найбільш важливу інформацію про аналізовану систему можна відобразити лінійною

моделлю, яка відповідає такому вибору системи координат у тому самому просторі, що забезпечує максимальні дисперсії для проекцій досліджуваних об’єктів. Такий підхід є доцільним, якщо більшість вихідних ознак узгоджено впливає на властивість, що вивчається, і пригнічує вплив іррелевантних чинників на розподіл об’єктів.

Адекватну модель можна отримати також у випадку, коли кількість пов’язаних інформативних ознак невелика, але вплив інших чинників є неузгодженим. У цьому разі не порушується однорідність еліпсоїда розсіювання, а лише зменшується його довгастість уздовж напрямку досліджуваної властивості. У факторному аналізі використовують також інші міри інформативності, що дають змогу визначити кількість істотних факторів.

Критерій Кайзера, або критерій власних чисел, запропонований американським психологом *Генрі Феліксом Кайзером*, передбачає, що до моделі включають тільки фактори, для яких власні числа є не меншими, ніж одиниця. За змістом це означає, що таким факторам відповідає дисперсія, еквівалента принаймні дисперсії одної змінної. У противному випадку виокремлення фактора не має сенсу. Цей критерій іноді залишає в моделі занадто багато факторів.

Критерій кам’янистого осипу (критерій відсіювання) передбачає побудову графіка, де по осі абсцис відкладають порядковий номер власного числа, а по осі ординат – його значення. Згідно з *Р. Кеттелом* необхідно знайти точку найбільшого уповільнення спадання власних значень і враховувати лише фактори, яким відповідають власні числа, розташовані лівіше цієї точки. На відміну від попереднього цей критерій статистично необґрунтовані і часто залишає в моделі не всі істотні фактори. Втім у випадках, коли істотних факторів небагато, а кількість змінних є великою, обидва критерії є придатними для практичного застосування.

На практиці часто здійснюють розрахунки, використовуючи різні критерії, а потім обирають модель, що містить найбільшу кількість факторів, яким можна надати змістову інтерпретацію.

Критерії, що ґрунтуються на аналізі визначників вихідної та відтвореної кореляційної матриць, часто виявляються нестійкими. Критерії, які базуються на величині власних значень кореляційної матриці, у підсумку призводять до аналізу відсотка дисперсії, виділеної факторами. Усі загальні фактори, кількість яких дорівнює кількості параметрів, пояснюють 100% дисперсії. Якщо сума відсотків за

факторами перевищує 100%, це свідчить про отримання від'ємних власних значень і, відповідно, комплексних власних векторів, що може бути наслідком некоректної редукції вихідної кореляційної матриці. Доцільно здійснювати двохетапну процедуру аналізу. На першому етапі максимальну кількість факторів не задають. Після його проведення аналізують дисперсії, оцінюють приблизну кількість факторів і проводять повторний аналіз.

Метод головних факторів

Цей метод використовують для зменшення кількості змінних. У його основі лежить припущення, що не всі змінні, які вимірювали при дослідженні системи, є незалежними. Тому можливо формування нових змінних, що достатньо повно відображають наявну інформацію. На відміну від методу головних компонент, метод головних факторів ґрунтується не на дисперсійному критерії інформативності множини ознак, а на поясненні кореляцій, що існують між цими ознаками. Він враховує, що вихідні дані можуть містити грубі помилки, які у багатовимірному аналізі призводять до помилок інтерпретації. Тому метод головних факторів застосовують у більш складних випадках, зокрема за наявності сумісного прояву аналізованих й іррелевантних властивостей об'єктів, що є порівнянними за ступенем внутрішньої узгодженості, а також для виділення групи діагностичних показників із вихідної множини ознак.

Основну модель методу головних факторів записують у вигляді:

$$X^{*}=MF$$

Матриця X^{*} є матрицею нормовано-центрованих значень вихідних ознак, що має розмірність $n \times p$. У методі головних факторів припускають, що кожний елемент матриці X^{*} є результатом впливу m гіпотетичних *загальних факторів* та одного *характерного фактора*.

Характерні фактори вважають некорельованими один з одним, а також із загальними факторами. *Загальні фактори* пов'язані з істотними ваговими коефіцієнтами більше, ніж з одною ознакою. Ті з них, для яких істотними є всі вагові коефіцієнти, називають *генеральними факторами*.

Перший варіант методу, запропонований *Ч. Спірменом* на початку 1900 р. передбачав існування одного загального та одного характерного факторів. Пізніше у

1920 р. британський та американський психолог *Раймонд Бернард Кеттел* та американський психолог *Карл Джон Хользінгер* запропонували *біфакторну* модель, яка передбачала існування декількох, зазвичай двох, загальних факторів. Сучасний варіант методу головних факторів було запропоновано *Г. Томсоном.*

Інші методи факторного аналізу

У **методі максимуму правдоподібності**, який запропоновано *Д. Лоулі*, оцінювання загальностей до безпосереднього застосування алгоритму факторного аналізу не здійснюють. Їх визначають за результатами обчислень з умови повного відтворення кореляційної матриці. За будь-якої кількості факторів, що розглядаються, цей метод дає можливість відтворити її з точністю до похибки обчислень. Якщо кількість факторів дорівнює кількості параметрів, то оцінки загальностей будуть збігатися із загальностями нередукованої кореляційної матриці, тобто дорівнювати одиниці. Основним недоліком методу є його нестійкість при використанні окремих типів даних, зокрема даних, що містять однакові або лінійно залежні вектори. Це призводить до виродження матриці характерностей. У такому випадку можна спробувати зняти проблему шляхом виключення з розгляду лінійно залежних параметрів або застосування методу головних факторів.

У попередніх методах максимізується квадратичний критерій. На відміну від них, у **центроїдному методі**, розробленому американським психологом *Луїсом Леоном Терстоуном* у 1930р., максимізують модульний критерій. З погляду змістової інтерпретації ці критерії є еквівалентними.

Сучасні апроксимуючи методи виходять з припущення, що є певне початкове наближення, яке необхідно покращити. Крім розглянутих вище методів головних факторів, найбільшої правдоподібності й контрастних груп до них належать:

- **груповий метод** *Л. Гуттмана й П. Хорста*, що базується на попередньому відборі груп елементарних ознак;
- **метод мінімальних залишків** *Г. Хартмана*;
- **метод α-факторного аналізу**, запропонований *Г. Кайзером й І. Кеффрі* в 1965 р.;
- **метод канонічного факторного аналізу** *К. Рао*;
- **методи, що оптимізують**.

Всі ці методи дають змогу послідовно покращувати знайдені розв’язки на основі використання статистичних прийомів оцінювання випадкової величини або статистичних критеріїв й передбачають великий обсяг трудомістких обчислень [3].

Тема 6. Завдання та методи класифікації даних

Параметричні методи класифікації без навчання. Кластерний аналіз. Класифікація з навчанням. Приклади здійснення класифікації даних.

У загальному випадку ***класифікацією*** (**розпізнаванням образів**) називають поділ досліджуваної сукупності об’єктів на однорідні в певному розумінні групи (класи) або зарахування кожного із заданої множини об’єктів до деякого із заздалегідь відомих класів. При цьому вирізняють три групи завдань: *дискримінацію, кластеризацію й групування*. Останні дві групи є близькими за метою (поділ даних на класи або групи близьких у певному розумінні об’єктів), а також за алгоритмами. Але принципова різниця між ними полягає у тому, що у першому випадку межі класів є природними, а у другому – умовними й їх можна встановлювати суб’єктивно.

Параметричні методи класифікації без навчання

У методах класифікації без навчання програмна система на основі визначених нею самою критеріїв здійснює класифікацію певних об’єктів (образів). У деяких випадках можуть бути задані окремі параметри, але розподіл об’єктів за класами на основі цих параметрів виконується автоматично.

Для розв’язання задачі розщеплення суміші розподілів часто використовують **EM** (**Expectation – Maximization**) алгоритм, вперше запропонований в 1977 р. американськими статистиками *Артуром Демпстером, Неном Лайрдом і Дональдом Рубіном*. Цей алгоритм дає змогу визначати методом найбільшої правдоподібності параметри статистичних моделей, що містять певні приховані змінні. Він передбачає здійснення двох кроків на кожній ітерації. Перший крок (**Expectation**) полягає в обчисленні значення функції правдоподібності за умови, що задані деякі значення прихованих змінних. На другому кроці (**Maximization**) обчислюють значення параметрів, що максимізують функцію правдоподібності. Обчислення виконують до виконання заданих умов збіжності. Недоліками алгоритму є залежність результату від вибору початкового наближення (якщо функція правдоподібності не є унімодальною). Крім того, цей алгоритм не дає змоги визначити кількість компонент суміші. Для усунення цих недоліків пізніше було запропоновано різноманітні модифікації **EM** алгоритму: *медіанні*, *стохастичний* (**SEM**), *класифікаційний* (**CEM**), *узагальнений* (**GEM**), з додаванням компонент тощо.

Кластерний аналіз

Для формалізації задачі класифікації кожний об’єкт зручно інтерпретувати як точку в багатовимірному просторі ознак. Геометрична близькість точок у такому просторі відповідає близькості досліджуваних об’єктів з погляду досліджуваних властивостей (рис. 15). Наочне уявлення про зміст класифікації дає діаграма *Герцшпрунга – Ресселла* (рис. 16), яка є основою для однієї з найпоширеніших класифікацій зірок за поєднанням їх світності й кольору (*температури* або *спектрального класу*).

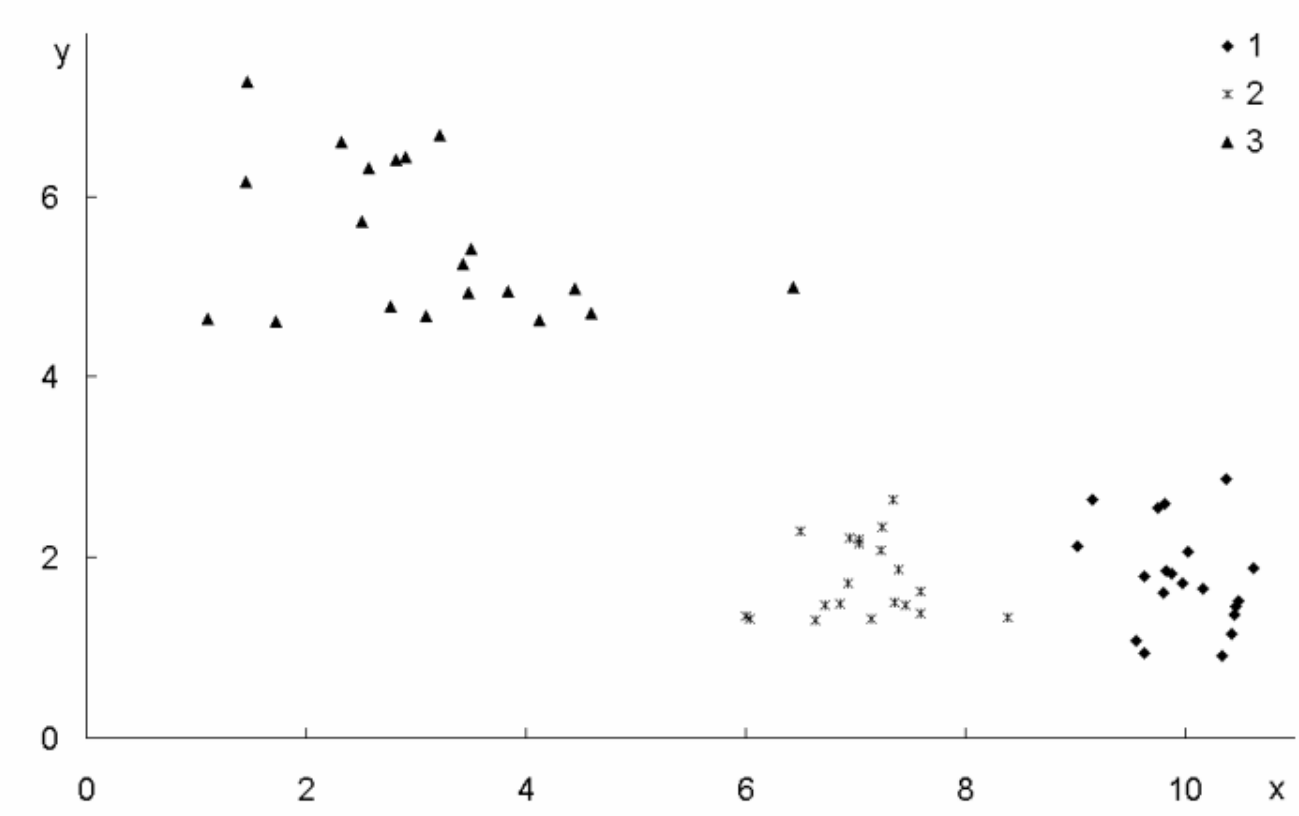


Рис. 15. Геометричне зображення сукупності об’єктів, що характеризуються двома ознаками й утворюють три кластери

Залежно від мети дослідження задачу класифікації можна сформулювати як розбиття аналізованих об’єктів на певну кількість груп, усередині яких вони розташовані на порівняно малій відстані один від одного, або як виявлення природного розшарування сукупності, що вивчається, на окремі кластери. Другу задачу можна також сформулювати як визначення областей підвищеної густини точок, що відповідають наявним спостереженням. Перша задача завжди має розв’язок, а друга може не мати розв’язку. Це відповідає відсутності природного розшарування досліджуваних об’єктів (наприклад, вони утворюють один кластер або

відповідні точки рівномірно заповнюють весь простір ознак).

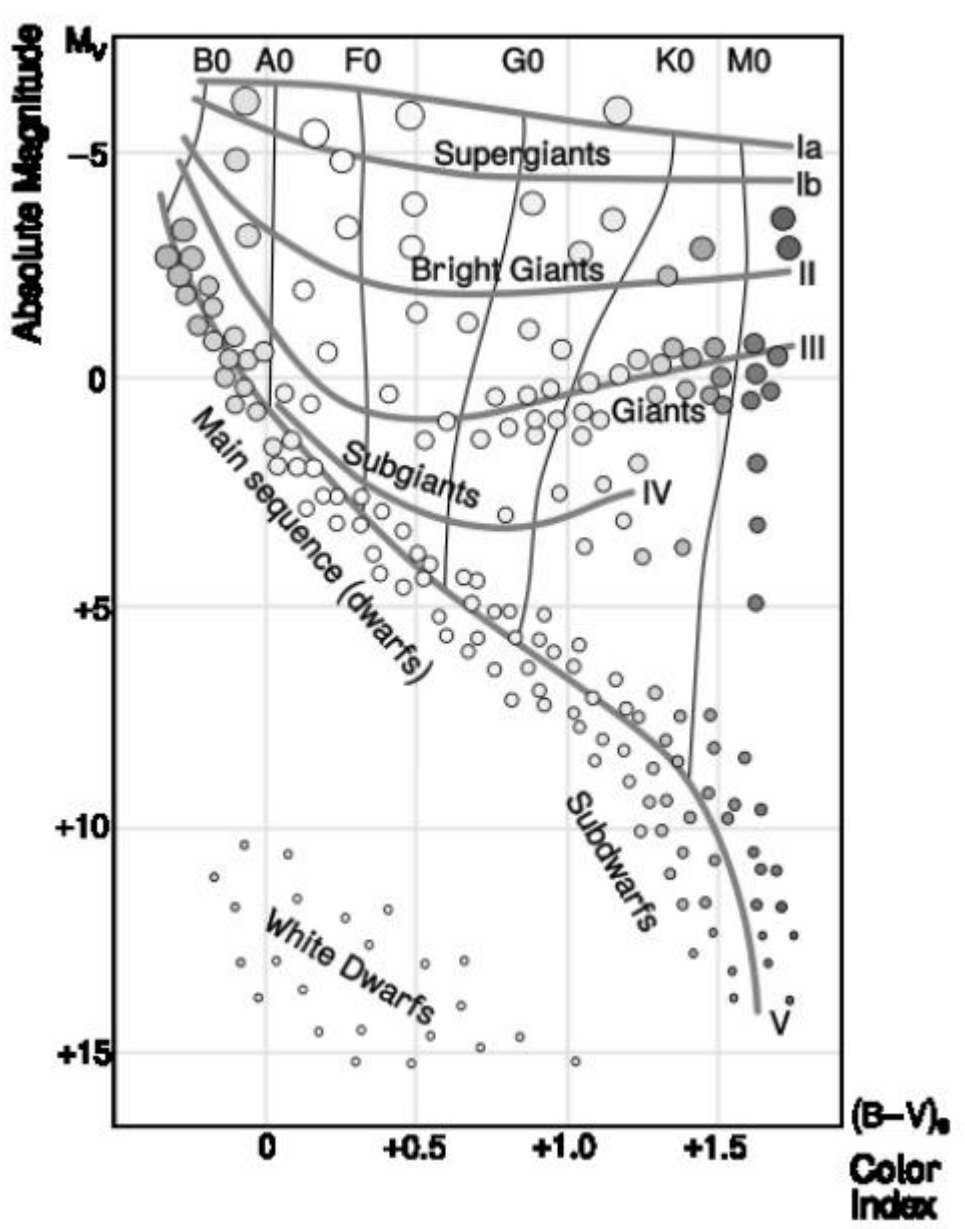


Рис. 16. Діаграма Е. Герцшпрунга – Г. Ресселла

Класичними непараметричними методами класифікації без навчання є методи **кластерного аналізу** (**таксономії**). За їх допомогою вирішують проблему такого розбиття (*класифікації, кластеризації*) множини об’єктів, за якого всі об’єкти, що належать до одного класу, були б більш подібними один до одного, ніж до об’єктів інших класів. З формальної точки зору, основне завдання методів кластерного аналізу можна сформулювати, як визначення класів еквівалентності й рознесення за ними досліджуваних об’єктів. Під класом, як правило, розуміють генеральну сукупність, що описується *одномодальною* функцією щільності ймовірності ***f(X)*** або, у випадку дискретних ознак, – *одномодальним* полігоном імовірностей. Номери

класів не мають змістового навантаження й використовуються лише для того, щоб відрізняти їх один від одного. Для формування кластерів застосовують міри подібності та відмінності даних, які можуть бути поділені на три основних види:

– **міри подібності (відмінності) типу "відстань"** (при їх застосуванні об'єкти вважають тим більш подібними один до одного, чим меншою є відстань між ними);

– **міри подібності типу "зв'язок"** (у цьому випадку об'єкти вважають тим більш подібними, чим сильнішим є зв'язок між ними);

– **інформаційна статистика**.

Найпоширенішими методами кластерного аналізу є:

– **ієрархічні методи** (ближнього зв'язку, середнього зв'язку *Кінга, Уорда*, далекого зв'язку);

– **ітеративні методи групування** (метод *k*-середніх *Мак-Куїна*);

– **алгоритми типу розрізування графа** (кореляційних плеяд *Терентьєва*, вроцлавська таксономія).

Ієрархічні (*агломеративні* та *дивізимні*) методи призначені переважно для побудови ієрархічних дерев відносно невеликих за обсягом сукупностей. Іноді їх використовують також для задач класифікації перших двох типів. У цьому випадку реалізацію ієрархічного алгоритму продовжують до досягнення кількості класів, яка дорівнює заздалегідь заданому числу k, або до досягнення екстремуму одного з критеріїв якості розбиття. Перевагами ієрархічних методів є можливість більш повного і тонкого аналізу структури досліджуваної сукупності порівняно з іншими методами, а також наочність подання результатів кластеризації. Їх основними недоліками є громіздкість обчислювальної процедури, яка пов'язана з перерахунком усієї матриці відстаней на кожному кроці, а також "скінченна неоптимальність" гранично оптимальних алгоритмів у багатьох випадках [5,7] .

Результати ієрархічних методів кластерного аналізу стають більш наочними, якщо їх подати у вигляді **дендрограми** (дендограми). Типовий вигляд дендрограми наведено на рис. 17.

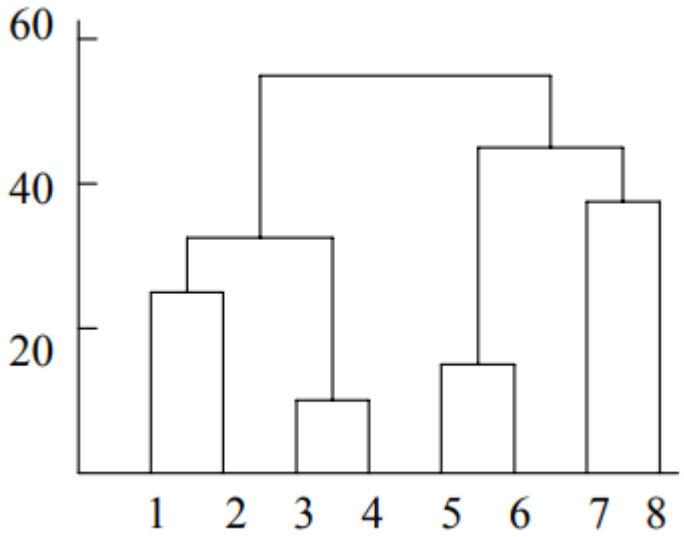


Рис. 17. Приклад дендрограми

Пари об’єктів при побудові дендрограми з’єднують згідно з рівнем зв’язку, який відкладають по вісі ординат. Задаючи кількість кластерів, наприклад ***n*** = 3, знаходять, на якому рівні кількість перетинів горизонтальної лінії, яка відповідає рівню зв’язку, і вертикальних ліній, що відповідають об’єктам, дорівнює трьом. У нашому випадку такій кількості кластерів відповідає рівень зв’язку, що знаходиться приблизно в межах від 38 до 45 одиниць.

Типовим представником (найбільш відомим і поширеним) є **Information Management System** (IMS) фірми **IBM**. Перша версія з'явилася в 1968 р. **Time-Shared Date Management System** (TDMS) компанії *Development Corporation; Mark IV Multi - Access Retrieval System* компанії *Control Data Corporation; System 2000* розробки *SAS-Institute*; Сервери каталогів, такі, як *LDAP і Active Directory* (допускають чітке уявлення у вигляді дерева) За принципом ієрархічної БД побудовані ієрархічні файлові системи та ***Реєстр Windows. InterSystems Cach Google App Engine DataStore API.***

Класифікація з навчанням. Приклади здійснення класифікації даних.

Як приклад здійснення кластерного аналізу розглянемо таку задачу. Сформуємо вибірку, що містить чотири класи об’єктів, кожний з яких характеризується трьома кількісними ознаками. Значення ознак сформуємо як нормально розподілені випадкові величини з параметрами, що наведено у табл. 7.

Таблиця 7

Параметри розподілу ознак для прикладу кластерного аналізу

Номер кластера	x_{1cp}	s_1	x_{2cp}	s_2	x_{3cp}	s_3	n
1	2	0,2	100	10	49	5	20
2	3	0,4	83	8	78	7	15
3	1	0,3	88	9	32	4	25
4	5	0,5	53	8	37	5	30

Занесемо отримані значення до таблиці вихідних даних пакету SPSS (рис. 18).

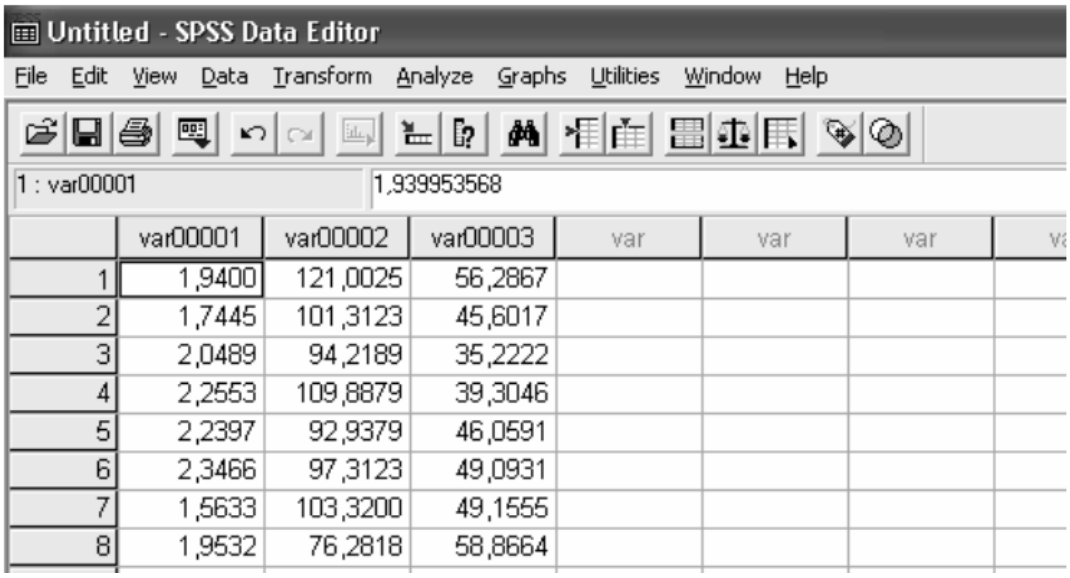


Рис. 18. Вихідні дані для кластерного аналізу

Далі обираємо в меню: *Analyze/Classify/K-means Cluster …*

При цьому з'являється діалогове вікно, показане на рис. 19.

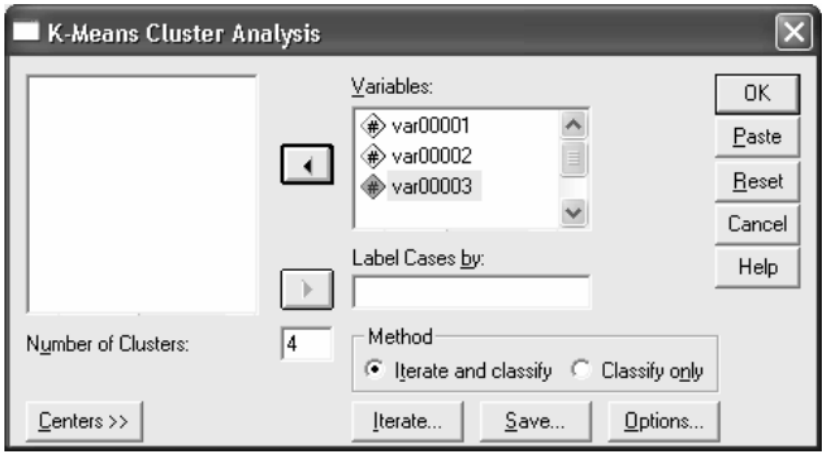


Рис. 19. Діалогове вікно кластерного аналізу методом *k*-середніх

У цьому вікні необхідно вказати змінні, що характеризують досліджувані об'єкти, кількість кластерів, що потрібно виокремити та метод кластеризації. Крім того можна встановити додаткові параметри процедури. За допомогою кнопки *"Centers"* вказуємо, чи потрібно зчитати з окремого файлу початкові значення

центрів кластерів, або записати до файлу кінцеві значення центрів. За допомогою кнопки *"Iterate"* відкриваємо вікно установки параметрів ітераційної процедури.

За допомогою кнопки *"Options"* відкриваємо діалогове вікно, в якому вказуємо додаткові параметри процедури кластеризації: необхідність виводу початкових значень центрів кластерів, таблиці **ANOVA** й інформації про кількість об'єктів у кожному кластері, а також спосіб обробки пропущених значень.

Final Cluster Centers

	Cluster			
	1	2	3	4
VAR00001	1,9177	2,8655	1,0529	5,0436
VAR00002	100,2400	82,0074	88,0278	53,6599
VAR00003	49,2148	74,1121	33,8212	38,5363

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
VAR00001	82,102	3	,165	86	499,014	,000
VAR00002	9716,806	3	63,743	86	152,438	,000
VAR00003	6156,787	3	28,537	86	215,747	,000

Distances between Final Cluster Centers

Cluster	1	2	3	4
1		30,874	19,668	47,891
2	30,874		40,778	45,541
3	19,668	40,778		34,919
4	47,891	45,541	34,919	

Number of Cases in each Cluster

Cluster	1	17,000
	2	16,000
	3	27,000
	4	30,000
Valid		90,000
Missing		,000

Рис. 20. Результати кластеризації

З наведених даних бачимо, що отримані результати дещо відрізняються від заданих (при цьому слід мати на увазі, що номера вихідних кластерів можуть не збігатися з номерами, наданими кластерам при використанні процедури кластеризації методом *k*-середніх).

Тема 7. Методи побудови й дослідження регресійних моделей

Загальна характеристика методів і задач регресійного аналізу. Лінійні однофакторні моделі. Поліноміальні моделі. Однофакторні моделі інших типів. Лінійні багатофакторні моделі. Інші типи багатофакторних моделей. Перевірка адекватності регресійних моделей.

Завданням дослідження складних систем і процесів часто є перевірка наявності й встановлення типу зв'язку між незалежними змінними *xi* (предикторами, факторами), значення яких можуть змінюватися дослідником і мають певну заздалегідь задану похибку, та залежною змінною (відгуком). Розв'язання таких завдань є предметом регресійного аналізу. Термін **"Регресія"** вперше був уведений *Ф. Гальтоном* наприкінці XIX ст. На практиці завдання регресійного аналізу зазвичай формулюють так: необхідно підібрати достатньо просту функцію, що в певному розумінні найкращим чином описує наявну сукупність емпіричних даних [4].

Класичний регресійний аналіз включає методи побудови математичних моделей досліджуваних систем, методи визначення параметрів цих моделей і перевірки їх адекватності. Він припускає, що регресія є лінійною комбінацією лінійно незалежних базисних функцій від факторів з невідомими коефіцієнтами (параметрами). Фактори й параметри є детермінованими, а відгуки – *рівноточними* (тобто мають однакові дисперсії) *некорельованими* випадковими величинами. Передбачається також, що всі змінні вимірюють у неперервних числових шкалах. Звичайна процедура класичного регресійного аналізу є такою. Спочатку обирають гіпотетичну модель, тобто формулюють гіпотези про фактори, які суттєво впливають на досліджувану характеристику системи, і тип залежності відгуку від факторів. Потім за наявними емпіричними даними про залежність відгуку від факторів оцінюють параметри обраної моделі. Далі за статистичними критеріями перевіряють її адекватність.

При побудові регресійних моделей реальних систем і процесів вказані вище припущення виконуються не завжди. У більшості випадків їх невиконання призводить до некоректності застосування процедур класичного регресійного

аналізу і потребує застосування більш складних методів аналізу емпіричних даних.

Постулат про *рівноточність* і *некорельованість* відгуків не є обов'язковим. У випадку його невиконання процедура побудови регресійної моделі певною мірою змінюється, але суттєво не ускладнюється. Більш складною проблемою є вибір моделі та її незалежних змінних. У класичному регресійному аналізі припускають, що набір факторів задається однозначно, всі суттєві змінні наявні в моделі й немає ніяких альтернативних способів обрання факторів. На практиці це припущення не виконується. Тому виникає необхідність розробки формальних та неформальних процедур перетворення й порівняння моделей. Для пошуку оптимальних формальних перетворень використовують методи факторного та **дискримінантного** аналізу. На сьогодні розроблено комп'ютеризовані технології послідовної побудови регресійних моделей.

Фактори в класичному регресійному аналізі вважають **детермінованими**, тобто вважається, що дослідник має про них всю необхідну інформацію з абсолютною точністю. На практиці це припущення часто не виконується. Відмова від детермінованості незалежних змінних зумовлює необхідність застосування моделей кореляційного аналізу. В окремих випадках можна використовувати компромісні методи **конфлюентного** аналізу, які передбачають можливість нормально розподіленого та усіченого розкиду значень факторів. Якщо ця умова виконується, побудову моделі можна звести до багаторазового розв'язування регресійної задачі. Відмова від припущення про детермінованість параметрів моделей у регресійному аналізі призводить до суттєвих ускладнень, оскільки порушує його статистичні основи. Але на практиці це припущення виконується не завжди.

У деяких випадках можна вважати параметри випадковими величинами із заданими законами розподілу. Тоді як оцінки параметрів можна брати їх умовні математичні сподівання для відгуків, що спостерігалися. Умовні розподіли та математичні сподівання розраховують за узагальненою формулою *Байєса*, тому відповідні методи називають **байєсівським регресійним аналізом**.

Регресійні моделі часто використовують для опису процесів, що розвиваються у часі. У певних випадках це зумовлює необхідність переходу від випадкових значень відгуків до випадкових послідовностей, випадкових процесів або випадкових полів.

Однією з поширених і найпростіших моделей такого типу є модель **авторегресії**, згідно з якою відгук залежить не тільки від факторів, але також і від часу. Якщо останню залежність можна виявити, то проблема зводиться до стандартної задачі побудови регресії для модифікованого відгуку. В інших випадках необхідно використовувати більш складні прийоми.

Часто як попередній етап регресійного аналізу рекомендують за допомогою методів кореляційного аналізу перевіряти наявність значущого зв'язку між досліджуваними змінними. Але при цьому слід ураховувати, що звичайні методи кореляційного аналізу дають змогу перевіряти лише гіпотезу про наявність лінійного зв'язку. Якщо зв'язок є, але він нелінійний, висновки, отримані за допомогою кореляційного аналізу, можуть бути помилковими [5,6].

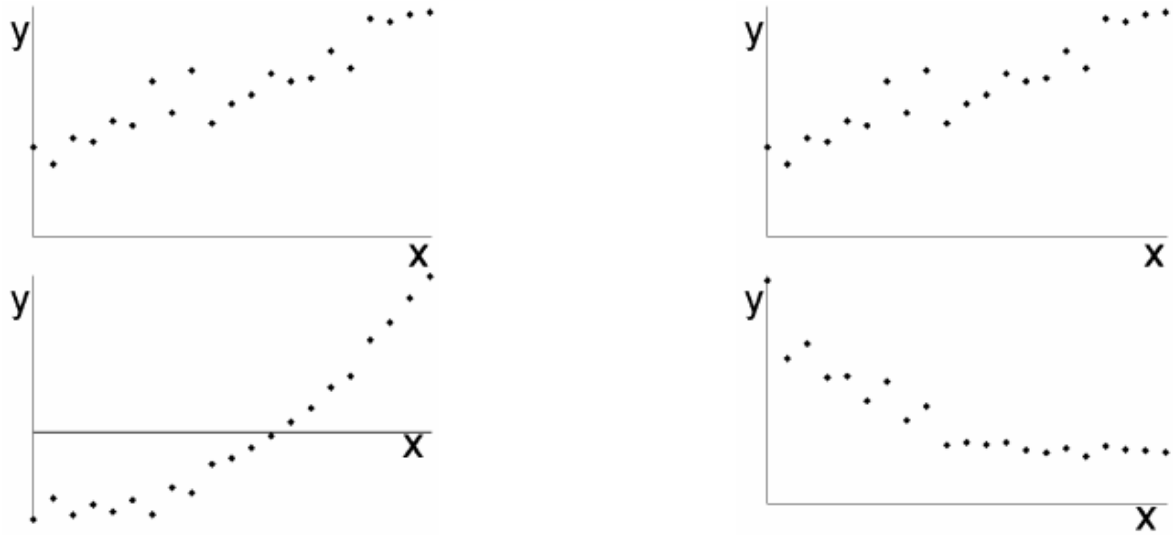


Рис. 21. Приклади даних, для яких треба побудувати регресійні моделі

При використанні регресійних моделей типу полінома, оберненого полінома, тригонометричного ряду та деяких інших слід враховувати, що, збільшуючи кількість членів ряду, можна одержати скільки завгодно близькі до нуля значення функціоналів (рис.21). Проте це не завжди свідчить про якість апроксимації, оскільки ці функціонали не дають інформації про ступінь наближення моделі до емпіричної залежності у проміжках між наявними точками.

Лінійні однофакторні моделі

Найпростішим для аналізу і найбільш дослідженим є випадок лінійної кореляційної залежності між двома змінними X та Y. Наявність лінійного зв'язку можна перевірити, розрахувавши *коефіцієнт парної кореляції Пірсона*.

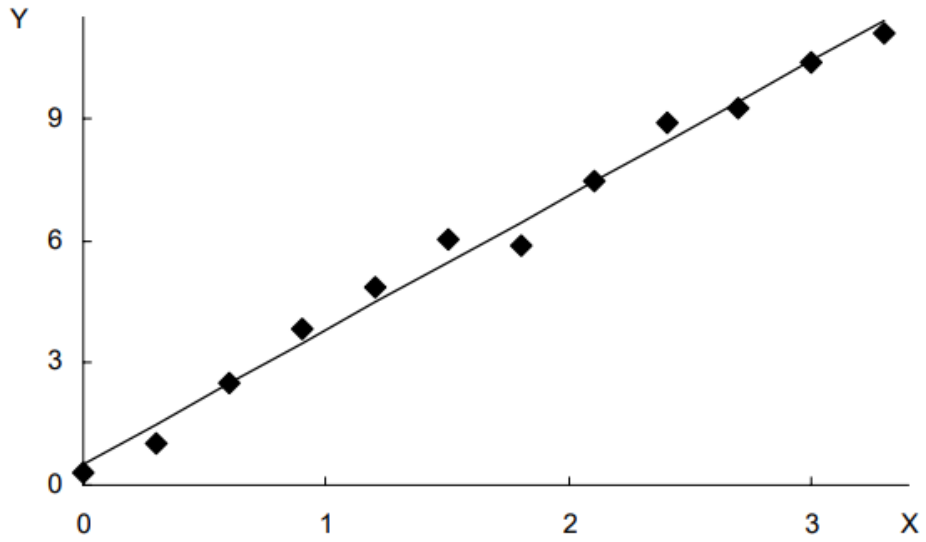


Рис. 22. Приклад лінійної моделі

Незважаючи на те, що, як правило, реальні залежності відгуків від факторів є нелінійними, розглянутий випадок широко використовують у практиці побудови регресійних моделей. Це пов'язано з трьома основними причинами.

По-перше, він є найбільш простим і дослідженим. Зокрема, для нього достатньо повно розроблені процедури визначення статистичних характеристик одержуваних оцінок параметрів (дисперсії, довірчих інтервалів тощо) та перевірки адекватності моделей.

По-друге, у багатьох випадках складні залежності можна подати як набір лінійних (на малих відрізках змінювання факторів) залежностей.

По-третє, нелінійні залежності у деяких випадках можна перетворити до лінійного вигляду шляхом заміни змінних. Деякі приклади такого перетворення наведено у табл. 8.

Таблиця 8

Приклади лінеаризації нелінійних залежностей

Вихідна залежність	Лінеаризована залежність	Нові змінні
$z = \alpha_0 exp\,(-\alpha_1 x)$	$\ln z = \ln \alpha_0 - \alpha_1 x$	x, $\ln z$
$z = \alpha_0[1 - exp(-\alpha_1 x)]$	$\ln \frac{\alpha_0}{\alpha_0 - z} = \alpha_1 x$	x, $\ln \frac{\alpha_0}{\alpha_0 - z}$
$z = \alpha_0 exp(-\alpha_1/x)$	$\ln z = \ln \alpha_0 - \alpha_1/x$	$1/x$, $\ln z$
$z = z = \alpha_0 x^{\alpha_1}$	$\ln z = \ln \alpha_0 + \alpha_1 \ln x$	$\ln x$, $\ln z$
$z = \alpha_0 x + \alpha_1 x^2$	$z/x = \alpha_0 + \alpha_1 x$	x, z/x
$z = \alpha_0 \sin(\alpha_1 x)$	$\arcsin (z/\alpha_0) = \alpha_1 x$	x, $\arcsin (z/\alpha_0)$

На рис. 23 наведено графіки емпіричних даних і відповідних регресійних моделей, побудованих у вигляді тригонометричних рядів, для *m* рівних 2, 3, 4 й 5. Видно, що зі збільшенням кількості членів тригонометричного ряду різниця між моделлю та емпіричними точками зменшується. Добре видно також, що найбільшу похибку модель дає поблизу меж відрізка, на якому визначені емпіричні дані.

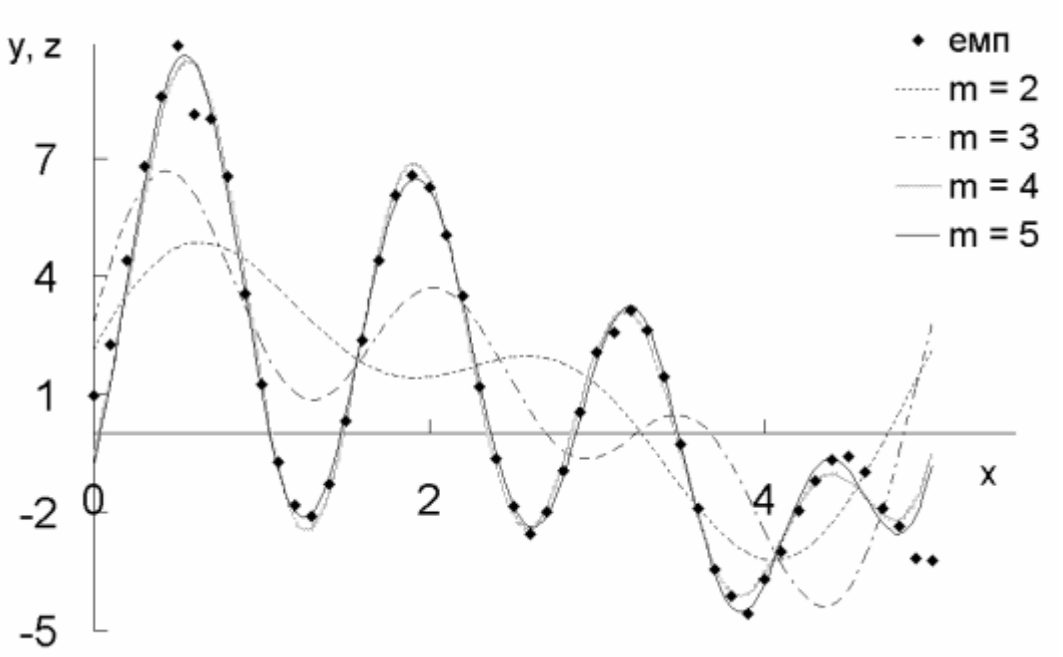


Рис. 23. Графіки вихідних даних і побудованої тригонометричної регресійної моделі

Перевірка адекватності регресійних моделей Основні методи перевірки адекватності регресійних моделей ґрунтуються на таких трьох властивостях їх залишків. По-перше, для адекватної моделі дисперсія залишків має бути близькою до дисперсії емпіричних точок. При цьому припускають, що дисперсії всіх емпіричних точок є однаковими. У випадку, коли для кожної точки здійснюють декілька вимірювань значення відгуку, останнє припущення можна перевірити за допомогою критеріїв *Кокрена або Бартлетта*. Причиною неадекватності при невиконанні цієї властивості є використання надмірно спрощених або ускладнених регресійних моделей. Відомо, наприклад, що за наявності n емпіричних точок можна побудувати поліноміальну модель *n – 1* порядку, яка пройде строго через всі ці точки. Але використовувати такий поліном як регресійну модель за наявності похибок

емпіричних даних, очевидно, недоцільно. З іншого боку, якщо степінь полінома буде надто малим, то він не відтворюватиме істотних рис досліджуваної залежності, тому існує певне оптимальне значення ступеня такого полінома.

Невиконання першої умови свідчить про надмірну спрощеність моделі, зокрема про необхідність збільшення порядку поліноміальної моделі.

Невиконання другої умови є свідченням того, що модель треба спростити, наприклад зменшити порядок полінома. У деяких випадках друга умова може не виконуватися навіть для однофакторних лінійних моделей. Найчастіше це може бути наслідком свідомого підганяння емпіричних даних під заздалегідь задану модель. Це часто роблять у навчальних задачах, але на практиці такий результат свідчить про навмисне викривлення первинних даних.

Іншою причиною може бути неправильна (завищена) оцінка похибки емпіричних даних. Це може бути пов’язано, зокрема, з нехтуванням зміною дисперсії емпіричних даних при їх попередній обробці. Інша властивість залишків, яку перевіряють при визначенні адекватності моделі, полягає в тому, що вони мають підпорядковуватися нормальному закону розподілу з нульовим математичним сподіванням і однаковими дисперсіями.

Перевірку цих властивостей можна здійснити за допомогою критеріїв, що описані у темі 2. На рис. 24 показано деякі типові випадки порушення вказаних властивостей, які можуть бути виявлені при візуальному аналізі ряду залишків [7,8].

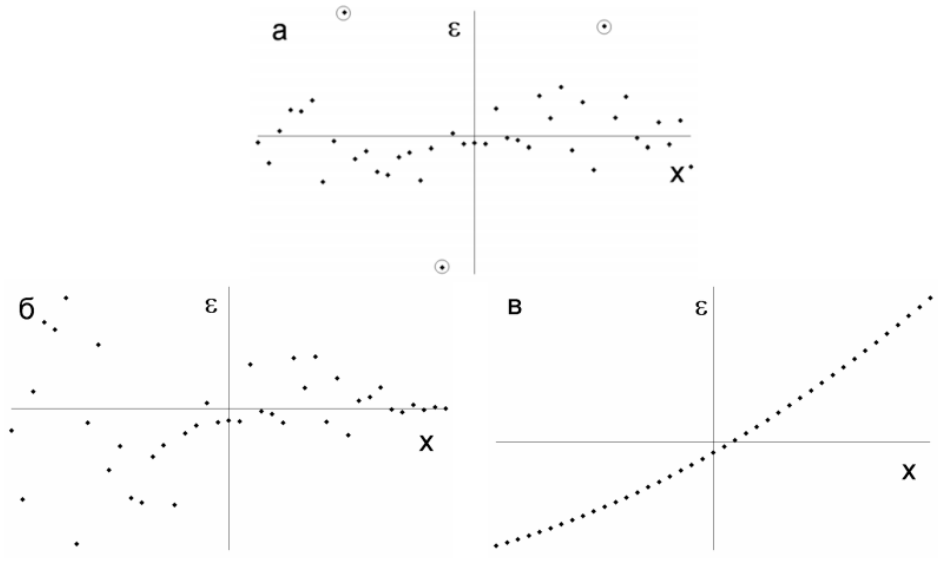


Рис. 24. Приклади порушення властивостей залишків неадекватних моделей

У випадку а) на графіку є так звані викиди – точки, що аномально сильно відхиляються від середнього значення.

У випадку б) дисперсія залишків помітно зменшується при зміщенні вправо.

У випадку в) ряд залишків не є випадковим, що свідчить про наявність неврахованих істотних закономірностей у моделі.

Наявність автокореляції вищих порядків перевіряють шляхом дослідження **автокореляційної функції**. Про наявність автокореляції в цьому випадку свідчить збільшення абсолютних значень коефіцієнта автокореляції при певних значеннях параметра зсуву. На рис 25 показано приклади **автокореляційних** функцій для ряду, що є білим шумом, (*ліворуч*) та рядом, який змінюється за синусоїдальним законом, (*праворуч)* [7].

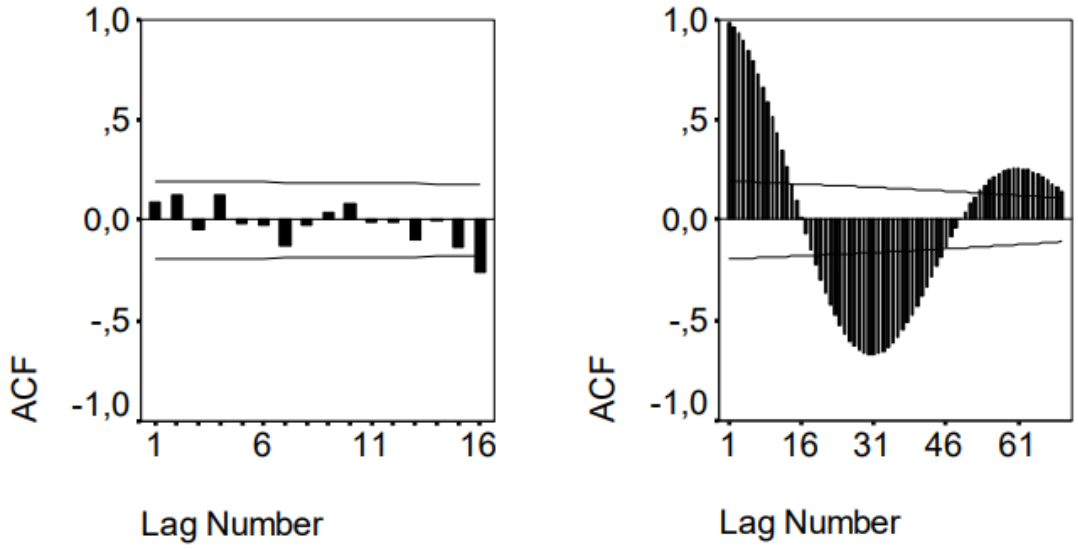


Рис. 25. Приклади автокореляційних функцій деяких рядів

Горизонтальними лініями на цих графіках показано довірчі інтервали для нульових значень коефіцієнта автокореляції. З наведених графіків добре видно, що у першому випадку автокореляція є практично відсутньою, а в другому – для певних значень параметра зсуву спостерігається істотна додатна або від'ємна автокореляція.

Найпростішим випадком є побудова одновимірної лінійної регресійної моделі. Її параметри можна визначити безпосередньо за формулами але в електронних таблицях ***MS Excel*** це можна зробити за допомогою вбудованих формул та пакета аналізу.

Для побудови лінійної моделі можна скористатися функцією “ЛИНЕЙ()”. На рис. 26 показано діалогове вікно задання її параметрів. Розглянемо такий приклад. Нехай ми маємо дві пов’язані вибірки обсягом по 41 елементу. Елементами першої вибірки xi є числа від –2 до 2, взяті у порядку зростання з кроком 0,1.

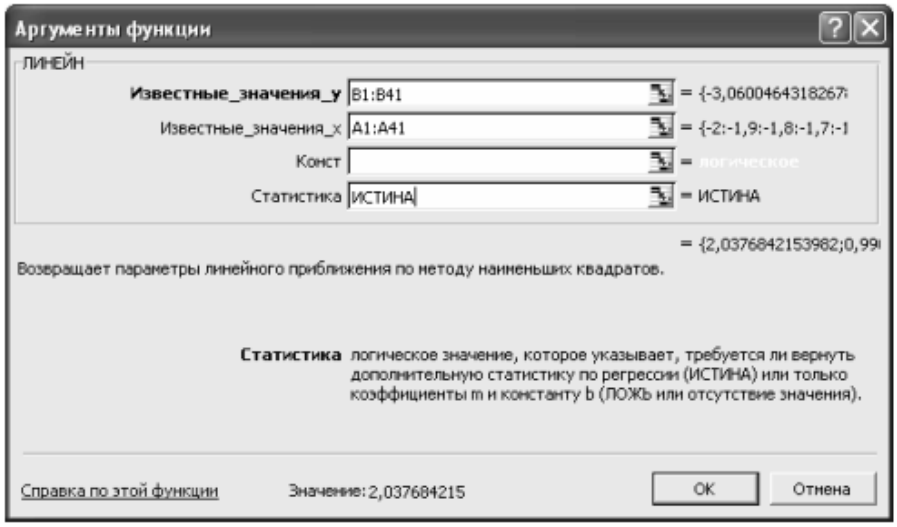
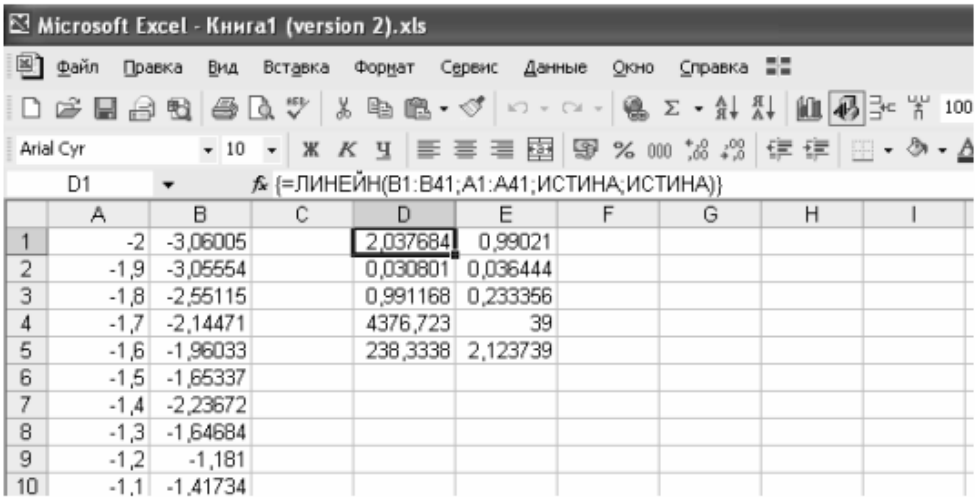


Рис. 26. Діалогове вікно задання параметрів функції “ЛИНЕЙ()”

Уведення формули необхідно здійснювати як формулу масиву. Для цього потрібно виділити на робочому аркуші масив суміжних комірок обсягом 2*5, записати формулу й після цього натиснути клавішу F2, а потім одночасно клавіші *Ctrl+Shift+Enter*. При цьому отримуємо масив результатів, наведений на рис. 27.



	A	B	C	D	E	F	G	H	I
1	-2	-3,06005		2,037684	0,99021				
2	-1,9	-3,05554		0,030801	0,036444				
3	-1,8	-2,55115		0,991168	0,233356				
4	-1,7	-2,14471		4376,723	39				
5	-1,6	-1,96033		238,3338	2,123739				
6	-1,5	-1,65337							
7	-1,4	-2,23672							
8	-1,3	-1,64684							
9	-1,2	-1,181							
10	-1,1	-1,41734							

Рис. 27. Результати обчислення параметрів лінійної регресії

Іншим варіантом побудови регресійної моделі в електронних таблицях ***MS Excel*** є застосування пакету аналізу. Для цього обираємо у головному меню: *Сервіс/Аналіз даних/Регресія.* Після цього відкривається діалогове вікно (рис. 28). У цьому вікні позначаємо посилання на комірки, де містяться значення змінних x, y. Позначку *“Метки”* робимо у випадку, коли перший стовпчик або перший рядок

вхідних даних містять заголовки. Позначку *“Константа – ноль”* робимо у випадку, коли вільний член моделі дорівнює нулю. У комірці *“Уровень надежності”* задаємо довірчий рівень (за умовчанням він дорівнює 0,95). Далі позначаємо, куди саме слід виводити результати, а також необхідність виводу статистичних характеристик моделі та побудови графіку нормальної імовірності [8] . Результати для тих самих вихідних даних, що і у попередньому випадку, показано на рис. 29.

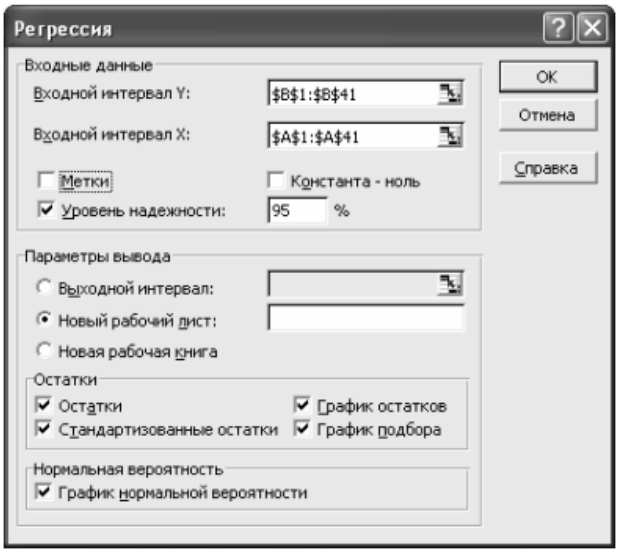


Рис. 28. Діалогове вікно побудови регресійної моделі в пакеті аналізу

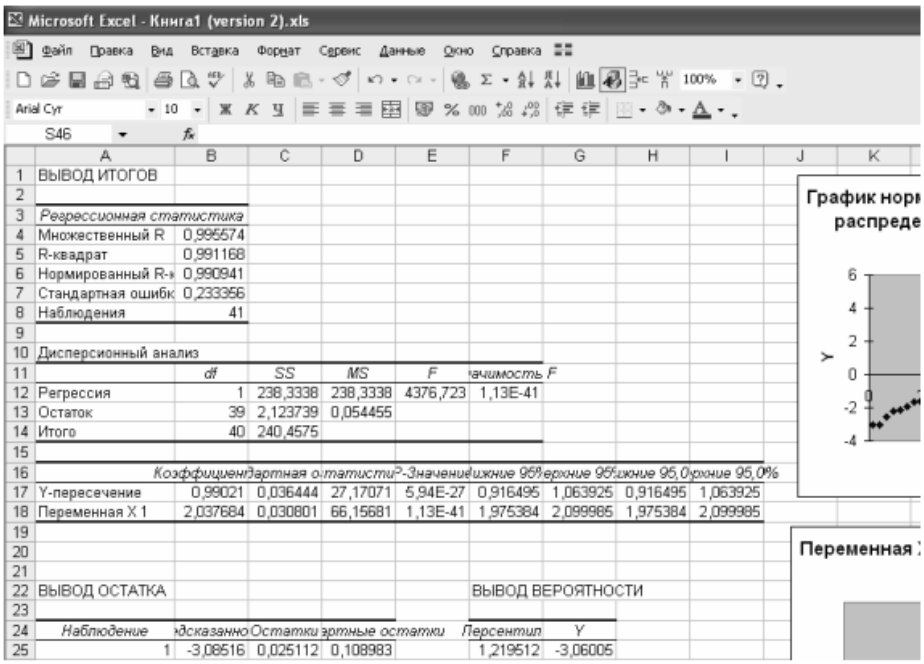
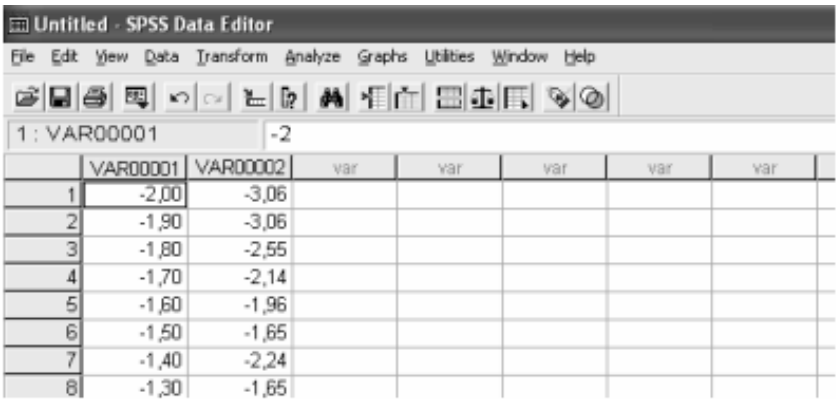


Рис. 29. Результати підбору параметрів регресійної моделі

Бачимо, що основні параметри є такими самими, що й у попередньому випадку. Проте слід зазначити, що за допомогою пакету аналізу ми можемо отримати більш докладну інформацію про властивості моделі. Крім того, використання пакету аналізу є простішим. Зокрема, воно не передбачає необхідності заздалегідь розраховувати й виокремлювати на робочому аркуші комірки для виводу результатів.

У пакеті **SPSS** також є різні засоби побудови регресійних моделей. Для побудови лінійної моделі заносимо дані аркуш даних (рис. 30) й використовуємо пункти меню: *Analyze/Regression/Linear* [8].



	VAR00001	VAR00002	var	var	var	var	var
1	-2,00	-3,06					
2	-1,90	-3,06					
3	-1,80	-2,55					
4	-1,70	-2,14					
5	-1,60	-1,96					
6	-1,50	-1,65					
7	-1,40	-2,24					
8	-1,30	-1,65					

Рис. 30. Аркуш даних пакету SPSS при побудові регресійної моделі

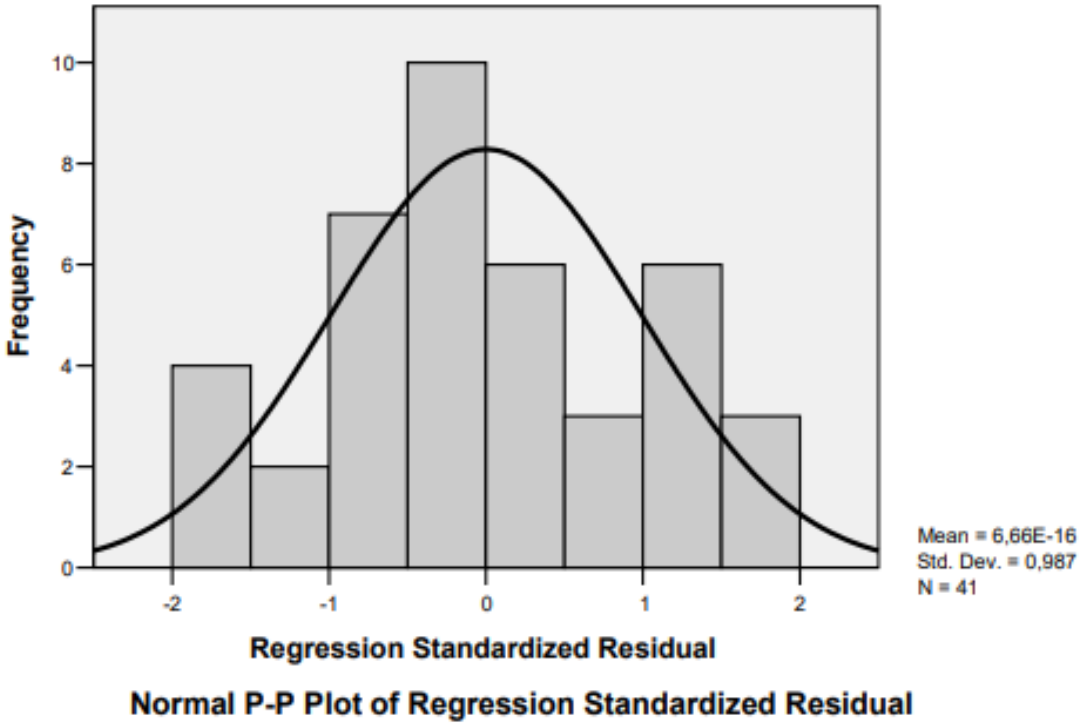
Деякі результати наведено на рис. 31, 32.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-3,0852	5,0656	,9902	2,44097	41
Std. Predicted Value	-1,670	1,670	,000	1,000	41
Standard Error of Predicted Value	,036	,072	,050	,011	41
Adjusted Predicted Value	-3,0878	5,1024	,9899	2,44293	41
Residual	-,39495	,45477	,00000	,23042	41
Std. Residual	-1,692	1,949	,000	,987	41
Stud. Residual	-1,717	1,984	,001	1,013	41
Deleted Residual	-,40664	,47153	,00029	,24275	41
Stud. Deleted Residual	-1,763	2,066	,004	1,031	41
Mahal. Distance	,000	2,787	,976	,883	41
Cook's Distance	,000	,132	,027	,034	41
Centered Leverage Value	,000	,070	,024	,022	41

a. Dependent Variable: VAR00002

Рис. 31. Результати побудови лінійної регресійної моделі у пакеті SPSS



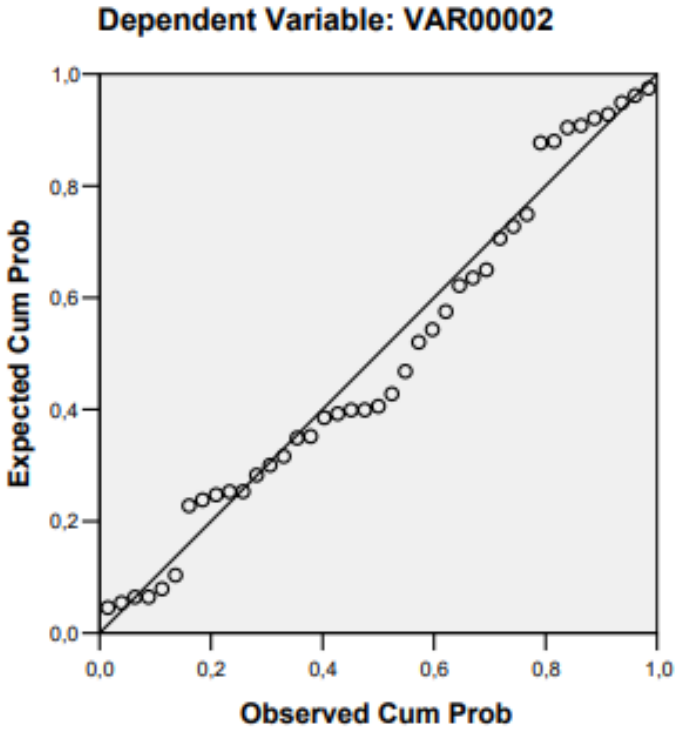


Рис. 32. Деякі графіки результатів побудови лінійної регресійної моделі

Бачимо, що вони збігаються з результатами, отриманими в електронних таблицях **MS Excel**, а також з вихідними даними. Але слід зазначити, що у пакеті **SPSS** ми маємо можливість отримати значно більше статистичних даних стосовно якості побудованої моделі.

Для побудови спеціальних моделей можна також використовувати спеціальні функції пакету **MathCad** [9] **.**

Результати побудови моделей показано на рис. 33, 34.

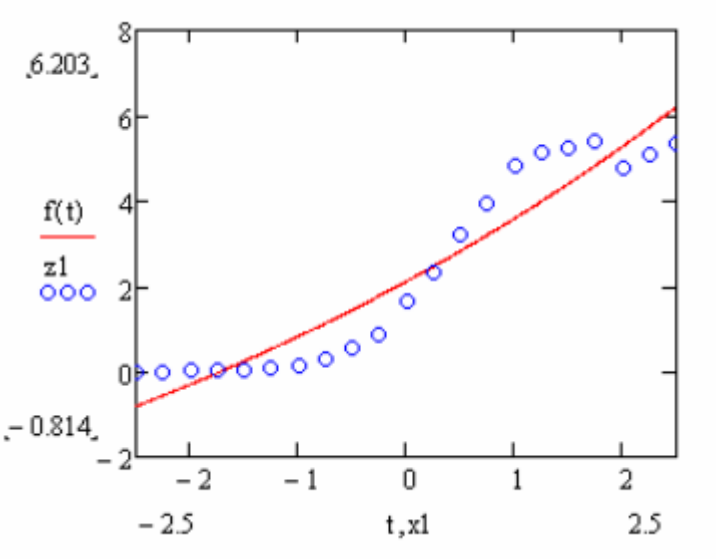


Рис. 33. Результат побудови експоненціальної регресійної моделі

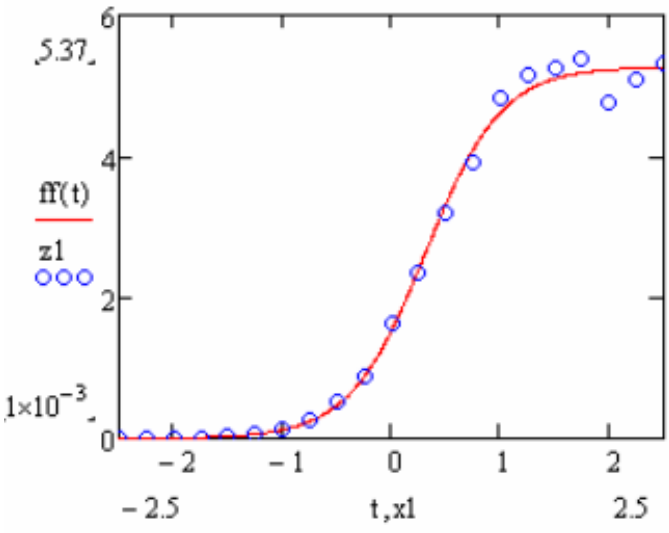


Рис. 34. Результат побудови логістичної регресійної моделі

Бачимо, що логістична модель значно краще відображає емпіричні дані, що у даному випадку є цілком природним оскільки вихідні дані побудовано саме на основі логістичної моделі. Дослідження впливу початкових значень параметрів моделей показує, що навіть для досить істотних їх відхилень від правильних значень, результат підбору параметрів моделей зазвичай є одним і тим самим. Але для окремих наборів вихідних значень алгоритм підбору параметрів не збігається.

Тема 8. Візуальне представлення даних.

Основні види та способи представлення даних, їхні особливості, переваги і недоліки. Вибір засобів та інструментів представлення даних.

Що таке візуалізація даних?

Візуалізація даних - це процес представлення даних та інформації у графічному вигляді для ілюстрації важливої інформації. Чому візуалізація даних є такою затребуваною технікою для багатьох компаній та осіб, які працюють в організаціях, що працюють з даними? Тому що зображення дійсно варте тисячі слів для тих, хто намагається швидко осмислити щось і прийняти швидке рішення стосовно цього.

Візуалізація даних дозволяє нам спростити складні ідеї та дані, зрозуміти тенденції та розпізнати закономірності у найпростіший спосіб. Сучасний світ висуває високий попит на будь-які технології чи процеси, які можуть допомогти перетворити складне на просте і дієве. Цей попит спричинив вибух можливостей візуалізації даних. Оскільки дані визначають більшу частину ділового світу, візуалізація даних зайняла своє місце серед найважливіших рішень, до яких звертається бізнес.

Хоча всі ми добре знайомі з круговими діаграмами або гістограмами, які використовуються для пояснення простих наборів даних, візуалізація даних може варіюватися від простого до складного. Візуалізація даних може також стосуватися інфографіки, ілюстрацій, відео-анімації, діаграм, схем, графіків тощо. Команди використовують візуалізацію даних для низки завдань, таких як ілюстрування ідей, генерування ідей та стратегій, а також концептуалізація даних у презентації для зацікавлених сторін, які не так добре орієнтуються в контексті, як дослідники даних [9].

Навіщо потрібна візуалізація даних?

Будемо відвертими. Не всі знаходять дані особливо захопливими. Проте всі люблять гарні історії. Люди - дуже візуальні істоти. Візуалізація даних дозволяє користувачам розповісти історію у візуальний спосіб, який може зацікавити й залучити тих, кому можуть набриднути прості цифри. Це найпростіший спосіб споживати складну інформацію. Це непросте завдання - вміти виявляти

закономірності, помічати тенденції та робити швидкі висновки про макро- та мікросередовище. Проте візуалізація даних дає нам таку можливість.

Сфери застосування візуалізації даних безмежні. Мабуть, мало знайдеться галузей чи ніш, які б не використовували візуалізацію даних та не отримували вигоду від її інсайтів. Майже кожен бізнес шукає спосіб виявити фактори, які впливають на рішення клієнтів, знайти свої слабкі місця, перетворити складні концепції на легко доступні для обговорення та прийняття рішень, робити прогнози, випереджати тренди, доносити до клієнтів інформацію, яка впливає на рішення про покупку, через маркетингові канали тощо. Завдяки цьому навіть ті, хто технічно не є науковцями з даних, стають експертами у своїх відділах з використанням візуалізації даних для ефективного виконання своїх обов'язків.

Чому так багато компаній та приватних осіб звертаються до візуалізації даних, щоб вразити клієнтів статистикою, прояснити концепції для акціонерів, висвітлити незрозумілі ситуації для колег по команді тощо? Існує кілька ключових переваг візуалізації даних у кожному аспекті роботи. По суті, візуалізація даних може дати нам перспективи щодо даних у легко засвоюваній формі. Завдяки цій здатності швидко засвоювати дані, компанії можуть швидше приймати рішення у світі, який вимагає оперативного реагування, щоб залишатися попереду.

Візуальна інформація краще сприймається і дозволяє швидко і ефективно донести до глядача власні думки та ідеї. Фізіологічно, сприйняття візуальної інформації є основною для людини. Є численні дослідження, які підтверджують, що:

- 90% інформації людина сприймає через зір
- 70% сенсорних рецепторів знаходяться в очах
- близько половини нейронів головного мозку людини задіяні в обробці візуальної інформації
- на 19% менше при роботі з візуальними даними використовується когнітивна функція мозку, що відповідає за обробку та аналіз інформації
- на 17% вище продуктивність людини, що працює з візуальною інформацією
- на 4,5% краще згадуються деталі візуальної інформації

Якщо попросити читача згадати назви материків, в більшості випадків у голові виникне саме ця картинка [9] .

- в 60000 разів швидше сприймається візуальна інформація в порівнянні з текстовою

На графіку читач швидше знайде мінімальне і максимальне значення

- 10% людина запам’ятовує з почутого, 20% – з прочитаного, і 80% – з побаченого і зробленого
- на 300% краще людина виконує інструкцію, якщо вона містить ілюстрації

Очевидно, що людина схильна обробляти саме візуальну інформацію. Крім прекрасної обробки нашим мозком, візуалізація даних має кілька переваг:

- Акцентування уваги на різних аспектах даних
- Аналіз великого набору даних зі складною структурою
- Зменшення інформаційного перевантаження людини і утримання його уваги
- Однозначність і ясність виведених даних
- Виділення взаємозв’язків і відносин, що містяться в інформації
- На графіку легко можна помітити важливі дані
- естетична привабливість
- Естетично привабливі графіки роблять подачу даних ефектною і такою, що запам’ятовується

Залежно від мети і даних можна вибрати найбільш підходящий їм графік. Найкраще уникати розмаїтості заради розмаїтості і вибирати за принципом “чим простіше, тим краще”. Тільки для специфічних даних використовувати специфічні

типи діаграм, в інших же випадках добре підійдуть найпоширеніші графіки [10] :

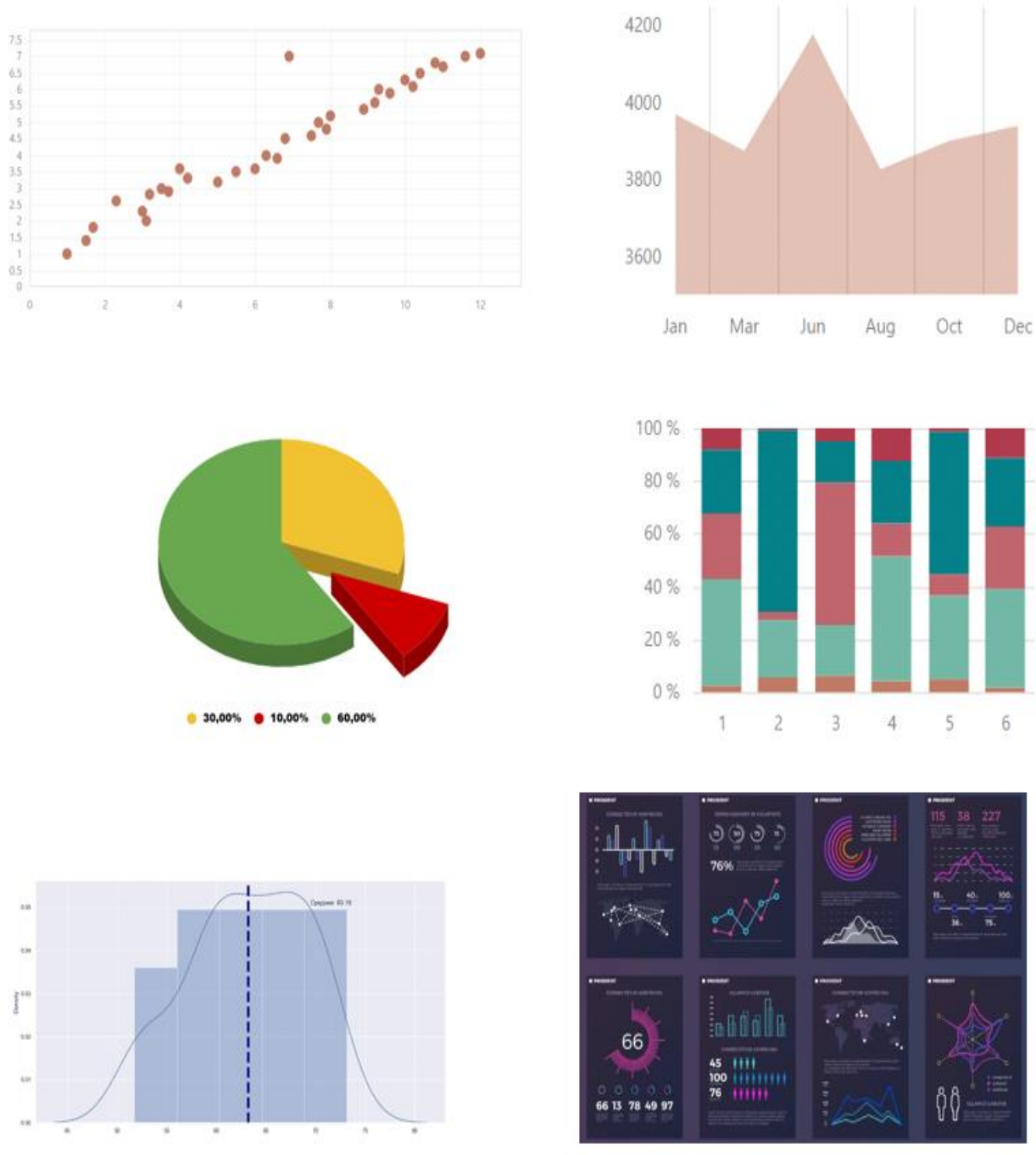


Рис. 35. Приклади візуалізації даних

Тільки для специфічних даних використовувати специфічні типи діаграм, в інших же випадках добре підійдуть найпоширеніші графіки:

- лінійний (***line***)
- з областями (***area***)
- колонки і гістограми (***bar***)
- кругова діаграма (***pie, doughnut***)
- полярний графік (***radar***)
- точковий графік (***scatter, bubble***)
- карти (***map***)
- дерева (***tree, mental map, tree map***)
- тимчасові діаграми (***time line, gantt, waterfall***).

При виборі відповідного графіка можна керуватися наступною таблицею 9, складеної на основі діаграми із книги “Говори мовою діаграм” Джина Желязни [10] :

Таблиця 9

Метод побудови графіків та діаграм Джина Желязни

ціль візуалізації/ тип даних	відносини даних	розподілення даних	порівняння даних	компазиція даних
неперервні числові	line, area, scatter, bubble	scatter, bubble	line, area, radar	stacked line, full stacked line, stacked area, full stacked area
неперервні часові	line, area, radar, scatter, bubble	time line, gant, waterfall, radar		gant, stacked line, full stacked line, stacked area, full stacked area
дескретні	bar, scatter, bubble		bar, pie, doughnut	pie, doughnut, stacked bar, full stacked bar
географічні	map, line, area	map, scatter	map, bar	map, stacked bar, full stacked bar
логічні	tree, mental tree		tree map	

Вибір засобів та інструментів представлення даних

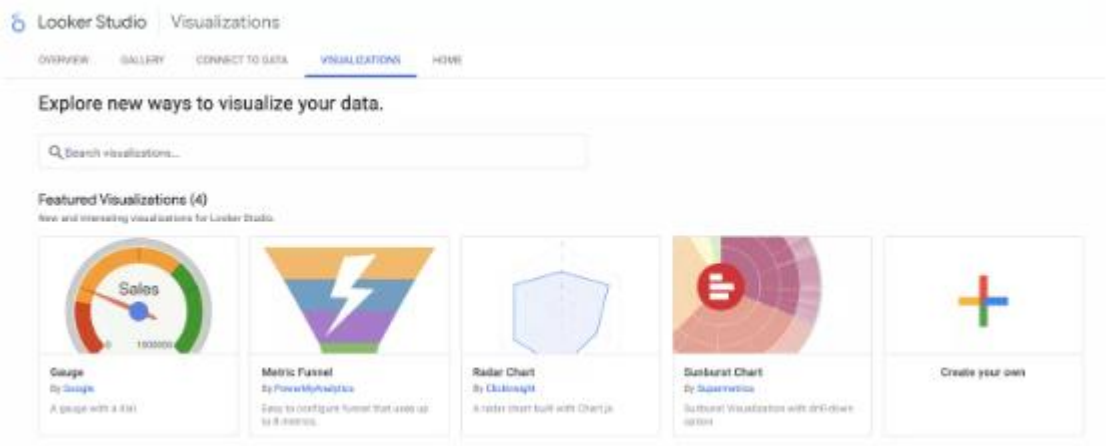
Прості засоби візуалізації включено до сучасних електронних таблиць. Вони не охоплюють всього різноманіття технік, але для простих задач і оперативного прототипування цілком годяться.

Ось список найкращих інструментів візуалізації даних.

- **Tableau**: Потужний інструмент для створення карт і візуалізацій для клієнтів.
- **Power BI**: інструмент, що використовує бізнес-аналітику для створення гнучкої інформаційної структури, заснованої на даних.
- **DataWrapper**: Інструмент для збагачення історій картами, діаграмами та графіками.
- **Google Charts**: Безкоштовний інструмент візуалізації даних, що використовується для створення інтерактивних діаграм, які пов'язують дані між собою.
- **Plecto**: Інструмент, який створює інформаційні панелі для співробітників, щоб мотивувати їх та тримати в тонусі.
- **RawGraphs**: Інструмент, який допомагає спростити складні дані за допомогою привабливого візуального представлення.
- **PolyMaps**: Безкоштовна бібліотека JavaScript з відкритим вихідним кодом, яка використовується для створення динамічних карт, щоб ви могли відображати свої дані за допомогою різних стилів візуальної презентації [9].

Google Looker Studio

Це простий у використанні інструмент зі стильним дизайном. Допомагає легко створювати інтерактивні діаграми та звіти.



Переваги

1. Безкоштовний інструмент, що вагомо для стартапів і невеликих бізнесів.
2. Має інтеграцію з Google, тому легко працює з іншими Google-сервісами.
3. Надає можливість автоматичного оновлення даних, що спрощує підтримку актуальності інформації.

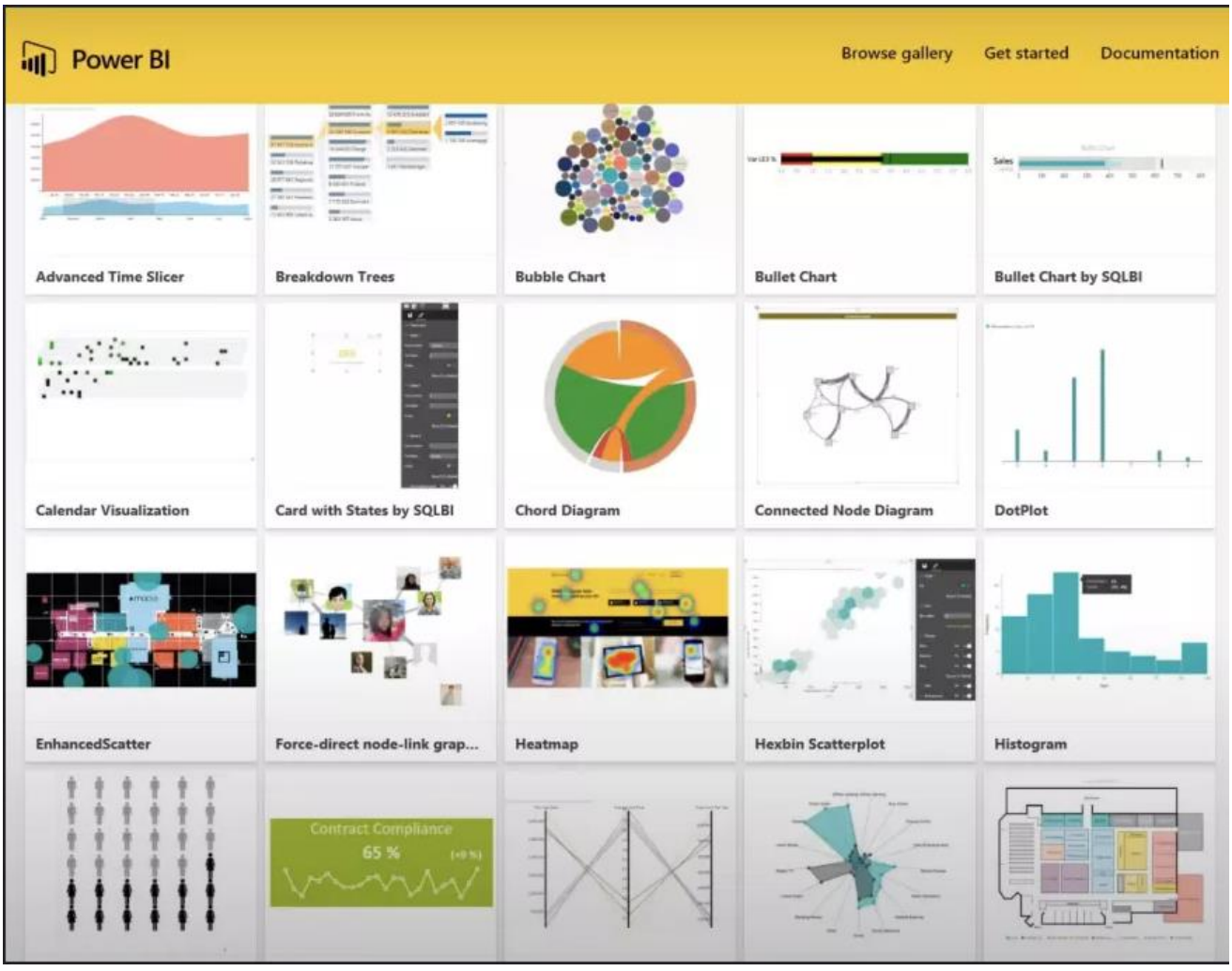
Недоліки

1. Повільний при обробці об’ємних даних.

Microsoft Power BI (PBI)

Пропонує широкий спектр функцій:

- створення діаграм і власних метрик;
- аналіз великих обсягів даних;
- використання двох мов обробки даних: DAX (Data Analysis Expressions) і M (Power Query Formula Language).



Переваги

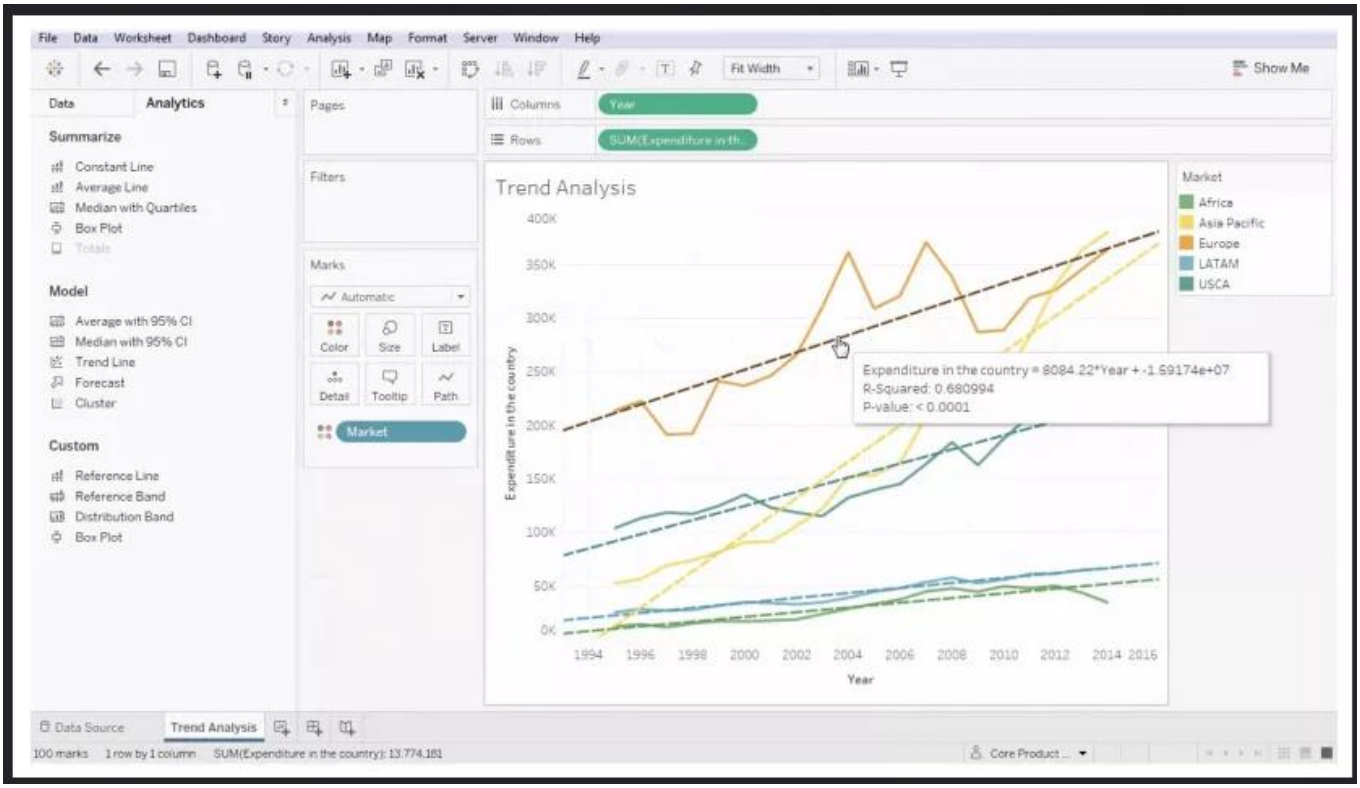
1. Інтеграція з Microsoft. Гарантує гармонійну роботу з Excel, Azure тощо.
2. Функціональні можливості для глибокого аналізу.

Недоліки

1. Виникають труднощі в роботі з пристроїв Mac.
2. Не для новачків. Вивчення інструменту вимагає часу і зусиль.

Tableau

Цей потужний інструмент дозволяє створювати креативні діаграми та глибше досліджувати дані завдяки інтерактивним елементам. Tableau підходить для поціновувачів високої якості, що не бояться складнощів.



Переваги

1. Інтерактивні можливості інструменту дозволяють створювати якісні діаграми.
2. Велика спільнота людей, що порадять і допоможуть.

Недоліки

1. Висока ціна.
2. Потрібен час для освоєння функцій.

Google Data Studio, Microsoft Power BI і Tableau — лідери на ринку, заслужено визнані й популярні. Залежно від потреб і завдань проєкту, доповнюйте їх додатковими інструментами.

1. **QlikSense**. Потужне програмне забезпечення для візуалізації та бізнес-аналітики, що приваблює інтуїтивно зрозумілим інтерфейсом. Спрямоване на самостійних користувачів.

2. QlikView. Розрахований на досвідчених аналітиків, оскільки вимагає глибокого розуміння процесів моделювання даних. Відзначається хорошою продуктивністю при роботі з великими об'ємами інформації і складними аналітичними завданнями.

Недоліки:

- робота з QlikView вимагає більше часу та зусиль, у порівнянні з іншими інструментами;
- інструмент дозволяє користувачам ділитися поточними сеансами з гостями, але є обмеження на кількість користувачів, що одночасно використовують спільний сеанс;
- велика кількість користувачів, що діляться одним сеансом, призведе до зниження продуктивності.

3. **Infogram**. Простий інструмент з великим вибором шаблонів. Його інтерактивні можливості дозволяють створювати прості та привабливі візуалізації. Найкраще підходить для узагальнення і пояснення результатів аналізу даних.

Недоліки:

- низка стандартних налаштувань порушують правила ефективного створення діаграм;
- у безкоштовній версії відсутні аналітичні функції.

4.Adobe Analytics. Інструмент для аналізу даних з гнучкою сегментацією, повним контролем над даними й широкими можливостями для інтеграцій. Adobe Analytics перетворює аналітику в захопливий процес вдосконалення продуктивності та вивчення даних.

Недоліки:

- висока вартість, що залежить від розміру даних і пакету
- складність у вивченні інструменту.

5. **Sisense**. Хороший вибір для складних проєктів, завдяки вражаючим аналітичним можливостям і високій продуктивності. Компанія пропонує змінні тарифи і стягує плату лише за необхідні послуги.

Недоліки:

- висока вартість, що залежить від конкретних потреб і об’єму даних, а саме
- для запуску програми потрібен потужний комп’ютер;
- будьте готовими до складнощів у налаштуваннях.

6. **Grafana.** Інструмент для ефективного керування даними з відкритим вихідним кодом і гнучким інтерфейсом.

Недоліки:

- складність налаштування;
- обмеженість аналітичних можливостей.

7. **Cognos Analytics**. Інструмент, представлений IBM, має потужні аналітичні функції та інтеграції з іншими системами.

Недоліки:

- складність налаштування й навчання;
- вартість

8. **Excel і Google Sheets**. Доступні інструменти, котрі приваблюють простотою використання.

Недоліки:

- обмеження в аспектах візуалізації, продуктивності й автоматизації;
- не завжди задовольняють потреби в складних аналітичних завданнях і при обробці великих обсягів даних.

STATISTICA – універсальний пакет статистичного аналізу, в якому реалізовані основні математичні методи аналізу даних.

Розробником пакету є фірма StatSoft, Inc (США). У 2014 р. ця фірма була поглинута корпорацією Dell, яка включила пакет STATISTICA до складу власної лінійки програмного забезпечення проблематики великих даних.

STATISTICA дозволяє проводити різні процедури (модулі) обробки статистичних даних (в термінології програми – аналізи):

1. Розрахунок описових статистик.
2. Аналіз динамічних рядів й прогнозування.
3. Множинна регресія.
4. Дискримінантний аналіз.
5. Аналіз відповідностей.
6. Кластерний аналіз.
7. Факторний аналіз.
8. Дисперсійний аналіз і та ін.

Крім загальних статистичних і графічних засобів STATISTICA має спеціалізовані модулі: для проведення соціологічних або біомедичних досліджень, вирішення технічних і, що дуже важливо, промислових завдань: карти контролю якості, аналіз процесів і планування експерименту [9] .

IBM SPSS Statistics від компанії IBM – це аналітичне програмне забезпечення, яке дозволяє проводити просунутий статистичний аналіз ділових даних, охоплюючи вирішення всіх завдань від планування та збору даних до безпосереднього аналізу та побудови бізнес-звітності.

IBM SPSS Statistics призначена для статистичного аналізу даних і дозволяє отримувати корисну інформацію з досліджуваних даних. Програмне забезпечення IBM SPSS Statistics використовується багатьма компаніями, дослідницькими центрами та незалежними аналітичними агентствами для вирішення специфічних власних специфічних бізнес-завдань, забезпечуючи вироблення якісних рішень.

Просунуті статистичні процедури та візуалізація в системі IBM SPSS Statistics забезпечують надійну, зручну та інтегровану платформу для розуміння предметних даних та вирішення складних бізнес- та дослідницьких завдань, забезпечуючи загалом збільшення доходів, перевагу над конкурентами, проведення досліджень та прийняття якісних рішень на базі фактичних даних. Програмне забезпечення підтримує такі методи статистичних досліджень: регресійний аналіз, дерева рішень, прогнозування, нейронні мережі, категоризація, комбінаційний аналіз, складні вибірки та інші.

Серед важливих особливостей інформаційно-аналітичної системи SPSS Statistics можна назвати:

- система забезпечує індивідуальне налаштування функцій та інтерфейсів для різних рівнів кваліфікації та функціональних обов’язків;

- SPSS Statistics охоплює всі основні частини комплексного аналітичного процесу - від підготовки даних та управління ними до аналізу та звітності;

- надає готові до використання шаблони графіків та звітів, дозволяючи легко переводити отримані результати у презентабельний та зрозумілий вигляд для інших зацікавлених осіб.

Програма **SPSS Statistics** має такі переваги:

- забезпечує швидке розуміння великих та складних наборів даних за допомогою сучасних статистичних процедур, які допомагають забезпечити високу точність та якість прийняття рішень;

- дозволяє використовувати програмні розширення мовами програмування Python і R для інтеграції з програмним забезпеченням з відкритим вихідним кодом;

- полегшує вибір та керування програмним забезпеченням, надаючи гнучкі варіанти розгортання.

Функції IBM SPSS Statistics:

- Звітність та аналітика.
- Імпорт/експорт даних.
- Статистичний аналіз.
- Прогнозування та передбачувана аналітика.
- Інтерактивна аналітична обробка (OLAP).
- Конектори для джерел даних.
- Індикація трендів та проблем.
- Інтелектуальний аналіз даних (ІАД).
- Візуалізація даних.
- Аналіз великих даних.

Інформаційні технології перевірки етичності досліджень. Сервіс *UNICHECK*.

Щодо можливих шляхів виявлення наявності плагіату є два способи:

- ручний пошук, що здійснюється безпосередньо викладачами, науковцями, редакторами, читачами журналів;
- автоматичний пошук за допомогою комп'ютерної техніки та програмних засобів.

Наведемо характеристики поширених он-лайн ресурсів для пошуку текстового плагіату (таблиця 11).

Таблиця 10

Поширені он-лайн ресурси для пошуку текстового плагіату

№	Назва ресурсу	Доступ	Типи перевірки			Обсяг тексту для перевірки
			Текстові фрагменти	Файли	URL	
1	Duplichecker	вільний	+	+ (до 50 кБ)	+ (за умови реєстрації)	1500 слів
2	PaperRater	вільний	+			необмежений
3	Plagiarisma.Net	вільний	+	+	+	2000 символів
4	Plagium	платний	+	+	+	до 25000 символів безкоштовно
5	PlagTracker	вільний	+	+ (за умови реєстрації та придбання Premium акаунта)		до 5000 слів безкоштовно
6	SeeSources	вільний (1 перевірка безкоштовно)	+ (до 300 кБ)			до 1000 слів
7	PlagScan	передплата	+	+		
8	Plagiarism Detector	вільний	+	+		
9	Docol©c	передплата (демо-режим – безкоштовно)	+			

Додаткові можливості он-лайн ресурсів для пошуку текстового плагіату:

- ***Duplichecker*** – порівняльний пошук плагіату в двох введених текстових фрагментах або за URL посиланнями на дві веб-сторінки;
- ***PaperRater*** – містить автоматизований літературний редактор для перевірки граматики, правопису та стилістики;
- ***Plagiarisma.Net*** – підтримується понад 190 мов; є можливість окремого пошуку плагіату по базах даних Google Scholar та Google Books;
- ***Plagium*** – 6 європейських мов;
- ***PlagScan*** – є можливість одночасної перевірки декількох документів та завантаження їх одним .zip архівом;
- ***Plagiarism Detector*** – автоматично сканує текстові документи, екстрагує дані та порівнює їх з доступними джерелами в мережі Інтернет;
- ***Docol©c*** – безкоштовно можна працювати в демо-режимі, попередньо зареєструвавшись на сайті
- ***Grammarly*** – додатково доступний для завантаження безкоштовний інструмент Grammarly Lite для веб-браузерів **Chrome, Firefox і Safari**, що дозволяє здійснювати поточну перевірку орфографії під час роботи в мережі Інтернет, виявляти запозичені тексти, дозволяє оформити їх в цитати.

До недоліків розглянутих он-лайн ресурсів для пошуку текстового плагіату слід віднести наступні:

- ***Duplichecker*** – обмежена кількість перевірок (1 в день для незареєстрованих користувачів і до 50 в день – для зареєстрованих);
- ***PaperRater*** – звіт стислий, не містить статистики, не відображені запозичені франменти тексту;

- ***Plagiarisma.Net*** – обмежена кількість перевірок (3 в день); для безлімітного користування слід придбати пакет Premium;
- ***Plagium*** – платний доступ;
- ***PlagTracker*** – завантаження файлів формату *.doc та *.txt, а також текстів понад 5000 слів лише за умови реєстрації та придбання Premium акаунта;
- ***SeeSources*** – обмежена кількість безкоштовних перевірок; обмеження за розміром завантажуваних файлів до 300kB та текстових фрагментів до 1000 слів;
- ***Plagiarism Detector*** – програма розбиває текст на абзаци, кожен з яких перевіряється окремо;
- ***Docol©c*** – необхідна реєстрація на сайті та передплата ліцензійного доступу, розмір якої залежить від запланованої кількості сторінок, що перевірятимуться за рік;
- ***Grammarly*** – платний доступ; лише англомовні тексти; не підтримується формат *.pdf; є обмеження по обсягу тексту [11] .

Рекомендації:

1. Візуалізація робить дані доступнішими, сприяє запам’ятовуванню та дозволяє виявляти тренди й аномалії.
2. Підводні камені візуалізації: перекручення інформації, перенасиченість елементами, довгий час на навчання і висока вартість програмного забезпечення.
3. Сучасні інструменти на кшталт **Looker Studio**, **Microsoft Power BI та Tanleau** перетворюють сухі цифри на візуальні історії. Потрібно правильно обрати інструмент, зважаючи на його сильні сторони.
4. Зведені таблиці, графіки, картограми тощо допомагають бачити дані, шукати інсайти та можливості для росту. Важливо правильно обрати тип графіки залежно від цілі її створення.
5. Уникайте зайвого візуального шуму і зменшіть кількість графічного сміття, щоб отримати чіткі та лаконічні візуалізації. Кольорова гама і контраст грають важливу роль у легкому сприйнятті графічної інформації.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. В.Є. Бахрушин. Методи аналізу даних : навчальний посібник для студентів /. – Запоріжжя : КПУ, 2011. – 268 с. http://kist.ntu.edu.ua/textPhD/metDataManing.pdf
2. Василенко О. А., Сенча І. А. Математично-статистичні методи аналізу у прикладних дослідженнях: навч. посіб. Одеса:ОНАЗ ім. О. С. Поп, 2011.166 с.
3. Математичні та статистичні методи аналізу соціологічної інформації : Практикум / Л. Ф. Панченко; КПІ ім. І.Сікорського, 2018. – 289 с.
4. Майборода Р.Є. Регресія: лінійні моделі: навч. посіб. Київ: ВПЦ «Київ. ун-т», 2007. 296 с.
5. Бахрушин В. Є. Математичне моделювання. Запоріжжя : ГУ "ЗІДМУ", 2003. 138 с.
6. Zaselskiy, I., Sokur, M., Biletskyi, V., Fyk, M., Fyk, O. The study of the lining layer abrasing wear in the semiautogenous grinding mill/E3S Web of Conferences, 2020, 166, 03008 https://www.scopus.com/record/display.uri?eid=2-s2.0-85084958406&origin=inward&txGid=8e5ff62c9555178b9d32a8697a968393
7. Popolov, D., Shved, S., Zaselskiy, I., Pelykh, I., Studying of movement kinematics of dynamically active sieve/Mechanics and Mechanical Engineering, *2019, 23(1), p. 94–97.* http://www.scopus.com/inward/record.url?eid=2-s2.0-85069804653&partnerID=MN8TOARS.
8. Іващенко П. О., Семеняк І. В., Іванов В. В. Багатовимірний статистичний аналіз. Харків : Основа, 1992. 144 с.
9. Грицюк П. М., Остапчук О. П. Аналіз даних : навч. посіб. Рівне : НУВГП, 2008. 218 с. 2. Методи інтерпретації емпіричних даних. Stud. URL: https://stud.com.ua/144071/psihologiya/metodi_interpretatsiyi_empirichnih_danih
10. Учасники проектів Вікімедіа. Аналіз даних – Вікіпедія. Вікіпедія. https://uk.wikipedia.org/wiki/%D0%90%D0%BD%D0%B0%D0%BB%D1%96%D0%B7_%D0%B4%D0%B0%D0%BD%D0%B8%D1%85
11. Iryna Kryvenko. Факторний аналіз - обчислення, 2021. YouTube. URL: https://www.youtube.com/watch?v=5v4s1dFTJ-k