МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Національний аерокосмічний університет ім. М. Є. Жуковського
«Харківський авіаційний інститут»

С. С. Курєннов, К. П. Барахов

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ

Навчальний посібник

Харків «ХАІ» 2024

УДК 519.2(075.8)
К93

Рецензенти: д-р техн. наук, проф. В. А. Ванін,
д-р фіз.-мат. наук, проф. О. В. Макарічев

Курєннов, С. С.

К93 Інтелектуальний аналіз даних [Електронний ресурс] : навч. посіб. / С. С. Курєннов, К. П. Барахов. – Харків : Нац. аерокосм. ун-т ім. М. Є. Жуковського «Харків. авіац. ін-т», 2024. – 132 с.

Викладено основи інтелектуального аналізу даних. Розглянуто основні поняття й методологічні принципи аналізу даних.

Матеріал подано на рівні, доступному для студентів, які ознайомлені з курсом вищої математики. Методи, розглянуті в посібнику, проілюстровано значною кількістю прикладів.

Для студентів спеціальностей таких галузей знань: 05 «Соціальні та поведінкові науки», 07 «Управління та адміністрування», 11 «Математика та статистика», 12 «Інформаційні технології».

Іл. 67. Табл. 4. Бібліогр.: 9 назв

УДК 519.2(075.8)

ЗМІСТ

1. ВСТУП У BIG DATA

1.1. Предмет і завдання аналізу даних

Більшість організацій під час своєї діяльності накопичують величезні обсяги даних, але головне, що вони хочуть від них отримати, – це корисну інформацію. Як можна дізнатися з даних про те, що є вигіднішим для клієнтів організації, як ефективно розмістити ресурси або як мінімізувати втрати? Для вирішення цих проблем призначено новітні технології інтелектуального аналізу, у яких для знаходження моделей і відносин, прихованих в середовищі даних, використовуються моделі, які не можуть бути знайдені звичайними методами.

Модель – це абстрактне подання реальності. Існують два види моделей: прогнозувальні й описові. Перші моделі використовують один набір даних з відомими результатами для побудови моделей, які явним чином прогнозують результати для інших наборів, а другі – описують залежності в наявних даних, які, своєю чергою, використовуються для прийняття рішень або дій. Звичайно ж, компанія, що давно знаходиться на ринку і знає своїх клієнтів, користується безліччю моделей. Технології інтелектуального аналізу можуть не тільки підтвердити ці емпіричні спостереження, але й знайти нові, невідомі раніше моделі. Спочатку це може дати користувачеві лише невелику перевагу, але якщо аналіз об'єднати по кожному товару і кожному клієнтові, то відбувається великий відрив від тих, хто не використовує таких технологій. З іншого боку, за допомогою методів інтелектуального аналізу можна знайти таку модель, яка приведе до радикального поліпшення фінансового стану компанії та її положення на ринку.

Термін Data Mining складається з двох понять: пошуку цінної інформації у великій базі даних (Data) і видобутку гірської руди (Mining). Обидва процеси потребують або просіювання величезної кількості сирого матеріалу, або розумного дослідження й пошуку корисних цінностей. Найчастіше Data Mining переводиться як видобуток даних, витягання інформації, розкопка даних, інтелектуальний аналіз даних, засоби пошуку закономірностей, витягання знань, аналіз шаблонів, «витягання зерен знань з гір даних», розкопка знань в базах даних, інформаційна проходка даних, «промивання» даних. Поняття «виявлення знань у базах даних» (Knowledge Discovery in Databases, KDD) можна вважати синонімом Data Mining.

В основу сучасної технології Data Mining покладено концепцію шаблонів (патернів), що відображають фрагменти багатоаспектних взаємозв’язків у даних. Ці шаблони є закономірностями, властивими підвибіркам даних, які можуть бути компактно виражені в зрозумілій людині формі. Пошук шаблонів проводиться методами, не обмеженими

апріорними припущеннями про структуру вибірки і видом розподілів значень аналізованих показників. Важливе положення Data Mining – нетривіальність шуканих шаблонів. Це означає, що знайдені шаблони мають відображати неочевидні, несподівані (unexpected) регулярності в даних, такі, що являють собою так звані приховані знання (hidden Knowledge). До суспільства прийшло розуміння, що неякісні дані (raw Data) містять глибинний пласт знань, при грамотній розкопці якого можуть бути виявлені справжні «самородки» (рис. 1.1).

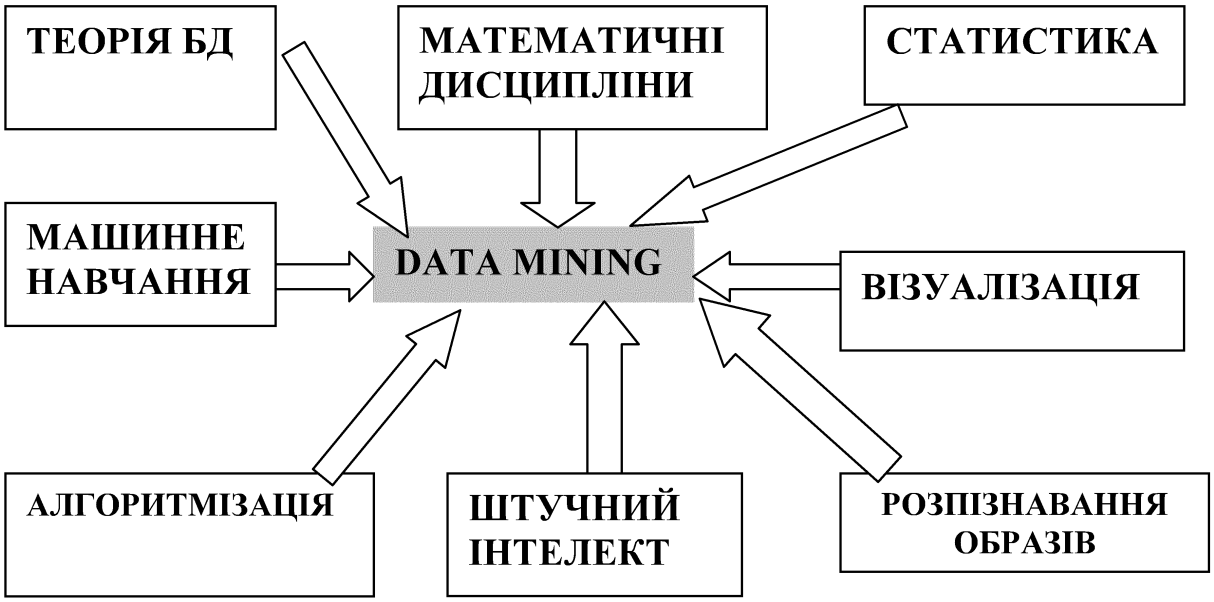


Рис. 1.1. Зв’язок Data Mining з іншими напрямами науки

Суть й мету технології Data Mining можна охарактеризувати так: це технологія, призначена для пошуку у великих обсягах даних неочевидних, об'єктивних і корисних на практиці закономірностей. **«Неочевидних»** – це означає, що знайдені закономірності не виявляються стандартними методами обробки інформації або експертним шляхом. **«Об'єктивних»** – це означає, що виявлені закономірності повністю відповідатимуть дійсності, на відміну від експертної думки, яка завжди є суб'єктивною. **«Практично корисних»** – це означає, що висновки мають конкретне значення, якому можна знайти практичне застосування.

Під загальним терміном «Big Data» розуміють будь-які набори даних, достатньо великі й складні для того, щоб їх можна було обробити традиційними засобами роботи з даними (наприклад, системами курування базами даних). У концепцію **Data Science** входить використання методів аналізу даних і витяг з них наявної інформації.

Характеристики великих даних часто називають «трьома V»:

– обсяг (Volume);

– різноманітність (Variety) – наскільки відрізняються один від одного різні типи даних;
– швидкість (Velocity) – з якою швидкістю надходять нові дані.

Часто ці характеристики доповнюються «четвертим V» – **достовірністю** (Veracity) – наскільки дані є точними. Перелічені властивості виявляються у всіх аспектах: зборі даних, зберіганні й обслуговуванні, пошуку, обміні, візуалізації.

Data Science – це певне розширення статистики, яке додає методи computer science в репертуар статистики.

1.2. Кому це потрібно?

Сфера застосування *Data Mining* нічим не обмежена – вона скрізь, де є які-небудь дані. Але насамперед методи Data Mining сьогодні, м'яко кажучи, заінтригували комерційні підприємства, що розгортають проєкти на основі інформаційних сховищ даних (*Data Warehousing*). Досвід багатьох таких підприємств свідчить про те, що економічна віддача від використання *Data Mining* може становити 100 %. А в деяких напрямах науки й техніки *Data Mining* є єдиним методом оброблення даних.

Банківська справа. Досягнення технології Data Mining використовуються в банківській справі для вирішення таких поширених завдань:

1. *Виявлення шахрайства з кредитними картками*. Шляхом аналізу минулих транзакцій, які згодом виявилися шахрайськими, банк визначає деякі стереотипи такого шахрайства.
2. *Сегментація клієнтів*. Розбиттям клієнтів на різні категорії банки роблять свою маркетингову політику більш цілеспрямованою й результативною, пропонуючи різні види послуг різним групам клієнтів.
3. *Прогнозування змін клієнтури*. За допомогою Data Mining банки будують прогнозні моделі цінності своїх клієнтів і відповідним чином обслуговують кожну категорію.

Телекомунікації. В області телекомунікацій за допомогою методів Data Mining компанії більш енергійно просувають свої програми маркетингу й ціноутворення для утримання наявних клієнтів і залучення нових. Серед типових заходів зазначимо такі:

1. *Аналіз записів* з докладними характеристиками викликів. Призначення такого аналізу – виявлення категорій клієнтів зі схожими стереотипами користування послугами компанії і розроблення привабливих наборів цін і послуг.

2. *Виявлення лояльності клієнтів*. Data Mining можна використовувати для визначення характеристик клієнтів, які, один раз скориставшись послугами компанії, з великою часткою ймовірності залишаться їй вірними. Унаслідок цього кошти, які виділяються на маркетинг, можна витрачати там, де віддача є найбільшою.

Страхування. Страхові компанії протягом кількох років накопичують великі обсяги даних. Тут велике поле діяльності для методів Data Mining:

1. *Виявлення шахрайства*. Страхові компанії можуть знизити рівень шахрайства, відшукуючи певні стереотипи в заявах про виплату страхового відшкодування, що характеризують взаємини між юристами, лікарями і заявниками.
2. *Аналіз ризику*. Шляхом виявлення поєднань факторів, пов'язаних з оплаченими заявами, страховики можуть зменшити свої втрати за зобов'язаннями.

Сферами застосування Data Mining також є медицина, молекулярна генетика, астрономія (пошук екзопланет і позаземного життя), навігація, оптимізація руху транспорту, соціологія, політична агітація (наприклад, за допомогою аналізу соціальних мереж) тощо.

1.3. Типи даних

У Data Science і Data Mining існує багато різних типів даних, для кожного з яких застосовуються власні інструменти й методи. Класифікувати дані можна, наприклад, таким чином:

- *структуровані;*
- *неструктуровані;*
- *природною мовою (книги, статті тощо);*
- *машинні;*
- *графові;*
- *аудіо- і відеодані, графіка;*
- *потокові.*

Структуровані дані залежать від моделі й зберігаються у фіксованому полі всередині запису. Такі дані зберігаються в таблицях, базах даних тощо. Цей тип даних є одним з основних засобів зберігання даних, тому що є зручним і для людини, і для машин.

Неструктуровані дані важко підігнати під конкретну модель даних. Один з прикладів таких даних – повідомлення електронної пошти. Вони мають структуровані елементи (відправник, звернення, тіло повідомлення, …), але є велика кількість засобів реалізації цих елементів. Це також може бути прикладом даних природною мовою.

Оброблення даних природною мовою загалом є складним і потребує як знань лінгвістики, так і спеціальних знань Data Science. Це новий напрям у науці даних.

Машинні дані – інформація, яку генерують процесори, датчики, комп'ютери. Ці дані є основним джерелом інформації. Прикладами таких даних є журнали вебсерверів, записи телефонних дзвінків, журнали мережних подій, телеметрія.

Аудіо- і відеодані, графіка – це тривіальні, з точки зору людини, дані (наприклад, розпізнавання того, що зображено на картинці), що є складними для комп'ютера. Наприклад, існують системи, які дають змогу за допомогою аналізу зображень з відеокамер на футбольному стадіоні відобразити рух м'яча та кожного з гравців, визначити його активність, помилки або внесок у перемогу.

Потокові дані можуть набувати різних форм і характеризуються тим, що не завантажуються великими масивами, а є розгорнутими в часі. Приклади таких даних – інтернет-трафік, дані з радіолокаторів, дані GPS-моніторингу переміщення автомобілів і літаків тощо.

Ще один тип класифікації даних:

1. *Безперервні дані (continuous)* – дані, які можуть набувати будь-якого значення в інтервалі.
   Синоніми: інтервал, число з плаваючою точкою, числове значення.
2. *Дискретні дані (discrete)* – дані, які можуть набувати тільки цілочислових значень, зокрема кількісних.
   Синоніми: ціле число, кількість.
3. *Категоріальні дані (categorical)* – дані, які можуть набувати тільки певного набору значень, зокрема набору можливих категорій.
   Синоніми: перерахунки, перелічувані дані, фактори, іменовані дані.
4. *Двійкові дані (binary)* – особливий випадок категоріальних даних усього з двома категоріями значень (0/1, істина/неправда).
   Синоніми: дихотомічні, логічні, індикатори, логічне значення.
5. *Порядкові дані (ordinal)* – категоріальні дані з явно вираженою впорядкованістю.
   Синоніми: порядковий фактор.

У програмних системах науки про дані, таких як R і Python, ці типи даних використовуються для підвищення швидкості розрахунків. А більш важливим є те, що тип даних змінної визначає спосіб роботи програмної системи з обчисленнями для цієї змінної.

1.4. Процес Data Science

Процес Data Science зазвичай складається з шести етапів (рис. 1.2).

У проєктному завданні до дослідження вказано, що саме треба досліджувати, яку користь це дасть замовнику, які ресурси для цього

потрібні. На першому етапі створюється календарний план та описуються вихідні результати тощо.

Рис. 1.2. Етапи процесу Data Science

На другому етапі відбувається збір даних. У проєктному завданні вказано, які саме дані потрібні і як їх знайти. На цьому етапі здійснюється перевірка існування даних, їх якості й доступності. Дані можуть надаватися третіми сторонами й мати різні форми подання, від таблиць Excel до різного роду баз даних. Це – найтриваліший етап, який може займати від 50 до 85 % часу всього процесу знаходження нового знання. Процес збору даних характеризується наявністю численних похибок.

На третьому етапі дослідник підвищує якість даних і підготовляє їх для подальшого застосування. Цей етап складається з трьох фаз: **очищення даних**, коли видаляються некоректні значення, усуваються розбіжності між джерелами; **інтеграція даних**, коли розширюється інформація шляхом об’єднання інформації із різних джерел; **перетворення даних**, коли гарантується, що дані мають відповідний формат для застосування в моделі.

Дослідження (розвідка) даних спрямоване на досягнення більш глибокого розуміння даних. На цьому етапі намагаються зрозуміти, як змінні взаємодіють між собою, оцінюють розподіл змінних, визначають наявність викидів. Тут застосовується описова статистика, візуальні методи й просте моделювання. Цей етап часто називають EDA (Exploratory Data Analysis).

Моделювання даних застосовується для пошуку відповіді на певні запитання і досягнення мети дослідження. У процесі моделювання застосовуються методи математичної статистики, дослідження операцій, машинного навчання тощо. Побудова моделі є ітеративним процесом, під

час якого дослідник вибирає змінні для моделі, застосовує модель і проводить діагностику моделі.

Наприкінці результати дослідження мають бути подані замовнику. Їх можна подати в різних формах – від презентацій до наукових звітів. Іноді результати необхідно подавати у формі програмного продукту, тому що замовник планує застосувати розроблену модель на інших даних або використати отримані результати в наступному проєкті. На цьому етапі дослідник може виступити в ролі людини, яка впливає на прийняття рішень.

Більш детально етапи роботи Data Science показано рис. 1.3.

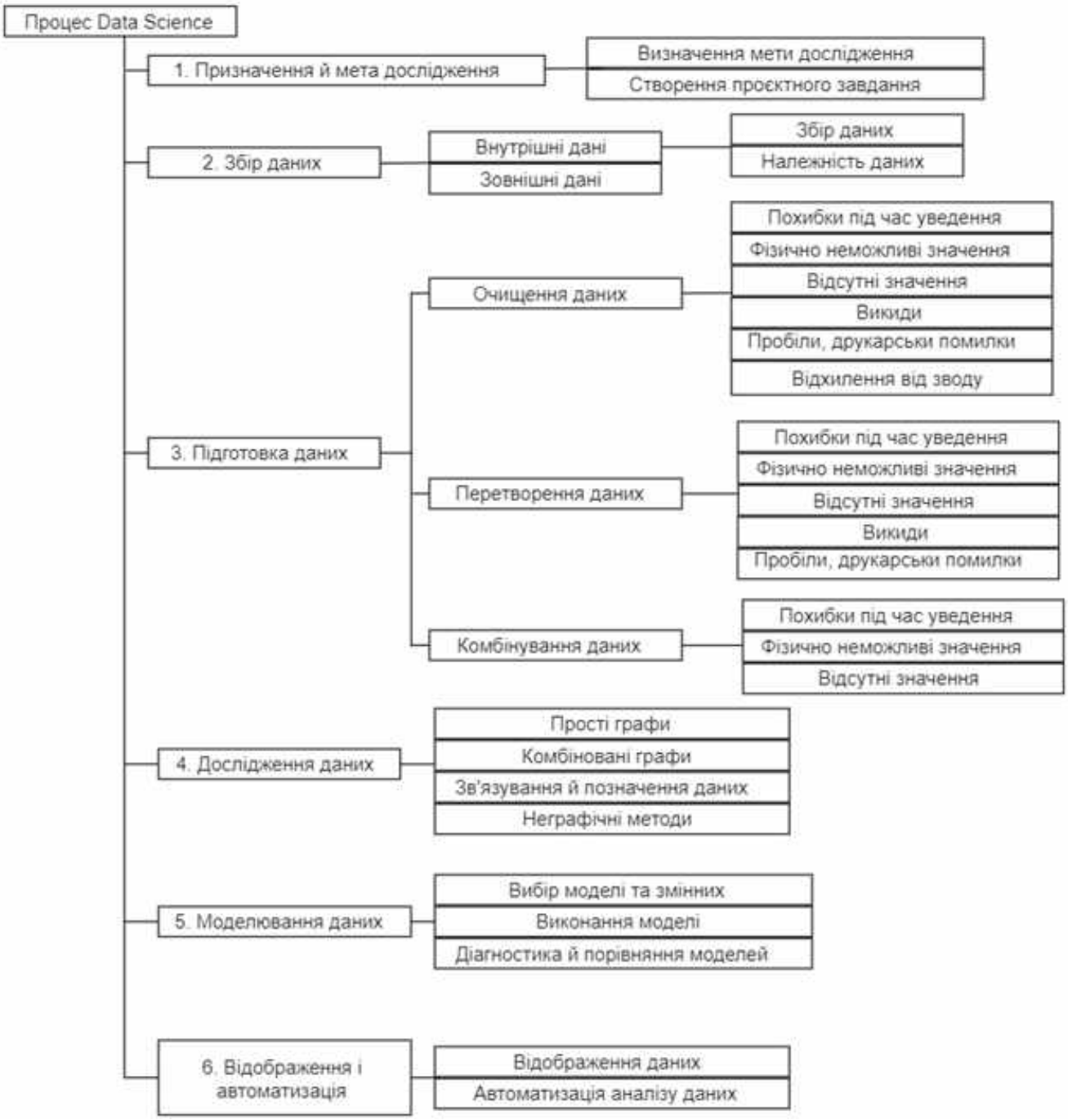


Рис. 1.3. Розгорнутий вигляд процесу Data Science

Хоча процес має лінійний характер, кожен крок не обов'язково веде безпосередньо до наступного кроку. Створення моделі інтелектуального аналізу даних є динамічним ітеративним процесом. Виконавши огляд даних, користувач може виявити, що наявних даних недостатньо для створення необхідних моделей інтелектуального аналізу даних, що, відповідно, веде до необхідності пошуку додаткових даних. Можна розробити декілька моделей і зрозуміти, що вони не вирішують поставленого завдання. Отже, потребується змінення характеристик завдання. Може виникнути необхідність в оновленні вже створених моделей з використанням нових даних, що надійшли. Таким чином, важливо розуміти, що створення моделі інтелектуального аналізу даних є процесом і що кожен крок такого процесу може бути повторений стільки разів, скільки необхідно для створення ефективної моделі й отримання результату.

1.5. Основні типи завдань

Можна виокремити принаймні шість методів виявлення й аналізу знань:

- класифікація;
- регресія;
- прогнозування часових послідовностей (рядів);
- кластеризація;
- асоціація;
- послідовність.

Перші три методи використовуються переважно для прогнозування, тоді як останні – є зручними для опису наявних закономірностей у даних.

Класифікація – найпоширеніша модель інтелектуального аналізу даних. З її допомогою виявляються ознаки, що характеризують групу, до якої належить той або інший об'єкт. Це робиться шляхом аналізу вже класифікованих об'єктів і формулювання деякого набору правил. Наприклад, у багатьох видах бізнесу проблемою є втрата постійних клієнтів. Класифікація допомагає виявити характеристики «нестійких» покупців і створити модель, яка прогнозує, хто саме схильний піти до іншого постачальника. Використовуючи її, можна визначити ефективні види знижок та інші вигідні пропозиції, що діють для різних покупців. Завдяки цьому можна утримати клієнтів, витративши стільки грошей, скільки необхідно.

Певний ефективний класифікатор може використовуватися для класифікації нових записів у базі даних у наявні класи, і в цьому випадку він набуває характеру прогнозування. Наприклад, класифікатор, що вміє ідентифікувати ризики повернення позики, може бути використаний для прийняття рішення залежно від величини ризику надання позики певному

клієнтові, тобто класифікатор використовується для прогнозування ймовірності повернення позики.

Регресійний аналіз використовується, коли залежності між змінними можуть бути виражені кількісно у вигляді деякої комбінації цих змінних. Отримана комбінація використовується для прогнозування значення, якого може набувати цільова (залежна) змінна, що обчислюється на заданому наборі значень вхідних (незалежних) змінних. У простому випадку для цього використовуються стандартні статистичні методи, такі як лінійна регресія, але більшість реальних моделей не укладаються в її межі. Наприклад, розміри продажів або фондові ціни є складними для прогнозування, оскільки можуть залежати від комплексу відношень змінних.

Прогнозування часових послідовностей. Основою для будь-яких систем прогнозування є історична інформація, що зберігається в інформаційних сховищах у вигляді часових рядів. Якщо можна побудувати математичну модель і знайти шаблони, що адекватно відображують цю динаміку, то існує ймовірність, що з їх допомогою можна прогнозувати й поводження системи в майбутньому. Завдяки прогнозуванню часових послідовностей на основі аналізу поводження часових рядів можна оцінювати майбутні значення прогнозованих змінних. Ці моделі мають містити особливі ознаки часу: ієрархію періодів (місяць – квартал – рік), особливі відрізки часу (п’яти-, шести- або семиденний робочий тиждень), сезонність, свята тощо.

Прикладами задач є прогнози курсу валют, ціни на нафту, змін клімату, обсягів продажу товарів за певний період, параметрів економіки тощо.

Кластеризація відрізняється від класифікації тим, що класи заздалегідь не задано і за допомогою моделі кластеризації засоби інтелектуальних обчислень автоматично створюють однорідні групи даних.

Асоціація належить до аналізу структури й застосовується, коли декілька подій є взаємозв’язаними. Класичним прикладом аналізу структури покупок є подання придбання якої-небудь кількості товарів як одиничної економічної операції (транзакції). Оскільки велика кількість покупок здійснюється в супермаркетах, а покупці для зручності використовують корзини, куди і складають весь товар, то для знаходження асоціацій використовується аналіз вмісту корзини. Метою підходу є знаходження трендів (однакових ділянок) серед великої кількості транзакцій, які можна використовувати для пояснення поведінки покупців. Така інформація може бути використана для регулювання запасів, зміни розміщення товарів на території магазину і прийняття рішення щодо проведення рекламної кампанії для збільшення продажів або для просування певного виду продукції. Наприклад, дослідження, проведене в супермаркеті, може показати, що 65 % людей купують картопляні чіпси,

беруть також і кока-колу, а за наявності знижки за такий комплект кока-колу купують у 85 % випадків. Маючи такі дані, менеджерам легко оцінити, наскільки дієвою є надана знижка.

Хоча цей підхід прийшов виключно з роздрібної торгівлі, він може також застосовуватися у фінансовій сфері для аналізу портфеля цінних паперів і знаходження наборів фінансових послуг, які клієнти часто купують разом. Це може використовуватися для створення деякого набору послуг як частини кампанії зі стимулювання продажів.

Послідовність має місце, якщо існує ланцюжок зв'язаних у часі подій. Предметом традиційного аналізу структури покупок є набір товарів, які являють собою одну транзакцію. Варіант такого аналізу застосовується, якщо існує додаткова інформація (номер кредитної карти клієнта або номер його банківського рахунку) для скріплення різних покупок в єдину часову серію. У такій ситуації важливим є не лише співіснування даних усередині однієї транзакції, але й порядок, у якому ці дані виникають у різних транзакціях, і час між цими транзакціями. Правила, що встановлюють ці зв’язки, можуть бути використані для визначення типового набору попередніх продажів, які можуть привести до подальших продажів певного товару. Після купівлі будинку в 45 % випадків протягом місяця купується й нова кухонна плита, а протягом наступних двох тижнів 60 % новоселів обзаводяться холодильником.

2. ОСНОВИ ТЕОРІЇ ЙМОВІРНОСТЕЙ

2.1. Вступ

Розглянемо просту задачу: монету підкидають п’ять разів; припустимо, що два рази випав аверс монети, а три рази – реверс. Що тут є ймовірністю, а що – частотою події?

По-перше, *імовірність* – це теоретична міра, що характеризує частоту появи випадкової події. Імовірність розраховують виходячи з міркувань про симетрію (монети, грального кубика тощо), рівномірність випадкового розташування елементів (наприклад, карт у колоді) тощо.

У цьому випадку, якщо вважати монету симетричною, імовірність появи аверсу монети (і, відповідно, реверсу) дорівнює 0,5.

По-друге, поділивши кількість «успішних» появ аверсу на загальну кількість кидків, отримаємо $\frac{2}{5} = 0,4$ – **емпіричну частоту**, що є результатом певного експерименту, у цьому випадку – кидання монети.

Ці дві величини – імовірність та емпірична частота – тісно зв’язані одна з одною, але мають суттєву відмінність. **Імовірність** – це теоретична

величина, що не залежить від експерименту. Під час кожного кидка чи серії кидків уважаємо, що ймовірність появи аверсу становить 0,5. Водночас **емпірична частота** являє собою результат експерименту і є **випадковою величиною**. Отже, якщо провести ще кілька експериментів (5 разів підкинути монету), то результати, скоріш за все, будуть різними. Можемо не отримати аверс жодного разу або отримати аверс в одному, двох, трьох, чотирьох або у всіх п'яти кидках монети!

Інтуїтивно розуміємо: якщо збільшити кількість кидків монети до тисячі або до мільярда разів і провести експеримент, то відношення кількості появ аверсу до загальної кількості кидків буде наближатися до 0,5. Тобто ймовірність можна трактувати як граничний випадок емпіричної частоти, якщо кількість експериментів наближається до нескінченності.

2.2. Випадкові події. Аксіоми теорії ймовірностей

Нехай проводиться деякий експеримент із випадковим результатом. Позначимо через Ω множину всіх можливих результатів експерименту: $\Omega = \{\omega\}$. Елементи ω називають **елементарними подіями**, Ω – **простором елементарних подій**.

Підмножини $A \subset \Omega$ називають подіями. Саме множину Ω називають **вірогідною** подією.

Сумою двох подій A і B називають подію C, що полягає у виконанні хоча б однієї з подій A і B: $A \bigcup B = A + B = C$ (рис. 2.1).

Добутком двох подій A і B називають подію D, що полягає в сумісному виконанні подій A і B: $D = AB = A \bigcap B$ (рис. 2.2).

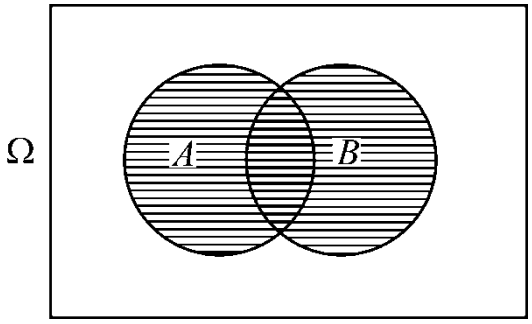


Рис. 2.1. Сума двох подій

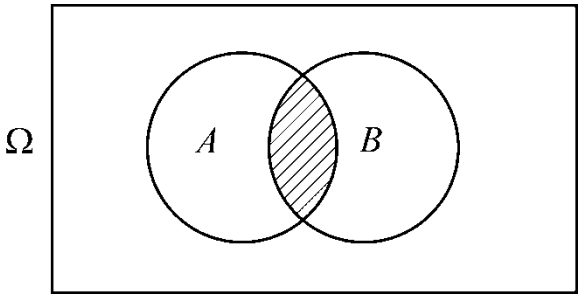


Рис. 2.2. Добуток двох подій

Протилежною відносно A подією $\bar{A}$ називають подію, що полягає в невиконанні події A: $\bar{A} = \Omega \setminus A$ (рис. 2.3).

Дві події A і B називають **несумісними**, якщо $AB = \varnothing$.

Нехай кожній події A ставиться у відповідність число $P(A)$ – імовірність події A. Для ймовірності виконуються аксіоми:

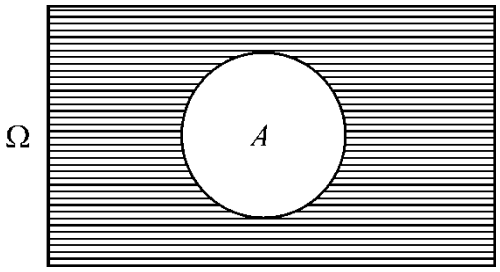


Рис. 2.3. Протилежна подія

1) імовірність будь-якої події A – це число між 0 і 1: $0 \le P(A) \le 1$;
2) якщо A і B – **несумісні події** $(AB = \varnothing)$, то $P(A+B) = P(A) + P(B)$ (цю аксіому називають **теоремою додавання** ймовірностей);

Зауваження. $P(\Omega) = 1$. Імовірність повної групи подій дорівнює одиниці, тобто в будь-якому разі відбудеться одна з випадкових подій.

Висновки:

1) якщо події $A_1, ..., A_n$ є несумісними й утворюють повну групу, тобто $\sum_{i=1}^{n} A_i = \Omega$, то $\sum_{i=1}^{n} P(A_i) = 1$, зокрема $P(A) + P(\overline{A}) = 1$ (сума ймовірностей протилежних подій дорівнює одиниці);
2) $P(A+B) = P(A) + P(B) - P(AB)$ (правило додавання, що виконується і для сумісних подій A та B).

2.3. Класичне означення ймовірності. Геометрична ймовірність

Нехай множина Ω результатів деякого експерименту є скінченною і складається з n елементів. Якщо при цьому всі n результатів виникають із однаковими можливостями, то експеримент називають **класичною** схемою. Випадок (результат) називають **сприятливим** для події A, якщо поява цього результату приводить до появи події A. Тоді

$$P(A) = \frac{m}{n},$$

де m – кількість результатів, сприятливих для події A; n – загальна кількість можливих результатів експерименту.

Якщо простір елементарних подій Ω – область на площині, A – підмножина Ω, то кажуть про **геометричну** схему, причому

$$P(A) = \frac{S(A)}{S(\Omega)}$$ – відношення площ.

Якщо Ω – множина на прямій, то $P(A)$ – відношення довжин; якщо $\Omega \subset R^3$, то $P(A)$ – відношення об'ємів.

Незважаючи на різні способи визначення в класичній і геометричній схемах, наведені ймовірності мають **спільні** властивості:

1) $0 \le P(A) \le 1$ (нульова ймовірність – подія не відбувається, імовірність, що дорівнює одиниці, – абсолютно вірогідна подія; ці граничні випадки відкидаємо);

2) $P(\Omega) = 1$, $P(\bar{A}) = 1 - P(A)$;

3) якщо A і B – несумісні події, то $P(A + B) = P(A) + P(B)$.

Приклад 2.3.1. Підкидають гральну кість. Знайти ймовірності подій: A – випаде парне число; B – випаде число, кратне трьом; C – випаде більше двох очок.

Розв'язання. Ясно, що $\Omega = \{1,2,3,4,5,6\}$, $A = \{2,4,6\}$, $B = \{3,6\}$, $C = \{3,4,5,6\}$. За означенням

$$P(A) = \frac{m}{n} = \frac{3}{6} = \frac{1}{2},\ P(B) = \frac{m}{n} = \frac{2}{6} = \frac{1}{3},\ P(C) = \frac{m}{n} = \frac{4}{6} = \frac{2}{3}.$$

Приклад 2.3.2. Підкидають одночасно дві монети. Знайти ймовірність того, що хоча б на одній із них випаде герб.

Розв'язання. Простір Ω складається з чотирьох рівноймовірних випадків: «герб – герб», «герб – цифра», «цифра – герб», «цифра – цифра». Подія A – «хоча б один герб», протилежна подія $\bar{A}$ – «цифра – цифра», отже,

$$P(\bar{A}) = \frac{1}{4} \Rightarrow P(A) = 1 - \frac{1}{4} = \frac{3}{4}.$$

Приклад 2.3.3. З колоди із 36 карт узяли чотири карти. Знайти ймовірності подій: A – усі чотири карти – тузи; B – усі чотири карти – пікової масті; C – два тузи, дві дами; D – серед чотирьох карт два тузи; E – серед чотирьох карт три червоні, одна чорна; F – хоча б одна чорна карта.

Розв'язання. Кількість рівноймовірних результатів цього експерименту дорівнює кількості способів узяти чотири елементи без повернення з 36, причому неважливо, у якому порядку. Це – **кількість сполучень**, або біноміальний коефіцієнт (коефіцієнт бінома Ньютона)

$$C_n^k = \frac{n!}{k!(n-k)!}.$$

У цьому випадку $n = 36$, $k = 4$, $C_{36}^4 = \frac{36!}{4!32!} = \frac{33 \cdot 34 \cdot 35 \cdot 36}{1 \cdot 2 \cdot 3 \cdot 4} = 33 \cdot 17 \cdot 35 \cdot 3.$

Події A сприяє тільки один результат, коли всі чотири карти – тузи:

$$P(A) = \frac{1}{C_{36}^4}.$$

Для події B сприятливими є ті випадки, коли всі чотири карти вийнято з дев'яти наявних пік:

$$P(B)=\frac{C_9^4}{C_{36}^4}\ .$$

Подія C має ймовірність

$$P(C)=\frac{C_4^2C_4^2}{C_{36}^4}.$$

Пояснення. Два тузи можна вибрати шістьома способами ($C_4^2=\frac{4!}{2!2!}=6$), дві дами – також шістьома. На кожний із шести варіантів вибору тузів припадає шість варіантів вибору дам, ось чому в чисельнику – добуток.

Подія D: два тузи, дві інші карти – не тузи. На кожний з $C_4^2=6$ варіантів вибору тузів припадає C_{32}^2 варіантів вибору двох інших, що не є тузами, а в колоді «не тузів» – 36 – 4 = 32:

$$P(D)=\frac{C_4^2C_{32}^2}{C_{36}^4}.$$

Імовірність події E (треба взяти з 18 червоних три карти, з 18 чорних – одну):

$$P(E)=\frac{C_{18}^3C_{18}^1}{C_{36}^4}.$$

Імовірність події F простіше знайти через протилежну подію $\bar{F}$ – усі карти червоні:

$$P(\bar{F})=\frac{C_{18}^4}{C_{36}^4}\Rightarrow P(F)=1-\frac{C_{18}^4}{C_{36}^4}.$$

Приклад 2.3.4. У квадраті з вершинами $(0;0)$, $(0;1)$, $(1;0)$, $(1;1)$ навмання поставили точку $M(x,y)$ (рис. 2.4). Знайти ймовірності події

$$A=\left\{y\le x^{\frac{1}{4}}\right\}.$$

Розв'язання. Зобразимо на площині простір елементарних подій Ω та геометричні місця точок, координати яких задовольняють необхідному співвідношенню.

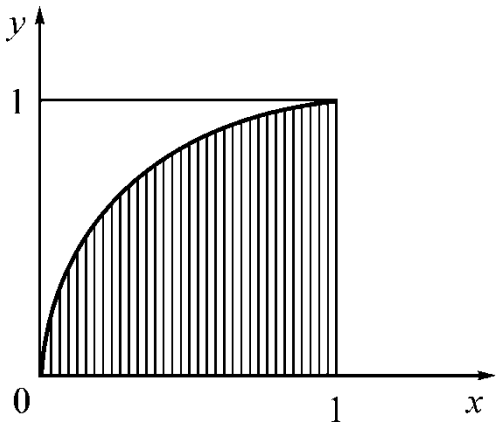


Рис. 2.4. Простір елементарних подій Ω та геометричні місця точок, координати яких задовольняють співвідношенню $A=\left\{ y\le x^{\frac{1}{4}} \right\}$

За формулою геометричної ймовірності $P\left(A\right)=\dfrac{S\left(A\right)}{S\left(\Omega\right)}=S\left(A\right)$, тому що $S\left(\Omega\right)=1$. Площа дорівнює інтегралу:

$$P\left(A\right)=\frac{\int\limits_0^1 x^{\frac{1}{4}}dx}{1}=\frac{4}{5}x^{\frac{5}{4}}\Big|_0^1=\frac{4}{5}.$$

2.4. Умовна ймовірність. Незалежність подій. Теорема множення

Умовною ймовірністю події B при виконанні події A називають величину $P\left(B\middle|A\right)=\dfrac{P\left(AB\right)}{P\left(A\right)}$ (передбачається, що $P\left(A\right)\neq 0$). Величину $P\left(B\middle|A\right)$ також називають імовірністю події B за умови, що подія A відбулася.

Цю ж формулу записують у вигляді

$$P\left(AB\right)=P(A)\cdot P(B\middle|A).$$

Кажуть, що подія B не залежить від події A, якщо

$$P\left(B\middle|A\right)=P(B)$$

(імовірність події В не залежить від того, відбулася подія A чи ні).

Якщо ймовірність появи одночасно двох подій дорівнює добутку ймовірностей цих подій $P\left(AB\right)=P(A)\cdot P(B)$, то ці події називають **незалежними**.

Кілька подій $A_1, \ldots, A_n$ називають незалежними, або незалежними в сукупності, якщо кожна з них не залежить від будь-якої комбінації інших. Для незалежних подій $P(A_1 A_2 \cdot \ldots \cdot A_n) = P(A_1) \cdot P(A_2) \cdot \ldots \cdot P(A_n)$.

2.5. Формула повної ймовірності

Припустимо, що подія A може відбутися тільки разом з однією з декількох **несумісних** подій (гіпотез) H_1, ..., H_n. Тоді

$$P(A) = P(H_1) \cdot P(A|H_1) + \ldots + P(H_n) \cdot P(A|H_n) ,$$

$$P(A) = \sum_{i=1}^{n} P(H_i) \cdot P(A|H_i).$$

Тут $P(A|H_i)$ – умовна ймовірність події A за умови, що відбулася подія H_i. Підкреслимо ще раз, що події H_i утворюють **повну групу несумісних гіпотез**.

Приклад 2.5.1. Партія складається зі 100 деталей одного виду: 20 деталей виготовлено на верстаті A, 30 – на верстаті B, 50 – на верстаті C. Деталі, виготовлені на верстаті A, виявляються такими, що мають відмінну якість з імовірністю 0,95, на верстаті B – 0,9 і на C – 0,8. Визначити ймовірність події Q – узята навмання деталь матиме відмінну якість.

Розв'язання. Тут гіпотеза H_1 полягає в тому, що взяту навмання деталь зроблено на верстаті A:

$$P(H_1) = \frac{20}{100} = 0,2.$$

За умовою задачі $P(Q|H_1) = 0,95$.

Аналогічно маємо

$$P(H_2) = \frac{30}{100} = 0,3,\ P(Q|H_2) = 0,9,\quad P(H_3) = \frac{50}{100} = 0,5,\ P(Q|H_3) = 0,8.$$

Тоді

$$P(Q) = P(H_1)P(Q|H_1) + P(H_2)P(Q|H_2) + P(H_3)P(Q|H_3) =$$
$$= 0,2 \cdot 0,95 + 0,3 \cdot 0,9 + 0,5 \cdot 0,8 = 0,86.$$

2.6. Теорема гіпотез. Формула Байєса

Нехай події (або гіпотези) H_1, H_2, … , H_n є несумісними й утворюють повну групу. Імовірності гіпотез $P(H_i)$ до проведення деякого

експерименту задано; їх називають апріорними. Тепер нехай експеримент проведено і **здійснилася** подія A. Потрібно знайти так звані апостеріорні ймовірності $P(H_1|A),\ldots, P(H_n|A)$.

З теореми множення й формули повної ймовірності випливає, що

$$P(H_i|A)=\frac{P(A|H_i)P(H_i)}{P(A)}=\frac{P(A|H_i)P(H_i)}{\sum_{j=1}^{n}P(A|H_j)P(H_j)}.$$

Звернемо увагу на той факт, що в знаменнику формули Байєса – **формула повної ймовірності**, а в чисельнику – один з доданків знаменника.

Зауваження. Імовірність події за формулою Байєса часто називають апостеріорною ймовірністю, на протилежність апріорній ймовірності, яку розраховують без урахування того, що деяка подія A відбулася.

На рис. 2.5 проілюстровано ідею формули Байєса. Окремі прямокутники – це гіпотези H_i. Подія A – це сукупність прямокутників, виділених темним кольором.

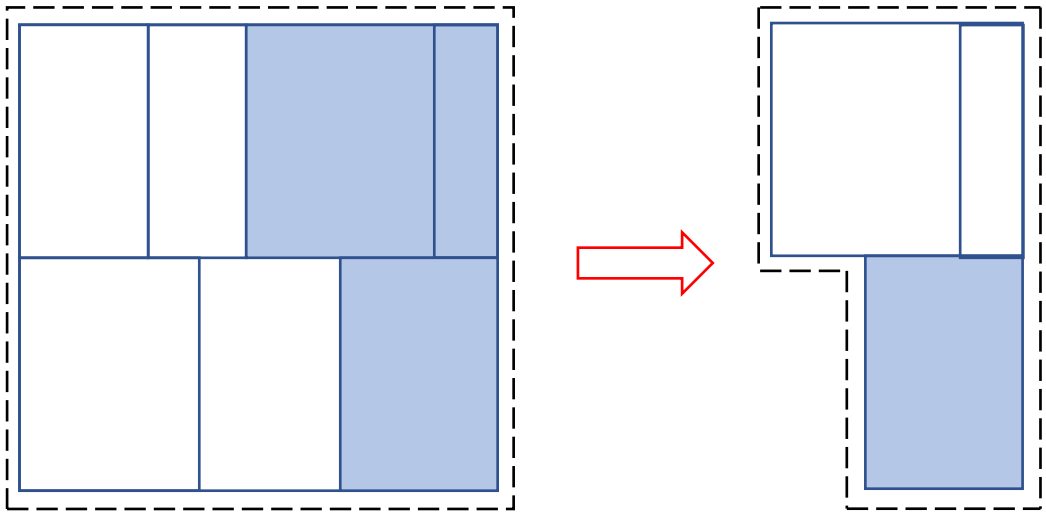


Рис. 2.5. Ілюстрація формули Байєса

Відношення площі темних прямокутників до загальної площі всіх прямокутників – це ймовірність події A. Відношення ж площі одного темного прямокутника до загальної площі темних прямокутників – це апостеріорна ймовірність гіпотези H_i. Вочевидь, апостеріорна ймовірність гіпотези H_i буде більшою за апріорну.

Приклад 2.6.1. У першій шухляді знаходилося три білі та дві чорні кулі, у другій – три білі та три чорні. З першої шухляди в другу навмання переклали дві кулі. Після цього з другої шухляди взяли навмання дві кулі, і

вони виявилися білими. Якою є ймовірність того, що з першої шухляди було перекладено дві білі кулі?

Розв'язання. Повна група несумісних гіпотез складається з трьох подій: H_1 – переклали дві білі кулі, H_2 – переклали одну білу і одну чорну кулі, H_3 – переклали дві чорні кулі. Очевидно, що

$$P(H_1)=\frac{C_3^2}{C_5^2}=\frac{3}{10},\ P(A|H_1)=\frac{C_5^2}{C_8^2}=\frac{10}{28},$$

тому що внаслідок події H_1 у другій шухляді буде п'ять білих і три чорні кулі. Аналогічно

$$P(H_2)=\frac{C_3^1C_2^1}{C_5^2}=\frac{6}{10},\ P(A|H_2)=\frac{C_4^2}{C_8^2}=\frac{6}{28},\ P(H_3)=\frac{C_2^2}{C_5^2}=\frac{1}{10},\ P(A|H_3)=\frac{C_3^2}{C_8^2}=\frac{3}{28}.$$

Завжди корисно перевірити, що $P(H_1)+P(H_2)+P(H_3)=1$. Маємо

$$P(H_1|A)=\frac{P(A|H_1)P(H_1)}{P(A|H_1)P(H_1)+P(A|H_2)P(H_2)+P(A|H_3)P(H_3)}=\frac{10}{13}.$$

Приклад 2.6.2. Нехай існує захворювання з частотою поширення серед населення 0,001, також існує метод діагностичного обстеження, який з імовірністю 0,9 виявляє хворого, але при цьому має ймовірність 0,01 помилкового результату – **помилкового виявлення захворювання у здорової людини**. Знайти ймовірність того, що людина здорова, якщо її було визнано хворою під час обстеження.

Розв'язання. Задача має прикладне значення, тому що показує важливість повторного аналізу при діагностуванні. А також вказує на те, що за певних умов діагноз не можна вважати прийнятним. Це є актуальним не тільки під час діагностування захворювань, а, наприклад, під час діагностування на наркотики.

Позначимо через X подію, що людина є хворою, через «X» – подію, що обстеження показало, що людина є хворою (тобто хворою в лапках, умовно), а через 3 – подію, що людина є здоровою. Тоді задані умови можна переписати таким чином:

- $P(«X»|X)=0,9$ – людину визнано хворою за аналізами, якщо вона є хворою;
- $P(«X»|3)=0,01$ – людину помилково визнано хворою за аналізами, якщо вона є здоровою;
- $P(X)=0,001$, тобто $P(3)=1-P(X)=0,999$.

Імовірність того, що людина є здоровою, якщо її було визнано хворою дорівнює умовній імовірності:

$$P\left(З\middle|«X»\right).$$

Щоб знайти цю ймовірність, спочатку знайдемо повну ймовірність того, що людину визначили як хвору. Для цього застосуємо формулу повної ймовірності:

$$P\left(«X»\right)=P\left(З\middle|«X»\right)P\left(З\right)+P\left(X\middle|«X»\right)P\left(X\right)=$$
$$=0,01\cdot 0,999+0,9\cdot 0,001=0,01089.$$

За формулою Байєса маємо

$$P\left(З\middle|«X»\right)=\frac{P\left(З\middle|«X»\right)P\left(З\right)}{P\left(«X»\right)}=\frac{0,01\cdot 0,999}{0,01089}\approx 0,917.$$

Іншими словами, 91,7 % людей, обстеження яких показало результат «хворий», насправді є здоровими. Причина цього полягає в тому, що за умовою задачі ймовірність помилкового результату хоч і мала, але на порядок більша від частки хворих в обстежуваній групі людей.

Якщо помилкові результати обстеження можна вважати випадковими, то повторне обстеження тієї ж людини буде давати такий результат, що не залежить від першого. У цьому випадку для зменшення частки хибнопозитивних результатів має сенс провести повторне обстеження людей, які отримали результат «хворий». Імовірність того, що людина є здоровою після отримання повторного результату «хворий», також можна обчислити за формулою Байєса:

$$P\left(З\middle|«X»,«X»\right)=\frac{0,999\cdot 0,01\cdot 0,01}{0,999\cdot 0,01\cdot 0,01+0,9\cdot 0,9\cdot 0,001}\approx 0,11.$$

Формула Байєса широко застосовується під час розв’язання задач в інформатиці, наприклад, під час створення так званого **наївного байєсова класифікатора** для фільтрації спаму. Для цього створюється база з певних слів, наприклад «копія», «спадок» тощо. І на великій базі даних розраховуються ймовірності того, що ці слова трапляються у спамі та не в спамі (хем). Наївність полягає в тому, що для простоти вважають, що слова трапляються в тексті незалежно, хоча побудова речень або зміст тексту потребує певної залежності в появі цих слів. Маючи інформацію про «спамовість» певних слів, можна отримати просту, а отже, і швидку формулу для розрахунку ймовірності того, що повідомлення є спамом.

Крім того, формула Байєса дає змогу швидко оцінити невідому ймовірність якоїсь події, про яку немає попередньої інформації. Щоб розрахувати ймовірність (точніше, емпіричну частоту) події за класичним підходом, необхідно провести велику кількість дослідів і дослідити велику кількість подій. За формулою Байєса можна отримати початкову оцінку невідомої ймовірності, а потім її коригувати за результатами того, які випадкові події відбуваються.

2.7. Байєсова фільтрація спаму

Першою відомою програмою, що фільтрує пошту з використанням байєсова класифікатора, була програма iFile Джейсона Ренні, випущена 1996 року. У програмі використовувалося сортування пошти по папках. Перша академічна публікація щодо наївної байєсової фільтрації спаму виникла 1998 року. Незабаром після цієї публікації було розгорнуто роботу зі створення комерційних фільтрів спаму. Однак 2002 року Пол Грем зміг значно зменшити кількість хибнопозитивних спрацьовувань до такої міри, що байєсів фільтр міг використовуватися як єдиний фільтр спаму.

Під час навчання фільтра для кожного слова у листах розраховується і зберігається його «вага» – **оцінка ймовірності** того, що лист з цим словом – спам. У найпростішому випадку як оцінка використовується частота «поява в спамі / поява всього». У більш складних випадках можливим є попереднє оброблення тексту: зведення слів до початкової форми, видалення службових слів, обчислення «ваги» для цілих фраз, транслітерація тощо.

Поштові байєсові фільтри ґрунтуються на теоремі Байєса, яка використовується кілька разів у контексті спаму:

- перший раз – для обчислення ймовірності, що повідомлення є спамом, а слово з'являється в цьому повідомленні;
- другий раз – для обчислення ймовірності, що повідомлення є спамом з урахуванням усіх його слів (або відповідних їх підмножин);
- іноді – третій раз – коли трапляються повідомлення з рідкісними словами.

Припустимо, що підозріле повідомлення містить слово «Replica». Більшість людей, які звикли отримувати електронного листа, знають, що це повідомлення, швидше за все, буде спамом, а точніше – пропозицією продати підроблені копії годинників відомих брендів. Програма виявлення спаму, однак, не «знає» таких фактів; усе, що вона може зробити, – обчислити ймовірності.

За формулою Байєса маємо

$$P(S|W) = \frac{P(W|S)P(S)}{P(W)} = \frac{P(W|S)P(S)}{P(W|S)P(S) + P(W|H)P(H)},$$

де $P(S)$ – імовірність того, що довільне повідомлення – спам;

$P(H)$ – імовірність того, що довільне повідомлення – не спам («Ham»);

$P(S|W)$ – імовірність того, що лист – спам за умови того, що в ньому є слово «Replica»;

$P(W|S)$ – імовірність наявності слова «Replica» у спамі;

$P(W|H)$ – імовірність наявності слова «Repliса» в повідомленнях класу «Ham».

Статистичні дослідження електронного листування свідчать про те, що ймовірність того, що будь-яке повідомлення буде спамом, становить 0,8 (або 80 %), тобто

$$P(S) = 0{,}8 \text{ і } P(H) = 0{,}2.$$

(Пам'ятаємо, що сума ймовірностей повної системи гіпотез дорівнює одиниці). Але заради спрощення можна припустити, що ймовірності спаму й хему є однаковими, що дорівнюють 50 % (як у жарті про те, яка ймовірність зустріти на вулиці динозавра – так або ні, тобто 50 %). У цьому випадку маємо

$$P(S|W) = \frac{P(W|S)}{P(W|S) + P(W|H)}.$$

Значення $P(S|W)$ має назву «**спамовість**» слова W, а ймовірність $P(W|S)$ дорівнює частоті повідомлень, які містять слово W та ідентифікуються як спам на етапі збору статистичних даних (навчання моделі).

Звичайно, визначення, чи є повідомлення спамом або хемом, що базується на наявності лише одного певного слова, схильне до помилок. Саме тому байєсові фільтри спаму намагаються розглянути кілька слів і комбінувати їх, щоб визначити повну ймовірність того, що повідомлення є спамом.

Програмні спам-фільтри, побудовані на принципах наївного байєсового класифікатора, роблять наївне припущення про те, що події, що відповідають наявності того чи іншого слова в електронному листі або повідомленні, є такими, що не залежать одна до одної. Це спрощення в загальному випадку є неправильним для природних мов, наприклад англійської, де ймовірність виявлення прикметника підвищується при наявності, наприклад, іменника. Виходячи з такого наївного припущення, для розв'язання задачі класифікації повідомлень лише на два класи – S і H (спам і хем, які є несумісними) – з теореми Байєса можна вивести таку формулу для оцінювання ймовірності спамовості всього повідомлення, що містить слова $W_1, W_2, ..., W_N$.

За теоремою Байєса

$$P(S|W_1, W_2, ..., W_N) = \frac{P(W_1, W_2, ..., W_N|S)P(S)}{P(W_1, W_2, ..., W_N)} =$$

$$= \frac{P(W_1|S) \cdot P(W_2|S) \cdot .... \cdot P(W_N|S)P(S)}{P(W_1, W_2, ..., W_N)}.$$

Тут припускаємо, що ймовірності наявності слів $W_1, W_2, ..., W_N$ є незалежними подіями. Якщо далі записати кожний доданок у чисельнику за формулою Байєса, то отримаємо

$$P(S|W_1, W_2, ..., W_N) = \frac{P(W_1|S) \cdot P(W_2|S) \cdot .... \cdot P(W_N|S) \cdot P(S)}{P(W_1, W_2, ..., W_N)} =$$

$$= \frac{\frac{P(S|W_1)P(W_1)}{P(S)} \cdot \frac{P(S|W_2)P(W_2)}{P(S)} \cdot .... \cdot \frac{P(S|W_N)P(W_N)}{P(S)} \cdot P(S)}{P(W_1, W_2, ..., W_N)} =$$

$$= \frac{P(S|W_1)P(W_1) \cdot P(S|W_2)P(W_2) \cdot .... \cdot P(S|W_N)P(W_N) \cdot P(S)^{1-N}}{P(W_1, W_2, ..., W_N)}.$$

Далі запишемо знаменник за формулою повної ймовірності:

$$P(S|W_1, W_2, ..., W_N) =$$

$$= \frac{P(S|W_1)P(W_1) \cdot P(S|W_2)P(W_2) \cdot .... \cdot P(S|W_N)P(W_N) \cdot P(S)^{1-N}}{P(W_1, W_2, ..., W_N)} =$$

$$= \frac{P(S|W_1)P(W_1) \cdot P(S|W_2)P(W_2) \cdot .... \cdot P(S|W_N)P(W_N) \cdot P(S)^{1-N}}{\left[P(W_1|S) \cdot ... \cdot P(W_N|S)\right] \cdot P(S) + \left[P(W_1|H) \cdot ... \cdot P(W_N|H)\right] \cdot P(H)} =$$

$$= \frac{\Pi_{i=1}^{N}\left[P(S|W_i)P(W_i)\right] \cdot P(S)^{1-N}}{\Pi_{i=1}^{N}\left[P(S|W_i)P(W_i)\right] \cdot P(S)^{1-N} + \Pi_{i=1}^{N}\left[P(H|W_i)P(W_i)\right] \cdot P(H)^{1-N}},$$

де $\Pi_{i=1}^{N}[....]$ – добуток елементів з індексами від 1 до N.

Спрощуючи вираз, отримаємо

$$\frac{\Pi_{i=1}^{N}\left[P(S|W_i)P(W_i)\right]}{\Pi_{i=1}^{N}\left[P(S|W_i)P(W_i)\right] + \Pi_{i=1}^{N}\left[P(H|W_i)P(W_i)\right]\frac{P(H)^{1-N}}{P(S)^{1-N}}}.$$

Якщо припустити ймовірності спаму й хему в будь-якому листі однаковими (тобто 0,5 і 0,5), то отримаємо вираз

$$P(S|W_1, W_2, ..., W_N) = \frac{p_1 p_2 ... p_N}{p_1 p_2 ... p_N + (1-p_1)(1-p_2)...(1-p_N)},$$

де $p_k = P(S|W_k)$ – умовна ймовірність того, що повідомлення є спамом за умови того, що воно містить слово W_k (наприклад, «Replica» або «Nigeria»).

Отриманий результат необхідно порівняти з деяким граничним значенням, наприклад 0,5. Якщо знайдена ймовірність є більшою за

граничне значення, то лист класифікується як спам, якщо меншою – як хем.

Цей метод є простим (алгоритми – елементарні), зручним (можна не застосовувати чорні списки і подібні штучні прийоми), ефективним (після навчання на досить великій вибірці відкидається 95…97 % спаму), причому в разі будь-яких помилок його можна донавчати. Загалом, є всі передумови для широкого використання цього методу, що й має місце на практиці – на його основі побудовано практично всі сучасні спам-фільтри.

Утім, метод має й принциповий недолік: метод базується на припущенні, що одні слова частіше трапляються в спамі, а інші – у звичайних листах, і є неефективним, якщо це припущення – неправильне. Як показує практика, такий спам навіть людина не може визначити «на око» – тільки прочитавши лист і зрозумівши його зміст. Існує метод **байесова отруєння**, що дає змогу вводити багато додаткового тексту, іноді ретельно підібраного, щоб «обдурити» фільтр.

Метод має ще один недолік, пов'язаний з реалізацією, що не є принциповим, – робота тільки з текстом. Знаючи про це обмеження, спамери почали вкладати рекламну інформацію в картинку, а тексту в листі або немає, або він не має сенсу. Отже, доводиться користуватися або засобами розпізнавання тексту («дорога» процедура, що застосовується тільки за гострої потреби), або старими методами фільтрації – «чорні списки» і регулярні вирази (оскільки такі листи часто мають стереотипну форму).

2.8. Схема Бернуллі

Послідовність незалежних випробувань називають **схемою Бернуллі**, якщо:

1) проводяться деякі випробування, результати яких не залежать один від одного;
2) кожне випробування має два можливих результати, що мають назви «успіх» і «невдача»;
3) імовірність появи успіху в кожному випробуванні є сталою і дорівнює p, а невдачі – відповідно, $q = 1 - p$.

Тоді ймовірність того, що успіх має місце у k випробуваннях із n випробувань (а невдача – у $(n-k)$ випробуваннях), визначається формулою

$$P_n(k) = C_n^k p^k q^{n-k}.$$

Цю формулу також називають **біноміальним розподілом** випадкової величини.

Приклад 2.8.1. Стрілок робить п'ять пострілів. Імовірність влучення при кожному пострілі дорівнює 0,9. Знайти ймовірність подій: A – перші три постріли невдалі, а останні два – успішні; B – із п'яти пострілів два успішні; C – із п'яти пострілів не менше двох – успішні.

Розв'язання. Подію A зручно записати у вигляді $\{---++\}$, тоді за теоремою множення ймовірностей маємо $P(A)=(1-0,9)^3(0,9)^2=0,00081$.

Імовірність події B знайдемо за формулою Бернуллі ($n=5$, $p=0,9$, $q=0,1$):

$$P(B)=P_5(2)=C_5^2(0,9)^2(0,1)^3=0,0081.$$

Звернемо увагу на те, що подія B відрізняється від події A тим, що у випадку B не має значення, у яких саме випробуваннях були успіхи, а важливою є лише їх кількість.

Імовірність події C можна знайти за теоремою додавання ймовірностей несумісних подій $\{2,3,4,5 \text{ успіхів}\}$. Але краще обчислити $P(\bar{C})=\{0 \text{ або } 1 \text{ успіх }\}=P_5(0)+P_5(1)=C_5^0(0,9)^0(0,1)^5+C_5^1(0,9)^1(0,1)^4=0,00046$, тоді $P(C)=0,99954$.

Приклад 2.8.2. Мисливець тричі стріляє в кабана, імовірність влучення при одному пострілі – 0,5. Якщо влучили три кулі, то кабан загине з імовірністю 1, якщо дві – з імовірністю 0,8, якщо одна – з імовірністю 0,3. Знайти ймовірність $P(A)$ загибелі кабана.

Розв'язання. Задача розв'язується за допомогою формули повної ймовірності та схеми Бернуллі. Нехай H_i – гіпотеза, що полягає в тому, що влучили i куль. Оскільки $P(A|H_0)=0$, гіпотезу H_0 відкидаємо. Далі маємо:

$$P(H_1)=P_3(1)=C_3^1\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^2=\frac{3}{8} \text{ і за умовою } P(A|H_1)=0,3;$$

$$P(H_2)=P_3(2)=C_3^2\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^2=\frac{3}{8} \text{ і за умовою } P(A|H_2)=0,8;$$

$$P(H_3)=P_3(3)=C_3^3\left(\frac{1}{2}\right)^3=\frac{1}{8} \text{ і за умовою } P(A|H_3)=1.$$

Отже, $P(A)=\frac{3}{8}\cdot 0,3+\frac{3}{8}\cdot 0,8+\frac{1}{8}=\frac{43}{80}$.

Приклад 2.8.3. Гравець підкидає монету доти, доки не випаде, наприклад, її аверс. Знайти ймовірності можливої кількості підкидань монети.

Розв'язання. Імовірність появи аверсу в кожному випробуванні $p=\frac{1}{2}$.

Тому легко зрозуміти таке:

а) імовірність того, що аверс випаде в першому випробуванні, $P_1 = \frac{1}{2}$;

б) імовірність появи аверсу в другому випробуванні є результатом двох незалежних подій – появи реверсу в першому випробуванні й аверсу в другому, тому $P_2 = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$;

в) поява аверсу в третьому випробуванні є ймовірною, якщо в перших двох випробуваннях випали реверси монети, тому $P_2 = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$, і т. д.

Загалом можна скласти формулу

$$P_n = \frac{1}{2^n}.$$

Зауваження. Стратегія збільшення ставки на кожному кроці іноді пропонується як безпрограшна стратегія гри в казино (наприклад, ставити лише на колір, червоний або чорний, збільшуючи ставку вдвічі доти, доки не відбудеться виграш). Якщо вважати ймовірності виграшу й програшу однаковими, що дорівнюють 0,5 (нехтуючи сектором «зеро»), то ймовірність серії із N програшів (або виграшів) дорівнює $\frac{1}{2^N}$. Якщо брати участь у досить великій кількості ігор, то такі серії будуть регулярно виникати. Але гра, як відомо, ведеться до розорення одного з гравців. Казино може собі дозволити програти більшу суму, ніж пересічний гравець. Тому, якщо гра триває довго, навіть за умов «справедливої» гри, коли ймовірності виграшу й програшу є однаковими, то гра буде вестися до розорення учасника, який має менше коштів. А це неминуче відбудеться, якщо грати достатньо довго. Тому казино завжди у виграші.

У наведених вище прикладах випадкова величина набувала значень, які можна порахувати. Хоча в останньому випадку їх кількість є нескінченною. Такі випадкові величини утворюють клас **дискретних випадкових величин**.

2.9. Дискретні випадкові величини

Під випадковою величиною розуміють величину, яка внаслідок експерименту з випадковим результатом набуває деякого значення. Якщо при цьому можливі значення випадкової величини ξ утворюють скінченну або зчисленну множину, то таку величину називають дискретною. Наприклад, при підкиданні грального кості величина ξ дорівнює кількості очок; можливі значення – 1, 2, …, 6. Інший приклад: величина ξ дорівнює

кількості влучень при десяти пострілах; можливі значення – 0, 1, ..., 10. Якщо величина ξ – кількість дзвінків в офіс протягом доби, то $\xi = 0, 1, \ldots$.

Щоб задати дискретну випадкову величину, потрібно вказати, яких значень $x_1, x_2, \ldots$ вона набуває і якими є **ймовірності**, що відповідають цим значенням:

$$p_i = P\{\xi = x_i\}.$$

Числа p_i утворюють скінченну або нескінченну множину і мають дві властивості:

1) $p_i > 0$;

2) $\sum_i p_i = 1.$

Сума може мати скінченну кількість доданків, тоді ряд розподілу випадкової величини ξ зручно записувати у вигляді таблиці:

x_i	x_1	...	x_n
p_i	p_1	...	p_n

Якщо множина можливих значень ξ є нескінченною, то сума ймовірностей утворює ряд $\sum_{i=1} p_i = 1.$

Для того щоб дати наочну характеристику (а також для подальших операцій, аналізу тощо), уводять числові характеристики розподілу:

1. **Математичне сподівання**

$$M\xi = \sum_{i=1}^{n} x_i p_i .$$

Це сума добутків усіх можливих значень і відповідних ймовірностей. Розмірність математичного сподівання збігається з розмірністю випадкової величини ξ. Математичне сподівання певною мірою **характеризує середнє значення** ξ, і є узагальненим поняттям середнього значення сукупності чисел у тому випадку, коли елементи множини значень цієї сукупності мають різну "вагу", ціну, важливість, пріоритет, що є характерним для значень випадкової змінної. Математичне сподівання також іноді називають сподіванням, середнім, середнім значенням або першим моментом.

2. **Дисперсія** випадкової величини

$$D\xi = M\left(\xi - M\xi\right)^2 = \sum_{i=1}^{n} \left(x_i - M\xi\right)^2 p_i ,$$

де $M\xi$ – математичне сподівання.

Дисперсія характеризує розсіювання величини ξ відносно її середнього значення. Існує інша формула для її обчислення:

$$D\xi = M\xi^2 - \left(M\xi\right)^2 = \sum_{i=1}^{n} x_i^2 p_i - \left(M\xi\right)^2.$$

Середньоквадратичним відхиленням випадкової величини ξ називають число $\sigma = \sqrt{D\xi}$.

Розмірність середньоквадратичного відхилення σ є такою ж, що й самої випадкової величини ξ.

Чим менше середньоквадратичне відхилення (і, відповідно, дисперсія), тим ближчими до середнього значення є випадкові величини. Простий приклад – постріли в мішень. Середнє значення координат пострілів різних стрілків може бути однаковим і збігатися з центром мішені. Але дисперсія є різною – усі кулі одного стрілка влучили в «яблучко» (дисперсія мала), другого – у «молоко» (дисперсія велика). Отже, одне лише середнє значення не може характеризувати розподіл, і для повної характеристики треба мати ще як мінімум дисперсію (існують й інші характеристики).

Деякі властивості математичного сподівання й дисперсії:

1) якщо C – стала, то $MC = C$, $DC = 0$;
2) $M\left(c\xi\right) = cM\xi$;
3) $D\left(c\xi\right) = c^2 D\xi$;
4) $M\left(\xi + C\right) = M\xi + C$;
5) $D\left(\xi + C\right) = D\xi$;
6) $M\left(\xi + \eta\right) = M\xi + M\eta$;
7) $D\xi \geq 0$ для будь-якої випадкової величини ξ.

Звернемо увагу на те, що нічого не сказано стосовно $D\left(\xi + \eta\right)$ та $M\left(\xi\eta\right)$, і це не випадково.

Приклад 2.9.1 (Петербурзький парадокс). Нехай казино проводить таку гру: вступаючи в гру, гравець платить деяку суму, а потім підкидає монету (імовірність кожного результату — 50 %) доти, доки не випаде «орел». При випаданні «орла» гра закінчується, а гравець отримує виграш, розрахований за такими правилами: якщо орел випав при першому підкиданні, та гравець отримує 2^0 гр. од., якщо при другому, то 2^1, і т. д.: при n-му підкиданні — 2^{n-1} гр. од. Іншими словами, виграш збільшується від підкидання до підкидання вдвічі, пробігаючи по ступенях двійки — 1, 2, 4, 8, 16, 32 і т. д. Який вступний внесок має взяти казино з гравця, щоб не бути в програші?

Розв'язання. Знайдемо математичне сподівання виграшу:

$$M\xi = 2^0 \frac{1}{2} + 2^1 \cdot \frac{1}{2} \cdot \frac{1}{2} + 2^2 \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} + ... = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + ... = \infty.$$

Таким чином, математичне сподівання виграшу дорівнює нескінченості, а це означає, що вартість гри, тобто сума, яку потрібно сплатити в казино, є нескінченою. Цей парадокс ілюструє розбіжність між теоретично оптимальною поведінкою гравця і «здоровим глуздом».

Існує декілька варіантів розв'язання парадоксу, зокрема обмеженість коштів банку. Ще один варіант розв'язання парадоксу полягає в тому, що людина внаслідок особливостей психології нехтує подіями, імовірність яких є малою, хоча вони можуть і дати великий виграш.

У Data Science зазвичай оперують величинами різної розмірності, що може призвести до проблем під час змінення даних, виключення зайвих даних, додавання нових параметрів у модель тощо. Тому над даними проводять операції **центрування** й **нормування**. Від величин віднімають середнє значення і різницю ділять на середньоквадратичне відхилення масиву даних. Отримані значення мають нульове математичне сподівання й дисперсію, яка дорівнює одиниці. Більш детально це розглянемо пізніше.

Крім того, слід зазначити, що існують ще такі числові характеристики випадкових величин:

1. **Медіана** – величина, розташована посередині ранжованого ряду (масиву даних), тобто величина в середині ряду величин, розташованих у зростальній або спадній послідовності. У загальному випадку за відсутності симетрії медіана не збігається з математичним сподіванням.

2. **Мода** – значення випадкової величини, що трапляється найчастіше в сукупності спостережень. Це таке значення x, у якому функція мас ймовірностей набуває максимального значення. Іноді трапляється більше однієї моди.

2.10. Граничні теореми Пуассона і Муавра – Лапласа

Безпосереднє обчислення ймовірностей $P_n(k)$ того, що в n незалежних випробуваннях відбудеться k успіхів, є занадто важким при досить великій кількості випробувань. Тому необхідно розраховувати біноміальні коефіцієнти, у які входять факторіали. А факторіал, як відомо, – це функція, що дуже швидко зростає. Наприклад, факторіал числа 100 перевищує кількість атомів у Всесвіті. Для спрощення розрахунків створено дві теореми, що застосовуються в певних випадках.

У випадку великих значень n та малих значень імовірності успіху p зручно користуватися граничною теоремою Пуассона: якщо $n \to \infty$, $p \to 0$ так, що їх добуток $np = \lambda$, то

$$P_n(k) \to \frac{\lambda^k}{k!} e^{-\lambda},$$

де $k!$ – факторіал числа.

Зрозуміло, що теорема Пуассона широко використовується для аналізу «рідких» подій. Графіки розподілу Пуассона для різних значень λ показано на рис. 2.6.

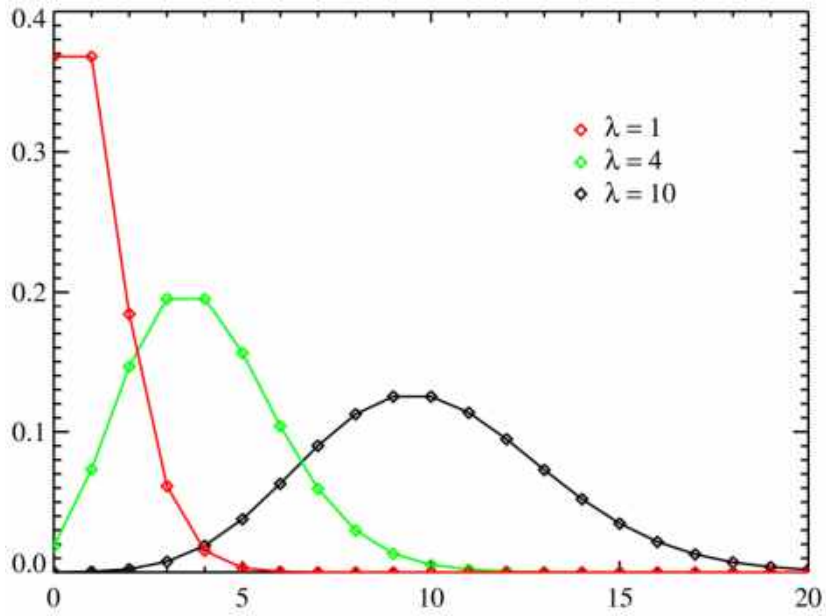


Рис. 2.6. Розподіл імовірностей Пуассона

Розподіл Пуассона має лише один параметр, що характеризує розподіл ймовірностей, – число λ, і всі числові характеристики розподілу залежать від λ:

- **математичне сподівання** дорівнює λ;
- **дисперсія** дорівнює λ.

Приклад 2.10.1. Незалежно один від одного працюють 1000 високонадійних елементів. Імовірність того, що один елемент зламається протягом місяця, дорівнює 0,002. Знайти ймовірності подій, що протягом місяця:

1) відбудеться дві поломки;
2) не відбудеться жодної поломки;
3) відбудеться хоча б дві поломки.

Знайти також найімовірнішу кількість поломок.

Розв'язання. 1. Оскільки $n = 1000 -$ велике число, а $p = 0,002 -$ мале, то $np = a = 2$, отже, маємо

$$P_n(k) = P_{1000}(2) \approx \frac{2^2}{2!} e^{-2} \approx 0,270671.$$

Найближче значення можна знайти, користуючись спеціальними таблицями, наведеними в багатьох підручниках з теорії ймовірностей.

2. Імовірність події «жодної поломки»:

$$P_n(0) \approx \frac{a^0}{0!} e^{-a} = e^{-2} \approx 0{,}13534 .$$

3. Імовірність хоча б двох поломок зручно знаходити, використовуючи протилежну подію «0 або одна поломка»:

$$P(\bar{A}) \approx \frac{a^0}{0!} e^{-a} + \frac{a^1}{1!} e^{-a} \approx 0{,}13534 + 0{,}27067 \approx 0{,}406 , \; P(A) \approx 0{,}594 .$$

Найімовірніша кількість поломок – одна або дві. Це знов-таки видно з таблиці: якщо $a = 2$, то при $k = 1$ та $k = 2$ імовірності $P_{1000}(k)$ матимуть однакові, найбільші з усіх, значення, а саме 0,270671.

Дві теореми **Муавра – Лапласа** дають змогу наближено визначати ймовірності $P_n(k)$, якщо кількість випробувань $n \to \infty$. Вони діють тим краще, чим ближчою до $1/2$ є ймовірність успіху p у кожному випробуванні. Нагадаємо, що $1 - p = q$.

Локальна теорема. Якщо ймовірність появи успіху p у n незалежних випробуваннях – стала, $0 < p < 1$, то при $n \to \infty$

$$P_n(k) \sim \frac{1}{\sqrt{npq}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} ,$$

де $x = \dfrac{k - np}{\sqrt{npq}}$ – центрована й нормована випадкова величина.

Інтегральна теорема. Якщо p – стала, $0 < p < 1$, то при $n \to \infty$

$$P\left\{ a \le \frac{k - np}{\sqrt{npq}} < b \right\} \to \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx ,$$

де k – кількість успіхів у n випробуваннях.

Теорема застосовується для наближеного обчислення ймовірності того, що кількість успіхів буде знаходитись між числами k_1 та k_2:

$$P\{k_1 \le k < k_2\} \approx \Phi\left(\frac{k_2 - np}{\sqrt{npq}} \right) - \Phi\left(\frac{k_1 - np}{\sqrt{npq}} \right),$$

де $\Phi(x) = \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_0^x e^{-\frac{u^2}{2}} du$.

Оскільки первісну неможливо знайти серед елементарних функцій, значення інтеграла наближено обчислюють за допомогою комп'ютерних програм або таблиць, наведених у підручниках з теорії ймовірностей.

Функція є непарною: $\Phi(-x) = -\Phi(x)$.

Приклад 2.10.2. Монету підкидають 100 разів $(n=100)$. Знайти:

а) імовірність того, що випадуть 50 гербів;

б) імовірність того, що кількість гербів, що випали, буде становити від 45 до 55 включно.

Розв'язання: а) точне значення ймовірності – $P_{100}(50)=$ $=C_{100}^{50}\cdot\left(\frac{1}{2}\right)^{50}\cdot\left(1-\frac{1}{2}\right)^{50}$, а наближене за локальною теоремою –

$$P_{100}(50)\approx\frac{1}{\sqrt{100\cdot\frac{1}{2}\cdot\frac{1}{2}}}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}=\frac{1}{5}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}},$$

отже, у цьому випадку $x=\frac{50-100\cdot\frac{1}{2}}{\sqrt{25}}=0$; за таблицями значень функції $\phi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$, маємо $\phi(x)=0,3989$, отже, $P_{100}(50)\approx\frac{1}{5}\cdot0,3989=0,08$;

б) $\frac{k_1-np}{\sqrt{npq}}=\frac{45-50}{\sqrt{25}}=-1$, $\frac{k_2-np}{\sqrt{npq}}=\frac{55-50}{\sqrt{25}}=1$, звідки

$$P\{45\le k<55\}\approx\Phi(1)-\Phi(-1)=2\Phi(1)=0,683.$$

Зауваження. Тут уперше застосовано **розподіл Гаусса**, який має найважливіше значення у математичній статистиці, далі будемо часто його використовувати.

Вище здійснено перехід від **дискретних випадкових величин** до **неперервних випадкових величин**. Ці два класи величин є інструментами аналізу даних. Якщо випадкова величина набуває лише скінченної кількості значень (наприклад, лише двох значень 0 та 1 або нескінченної кількості дискретних значень 0, 1, 2, 3, ….), то вона належить до дискретних випадкових величин. Якщо ж випадкова величина набуває значень у певному діапазоні континуума (який також може бути нескінченним), то вона належить до неперервних випадкових величин. Для спрощення задачі часто можна переходити від одного класу до іншого. Наприклад, якщо випадковою величиною є кількість пачок чаю, проданих за добу в супермаркеті, то немає сенсу розраховувати значення ймовірностей для всіх можливих обсягів продажів, оскільки це досить велика множина значень, а можна, без зниження точності, дискретний розподіл замінити неперервним розподілом, який характеризується лише двома або трьома параметрами.

Залишається незрозумілим, як формула $P_n(k)=C_n^k p^k q^{n-k}$ перетворюється на формулу $P_n(k)\sim\frac{1}{\sqrt{npq}}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$. Математичне доведення є занадто важким, і простіше розглянути на графіку, як трансформуються ймовірності зі збільшенням n при $p=\frac{1}{2}$ за формулою $P_n(k)=C_n^k p^k q^{n-k}$ (рис. 2.7).

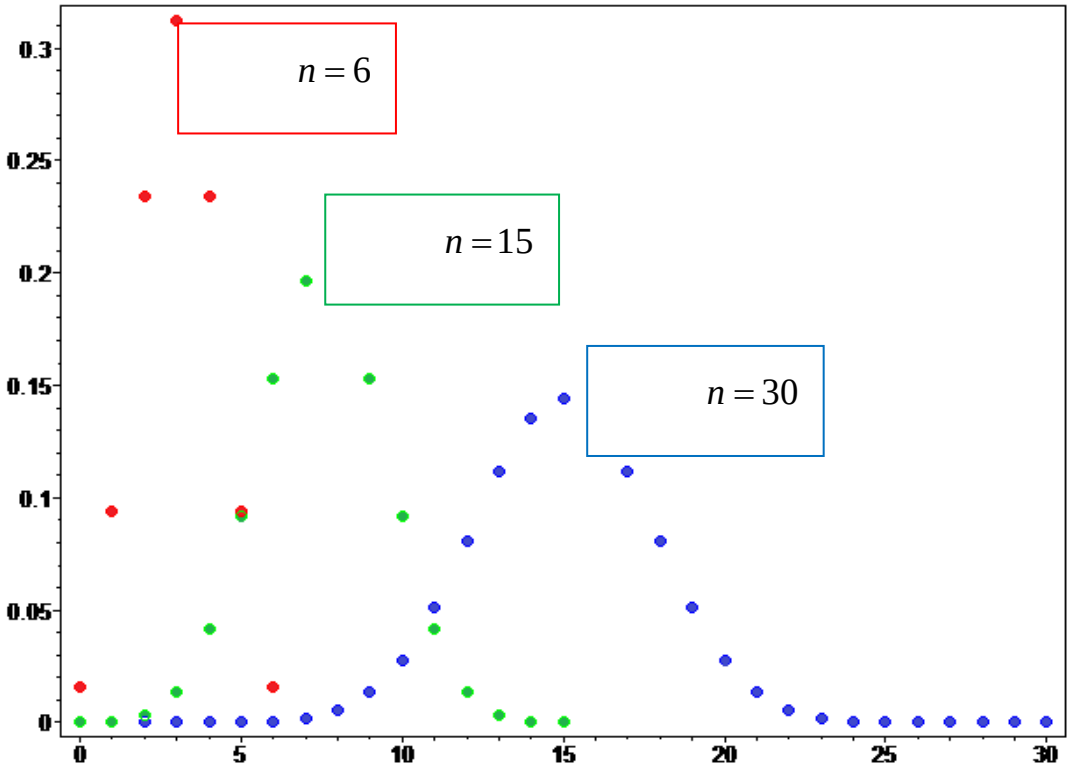


Рис. 2.7. Біноміальний розподіл ймовірностей

Як бачимо, зі збільшенням n графік наближається до певного вигляду, симетричного відносно $\frac{n}{2}$, тому що було взято ймовірність успіху в одному випробуванні $p=\frac{1}{2}$, а математичне сподівання становить саме $\frac{n}{2}$. І це – графік залежності

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-a)^2}{2\sigma^2}},$$

де $a=\frac{n}{2}$ – вісь симетрії. Значення і зміст параметра σ розглянемо нижче.

Резюме:

1. Випадкові величини, що є результатом певних ймовірнісних випробувань, можна поділити на дві групи – **дискретні** й **неперервні**. Дискретні величини можна подати у вигляді таблиці значень величини x_i і ймовірності її появи p_i або у вигляді формули для визначення ймовірності як функції дискретного індексу $P(k)$.

2. Розподіл найчастіше можна характеризувати обмеженою кількістю параметрів, серед яких – **математичне сподівання** й **дисперсія**. Існує також непараметрична статистика, але в цьому посібнику її розглядати не будемо.

3. При великій кількості випробувань біноміальний розподіл суми «успіхів» можна замінити розподілом Пуассона (при малій імовірності «успіху» в одному випробуванні) або розподілом Гаусса (при ймовірності «успіху», близькій до 0,5). Якщо ймовірність «успіху» у випробуванні є близькою до одиниці, то має сенс як «успіх» розглядати «невдачу» і застосовувати розподіл Пуассона.

2.11. Неперервні випадкові величини

Випадкову величину ξ називають неперервною, якщо існує невід'ємна функція $f(x)$, що задовольняє рівності

$$F(x) = \int_{-\infty}^{x} f(t)dt,$$

де $F(x)$ – функція розподілу.

Функція $f(x)$, яку називають **щільністю** розподілу ймовірності, задовольняє такі умови:

1) $f(x) \geq 0;$

2) $\int_{-\infty}^{\infty} f(x)dx = 1;$

3) $P\{a \leq \xi < b\} = \int_{a}^{b} f(x)dx.$

Опишемо **властивості** цієї функції: площа деякої криволінійної трапеції під графіком $f(x)$ з основою $x \in [a; b]$ дорівнює ймовірності, з якою випадкова величина потрапить в інтервал $x \in [a; b]$. З цього випливає, що якщо взяти інтервал $(-\infty; +\infty)$, то ймовірність того, що

випадкова величина потрапить у цей інтервал, дорівнює одиниці. Ще один важливий момент: питання про те, з якою ймовірністю випадкова величина набуде певного значення $x = c$, не має сенсу. Імовірність того, що неперервна величина набуде якогось конкретного значення, дорівнює нулю! Тут оперуємо проміжками, відрізками, у які величина потрапляє. А площа прямокутника, основа якого збігається в точку, є нульовою.

Наведемо формули для знаходження числових характеристик (за умови абсолютної збіжності інтегралів):

$$M\xi = \int_{-\infty}^{\infty} xf(x)dx$$ – математичне сподівання;

$$M\xi^2 = \int_{-\infty}^{\infty} x^2 f(x)dx;$$

$$D\xi = \int_{-\infty}^{\infty} (x - M\xi)^2 f(x)dx, \text{ або } D\xi = M\xi^2 - (M\xi)^2;\ \sigma = \sqrt{D\xi}.$$

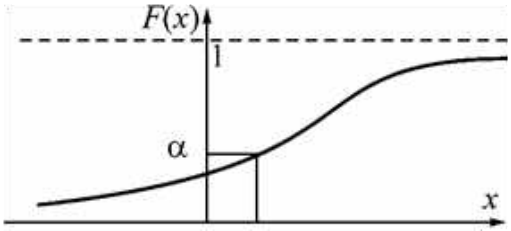


Рис. 2.8. α-квантиль

Число x_α, що визначається рівністю $F(x_\alpha) = \alpha$, називають **квантилем** рівня α. Це абсциса точки перетину графіка $F(x)$ та горизонтальної прямої $y = \alpha$ (рис. 2.8).

Прикладами неперервних випадкових величин є інтервали руху автобусів, час роботи приладу до першої поломки, відхилення точки приземлення літального апарата, інтервал між викликами «Екстреної допомоги», час обслуговування, зріст випадкового перехожого на вулиці, середня температура за добу тощо.

Приклад 2.11.1. Функція розподілу ξ має вигляд

$$F(x) = a + b\operatorname{arctg} x,\ -\infty < x < \infty$$

(розподіл Коші). Знайти a, b, щільність та ймовірність потрапляння величини ξ до множини $[-1; 1)$.

Розв'язання. За властивостями функції розподілу маємо

$$\begin{cases} F(-\infty) = 0 = a + b\operatorname{arctg}(-\infty); \\ F(+\infty) = 1 = a + b\operatorname{arctg}(+\infty), \end{cases} \text{ або } \begin{cases} a + b\dfrac{\pi}{2} = 1; \\ a - b\dfrac{\pi}{2} = 0, \end{cases}$$

звідки $a=\frac{1}{2}$, $b=\frac{1}{\pi}$. Щільність розподілу ймовірності – це похідна функції розподілу ймовірності $F(x)$:

$$f(x)=F'(x)=\frac{1}{\pi}\frac{1}{1+x^2},\ -\infty<x<\infty.$$

Імовірність події $-1\le\xi<1$ можна знайти так:

- інтегруванням щільності:

$$P\{-1\le\xi<1\}=\int_{-1}^{1}\frac{dx}{\pi\left(1+x^2\right)}=\frac{1}{\pi}\operatorname{arctg}x\Big|_{-1}^{1}=\frac{1}{2};$$

- безпосередньо з функції розподілу:

$$P\{-1\le\xi<1\}=F(+1)-F(-1)=\frac{1}{2}+\frac{1}{\pi}\operatorname{arctg}1-\left(\frac{1}{2}+\frac{1}{\pi}\operatorname{arctg}(-1)\right)=\frac{1}{2}.$$

Зауваження. Розподіл Коші належить до так званих розподілів з «важкими хвостами» (Heavy-tailed distribution), у яких щільність розподілу зменшується зі збільшенням аргументу повільніше, ніж експоненціальна функція. У деяких випадках, як і в цьому, швидкість зменшення є настільки малою, що розподіл не має ані дисперсії, ані математичного сподівання – відповідні інтеграли не збігаються, тобто

$$M\xi=\int_{-\infty}^{\infty}xf(x)dx=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{x}{1+x^2}dx=\frac{1}{2\pi}\ln x\Big|_{-\infty}^{\infty}=\infty-\infty,$$

$$M\xi^2=\int_{-\infty}^{\infty}x^2f(x)dx=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{x^2}{1+x^2}dx=\infty.$$

Для того щоб якось описати такі, а також будь-які інші розподіли поряд з математичним сподіванням і дисперсією, уводяться такі величини, як медіана та мода, які для симетричних розподілів збігаються з математичним сподіванням.

Медіана – це така точка $x=m$, яка поділяє ймовірність на дві рівні частини: $P(x>m)=P(x<m)=\frac{1}{2}$, тобто

$$\int_{-\infty}^{m}f(x)\,dx=\int_{m}^{\infty}f(x)\,dx=\frac{1}{2}.$$

Вочевидь, медіана поділяє площу під графіком функції $f(x)$ на дві рівні частини, що дорівнюють 0,5 (загальна проща під графіком дорівнює

одиниці за визначенням).

Мода – це найбільш імовірне значення випадкової величини. Для неперервних випадкових величин ця точка відповідає максимальному значенню функції розподілу $f(x)$. Зобразимо на рис. 2.9 ці точки.

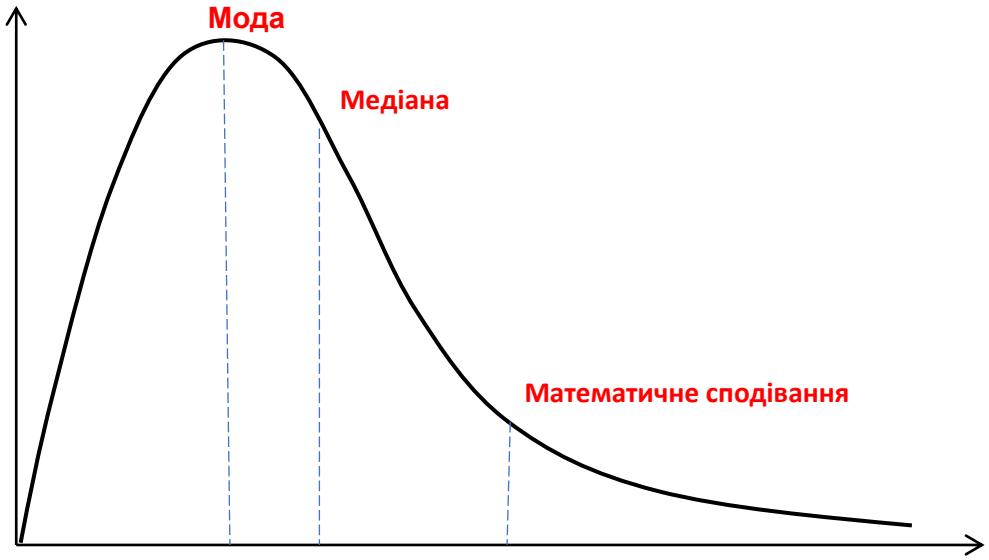


Рис. 2.9. Значення моди, медіани та математичного сподівання для неперервних випадкових величин

Унаслідок «важкого хвоста» математичне сподівання зміщується вправо. Можна сказати, що медіана поділяє площу під графіком на дві рівні за площею частини (кожна з яких дорівнює 0,5), а математичне сподівання – це центр ваги площі під графіком.

Приклад 2.11.2. Випадкова величина має щільність розподілу ймовірності

$$f(x)=\begin{cases} 0, & x\notin(-1;\,1), \\ \dfrac{c}{\sqrt{1-x^2}}, & x\in(-1;\,1) \end{cases}$$ (закон арксинуса).

Знайти константу c, величини $M\xi$, $D\xi$, імовірність події $-\frac{1}{2}\le\xi<\frac{1}{2}$, функцію розподілу $F(x)$ і квантиль рівня 0,75.

Розв'язання. Число c знайдемо з умови $\int\limits_{-1}^{1}\frac{c}{\sqrt{1-x^2}}dx=1$ (тому що поза інтервалом $(-1;\,1)$ щільність розподілу ймовірності дорівнює 0):

$$c\arcsin x\Big|_{-1}^{1}=1,\ c=\frac{1}{\pi}.$$

Розраховуємо ймовірність події:

$$P\left\{-\frac{1}{2}\le\xi<\frac{1}{2}\right\}=\int\limits_{-\frac{1}{2}}^{\frac{1}{2}}f(x)dx=\int\limits_{-\frac{1}{2}}^{\frac{1}{2}}\frac{dx}{\pi\sqrt{1-x^2}}=\frac{2}{\pi}\arcsin x\Big|_{0}^{\frac{1}{2}}=\frac{1}{3}$$

(під знаком інтеграла – парна функція).

Математичне сподівання

$$M\xi=\int\limits_{-1}^{1}x\frac{c}{\sqrt{1-x^2}}dx=0$$

(під знаком інтеграла – непарна функція).

Знайдемо $M\xi^2$ шляхом заміни змінної ($x=\sin t;\ dx=\cos t dt$):

$$M\xi^2=\frac{1}{\pi}\int\limits_{-1}^{1}\frac{x^2}{\sqrt{1-x^2}}dx=\frac{2}{\pi}\int\limits_{0}^{\frac{\pi}{2}}\frac{\sin^2 t\cos t dt}{\cos t}=\frac{1}{2}.$$

Визначимо функцію розподілу $F(x)$. Зазначимо, що щільність розподілу ймовірності відрізняється від 0 лише на інтервалі $(-1;\ 1)$. Якщо число $x\le-1$, то подія $\xi<x$ є неможливою, отже,

$$F(x)=P\{\xi<x\}=0,\ \text{якщо}\ x<-1.$$

Інтегрування дає такий же результат, оскільки інтеграл від щільності розподілу ймовірності у проміжку від –1 до 1 у цьому випадку дорівнює 1.

Залишилось розглянути випадок $x\in(-1;\ 1]$:

$$F(x)=P\{\xi<x\}=\int\limits_{-\infty}^{x}f(t)dt=\int\limits_{-\infty}^{-1}0dt+\int\limits_{-1}^{x}\frac{d(t)}{\pi\sqrt{1-t^2}}=\frac{1}{\pi}\arcsin t\Big|_{-1}^{x}=\frac{1}{\pi}\left(\arcsin x+\frac{\pi}{2}\right).$$

Таким чином,

$$F(x)=\begin{cases}0, & x\le1,\\ \dfrac{1}{\pi}\left(\arcsin x+\dfrac{\pi}{2}\right), & -1<x\le1,\\ 1, & x>1.\end{cases}$$

Корисно перевірити, що в точках ±1 функція – неперервна.

За означенням квантиля

$$F(x)=0,75=\frac{1}{2}+\frac{1}{\pi}\arcsin x,$$

звідки $\arcsin x=\frac{\pi}{4}$, $x=\frac{\sqrt{2}}{2}$.

Зауваження. Розподіл за законом арксинуса виникає, наприклад, якщо точка здійснює гармонійні коливання від -1 до 1 з центром у точці 0, і її координати фіксуємо у випадкові моменти часу. Тоді в центрі, де швидкість є максимальною, імовірність «піймати» точку буде мінімальною, а по краях, де швидкість зменшується, імовірність зафіксувати точку збільшується.

Приклад 2.11.3. Знайти $M\xi$ та $D\xi$, якщо ξ має **розподіл Лапласа** зі щільністю $f(x)=\frac{\lambda}{2}e^{-\lambda|x|}$, $-\infty<x<\infty$ $(\lambda>0)$.

Розв'язання. Зрозуміло, що внаслідок парності щільності

$$M\xi=\int_{-\infty}^{\infty}x\frac{\lambda}{2}e^{-\lambda|x|}dx=0.$$

Знайдемо $M\xi^2$:

$$M\xi^2=\int_{-\infty}^{\infty}x^2\frac{\lambda}{2}e^{-\lambda|x|}=\left|\text{користуємося парністю}\right|=\lambda\int_{0}^{\infty}x^2e^{-\lambda x}dx.$$

Інтеграл можна знайти, двічі зінтегрувавши частинами. Але цього можна не робити, якщо згадати перетворення Лапласа: оригіналу t^2 відповідає зображення $\frac{2}{p^3}$, що у наших позначеннях дає $M\xi^2=\lambda\frac{2}{\lambda^3}=\frac{2}{\lambda^2}$.

Зауваження. Розподіл Лапласа іноді називають подвійним експоненціальним розподілом, тому що його можна розглядати як два експоненціальні розподіли (з додатковим параметром місцеположення), з'єднані один з одним. Застосовується цей розподіл під час аналізу результатів прямого й оберненого дискретного перетворення Фур'є, яке своєю чергою застосовується у форматі зображень .jpg та під час диджиталізації аудіо.

Графік розподілу $f(x)=\frac{1}{2b}e^{-\frac{|x-\mu|}{b}}$, $-\infty<x<\infty$ $(b>0)$ зображено на рис. 2.10.

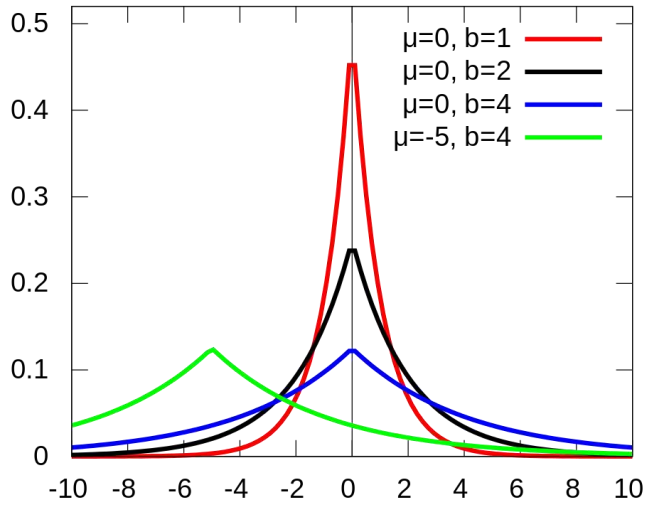


Рис. 2.10. Графік щільності розподілу ймовірності $f(x)=\frac{1}{2b}e^{-\frac{|x-\mu|}{b}}$ для деяких значень μ та b

Рівномірний розподіл. Випадкова величина ξ має неперервний рівномірний розподіл на відрізку $[a;b]$, якщо щільність її розподілу на цьому відрізку є сталою величиною:

$$f(x)=\begin{cases}\dfrac{1}{b-a}, \text{ якщо } x\in[a;b],\\ 0, \quad \text{якщо } x\notin[a;b].\end{cases}$$

Таку щільність називають прямокутною. Назва розподілу підкреслює, що ймовірність потрапляння величини ξ у будь-який інтервал $I\subset[a;b]$ залежить лише від довжини інтервалу, а не від його розташування.

Числові характеристики:

$$M\xi=\frac{a+b}{2},\ D\xi=\frac{(b-a)^2}{12}.$$

Про рівномірний розподіл величини ξ часто кажуть: «точку ξ вибрано навмання із $[a;b]$». Такий розподіл мають помилки округлення, випадкові числа, час, протягом якого пасажир чекатиме на зупинці, якщо автобуси під'їжджають з однаковим інтервалом.

Показниковий (експоненціальний) розподіл. Неперервна випадкова величина ξ має показниковий розподіл, якщо щільність її розподілу визначається формулою

$$f(x)=\begin{cases}0, & x\le 0,\\ \lambda e^{-\lambda x}, & x>0.\end{cases}$$

Числові характеристики: $M\xi = \frac{1}{\lambda}$, $D\xi = \frac{1}{\lambda^2}$. Графік $f(x)$ при деяких значеннях параметра λ зображено на рис. 2.11.

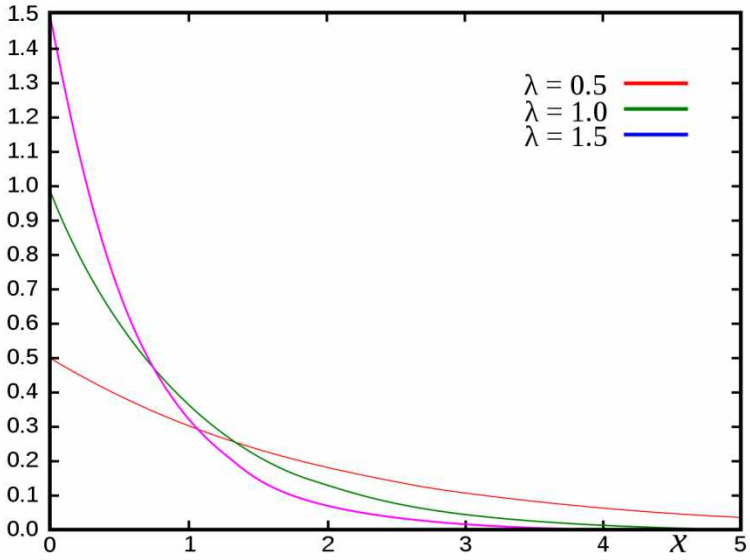


Рис. 2.11. Графіки щільності показникового розподілу ймовірності

Показниковий розподіл описує процес, у якому події відбуваються безперервно й незалежно одна від одної з **постійною середньою швидкістю**. Це окремий випадок гамма-розподілу. Це безперервний аналог геометричного розподілу, і його ключова властивість – **відсутність пам'яті** (тобто чергова подія не залежить від попередньої).

Показниковий розподіл у багатьох випадках можуть мати проміжок часу між двома послідовними поломками складної системи, проміжок часу між двома послідовними викликами на АТС, час очікування на обслуговування, час між спрацьовуваннями датчика Гейгера тощо. При певних припущеннях час між появами двох послідовних покупців у магазині буде випадковою величиною з експоненціальним розподілом і тому застосується в **теорії масового обслуговування** (теорії черг) і **теорії надійності**. Середній час очікування нового покупця дорівнює $\frac{1}{\lambda}$. Сам параметр λ тоді можна інтерпретувати як середню кількість нових покупців за одиницю часу.

Функція розподілу має вигляд

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Унаслідок того, що показниковий розподіл часто застосовується в задачах теорії надійності, разом з функцією розподілу розглядається також

так звана функція надійності $\bar{F}(x) = P\{\xi > x\} = e^{-\lambda x}$. Її зміст – імовірність того, що на проміжку часу x елемент або система працюють без відмови.

Розподіл Парето – це двопараметрична сім'я абсолютно неперервних розподілів. Названий на честь італійського інженера з цивільного будівництва, економіста й соціолога Вільфредо Парето. Це **степеневий розподіл** імовірностей, який використовується для опису соціальних, наукових, геофізичних, актуарних та багатьох інших спостережуваних явищ.

Спочатку розподіл Парето застосовувався для опису розподілу багатства серед суспільства, що відповідає тенденції, коли велика частина багатства зосереджується в руках невеликої частини населення. У розмовній версії розподіл Парето відомий як принцип Парето, або «правило 80—20», а іноді його також називають ефектом Матвія. За цим правилом, наприклад, 80 % багатства суспільства утримують 20 % його населення.

Як вже було сказано, цей вид розподілу залежить від двох параметрів. Функція щільності розподілу має вигляд

$$f(x) = \begin{cases} 0, & x < a, \\ \dfrac{ma^m}{x^m}, & x \geq a, \end{cases}$$

де a – деяке значення, з якого починаються ненульові значення $f(x)$; m – степінь, який не обов'язково має бути цілим числом.

На рис. 2.12 показано графіки розподілу з $a = 1$ при різних m.

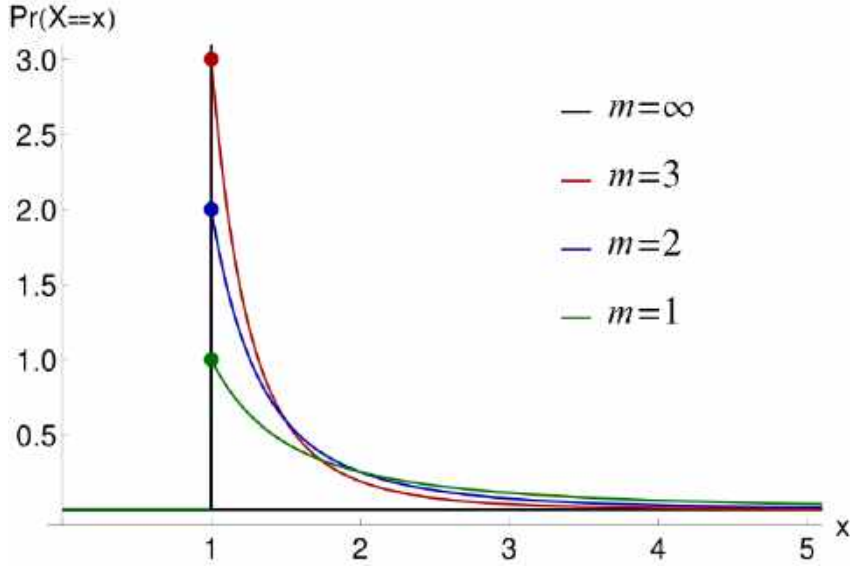


Рис. 2.12. Графіки щільності розподілу ймовірності при деяких значеннях m

Мода розподілу Парето дорівнює a, медіана – $a\sqrt[m]{2}$, а математичне сподівання $M\xi = \begin{cases} \infty, & m \le 1, \\ \dfrac{ma}{m-1}, & m > 1. \end{cases}$

Цікаві факти. Наведені приклади іноді розглядають як такі, що приблизно мають розподіл Парето:

- розмір населених пунктів (мало великих міст і багато селищ);
- розподіл розмірів файлів в інтернет-трафіку, у якому використовується протокол TCP (Transmission Control Protocol – протокол керування передаванням);
- величина запасів нафти в родовищах;
- розмір метеоритів;
- обсяг задач, які виносилися для розв'язування на суперкомп'ютерах (декілька великих, багато малих);
- **закон Ципфа** — лінгвостатистичний закон, згідно з яким відношення рангу слова в частотному словнику до частотності слова в мові є постійною величиною (константою). Інакше кажучи, якщо всі слова мови (або просто достатньо довгого тексту) упорядкувати за зменшенням частоти їх використання, то частотність *n*-го слова в такому списку буде приблизно обернено пропорційною його порядковому номеру *n* (так званому рангу цього слова).

Заміною змінної (якщо від значень випадкової величини взяти логарифм) степеневий розподіл можна звести до експоненціального розподілу.

Нормальний розподіл (розподіл Гаусса). Для нормального розподілу існує спеціальне позначення: $N\left(a;\sigma^2\right)$. Графік щільності випадкової величини має вигляд, який називають дзвоном (рис. 2.13).

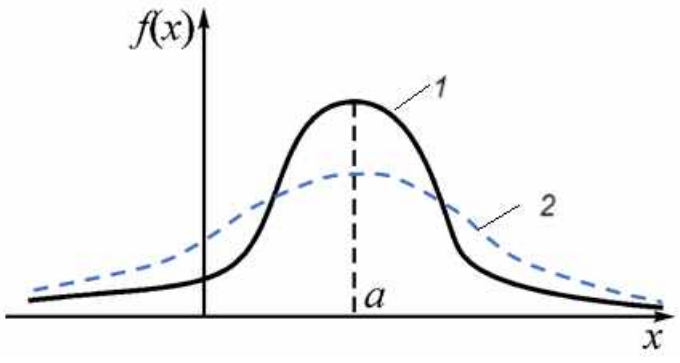


Рис. 2.13. Розподіл Гаусса

Формула розподілу Гаусса:

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-a)^2}{2\sigma^2}},\ -\infty<x<\infty,$$

де число a може бути довільним, $M\xi=a$; число $\sigma>0$ – це середньоквадратичне відхилення, $D\xi=\sigma^2$.

Величина σ характеризує розкид розподілу навколо математичного сподівання a. Чим менше σ, тим пік розподілу вище, а «хвости» тонше, тобто випадкова величина щільніше групується навколо математичного сподівання (графік 1 на рис. 2.13). Збільшення σ приводить до розширення графіка і зменшення його висоти (графік 2 на рис. 2.13).

Надзвичайно широке застосування нормального розподілу зумовлене тим, що саме такий розподіл мають **суми великої кількості незалежних випадкових величин**, тобто результат виникає під впливом великої кількості факторів, які діють незалежно і кожний з яких окремо впливає незначно (похибки вимірювань, розміри деталей, відхилення при пострілах тощо).

Для нормальної величини ймовірність потрапляння в інтервал $[A;B)$ можна обчислити за формулою

$$P\{A\le\xi<B\}=\Phi\left(\frac{B-a}{\sigma}\right)-\Phi\left(\frac{A-a}{\sigma}\right).$$

Нагадаємо, що $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{t^2}{2}}dt$ – функція Лапласа. Її значення можна знайти наближено за допомогою таблиць або ЕОМ (де вона позначається $Erf(x)$). Якщо $x<0,$ то згадаємо про непарність:

$$\Phi(-x)=-\Phi(x).$$

Таблиці закінчуються на значенні $\Phi(5)\approx 0,49999997$. Це пов'язано з тим, що $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}dx=1,$ а відповідний інтеграл від 0 до $+\infty$ дорівнює $\frac{1}{2}$.

Функція $e^{-\frac{x^2}{2}}$ швидко зменшується, тому майже вся площина зосереджена в проміжку від -5 до 5. Тому, наприклад, $\Phi(7)\approx\Phi(10)\approx 0,5.$

Зауваження 1. Показниковий (експоненціальний) розподіл характеризується одним параметром λ і тому є однопараметричним. Розподіл Гаусса характеризується двома параметрами a,σ і тому є двопараметричним. Існують й інші розподіли, що характеризуються трьома параметрами. Ще деякі важливі однопараметричні й двопараметричні

розподіли розглянемо пізніше в задачах математичної статистики.

Зауваження 2. Для нормального розподілу значення, що відрізняються від середнього на число, менше за одне стандартне відхилення, становлять 68,27 % популяції. Водночас значення, що відрізняються від середнього на два стандартних відхилення, становлять 95,45 %, а на три стандартних відхилення – 99,73 %. Це іноді називають «правилом трьох сигм», яке полягає в тому, що майже всі значення випадкової величини потрапляють в інтервал, який відхиляється від математичного сподівання на три σ, а ймовірність того, що значення випадкової величини відрізнятиметься від математичного сподівання більш ніж на три σ, є мізерно малою (рис. 2.14).

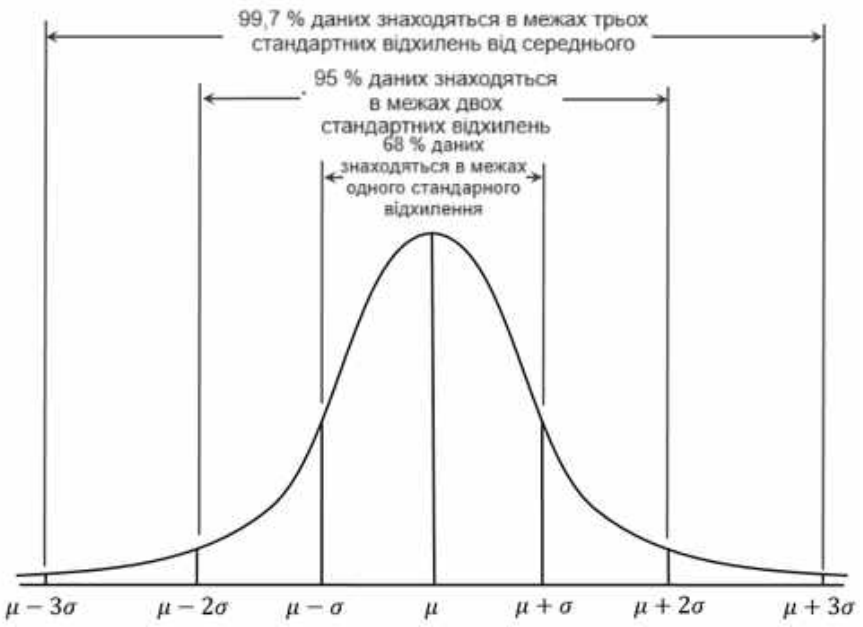


Рис. 2.14. Правило трьох сигм

З розподілом Гаусса тісно пов’язана **центральна гранична теорема**, яку розглянемо окремо.

2.12. Центральна гранична теорема

Центральна гранична теорема – теорема теорії ймовірностей про те, що розподіл суми незалежних однаково розподілених випадкових величин наближається до нормального. Ця теорема підкреслює особливість нормального розподілу в теорії ймовірностей.

Наприклад, отримано вибірку, яка містить велику кількість спостережень, кожне з яких було отримано випадковим чином і не залежить від інших спостережень, на основі значень цих спостережень розраховують арифметичне середнє. Якщо цю процедуру повторити багаторазово, то за центральною граничною теоремою розраховані

середні значення матимуть **нормальний розподіл**. Простим прикладом цього є багаторазове підкидання монети, коли ймовірність випадання заданої кількості гербів у всій послідовності подій наближається до нормальної кривої, а середнє знаходиться в середині загальної кількості випадання монети на кожен бік. (Граничне значення для нескінченної кількості підкидань дорівнює нормальному розподілу.)

Центральна гранична теорема має декілька варіантів. У загальній формі випадкові величини мають бути однаково розподілені. Середнє значення має нормальний розподіл також у випадку неоднаково розподілених величин, тобто не лише при незалежних спостереженнях, що буде відбуватися за умови виконання певних умов.

У перших версіях цієї теореми нормальний розподіл може використовуватися у вигляді апроксимації біноміального розподілу, що відомо як локальна теорема Муавра – Лапласа.

Таким чином, центральна гранична теорема встановлює умови виникнення нормального розподілу при додаванні великої кількості незалежних випадкових величин.

Якщо $\xi_1, \xi_2, \ldots, \xi_n$ – незалежні випадкові величини, які мають один і той же розподіл з математичним сподіванням m і дисперсією σ^2, то при збільшенні кількості величин n закон розподілу суми $\xi_1 + \xi_2 + \ldots + \xi_n$ буде наближатися до нормального:

$$P\left\{ A \le \frac{\xi_1 + \xi_2 + \ldots + \xi_n - nm}{\sigma\sqrt{n}} \le B \right\} \to \frac{1}{\sqrt{2\pi}} \int_A^B \mathrm{e}^{-\frac{x^2}{2}}\, dx .$$

Випадкова величина $\dfrac{\xi_1 + \xi_2 + \ldots + \xi_n - nm}{\sigma\sqrt{n}}$ є центрованою та нормованою, тобто внаслідок того, що із суми $\xi_1 + \xi_2 + \ldots + \xi_n$ відняли nm, математичне сподівання отриманої випадкової величини дорівнює нулю, а внаслідок того, що суму поділили на $\sigma\sqrt{n}$ (середньоквадратичне відхилення), дисперсія отриманої величини дорівнює одиниці.

Важливо зазначити, що математичне сподівання суми n випадкових величин з математичним сподіванням m буде дорівнювати nm, а середньоквадратичне відхилення – $\sigma\sqrt{n}$, тобто математичне сподівання збільшується пропорційно n, а середньоквадратичне відхилення – пропорційно $\sqrt{n}$.

Приклад 2.12.1. Гравець кидає 1000 разів гральний кубик. Якою є ймовірність того, що сума очок знаходиться в межах від 3450 до 3550?

Розв'язання. Спочатку дослідимо параметри розподілу випадкової величини – кількості очок у кожному кидку. Будемо вважати, що

ймовірності появи 1, 2, 3, 4, 5 і 6 очок є однаковими й дорівнюють $\frac{1}{6}$.

Таким чином, математичне сподівання

$$m=\frac{1}{6}\cdot 1+\frac{1}{6}\cdot 2+\frac{1}{6}\cdot 3+\frac{1}{6}\cdot 4+\frac{1}{6}\cdot 5+\frac{1}{6}\cdot 6=\frac{1}{6}\cdot\left(1+2+3+4+5+6\right)=\frac{21}{6}=3{,}5.$$

Дисперсію розрахуємо так:

$$D\xi=\sum_{n=1}^{6}x_i^2\,p_i-\left(m\right)^2=\frac{1}{6}\cdot\left(1^2+2^2+3^2+4^2+5^2+6^2\right)-3{,}5^2\approx 2{,}9167.$$

Середньоквадратичне відхилення

$$\sigma=\sqrt{2{,}9167}\approx 1{,}708.$$

Розрахуємо потрібні параметри:

$$mn=3{,}5\cdot 1000=3500,$$

$$\sigma\sqrt{n}=1{,}708\cdot\sqrt{1000}=54{,}006.$$

Далі застосуємо отримані результати:

$$P\left\{3450\le\xi_1+\xi_2+\ldots+\xi_{1000}\le 3550\right\}=$$

$$=P\left\{\frac{3450-3500}{54{,}006}\le\frac{\xi_1+\xi_2+\ldots+\xi_{1000}}{54{,}006}\le\frac{3550-3500}{54{,}006}\right\}=$$

$$=P\left\{-0{,}926\le\frac{\xi_1+\xi_2+\ldots+\xi_{1000}}{54{,}006}\le 0{,}926\right\}=\frac{1}{\sqrt{2\pi}}\int_{-0{,}926}^{0{,}926}\mathrm{e}^{-\frac{x^2}{2}}\,dx=$$

$$=2\left(\Phi\left(0{,}926\right)-\frac{1}{2}\right)=2\Phi\left(0{,}926\right)-1=2\cdot 0{,}823-1=0{,}646.$$

Тут для розрахунку інтеграла було застосовано відповідну таблицю (загалом для розрахунку функції Лапласа можна застосувати будь-яку математичну програму – Python або навіть EXCEL), а також ураховано те, що межі інтервалу є симетричними відносно математичного сподівання й відхиляються від нього лише на 50. Таким чином, імовірність того, що сумарне значення кількості очок в інтервалі від 3450 до 3550 дорівнює 0,646.

Зауваження. Розподіл Гаусса згідно з центральною граничною теоремою вважається універсальним розподілом, який можна застосувати в будь-якій ситуації. Чи це так?

Узагалі нормальний розподіл дійсно широко застосовується, але має й певні недоліки. Насамперед це симетричний розподіл, який описується двома параметрами. Але можливими є ситуації, коли симетричний розподіл неможливо застосувати. Крім того, нормальний розподіл має один «хвіст» у напрямку від'ємних величин, тобто випадкова величина може набувати від'ємних значень, хоча і з малою ймовірністю. Це,

зрозуміло, може суперечити фізичній суті величини. Наприклад, не може добова каса продажів у супермаркеті бути меншою від нуля. Або вміст спирту у винах, кислотність тощо.

Приклад 2.12.2. Нехай хтось здійснює постріли по мішені. Похибки влучання в центр мішені по осях x і y нехай будуть незалежними й такими, що підпорядковуються однаковому нормальному закону розподілу з нульовим математичним сподіванням, що збігається з центром мішені й центром системи координат, і середньоквадратичним відхиленням, що дорівнює, наприклад, 3. Необхідно дослідити розподіл відстаней між точками влучань та центром мішені.

Розв'язання. Згідно з правилом трьох сигм більшість значень координат будуть міститися в межах відрізків $[-9;9]$ як по осі x, так і по осі y.

Розрахуємо за цим законом координати 1500 точок і зобразимо точки на рис. 2.15.

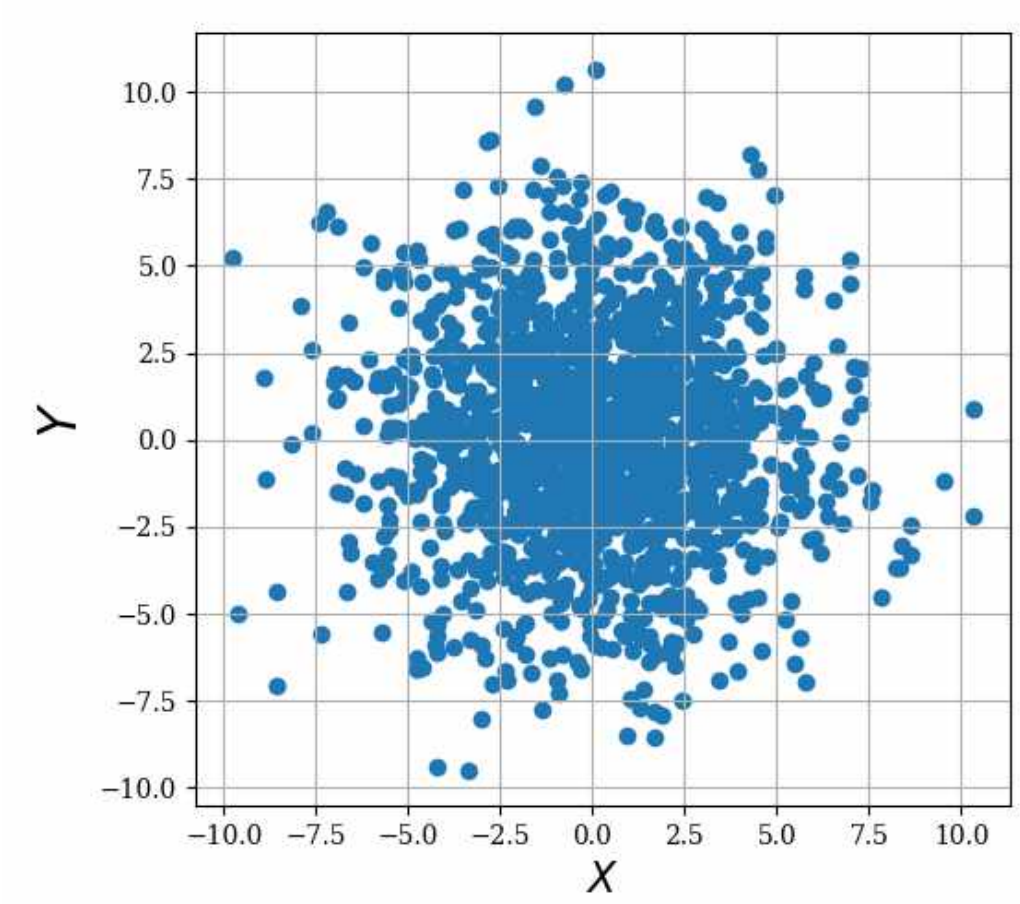


Рис. 2.15. Розподіл точок на площині

Покажемо на гістограмах розподіл координат (рис. 2.16). Для цього розіб'ємо інтервали, у яких містяться координати x і y, на певну кількість інтервалів (у цьому випадку на 15) і порахуємо, скільки точок потрапляє в

кожний інтервал. У Python це можна зробити з допомогою однієї команди plt.hist().

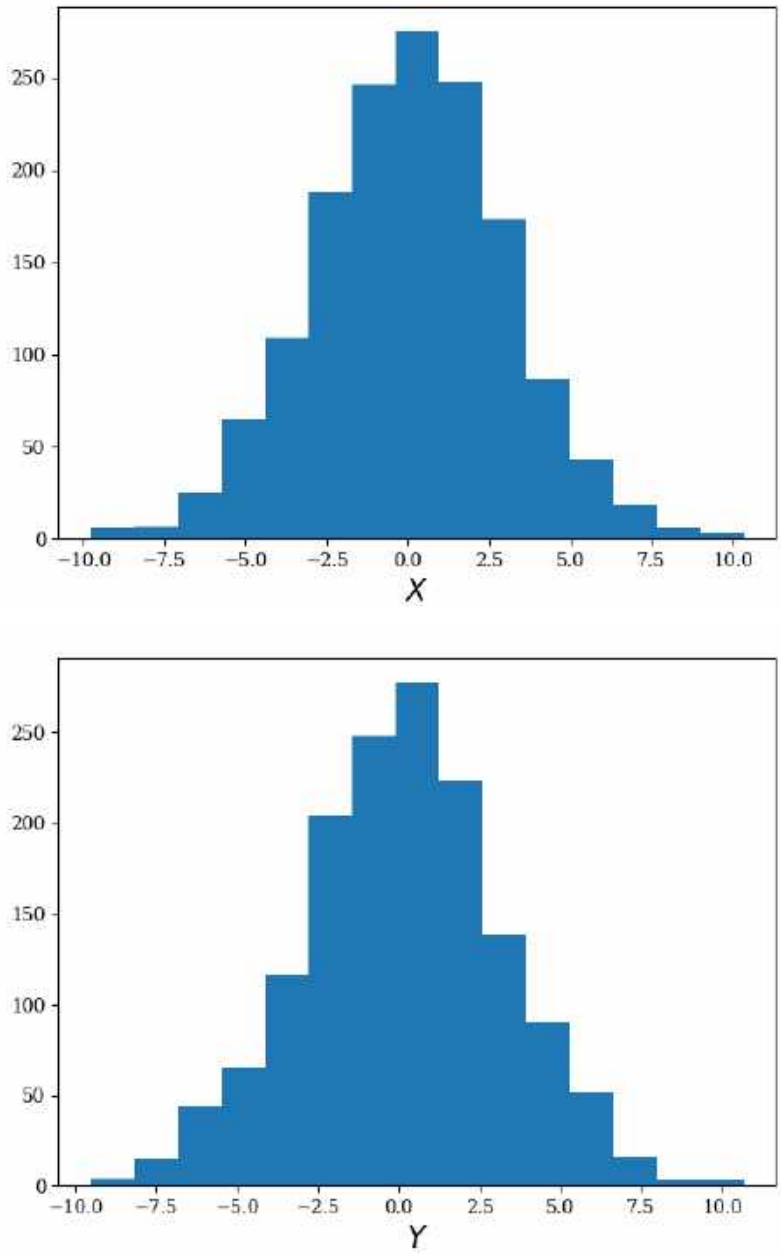


Рис. 2.16. Гістограми розподілу координат точок

Як бачимо, розподіли дуже близькі до теоретичного нормального розподілу з нульовим математичним сподіванням і середньоквадратичним відхиленням, що дорівнює 3.

Зараз необхідно визначити, за яким законом розподіляється відстань точок від початку координат

$$r = \sqrt{x^2 + y^2}\ .$$

Зрозуміло, що відстань не може бути меншою від нуля, тобто

розподіл Гаусса тут не буде узгоджуватися з емпіричними даними. Розрахуємо відстані й побудуємо гістограму розподілу (рис. 2.17).

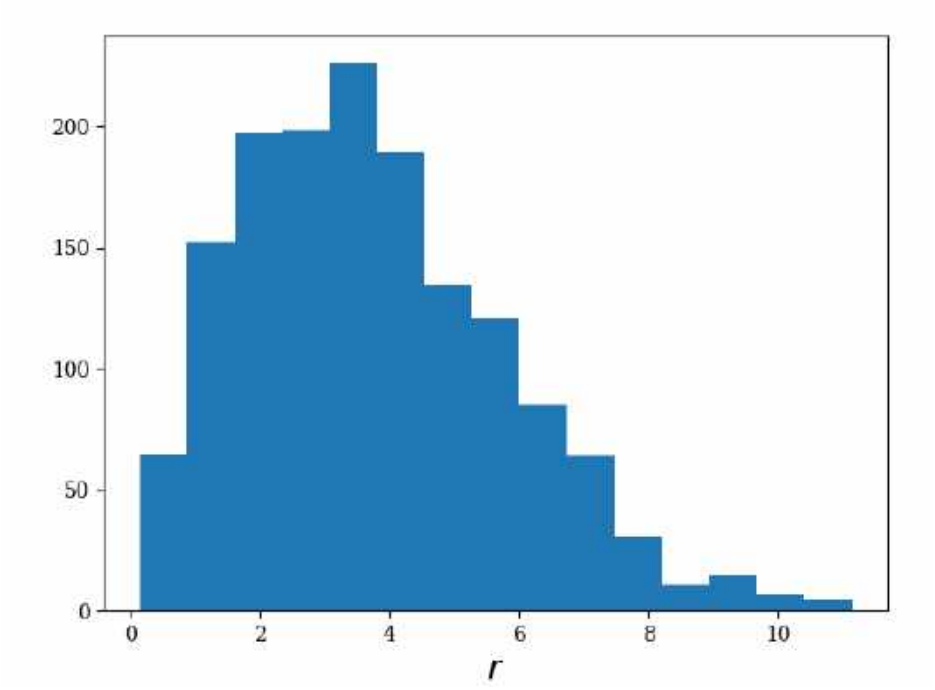


Рис. 2.17. Гістограми розподілу відстаней точок

Як бачимо, цей розподіл має лише один «хвіст», не є симетричним і починається з нуля.

Це так званий **розподіл Релея** (рис. 2.18)

$$f(x)=\frac{x}{a^2}\mathrm{e}^{-\frac{x^2}{2a^2}},\quad x\geq 0.$$

Уперше 1880 року цей розподіл увів Джон Вільям Стретт (лорд Релей) у зв'язку з задачею додавання гармонійних коливань з випадковими фазами. Належить до однопараметричних розподілів і має такі характеристики:

- математичне сподівання $a\sqrt{\frac{\pi}{2}}$;
- медіана $a\sqrt{\ln 4}$;
- мода a;
- дисперсія $\frac{4-\pi}{2}a^2$.

Якщо незалежні гавсівські випадкові величини мають ненульові математичні сподівання, неоднакові в загальному випадку, то розподіл Релея перетворюється на двопараметричний **розподіл Райса**.

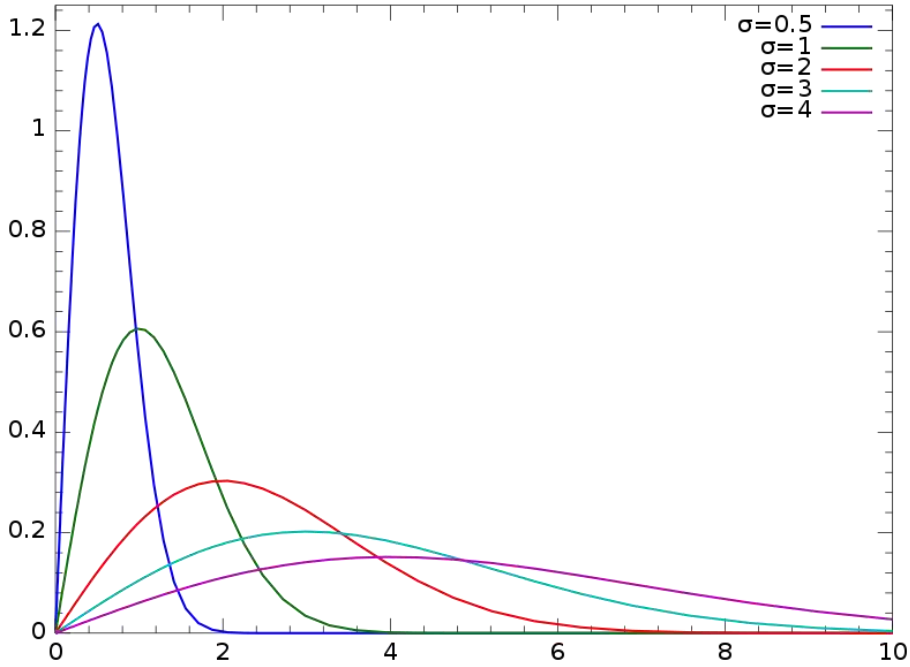


Рис. 2.18. Розподіл Релея

Зауваження:

- як бачимо, існує велика кількість розподілів випадкових величин, які класифікуються за кількістю параметрів у рівнянні;
- вигляд розподілу визначається типом випадкового процесу;
- заміна змінної (наприклад, якщо замість x узяти $\ln x$) дає змогу перейти від одного розподілу до іншого;
- наявні розподіли можуть бути базою для конструювання нових розподілів;
- у середовищах роботи з даними (Python, R тощо) вбудовано генератори випадкових величин за багатьма розподілами, побудова даних за певними розподілами застосовується під час тренування й перевірки моделей.

2.13. Системи випадкових величин

Розглянемо дві випадкові величини – ξ та η, що є визначеними на одному просторі елементарних подій. Пару (ξ, η) називають системою двох випадкових величин або двовимірним випадковим вектором.

Спочатку розглянемо випадок, коли обидві величини є дискретними, тобто множина можливих значень ξ – це $\{x_1, ..., x_n\}$, а η – це $\{y_1, ..., y_m\}$.

Сукупність імовірностей

$$p_{ij} = P\{\xi = x_i, \eta = y_j\}$$

являє собою так званий **сумісний розподіл** випадкових величин ξ, η, який зручно записати у вигляді таблиці:

ξ	η				
	b	y_2	...	y_m	
x_1	p_{11}	p_{12}	...	p_{1m}	$P\{x_1\}$
x_2	p_{21}	p_{22}	...	p_{2m}	...
...	...	...	...	...	...
x_n		p_{n2}	...	p_{nm}	...
	$P\{y_1\}$	...	...		

Кожне число в таблиці є невід'ємним: $p_{ij} \geq 0$. Зрозуміло, що сума всіх чисел дорівнює 1:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} = 1.$$

За цією таблицею можна легко знайти окремо розподіли величин ξ та η (так звані маргінальні розподіли), а саме: щоб знайти $P\{\xi = x_i\}$, достатньо знайти суму $P\{\xi = x_i\} = \sum_{j=1}^{m} p_{ij}$, тобто суму чисел у **рядку**; аналогічно знаходимо суму $P\{\eta = y_j\} = \sum_{i=1}^{n} p_{ij}$ – суму чисел у **стовпці**. У таблиці ці елементи виділено темною заливкою.

У випадку, коли для будь-яких x_i і y_j виконується рівність

$$P\{\xi = x_i, \eta = y_j\} = P\{\xi = x_i\} P\{\eta = y_j\},$$

величини ξ, η є **незалежними**.

Для системи (ξ, η) існує числова характеристика – коваріація

$$\operatorname{cov}(\xi, \eta) = M(\xi - M\xi)(\eta - M\eta),$$

яку можна обчислити також за формулою

$$\operatorname{cov}(\xi, \eta) = M\xi\eta - M\xi M\eta.$$

Коваріація має розмірність, що дорівнює добутку розмірностей величин ξ та η. Разом з тим існує безрозмірна характеристика – **коефіцієнт кореляції**

$$\rho = \frac{\operatorname{cov}(\xi, \eta)}{\sqrt{D\xi}\sqrt{D\eta}},$$

для якого завжди виконується нерівність $-1 \leq \rho \leq 1$.

Зауваження. Якщо величини ξ та η – **лінійно незалежні**, то $\operatorname{cov}(\xi, \eta) = 0$, $\rho = 0$.

Таким чином, наявність ненульової коваріації свідчить про **лінійну залежність** величин ξ та η.

Наведемо додатково деякі властивості математичного сподівання й дисперсії:

$$D(\xi+\eta)=D\xi+D\eta+2\operatorname{cov}(\xi,\eta),$$
$$M(\xi\eta)=M\xi M\eta+\operatorname{cov}(\xi,\eta).$$

Отже, дисперсія суми дорівнює сумі дисперсій лише у випадку **некорельованих** величин (зрозуміло, що це справджується і для **незалежних** величин).

Приклад 2.13.1. Сумісний розподіл ξ та η подано у вигляді таблиці:

ξ	η			
	–1	**0**	**1**	
0	0,1	0,2	0,1	0,4
1	0,2	0,3	0,1	0,6
	0,3	0,5	0,2	

Знайти ρ, вирішити питання про незалежність величин. Обчислити ймовірність випадкової події $\{\xi+\eta=0\}$.

Розв'язання. Знайдемо маргінальні розподіли, для чого спочатку обчислимо суми чисел у рядках: $P\{\xi=0\}=0,4;\{\xi=1\}=0,6.$ Очевидно, що $M\xi=0,6,\ M\xi^2=0,6,\ D\xi=0,6-0,36=0,24$.

Знаходимо суми чисел у стовпцях:

$$P\{\eta=-1\}=0,3,\quad P\{\eta=0\}=0,5,\quad P\{\eta=1\}=0,2.$$

Маємо $M\eta=-0,3+0,2=-0,1,\ M\eta^2=0,3+0,2=0,5,\ D\eta=0,49.$

Середнє значення добутку $M\xi\eta$ визначаємо так:

$$M\xi\eta=\sum_{i=1}^{n}\sum_{j=1}^{m}x_iy_jp_{ij}=0\cdot(-1)\cdot(0,1)+1\cdot(-1)\cdot0,2+0\cdot0,2+0\cdot0,3+0,1\cdot0,1+1\cdot1\cdot0,1=$$
$$=-0,2+0,1=-0,1.$$

Отже, $\operatorname{cov}(\xi,\eta)=M(\xi\eta)-M\xi M\eta=-0,04$, тоді

$$\rho=-\frac{0,04}{\sqrt{0,49}\sqrt{0,24}}\approx-0,117.$$

Зрозуміло, що $\rho\neq0$, звідки випливає, що ξ та η – залежні величини.

Визначаємо ймовірності випадкової події $\{\xi+\eta=0\}$. Зазначимо, що ця подія складається із двох несумісних подій $\{\xi+\eta=0\}=\{\xi=0,\eta=0\}\cup\{\xi=1,\eta=-1\}$, імовірності яких знаходимо з таблиці, і за теоремою додавання ймовірностей отримуємо

$$P\{\xi+\eta=0\}=P\{\xi=0,\eta=0\}+P\{\xi=1,\eta=-1\}=0,2+0,2=0,4.$$

У найширшому розумінні кореляція – це будь-яка **статистична асоціація**, хоча й зазвичай належить до ступеня лінійного зв'язку пари змінних. Відомі приклади залежних явищ містять кореляцію між зростанням батьків та їх потомства, а також кореляцію між ціною товару й кількістю товару, який споживачі готові купити.

Кореляції є корисними, тому що можуть указувати на передбачуваний взаємозв'язок, що можна використовувати на практиці, наприклад: електроенергетична компанія може виробляти менше електроенергії в «м'який» день унаслідок кореляції між попитом на електроенергію й погодними умовами. У цьому прикладі існує причиново-наслідковий зв'язок, оскільки екстремальні погодні умови змушують людей використовувати більше електроенергії для обігрівання або охолодження. Однак зазвичай наявності кореляції недостатньо, щоб зробити висновок про наявність причиново-наслідкового зв'язку (кажучи про кореляцію, не мають на увазі причиново-наслідковий зв'язок!).

Висловлюючись неформальною мовою, кореляція є синонімом залежності. Однак кореляція належить до будь-якого конкретного типу математичних операцій між тестованими змінними та відповідними очікуваними значеннями. По суті, кореляція – це міра того, як дві або більше змінні зв'язані одна з одною. Є кілька коефіцієнтів кореляції, якими визначається ступінь кореляції. Найбільш поширеним серед них є **коефіцієнт кореляції Пірсона**, що є чутливим тільки до лінійної залежності між двома змінними (яка може бути наявною, навіть якщо одна змінна є нелінійною функцією іншої). Інші коефіцієнти кореляції, наприклад коефіцієнт рангової **кореляції Спірмена**, є надійнішими, ніж коефіцієнти Пірсона, тобто чутливішими до нелінійних залежностей. Взаємна інформація також може застосовуватися для визначення залежності між двома змінними.

Фраза «кореляція не має на увазі причиново-наслідковий зв'язок» стосується неможливості законно вивести причиново-наслідковий зв'язок між двома подіями або змінними виключно на основі спостережуваної асоціації або кореляції між ними. Ідея про те, що «кореляція має на увазі причиновий зв'язок», є прикладом **логічної помилки** з сумнівною причиною, коли дві події, що відбуваються разом, уважаються такими, між якими є причиново-наслідковий зв'язок. Ця помилка також відома як латинський вислів cum hoc ergo propter hoc («з цим, отже, через це») і відрізняється від помилки, відомої як post hoc ergo propter hoc («після цього, отже, через це»), коли наступна подія розглядається як наслідок попередньої події, а злиття – помилкове злиття двох подій, ідей, баз даних тощо.

Як і у випадку з будь-якою логічною помилкою, визначення того, що аргумент є помилковим, не обов'язково означає, що отриманий висновок є помилковим. Було запропоновано статистичні методи, у яких

використовується кореляція як основа для перевірки гіпотез про причиновість, включаючи тест причиновості Грейнджера і конвергентне перехресне відображення.

Наведемо приклади нелогічного висновку зі знайденого кореляційного зв’язку.

Обернена причиновість (переплутано причину й наслідок)

1. *«Чим швидшим є спостережуване обертання вітряних млинів, тим сильнішим є спостережуваний вітер».*

Отже, вітер спричиняється обертанням вітряних млинів, простіше кажучи, вітряні млини, як указує їх назва, – це машини, що використовуються для виробництва вітру. У цьому прикладі кореляція (одночасність) між роботою вітряка і швидкістю вітру не означає, що вітер спричиняється вітряними млинами. Це, найімовірніше, навпаки, про що свідчить той факт, що вітру не потрібні вітряки для існування, тоді як вітряним млинам потрібен вітер для обертання. Вітер можна спостерігати в місцях, де немає вітряних млинів або не обертаються вітряні млини, і є вагомі підстави вважати, що вітер існував до винаходу вітряних млинів.

2. У деяких випадках може бути просто незрозуміло, що є причиною, а що – наслідком, наприклад: *«Діти, які багато дивляться телевізор, є найжорстокішими. Очевидно, що телебачення робить дітей більш жорстокими».*

Це могло легко бути й навпаки, тобто більш жорстокі діти більше люблять дивитися телевізор, ніж менш жорстокі.

3. Кореляція між вживанням рекреаційних наркотиків і психічними розладами може бути будь-якою: можливо, ліки спричиняють розлади, а, можливо, люди використовують наркотики для самолікування при вже наявних станах. За теорією «шлюзового наркотику» можна стверджувати, що вживання маріхуани призводить до вживання більш важких наркотиків, а вживання важких наркотиків може спричинити вживання маріхуани. Дійсно, у соціальних науках, де контрольовані експерименти часто не можуть бути використані для визначення напрямку причинового зв'язку, ця помилка може підживлювати давні наукові аргументи. Один з таких прикладів можна знайти в економіці освіти, сигналізації та людського капіталу: це може бути або те, що наявність вроджених здібностей дає змогу людині завершити освіту, або те, що завершення освіти розвиває її здібності.

Загальна причина (не враховано третій фактор, який впливає і на X, і на Y)

1. *«Чим більше пожежників гасять пожежу, тим більшим є збиток від пожежі».*

Звичайно, площа пожежі, місце пожежі й наявність у місці пожежі цінностей або людей впливають як на кількість пожежників, які її гасять, так і на обсяг збитків від пожежі. Велику пожежу гасять багато пожежників і

така пожежа призводить до великих збитків.

2. *«Маленькі діти, які сплять з увімкненим світлом, набагато частіше хворіють на короткозорість у більш пізньому віці. Тому сон з увімкненим світлом спричиняє короткозорість».*

Цей науковий висновок, отриманий унаслідок дослідження в Медичному центрі Університету Пенсільванії, було опубліковано 13 травня 1999 року, а дослідження висвітлювалося в популярній пресі. Однак більш пізні дослідження в Університеті штату Огайо не підтвердили залежність розвитку короткозорості від сну з увімкненим світлом. Дослідження дійсно виявили сильний зв'язок між батьківською короткозорістю й розвитком дитячої міопії, а також було зазначено, що короткозорі батьки з більшою ймовірністю залишали світло ввімкненим у спальні своїх дітей. У цьому випадку причиною обох станів є міопія батьків, а вищенаведений висновок є неправильним.

3. *«У міру збільшення продажів морозива різко збільшується кількість смертей від утоплення. Отже, споживання морозива спричиняє утоплення».*

У цьому прикладі не враховується важливість пори року й температури для продажу морозива. Морозиво продається в спекотні літні місяці значно частіше, ніж у холодну пору року, і саме в ці спекотні літні місяці люди з більшою ймовірністю будуть займатися водними видами спорту, такими як плавання. Збільшення кількості смертей від утоплення просто спричиняється великим впливом води, а не морозивом. Такий висновок є неправильним.

Двоспрямована причиновість (X спричиняє Y, а Y спричиняє X)

Прикладами є відношення хижак – жертва в екосистемі (зменшення кількості жертв призводить до вимирання хижаків, що зменшує споживання жертв і збільшення популяції); антропометрія спортсмена та його успіхи в спорті (успіхи будуть у генетично обдарованих, які мають видатну антропометрію, і навпаки, заняття спортом впливають на антропометричні дані).

Відношення між X та Y – випадкові

Прикладом є чергування лисих і волосатих російських лідерів: майже 200 років поспіль лисий (або явно лисуватий) державний лідер Росії змінюється на нелисого (волосатого) лідера, і навпаки.

3. ОСНОВИ МАТЕМАТИЧНОЇ СТАТИСТИКИ

3.1. Генеральна сукупність і вибірка

Предметом математичної статистики є вивчення випадкових величин за результатами спостережень, дослідів, випробовувань, що

повторюються. Однак у багатьох задачах, пов'язаних із випробовуваннями, що повторюються, неможливо провести всі можливі випробовування або спостереження так званої генеральної сукупності з фінансових, часових, принципових або інших причин, а можна опрацювати лише доступну, вибіркову частину цих випробовувань. Ця вибірка дає уявлення про частоту і склад потрібних ознак у генеральній сукупності. Іншими словами, вибірка – частина відібраних **випадковим чином** об'єктів (результати спостережень над обмеженою кількістю об'єктів). Елементи вибірки $x_1, x_2, ..., x_n$ будемо вважати незалежними однаково розподіленими випадковими величинами. Для того щоб за вибіркою можна було робити впевнені висновки, вибірка має бути репрезентативною. Репрезентативність вибірки забезпечується випадковістю відбору і достатньою кількістю елементів. Таким чином, завдання математичної статистики – оброблення результатів спостережень (вибіркової сукупності).

У популярному вислові помилково стверджується, що ера великих даних зводить нанівець потребу у вибірці. Насправді швидке поширення даних змінної якості та релевантності зміцнює потребу у вибірці як в інструменті ефективної роботи з різноманітними даними та мінімізації зміщення. Навіть у проєкті на основі великої кількості даних моделі прогнозування зазвичай розробляються й апробовуються за допомогою вибірок. Вибірки також використовуються в найрізноманітніших тестах (наприклад, у ціноутворенні, вебобробленні).

На рис. 3.1 зліва показано *популяцію*, яка в статистиці ймовірно підпорядковується базовому, але невідомому розподілу. Єдине, що є, – це вибірка даних та її емпіричний розподіл (справа). Для того щоб переміститися зліва направо, використовується процедура відбору (позначено стрілкою). Традиційна статистика зосереджена головним чином зліва, використовується теорія, що базується на серйозних припущеннях про характер популяції. Сучасна статистика перемістилася вправо, тут такі припущення не потребуються.

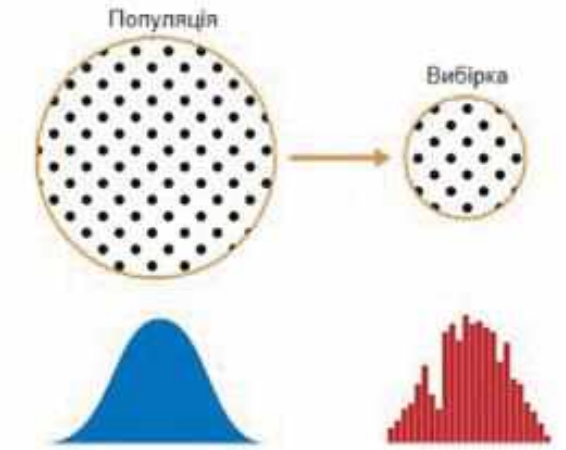


Рис. 3.1. Порівняння популяції (генеральної сукупності) із вибіркою

Загалом аналітики даних можуть не турбуватися про теоретичну природу лівого боку; замість цього їм слід приділяти основну увагу процедурам відбору й наявним даним. Правда, існують деякі суттєві винятки. Іноді дані генеруються з фізичного процесу, який може бути змодельований. Найпростіший приклад – підкидання монети, що підпорядковується біноміальному розподілу. Будь-яку реальну біноміальну ситуацію (купити чи не купити, шахрайство чи не шахрайство, натиснути чи не натиснути) можна ефективно змоделювати з допомогою монети (зрозуміло, з підібраною ймовірністю випадання «орла»). У цих випадках можна отримати додаткову інформацію, використовуючи поняття популяції.

Вибірка – це підмножина даних з великого набору даних, який у статистиці називають популяцією, або генеральною сукупністю. Популяція в статистиці – це не те саме, що популяція в біології, це великий заданий, але нерідко теоретичний або уявний набір даних.

Випадковий відбір – це процес, коли кожен доступний член популяції, що піддається відбору, має однакову можливість потрапити до вибірки при кожному вийманні. Результівна вибірка має назву «проста випадкова вибірка». Відбір може бути виконаний з поверненням, коли після кожного виймання спостереження кладуться назад у популяцію для можливого повторного відбору в майбутньому. Як альтернатива відбір може бути виконаний без повернення, і в цьому випадку один раз вибрані спостереження будуть недоступними для майбутніх виймань.

Якість даних часто має більше значення, ніж їх кількість, коли виконується оцінювання або створюється модель на основі вибірки. Якість даних у науці про дані пов’язана з повнотою, послідовністю формату, чистотою й точністю окремих точок даних. Зі статистики сюди додається поняття **репрезентативності**.

Класичний приклад – опитування, проведене 1936 року журналом «Літературний огляд» («Literary Digest»), коли було передбачено перемогу А. Лендона над Ф. Рузвельтом. Періодичне щоденне видання «Літературний огляд» опитало своїх передплатників, зареєстрованих у базі видання, і додатково інших людей (загалом понад 10 млн чоловік) і спрогнозувало нищівну перемогу А. Лендона. Джордж Геллап, засновник Інституту опитування громадської думки, проводив опитування кожні два тижні всього 2 тис. респондентів і точно спрогнозував перемогу Ф. Рузвельта. Відмінність полягала в тому, як вибиралися респонденти. Журнал «Літературний огляд» зробив ставку на кількість, мало звертаючи увагу на метод відбору, і виявилося, що були опитані люди з відносно високим соціально-економічним статусом (власні передплатники видання і ті, хто входили до списків маркетологів на підставі володіння предметами розкоші, такими як телефони й автомобілі). Результатом стала **зміщена вибірка**, що відрізнялася деяким змістовним невипадковим характером від

іншої, більш численної популяції, яку ця вибірка мала представляти. Термін «невипадковий» дуже важливий: чи будь-яка вибірка, включаючи випадкові вибірки, буде для популяції суворо репрезентативною. Зміщена вибірка виникає, коли відмінність стає змістовною, і очікується, що вона продовжиться відносно інших вибірок, що виймаються, таким же чином, що й перша.

Сьогодні існують різноманітні методи, що дають змогу досягати репрезентативності, але їх основою є **випадковий відбір**.

Випадковий відбір не завжди є простим, і належне визначення доступної популяції є ключем. Припустимо, що хочемо згенерувати репрезентативний профіль покупців, і потрібно провести їх пілотний статистичний огляд. Потребується репрезентативний огляд, але він трудомісткий.

Спочатку необхідно визначити, хто є покупцем. Можна вибрати всі записи покупців з сумою покупки більше 0. Чи варто включати всіх минулих покупців? Чи варто включати компенсації? Внутрішні покупки? Перекупників? Білінгового агента й покупця?

Далі слід визначити процедуру відбору, яка може полягати в тому, щоб «вибрати 100 покупців навмання». Там, де задіяно відбір з потоку (наприклад, транзакції покупців у реальному часі або відвідувачі вебсайту), особливої важливості набувають міркування стосовно часу (наприклад, відвідувач вебсайту о 10:00 у будній день може відрізнятися від відвідувача вебсайту о 22:00 у вихідні). Відомим є приклад похибки телефонного опитування з прогнозування результатів виборів у США, коли громадян США опитували з офісу в Каліфорнії, а на той час у Нью-Йорку вже була глибока ніч і люди не брали телефон. Це призвело до систематичної похибки.

У **стратифікованому відборі** популяція поділяється на **страти** і випадкові вибірки беруться з кожної страти. Політичні соціологи можуть спробувати з’ясувати електоральні уподобання білих, афроамериканців і латиноамериканців. Проста випадкова вибірка, узята з населення США, приведе до дуже невеликої кількості афроамериканців і латиноамериканців, і тому в стратифікованому відборі цим стратам може бути надано більшої ваги, щоб отримати еквівалентні розміри вибірок.

В еру великих обсягів даних викликає здивування, що іноді виявляється, що чим менше, тим краще. Час і зусилля, витрачені на випадковий відбір, не тільки зменшують зміщення, а й дають змогу приділяти більше уваги розвідці даних та їх якості. Наприклад, пропущені дані й викиди містять корисну інформацію. Пошук відсутніх значень або обчислення викидів у мільйонах записів може виявитися неприпустимо дорогим, але ця робота у вибірці, що складається з декількох тисяч записів, є цілком здійсненною. Відображення даних на графіках і ручне дослідження – практично безглузді, якщо даних занадто багато.

3.2. Варіаційний ряд. Вибіркові аналоги функції розподілу та щільності розподілу випадкових величин

Приклад 3.2.1. На касі супермаркету досліджували кількість покупців за хвилину. Спостереження протягом 30 хвилин дали такі результати: 1; 2; 4; 2; 2; 1; 3; 4; 0; 3; 0; 2; 2; 0; 2; 1; 4; 3; 3; 2; 1; 4; 1; 3; 0; 3; 2; 2; 1; 3. Знайти обсяг вибірки, її розмах. Скласти варіаційний і статистичний (дискретний варіаційний) ряди, знайти медіану.

Розв'язання. Розташувавши наведені вище дані в порядку неспадання (з усіма повтореннями), одержимо **варіаційний** (ранжований) ряд даних:

0; 0; 0; 0; 1; 1; 1; 1; 1; 1; 2; 2; 2; 2; 2; 2; 2; 2; 2; 3; 3; 3; 3; 3; 3; 3; 4; 4; 4; 4.

Розмах вибірки – це різниця між максимальним і мінімальним елементами вибірки: $R = 4 - 0 = 4$. У вибірці обсягом $n\,(n = 30)$ елемент 0 трапляється чотири рази $(n_1 = 4)$; елемент 1 – шість разів $(n_2 = 6)$ і т. д. Число n_i називають частотою елемента. Очевидно, що $\sum n_i = n$.

Статистичний (дискретний варіаційний) ряд запишемо у вигляді таблиці, перший рядок якої містить елементи вибірки x_i, а другий – їх частоти:

x_i	0	1	2	3	4
n_i	4	6	9	7	4

Якщо складено варіаційний ряд, то легко знайти **оцінку медіани** (вибірковий аналог медіани) $m\tilde{e}d$ – число, що ділить варіаційний ряд на дві частини, які містять однакову кількість елементів. Якщо обсяг вибірки $n = 2k + 1$ – непарне число, то $m\tilde{e}d = x_{k+1}$ (у варіаційному ряді!). Якщо $n = 2k$, то $m\tilde{e}d$ дорівнює півсумі середніх елементів. У цьому прикладі

$$m\tilde{e}d = \frac{x_{15} + x_{16}}{2} = \frac{2+2}{2} = 2.$$

Графічне зображення рядів за допомогою гістограм дає змогу отримати перше уявлення про щільності розподілу випадкових величин, що спостерігаються.

Для побудови гістограми в прямокутній системі координат на осі Ox відкладають відрізки завдовжки ℓ і на цих відрізках як на основах будують прямокутники з висотами $h_i = \frac{n_i}{n\ell}$, унаслідок чого одержують ступінчасту фігуру, що складається з прямокутників. Ширина всіх прямокутників є однаковою.

Приклад 3.2.2. Задано вибірку 1; 2; 2; 2; 1; 0; –1; 1; 3; 0; 1; 1.

Побудувати гістограму.

Розв'язання. Будуємо гістограму (рис. 3.2). Обсяг вибірки – $n=12$. Розіб'ємо інтервал $(-1,5;\ 3,5)$ точками $x_1=-1,5$, $x_2=-0,5$, $x_3=0,5$, $x_4=1,5$, $x_5=2,5$, $x_6=3,5$ на окремі інтервали завдовжки $\ell=1$. Висоти гістограми знайдемо за формулою $h_i=\dfrac{n_i}{n\ell}$, де n_i – кількість елементів вибірки, що потрапляють в i-й інтервал. Тоді $h_1=\dfrac{1}{12}$, $h_2=\dfrac{2}{12}$, $h_3=\dfrac{5}{12}$, $h_4=\dfrac{3}{12}$, $h_5=\dfrac{1}{12}$. Зазначимо, що площа гістограми дорівнює 1:

$$S=\ell\frac{n_1}{n\ell}+\ell\frac{n_2}{n\ell}+...+\ell\frac{n_k}{n\ell}=\frac{n_1+n_2+...+n_k}{n}=\frac{n}{n}=1.$$

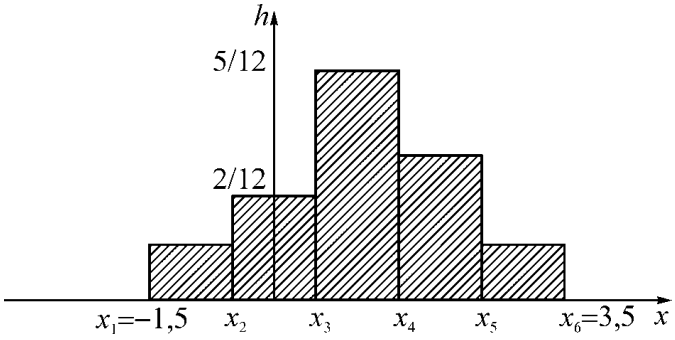


Рис. 3.2. Гістограма

Гістограма – це спосіб візуалізації частотної таблиці, де частотні інтервали відкладають на осі x, а кількість даних або відносну кількість (аналог ймовірності) – на осі y.

Зрозуміло, що вигляд гістограми залежить від кількості стовпців. Якщо взяти кількість стовпців занадто великою, то на гістограмі виникнуть пропуски. А через занадто малу кількість стовпців неможливо відобразити особливості розподілу даних. Тому рекомендується варіювати кількість стовпців залежно від кількості даних і від їх розподілу за величиною.

Емпірична (вибіркова) функція розподілу $F^*(x)=\dfrac{n_x}{n}$, де n – обсяг вибірки, а n_x – кількість вибіркових значень, менших за x. На відміну від вибіркової функції розподілу $F^*(x)$ функцію розподілу $F(x)$ генеральної сукупності називають теоретичною функцією розподілу. Відмінність функцій $F(x)$ та $F^*(x)$ полягає в тому, що теоретична функція розподілу визначає ймовірність події $\xi<x$ (де ξ – випадкова величина), а вибіркова – відносну частоту цієї події. При великих значеннях n функцію $F^*(x)$ можна використовувати як наближене значення функції $F(x)$. Властивості емпіричної функції розподілу:

– $0 \le F^*(x) \le 1$;

– $F^*(x)$ – неспадна;

– $F^*(-\infty) = 0$; $F^*(+\infty) = 1$.

Приклад 3.2.3. За посиланням http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv можна скачати файл з даними про 1599 вин, який містить 12 записів у кожному рядку – характеристики вина за 11 критеріями й середню оцінку цього вина, яку поставили дегустатори. Тобто записи (рядки) містять числові характеристики *fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, quality.*

Побудуємо гістограми розподілу певних ознак цих 1599 вин (рис. 3.3). Почнемо з якості вина (*quality*). Покажемо, як впливає на вигляд гістограми кількість стовпців у ній.

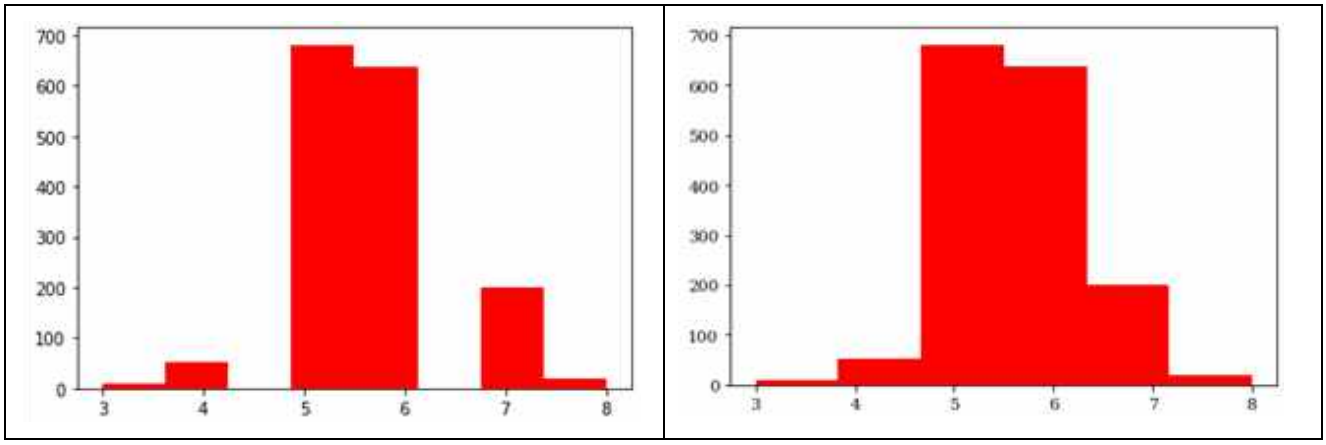


Рис. 3.3. Гістограма оцінок вина (8 та 6 стовпців)

Гістограми розподілу за деякими ознаками показано на рис. 3.4–3.8.

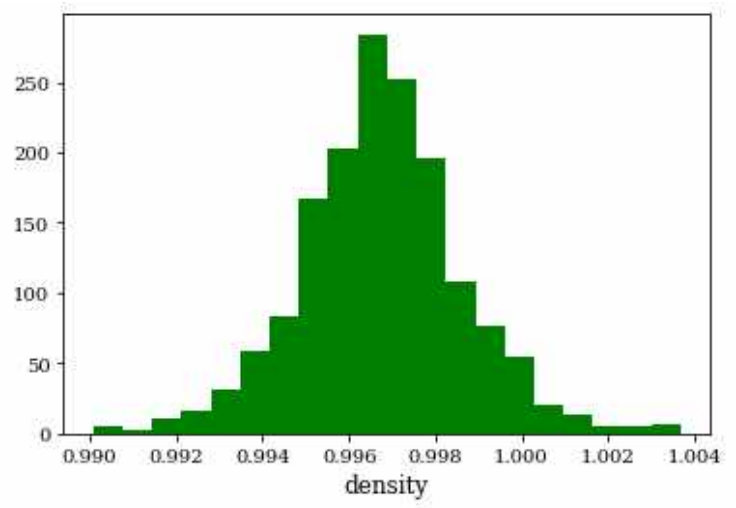


Рис. 3.4. Розподіл *density*

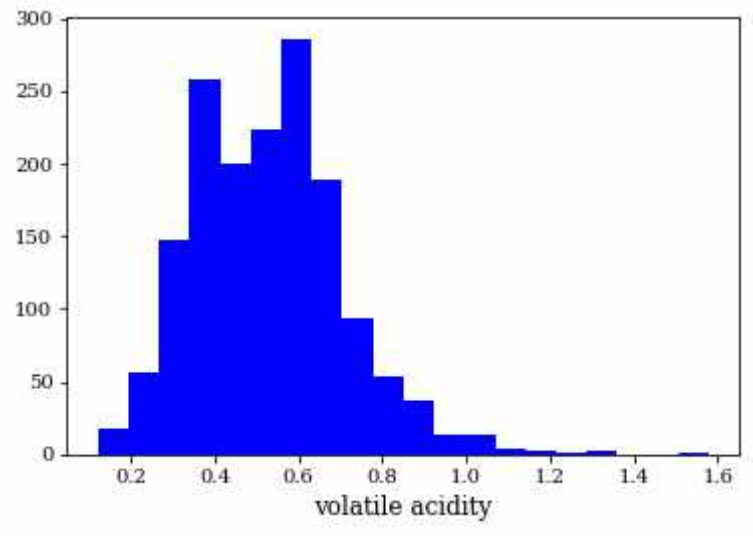


Рис. 3.5. Розподіл *volatile acidity*

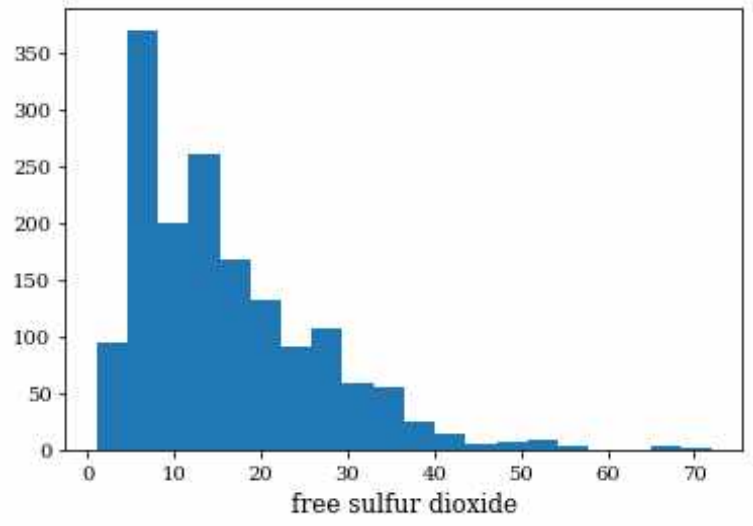


Рис. 3.6. Розподіл *free sulfur dioxide*

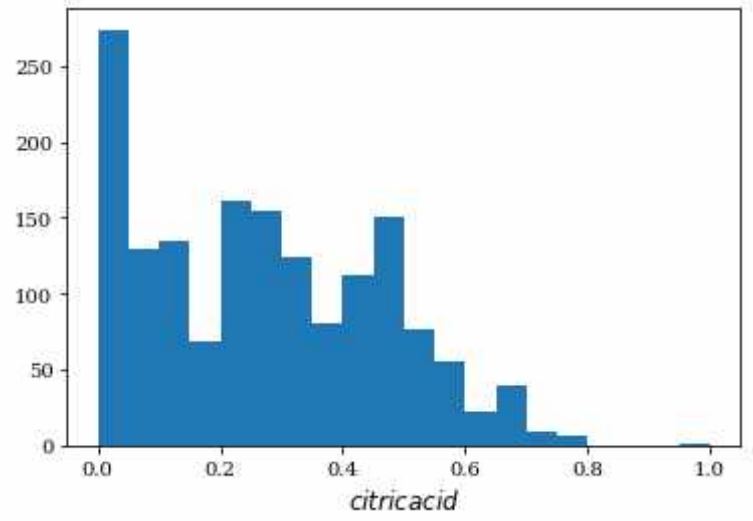


Рис. 3.7. Розподіл *citric acid*

При їх побудові було застосовано 20 стовпців. За командою *pylab.hist* (Python) автоматично розраховуються ширина інтервалу зміни ознаки й кількість потрапляння ознаки в заданий інтервал (кількість інтервалів задається).

Як бачимо, деякі ознаки мають розподіл, близький до розподілу Гаусса, а деякі – значно відрізняються від нормального розподілу. Деякі характеристики містять викиди.

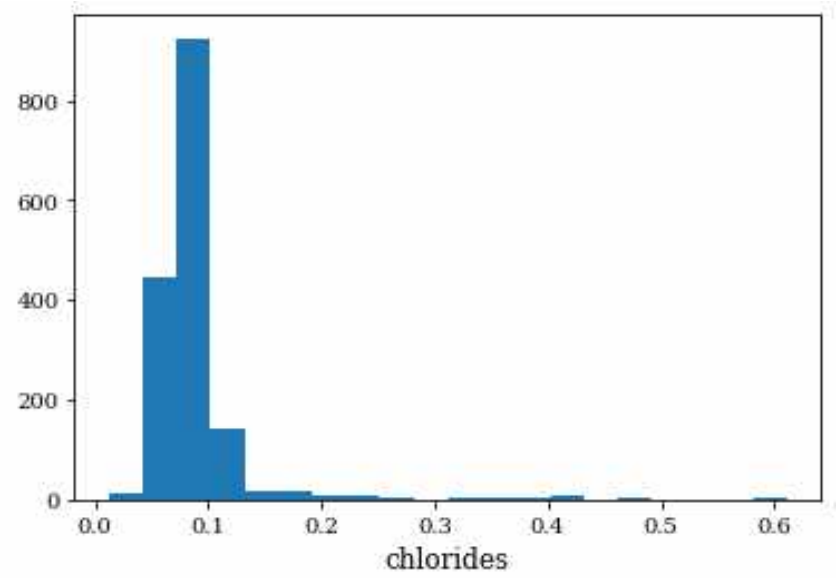


Рис. 3.8. Розподіл *chlorides*

Ще один інструмент аналізу – діаграми розсіювання, які дають змогу отримати розподіл двовимірної вибірки на площині. Покажемо деякі з них (рис. 3.9).

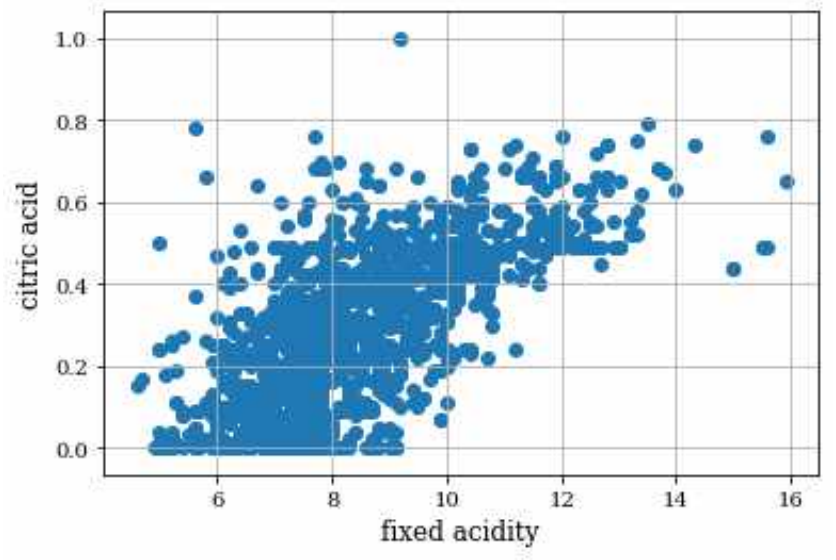
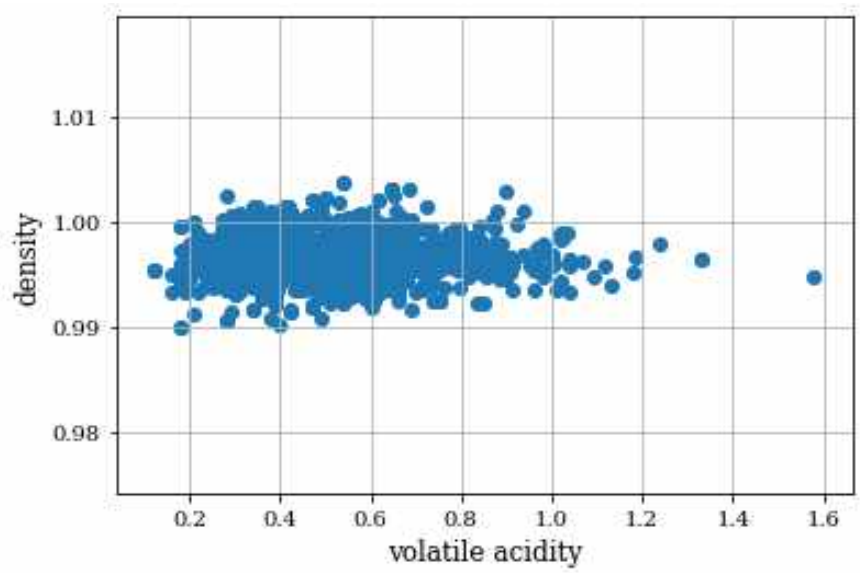
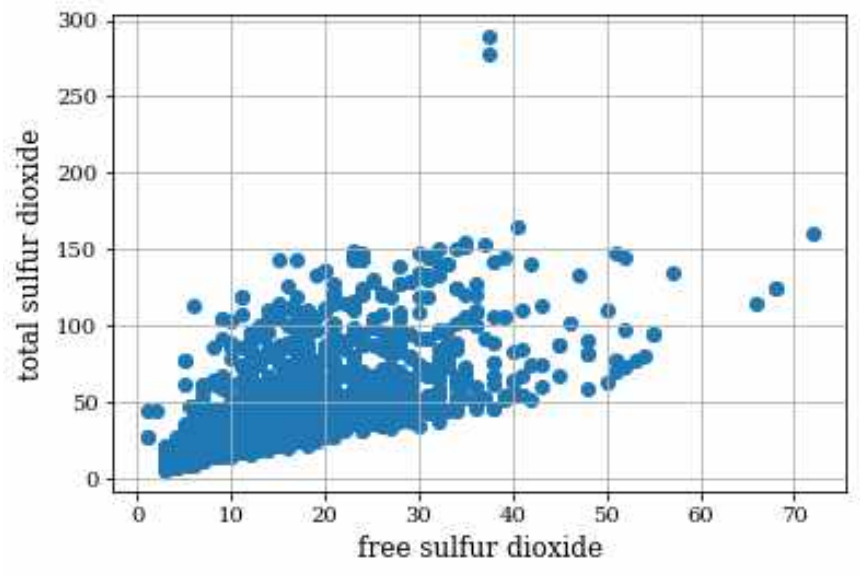


Рис. 3.9. Діаграми розсіювання





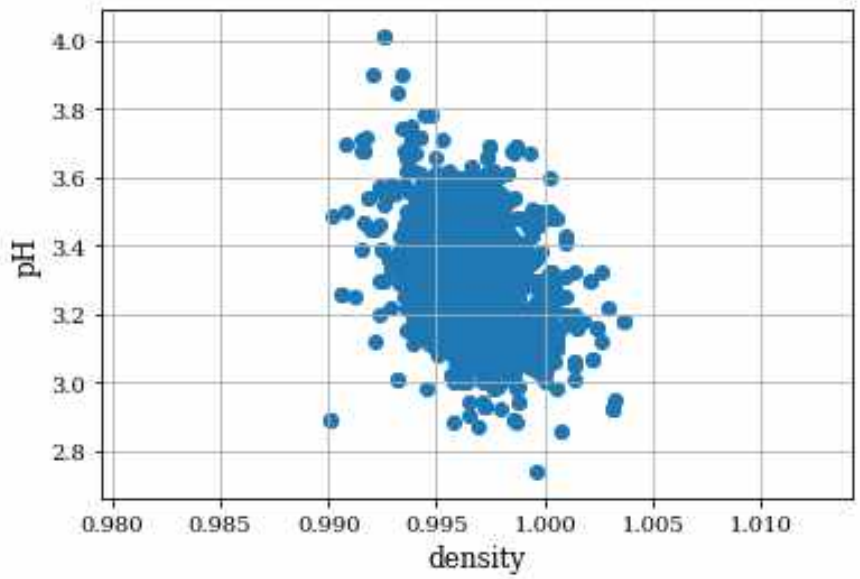


Рис. 3.9. Продовження

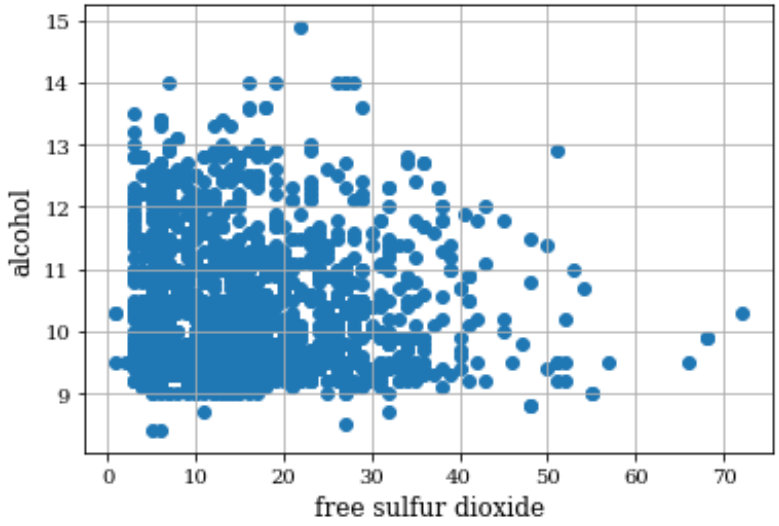


Рис. 3.9. Закінчення

Розглянувши ці діаграми, бачимо, що ознаки розсіюються за певними залежностями. Збільшення однієї ознаки приводить або до збільшення іншої (*fixed acidity – citric acid*), або до її зменшення (*free sulfur dioxide – alcohol*), або взагалі не впливає (*volatile acidity – density*)! Якщо обміркувати це, то ця картина є цілком логічною – збільшення вмісту лимонної кислоти у вині збільшує його загальну кислотність, збільшення вмісту алкоголю дає змогу не закріпляти вино від скисання за допомогою діоксиду сірки тощо. Це явище називають **кореляцією** ознак.

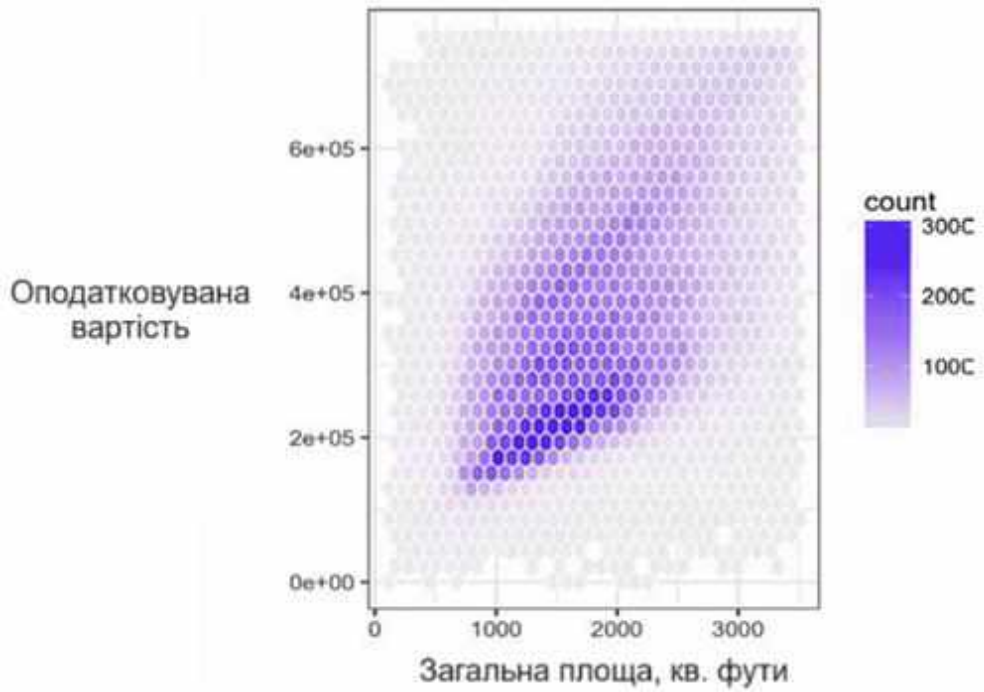


Рис. 3.10. Графік з шестикутною сіткою, що відображає залежність оподатковуваної вартості від загальної площі

Зауваження. Якщо точок на діаграми дуже багато, наприклад мільйон або мільярд, то вони будуть накладатися одна на одну, і картина розподілу буде незрозумілою. Вирішенням цього питання є застосування **теплових карт**: область значень поділяється у вигляді сітки (прямокутної або шестигранної), підраховується кількість потраплянь точок розподілу в кожний елемент сітки, після цього сітка розфарбовується, причому інтенсивність кольору є пропорційною до кількості потраплянь.

3.3. Точкове оцінювання параметрів випадкової величини. Властивості точкових оцінок

Нехай $x_1, x_2, \ldots, x_n$ – вибірка.

Величини $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ і $\tilde{D}x = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$ називають **точковими оцінками** (статистиками) математичного сподівання й дисперсії. Термін «точкова» означає, що оцінка являє собою число (точку на числовій осі). Оцінку математичного сподівання $\bar{x}$ – середнє арифметичне – позначають також $\bar{m}$ або $\bar{M}$.

Зазвичай необхідно мати якесь уявлення про центрування даних. Найчастіше для цього використовується середнє (або середнє арифметичне) значення, яке визначається як сума даних, поділена на їх кількість:

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

Ця величина є певним аналогом **математичного сподівання**, а відмінність полягає в тому, що математичне сподівання є фіксованою величиною, а **вибіркове середнє** – випадковою величиною, що залежить від вибірки. Зрозуміло, що якщо взяти іншу вибірку (навіть з наявної вибірки), то й вибіркове середнє, скоріш за все, також зміниться.

Вибіркове середнє – не єдина можлива характеристика середнього значення вибірки. Вибіркове середнє має таку негативну властивість, як чутливість до викидів. Наявність викидів – малої кількості значень статистичного ряду, що суттєво відрізняються від більшості елементів, – впливає на значення вибіркового середнього. Отже, вибіркове середнє не завжди може бути якісною характеристикою середнього значення вибірки.

Медіана є більш стійкою до викидів (тобто є робастною величиною). Медіана – це таке значення, при якому дві половини відсортованих даних знаходяться вище й нижче конкретного значення (50-й процентіль, інше кажучи). Для розрахунку медіани треба відсортувати значення

статистичного ряду, а це при великих обсягах може бути трудомістким процесом.

Різновидом середнього є **середнє зрізане**, яке обчислюється шляхом відкидання фіксованої кількості відсортованих значень з кожного кінця послідовності і визначення середнього арифметичного решти значень:

$$\overline{x} = \frac{\sum_{i=p+1}^{n-p} x_i}{n-2p},$$

де p – кількість відкинутих зліва і справа значень варіаційного ряду.

Середнє зрізане усуває вплив граничних значень. Наприклад, у міжнародних змаганнях зі стрибків у воду верхні й нижні бали п'яти суддів відкидаються, і підсумковим балом є середньоарифметичний бал трьох інших суддів. Такий підхід унеможливлює маніпуляції балом судді, можливо, щоб посприяти спортсмену зі своєї країни.

Середнє зрізане може вважатися компромісом між медіаною й середнім: воно є стійким до граничних значень у даних, але використовує більше даних для обчислення оцінки центрального положення.

Приклад 3.3.1. Розрахунок оцінок середньої величини ознаки *chlorides*, XC[4] з прикладу 3.2.3 (див. рис. 3.8) за допомогою Python.

Розв'язання. Виконаємо розрахунки:

- **вибіркове середнє**: np.mean(XC[4]) = 0.08746654158849279;
- **медіана**: np.median(XC[4]) = 0.079;
- **середнє зрізане**: stats.trim_mean(XC[4], 0.1) = 0.080234972677595.

Як бачимо, якщо застосувати 80 % вибірки, відкинувши по 10 % зліва й справа від неї, то отримаємо середнє значення, дуже близьке до медіани! Водночас вибіркове середнє через викиди зміщується в напрямку збільшення.

Центральне положення – це одна з розмірностей в узагальненні ознаки. Інша розмірність – варіабельність, або дисперсність – показує, значення даних щільно згруповані чи розкидані. Основою статистики є варіабельність: її вимірювання, зменшення, розрізнення довільної і реальної варіабельності, ідентифікація різних джерел реальної **варіабельності** й прийняття рішень в умовах її наявності.

Один зі способів вимірювання **варіабельності** полягає в оцінюванні типового значення відхилень. Усереднення самих відхилень замало, тому що від'ємні відхилення нейтралізують додатні. Фактично сума відхилень від середнього дорівнює нулю. Замість цього можна застосувати простий підхід – узяти середнє абсолютних значень відхилень від середнього значення

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

де $\bar{x}$ – середнє значення у вибірці, або вибіркове середнє.

Найвідомішими оцінками варіабельності є **дисперсія** і **стандартне відхилення**, що ґрунтуються на квадратичних відхиленнях. Дисперсія – це середнє квадратичних відхилень, а стандартне відхилення (s) – квадратний корінь з дисперсії:

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

Стандартне відхилення інтерпретується набагато простіше, ніж дисперсія, оскільки має таку саму одиницю вимірювання, що й вихідні дані. Однак, з огляду на його більш складну й інтуїтивно менш зрозумілу формулу, може здатися дивним, що в статистиці стандартному відхиленню віддається перевага порівняно із середнім абсолютним відхиленням. Це обумовлено статистичною теорією: математично робота з квадратичними значеннями є набагато зручнішою, ніж з абсолютними, особливо зі статистичними моделями.

Зауваження. У книгах зі статистики завжди так чи інакше обговорюється питання, чому у формулі дисперсії в знаменнику $n-1$ замість n, що приводить до поняття ступенів свободи. Ця відмінність не є важливою, оскільки значення n зазвичай є настільки великим, що вже не має значення, як буде виконуватися розподіл – на n чи $n-1$. Пояснення цього ґрунтується на передумові, що необхідно отримати оцінки популяції виходячи з вийнятої з неї вибірки. Якщо у формулі дисперсії застосувати інтуїтивно зрозумілий знаменник n, то істинні значення дисперсії і стандартного відхилення в популяції будуть недооціненими. Це називають зміщеною оцінкою. Однак, якщо поділити на $n-1$ замість n, то стандартне відхилення стає незміщеною оцінкою. Повне пояснення, чому використання n приводить до зміщеної оцінки, пов'язане з поняттям ступенів свободи, коли до уваги береться кількість обмежень під час обчислення оцінки. У цьому випадку існують $n-1$ ступенів свободи, оскільки існує одне обмеження: стандартне відхилення залежить від обчислення середнього у вибірці. Детальніше: коли розраховують вибіркове середнє, усі доданки в сумі є незалежними, а коли розраховують доданки у формулі дисперсії, лише $n-1$ доданків є незалежними, оскільки доданок n можна розрахувати, знаючи $n-1$ членів x_i і вибіркове середнє $\bar{x}$. У більшості завдань аналітикам даних не потрібно турбуватися з приводу ступенів свободи, але в окремих випадках це поняття має

особливе значення.

Дисперсія, стандартне відхилення, середнє абсолютне відхилення й медіанне абсолютне відхилення від медіани не є еквівалентними оцінками, навіть у випадку, коли дані надходять з нормального розподілу. У реальності стандартне відхилення завжди є більшим від середнього абсолютного відхилення, яке, своєю чергою, є більшим від **медіанного абсолютного відхилення (MAB).**

MAB – це медіана статистичного ряду

$$|x_1 - m|, |x_2 - m|, ..., |x_N - m|,$$

де m – медіана ряду $x_1, x_2, ..., x_N$.

Іноді медіанне абсолютне відхилення помножують на постійний поправковий коефіцієнт (який часто зводиться до 1,4826), щоб у разі нормального розподілу звести медіанне абсолютне відхилення до тієї самої шкали вимірювання, що й стандартне відхилення.

Для оцінювання параметрів (одержання точкових оцінок) за вибірковими даними найчастіше використовують метод максимальної правдоподібності (ММП). Цей метод може бути застосований у випадках, коли вигляд функції розподілу є відомим, а значення параметрів, що входять у цю модель, є невідомими. Оцінки невідомих параметрів у цьому випадку дорівнюють значенням, при яких одержана вибірка має максимальну ймовірність появи, тобто як оцінки знаходять значення, що максимізують функцію максимальної правдоподібності.

Приклад 3.3.2. Задано інтервальний варіаційний ряд (згрупована вибірка):

Інтервали	1…3	3…5	5…7	7…9	9…11	11…13
Частоти n_k	1	2	4	2	1	1

Знайти точкові оцінки математичного сподівання й дисперсії.

Розв'язання. Якщо вибірка є згрупованою, то оцінку математичного сподівання знаходять за формулою $\overline{x} = \frac{\sum_{k=1} x_k^* n_k}{n}$, де x_k^* – середина k-го інтервалу:

$$\overline{x} = \frac{2 \cdot 1 + 4 \cdot 2 + 6 \cdot 4 + 8 \cdot 2 + 10 \cdot 1 + 12 \cdot 1}{11} = \frac{72}{11}.$$

Відповідно

$$D^* = \frac{1}{n-1} \sum_{k=1} \left(x_k^* - \overline{x} \right)^2 n_k = \frac{1}{10} \left(\left(2 - \frac{72}{11} \right)^2 + 2 \left(4 - \frac{72}{11} \right)^2 + 4 \left(6 - \frac{72}{11} \right)^2 + \right.$$

$$+2\left(8-\frac{72}{11}\right)^2+\left(10-\frac{72}{11}\right)^2+\left(12-\frac{72}{11}\right)^2\approx 8{,}073.$$

3.4. Інтервальне оцінювання математичного сподівання

Точкове оцінювання вибіркового середнього дає оцінку математичного сподівання невідомого розподілу, що досліджується. Але вибіркове середнє є випадковою величиною

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

тобто, якщо взяти іншу вибірку, то майже достовірно можна отримати інше значення вибіркового середнього. Точність оцінки випадкової величини визначається її **дисперсією**. Як же оцінити дисперсію розподілу нової випадкової величини – вибіркового середнього?

Відповідь на це непросте запитання дає **центральна гранична теорема**, згідно з якою, якщо випадкова величина складається з великої кількості рівнозначних випадкових величин з однаковим розподілом, то її розподіл наближається до розподілу Гаусса. Вибіркове середнє саме так і визначається: як сума незалежних випадкових величин (елементів вибірки), яку поділено на кількість елементів.

Із властивості дисперсії

$$D\left[\frac{x_1+x_2+\ldots+x_n}{n}\right]=D\left[\frac{x_1}{n}+\frac{x_2}{n}+\ldots\frac{x_n}{n}\right]=\underbrace{\frac{D[x]}{n^2}+\ldots\frac{D[x]}{n^2}}_{n}=n\frac{D[x]}{n^2}=\frac{D[x]}{n}$$

випливає, що дисперсія вибіркового середнього

$$D_M=\frac{D}{n}.$$

Якщо дисперсія розподілу є невідомою, то для її оцінювання можна брати дисперсію вибірки

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

Якщо вибірка є достатньо великою (сотні й тисячі елементів), то можна вважати, що дисперсія вибіркового середнього $\overline{D}=\frac{s^2}{n}$, а сам розподіл є розподілом Гаусса. Ця гіпотеза дає можливість увести таке

поняття, як **довірчий інтервал** – інтервал значень, у який з певною заданою ймовірністю потрапляє досліджувана випадкова величина.

У випадку математичного сподівання досліджуваного розподілу маємо

$$P\left(\theta_1 \le M \le \theta_2\right) = \alpha,$$

де M – математичне сподівання; θ_1, θ_2 – межі довірчого інтервалу; α – довірча ймовірність, яку або задають самостійно, або беруть з технічного завдання дослідження.

Якщо дисперсія є відомою, а випадкова величина – вибіркове середнє – підпорядковується розподілу Гаусса, то можна ввести нову випадкову величину

$$\frac{x - \overline{x}}{\sqrt{D_M}},$$

яка має розподіл Гаусса з нульовим математичним сподіванням та одиничною дисперсією (рис. 3.11).

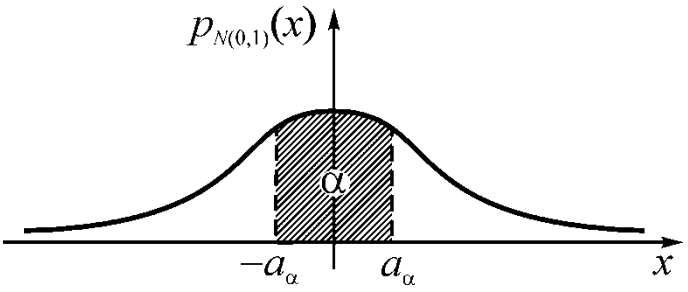


Рис. 3.11. Довірчий інтервал

У цьому випадку довірчий інтервал зазвичай має вигляд $\left[-a_\alpha, a_\alpha\right]$. Безумовно, можна побудувати й несиметричний інтервал, але він буде ширшим за симетричний. Кожен з двох «хвостів» має площу $\frac{1}{2}(1-\alpha)$. Величину a_α можна знайти або з таблиць функції Лапласа, або шляхом застосування будь-яких математичних програм, мов програмування тощо. Визначивши a_α із формули

$$P\left(-a_\alpha \le \frac{x - \overline{x}}{\sqrt{D_M}} \le a_\alpha\right) = \alpha,$$

знаходимо вибірковий інтервал для математичного сподівання:

$$\overline{x} - a_\alpha\sqrt{D_M} < M < \overline{x} + a_\alpha\sqrt{D_M}\ .$$

Оскільки $D_M = \sqrt{\frac{D}{n}}$, отримуємо

$$\bar{x} - a_\alpha \sqrt{\frac{D}{n}} < M < \bar{x} + a_\alpha \sqrt{\frac{D}{n}},$$

де D – дисперсія випадкової величини x_i, яка за умов великої вибірки наближається до вибіркової дисперсії: $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$.

Зауваження. При такому підході для визначення довірчого інтервалу математичного сподівання за вибіркою дисперсію вважали відомою й достовірно визначеною за вибіркою. Але це можливо лише для дуже великих вибірок. Якщо вибірка є невеликою, то той факт, що дисперсію визначено неточно, а лише за даними вибірки, і вона являє собою лише точкову оцінку, потребує суттєвих коректив у роботі.

Вільям Госсет (1876–1937) з'ясував, що наведена вище оцінка довірчого інтервалу у випадку малих вибірок дає занадто оптимістичні результати. Працюючи після закінчення університету 1899 року на пивоварній компанії Arthur Guinness Son & Co у Дубліні, він застосував свої знання в області статистики як при варінні пива, так і на полях – для виведення найурожайнішого сорту ячменю. В. Госсет здобув ці знання шляхом вивчення, методом проб і помилок, провівши два роки (1906–1907 рр.) у біометричній лабораторії Карла Пірсона. В. Госсет і К. Пірсон мали добрі стосунки, і К. Пірсон допомагав В. Госсету в математичній частині його досліджень. Так, К. Пірсон був причетний до публікацій 1908 року (що прославили Стьюдента), але надавав мало значення цьому відкриттю. Дослідження були спрямовані на потреби пивоварної компанії й проводилися з малою кількістю спостережень. Гіннес заборонив своїм працівникам публікацію будь-яких матеріалів незалежно від інформації, що містилася в них. Це означало, що В. Госсет не міг опублікувати свої роботи під своїм ім'ям, тому вибрав собі псевдонім Стьюдент, щоб приховати себе від роботодавця. Його найважливіше відкриття отримало назву **«розподіл Стьюдента»**.

У теорії ймовірностей і статистиці t-розподіл (або t-розподіл Стьюдента) – різновид розподілу ймовірностей, що виникає в задачі оцінювання прогнозованого значення нормально розподіленої популяції, коли розмір вибірки є малим. Цей розподіл є основою популярного t-тесту Стьюдента статистичної значущості різниці математичних сподівань двох вибірок і довірчого інтервалу різниці прогнозованих значень двох вибірок.

Зазвичай розглядають центрований розподіл Стьюдента з нульовим математичним сподіванням. Функція щільності ймовірності розподілу

Стьюдента залежить також від дискретної величини – кількості **ступенів свободи** k: $f(k,x)$. При великих значеннях k (порядку 100 і більше) розподіл Стьюдента наближається до розподілу Гаусса. Але при малих значеннях k розподіл Стьюдента має більш важкі «хвости» (рис. 3.12).

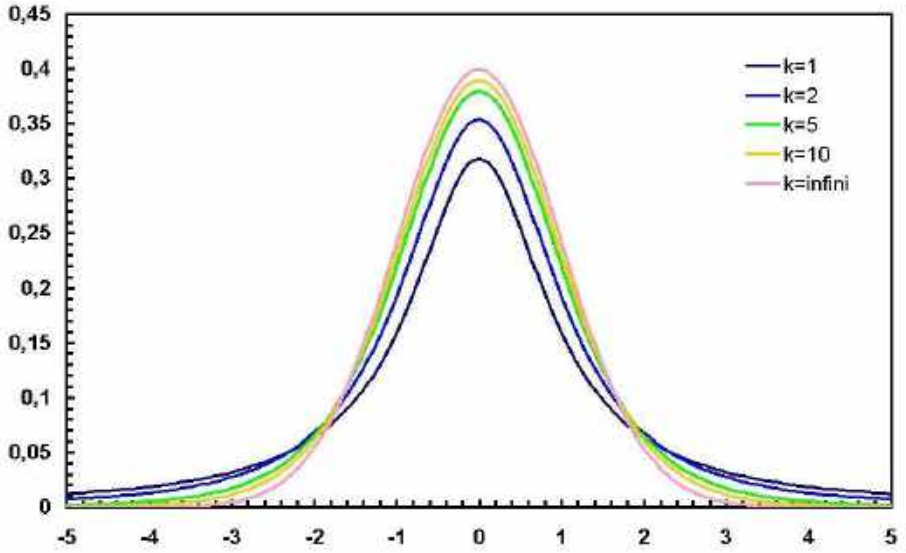


Рис. 3.12. Розподіл Стьюдента

Розподіл Стьюдента визначається складною формулою, у яку входить спеціальна функція (гамма-функція), і інтеграл від неї в елементарних функціях не береться. Тому ця функція й інтеграл від неї є табульованими і входять до всіх математичних програм та програм і мов роботи з даними (R, Python).

Довірчий інтервал для математичного сподівання нормального розподілу при невідомій дисперсії має вигляд

$$\bar{x} - t_{\alpha,n-1}\sqrt{\frac{D^*}{n}} < M < \bar{x} + t_{\alpha,n-1}\sqrt{\frac{D^*}{n}},$$

де $D^* = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ – незміщена оцінка дисперсії; $t_{\alpha,n-1}$ – число, знайдене з таблиці розподілу Стьюдента і таке, що

$$P\left(\left|\frac{\bar{x} - M}{\sqrt{\frac{D^*}{n}}}\right| < t_{\alpha,n-1}\right) = \alpha.$$

Величина $\frac{\overline{x}-M}{\sqrt{\frac{D^*}{n}}}$ має розподіл Стьюдента (t-розподіл) з $(n-1)$ ступенями свободи.

Приклад 3.4.1. За результатами спостережень випадкової величини складено дискретний варіаційний ряд (x_i – значення випадкової величини, n_i – кількість появ величини у вибірці):

x_i	9,4	9,8	10	10,6	11	11,4	12	12,4	$\alpha = 0,2$
n_i	3	5	5	6	6	6	4	1	

1. Знайти обсяг вибірки та її розмах.
2. Скласти інтервальний варіаційний ряд й побудувати гістограму.
3. Знайти точкові оцінки математичного сподівання й дисперсії.
4. Уважаючи, що генеральна сукупність розподілена за нормальним законом, знайти довірчі інтервали для математичного сподівання з довірчою ймовірністю $1-\alpha$.

Розв'язання:

1. Обсяг вибірки $n = \sum n_i = 36$; розмах $R = \max(x_i) - \min(x_i) = 3$.

2. Розіб'ємо інтервал (9,3; 12,4) на п'ять однакових інтервалів завдовжки $l = 0,62$ і складемо інтервальний варіаційний ряд. Висоту стовпців гістограми обчислюємо за формулою $h_i = \frac{n_i}{nl}$.

Кількість потраплянь n_i випадкової величини в ці інтервали:

$9,3 < x \le 9,92$	8
$9,92 < x \le 10,54$	5
$10,54 < x \le 11,16$	12
$11,16 < x \le 11,78$	6
$11,78 < x \le 12,4$	5

Гістограму зображено на рис. 3.13.

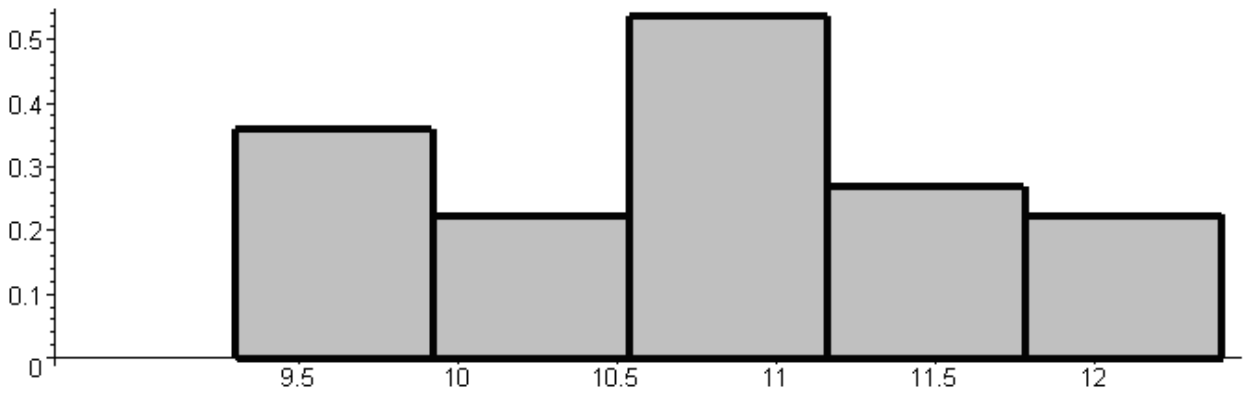


Рис. 3.13. Гістограма розподілу

3. Точкова оцінка математичного сподівання $\overline{x} = \frac{1}{n}\sum x_i n_i = 10,71$.

Незміщена оцінка дисперсії $s^2 = D^* = \frac{1}{n-1}\sum \left(x_i - \overline{x}\right)^2 n_i = 0,7073$.

4. Довірчий інтервал для математичного сподівання:

$$\overline{x} - t_{\alpha,n-1}\sqrt{\frac{D^*}{n}} < M < \overline{x} + t_{\alpha,n-1}\sqrt{\frac{D^*}{n}},$$

$$10,71 - t_{0,9;35}\sqrt{\frac{0,7073}{36}} < M < 10,71 + t_{0,9;35}\sqrt{\frac{0,7073}{36}}.$$

Тут $t_{0,9;35} = 1,3062$ – квантиль розподілу Стьюдента (t-розподілу), який має $n-1=35$ ступенів свободи. Пояснимо значення 0,9 і чому воно відрізняється від 0,8, заданого за умовою задачі (рис. 3.14).

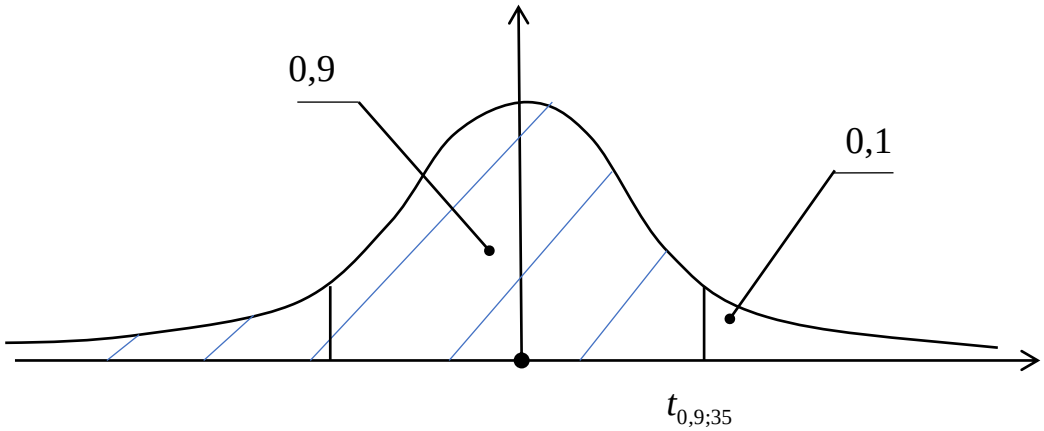


Рис. 3.14. Розподіл Стьюдента

Площа кожного «хвоста» дорівнює 0,1, площа інтервалу $\left[-t_{0,9;35};\ t_{0,9;35}\right]=0,8=\alpha$, а площа лівої частини без одного правого «хвоста» відповідно 0,9.

Таким чином, отримуємо

$$10,71-1,3062\sqrt{\frac{0,7073}{36}}<M<10,71+1,3062\sqrt{\frac{0,7073}{36}},$$

$$10,52802<M<10,8942.$$

Висновок: у цей інтервал математичне сподівання M генеральної сукупності (популяції) потрапляє з імовірністю 0,8.

Зауваження. Як бачимо, у формулі довірчого інтервалу при фіксованій вибірковій дисперсії $s^2=D^*$ зі збільшенням обсягу вибірки n довірчий інтервал зменшується. Це зумовлено насамперед тим, що під коренем у знаменнику стоїть n. Крім того, зі збільшенням n дещо зменшується величина $t_{\alpha,n-1}$, при великих значеннях n ця величина стабілізується. Але зі збільшенням довірчої ймовірності α величина $t_{\alpha,n-1}$ стрімко збільшується. Якщо взяти $\alpha=1$, то довірчий інтервал буде таким: $\left(-\infty;+\infty\right)$. Це зрозуміло – завжди існує ймовірність того, що випадкова величина, розподіл якої має «хвости», набуде екстремального значення, і щоб охопити цю малу ймовірність, потрібно розширювати довірчий інтервал. Тому зазвичай оперують зі значенням α, які дорівнюють 0,75; 0,9; 0,95; 0,99 тощо.

Значення $t_{\alpha,n-1}$ у різних програмах розраховується за різними принципами й потребує вказування або значень площ кожного з «хвостів», або площі одного «хвоста», або довірчої імовірності й умови симетрії, і тому треба читати мануал до певної функції розрахунку t-тесту. У мові програмування Python синтаксис має такий вигляд:

stats.t(df=n).ppf((alpha1, alpha2))
df=n – кількість ступенів свободи;

(alpha1, alpha2) – кортеж, де alpha1 і alpha2 – це ймовірність $\int_{-\infty}^{a}$, у наведеній вище задачі (0,1; 0,9), де 0,1 – площа лівого «хвоста», а 0,9 – площа від мінус нескінченності до початку правого «хвоста»;

stats.t(df=35).ppf((0.1, 0.9)) = array([-1.3062118, 1.3062118]).

3.5. Метод Bootstrap

Один з простих та ефективних способів оцінювання вибіркового розподілу статистики або модельних параметрів полягає в тому, щоб

брати додаткові вибірки з поверненням із самої вибірки й повторно обчислювати статистику або модель для кожної повторної вибірки. Цю процедуру називають бутстрапом (від англ. Bootstrap – розкручування, самонастроювання). Бутстрап не пов'язаний з якими-небудь припущеннями про нормальний розподіл даних або вибіркової статистики.

Бутстрапівська вибірка (bootstrap sample) – вибірка, узята з **поверненням** з набору спостережуваних даних.

Процес бутстрапування можна концептуально подати як повторення вихідної вибірки тисячі або мільйони разів з тим, щоб отримати гіпотетичну популяцію, яка втілює все знання, виходячи з оригінальної вибірки (вона просто є більшою). Потім з цієї гіпотетичної популяції можна виймати вибірки для оцінювання вибіркового розподілу.

На практиці немає необхідності фактично повторювати вибірку величезну кількість разів. Просто після кожного виймання кожне спостереження повертається назад, тобто виконується відбір з поверненням. Тим самим ефективно створюється нескінченна популяція, у якій імовірність елемента, що виймається, є незмінною від виймання до виймання.

Ключові ідеї для бутстрапування:

- бутстрап (відбір з набору даних з поверненням) є потужним інструментом для визначення варіабельності вибіркової статистики;
- бутстрап може застосовуватися однаково в різних обставинах без загального аналізу математичних наближень вибіркових розподілів;
- цей метод також дає змогу виконувати оцінювання вибіркових розподілів для статистик, де математичного наближення не розроблено.

З урахуванням вибірки обсягом n і цільової вибіркової статистики алгоритм для визначення бутстрапівського довірчого інтервалу буде таким:

1. Вийняти вибіркове значення, записати його і повернути назад.
2. Повторити n разів, створивши вибірку того ж обсягу, що й задана.
3. Записати вибіркове середнє.
4. Повторити кроки 1–3 достатньо велику кількість разів R.
5. Для знаходження довірчого інтервалу середнього значення треба:

- ранжувати масив значень вибіркових середніх за збільшенням;

- відкинути від нього зліва і справа потрібну однакову кількість значень, залишивши в центрі $\alpha \cdot 100\%$ вибірки;

- точками відсікання є точки довірчого інтервалу.

Кількість ітерацій R процесу бутстрапування визначається довільно. Чим більше виконується ітерацій, тим точнішою є оцінка стандартної

помилки або довірчого інтервалу. Результатом цієї процедури є бутстрапівський набір вибіркових статистик або оцінних модельних параметрів, які далі можна дослідити, щоб побачити, наскільки вони є мінливими.

Бутстрап не компенсує малого обсягу вибірки, не створює нових даних і не заповнює «дірки» у наявному наборі даних, а просто повідомляє про те, як поведуть себе численні додаткові вибірки, коли вони будуть вийматися з популяції.

Розглянемо застосування методу Bootstrap на прикладі 3.4.1, який було розглянуто раніше:

1. За даними таблиці

x_i	9,4	9,8	10	10,6	11	11,4	12	12,4
n_i	3	5	5	6	6	6	4	1

було створено «бокс», який містив усі 36 даних елементів.

2. Із «боксу» у циклі випадково виймався з поверненням 36 разів один елемент і створювалася чергова вибірка з 36 елементів. Визначалося математичне сподівання чергової вибірки й записувалося у масив.

3. Пункт 2 виконувався 1000 разів, унаслідок чого отримали масив з 1000 значень вибіркового середнього.

4. У створеній вибірці з 1000 елементів відсікались «хвости» крайніх значень (великі й малі), залилось $\alpha \cdot 100\%$ початкового масиву (за умовами задачі відсікалося 10 % ранжованої вибірки зліва і 10 % справа). Межі відсікання є межами довірчого інтервалу.

Для ранжування й відсікання «хвостів» у Python є функція **stats.trimboth(my, tails)**, яка відсікає з обох боків неранжованого масиву **my** долю **tails** (**0<tails<0.5**) і повертає нову вибірку меншої розмірності.

Зрозуміло, якщо кілька разів виконати цю процедуру, то буде отримано різні вибірки і, відповідно, різні вибірки математичного сподівання, унаслідок чого межі довірчого інтервалу будуть дещо іншими.

На рис. 3.15 показано результати багаторазового застосування алгоритму: червона лінія – вибіркове середнє $\overline{x} = \frac{1}{n}\sum x_i n_i = 10,71$; зелені лінії – теоретичні межі довірчого інтервалу, розраховані за розподілом Стьюдента; сині лінії – межі довірчого інтервалу, розраховані за методом Bootstrap.

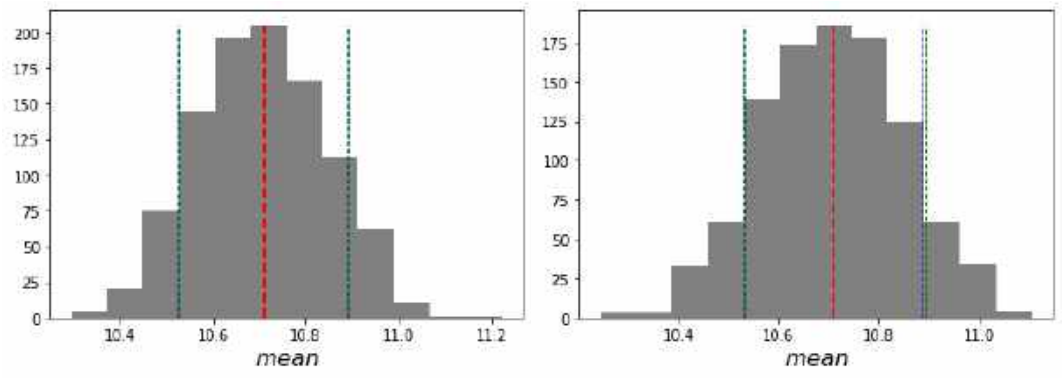


Рис. 3.15. Результати розрахунків довірчого інтервалу за методом Bootstrap

Нагадаємо, що вище було отримано довірчий інтервал для математичного сподівання $[10.52802;\ 10.8942]$ із застосуванням t-розподілу (розподілу Стьюдента). Застосування методу Bootstrap дало в першому прогоні межі $[10.5333;\ 10.8888]$, у другому – $[10.5388;\ 10.8833]$. Як бачимо, різниця становить соті та тисячні. Це дає змогу оцінити точність методу: 0,05 %. Звичайно, збільшення кількості ітерації з 1000 до декількох тисяч приводить до підвищення точності, але це не має значення з практичної точки зору.

3.6. Інтервальне оцінювання дисперсії розподілу

Дисперсія вибірки $D^* = s^2 = \dfrac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n-1}$, яка є оцінкою дисперсії певного розподілу генеральної сукупності (популяції), також є випадковою величиною. Але на відміну від вибіркового середнього всі доданки у формулі є невід'ємними, тобто розподіл цієї випадкової величини починається з нуля і має лише один «хвіст» справа. Це так званий розподіл хі-квадрат $\left(\chi^2\right)$, який також залежить від кількості ступенів свободи, як і розподіл Стьюдента:

$$\chi_n^2 = \xi_1^2 + \xi_2^2 + \ldots + \xi_n^2,$$

де ξ_i^2 – незалежні нормально розподілені величини, які мають нульове математичне сподівання й одиничну дисперсію; n – кількість ступенів свободи.

Графік функції χ^2 для різних ступенів свободи зображено на рис. 3.16.

При великих значеннях n розподіл χ^2 наближається до нормального розподілу з відповідними математичним сподіванням (ненульовим) і

дисперсією. Формула для розрахунку функції χ^2 та інтегралів від неї – складна, не будемо її використовувати.

При класичному статистичному підході довірчий інтервал для дисперсії розраховується таким чином:

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\chi^2_{\frac{1+\alpha}{2}}(n-1)} < D < \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\chi^2_{\frac{1-\alpha}{2}}(n-1)},$$

де $\bar{x}$ – вибіркове середнє; $n-1$ – кількість ступенів свободи; α – довірча ймовірність.

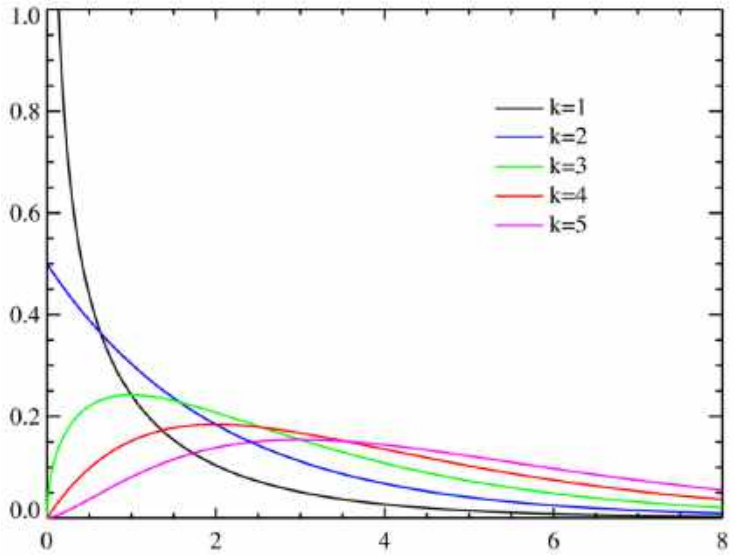


Рис. 3.16. Розподіл χ^2

Тут також зазвичай відкидають зліва і справа ділянки, щоб отримати довірчий інтервал найменшої ширини в центральній частині розподілу (рис. 3.17).

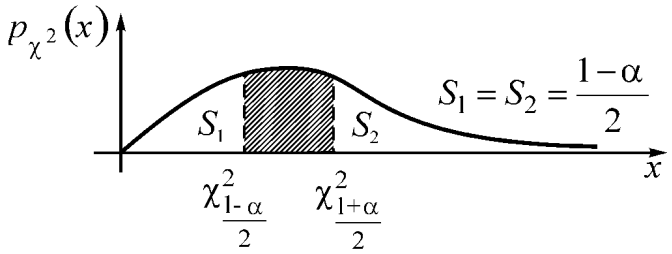


Рис. 3.17. Довірчий інтервал розподілу χ^2

У системах роботи зі статистичними даними є відповідні функції розрахунку χ^2, аргументи яких – кількість ступенів свободи й довірча ймовірність.

Але сучасний метод Bootstrap дає змогу розрахувати довірчий

інтервал для дисперсії без застосування розподілу χ^2 гіпотези про нормальний розподіл генеральної сукупності. Завдяки цьому він є кращим і більш гнучким та універсальним.

Розглянемо застосування методу Bootstrap на прикладі 3.4.1, уважаючи, що $\alpha = 0{,}9$.

1. За даними таблиці

x_i	9,4	9,8	10	10,6	11	11,4	12	12,4
n_i	3	5	5	6	6	6	4	1

було створено «бокс», який містив усі 36 даних елементів.

2. Із «боксу» у циклі випадково виймався з поверненням 36 разів один елемент і створювалася чергова вибірка з 36 елементів. Визначалася дисперсія чергової вибірки й записувалася в масив.

3. Пункт 2 виконувався 1000 разів, унаслідок чого отримали масив з 1000 значень вибіркової дисперсії.

4. У створеній вибірці з 1000 елементів відсікалися «хвости» крайніх значень (великі й малі), залилось $\alpha \cdot 100\%$ початкового масиву (за умовами задачі відсікалось 5% ранжованої вибірки дисперсії зліва і 5% справа). Межі відсікання є межами довірчого інтервалу дисперсії.

На рис. 3.18 показано типовий результат прогону алгоритму: червона лінія – вибіркова дисперсія $s^2 = D^* = \frac{1}{n-1}\sum\left(x_i - \overline{x}\right)^2 n_i = 0{,}7073$; сині лінії – межі довірчого інтервалу з довірчою ймовірністю $\alpha = 0{,}9$.

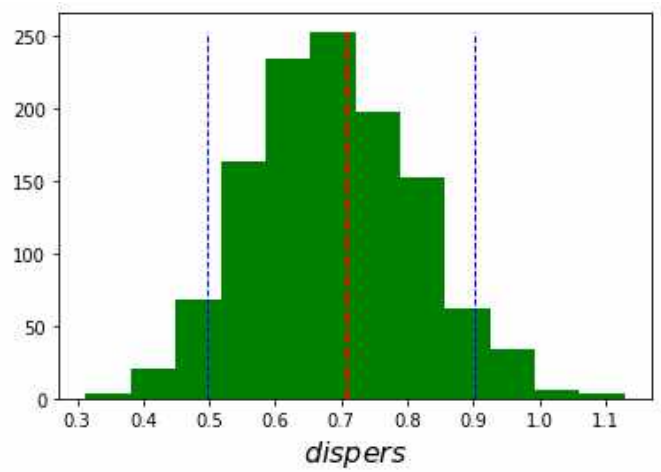


Рис. 3.18. Гістограма розподілу дисперсії і довірчі інтервали

Кілька прогонів дають такі межі довірчого інтервалу дисперсії:

$$[0{,}50168;\ 0{,}880635];$$

$$[0{,}49663;\ 0{,}90044];$$

$$[0{,}49406;\ 0{,}86959];$$
$$[0{,}5037;\ 0{,}896].$$

Отже, результати є стабільними і різняться між прогонами менш ніж на 2%.

3.7. Графік Q–Q

У попередніх розділах було розглянуто, як можна оцінити деякі дані, визначивши точкові й інтервальні оцінки математичного сподівання й дисперсії невідомого розподілу та вибірки. Якщо відомо, що вибірка підпорядковується нормальному закону розподілу, то, знайшовши ці два параметри, можна визначити все необхідне для опису генеральної сукупності (популяції). Але не завжди можна точно визначити вид розподілу.

Наприклад, маємо деякий набір з 700 чисел, максимальне з яких – 18,455, а мінімальне – 8,84. Вибіркове середнє (точкова оцінка математичного сподівання) $\overline{x} = 3{,}039$, медіана дорівнює 2,932, середньоквадратичне відхилення $s = 2{,}767$, середньоквадратичне відхилення вибірки без викидів (10 % вибірки відкинуто) $s = 2{,}026$. Як бачимо, дисперсія суттєво відрізняється через наявність викидів на кінцях розподілу.

Гістограму розподілу (20 стовпців) і графіки відповідних розподілів Гаусса ($M = 3{,}034$, $\sigma_1 = 2{,}767$ та $\sigma_2 = 2{,}026$) показано на рис. 3.19.

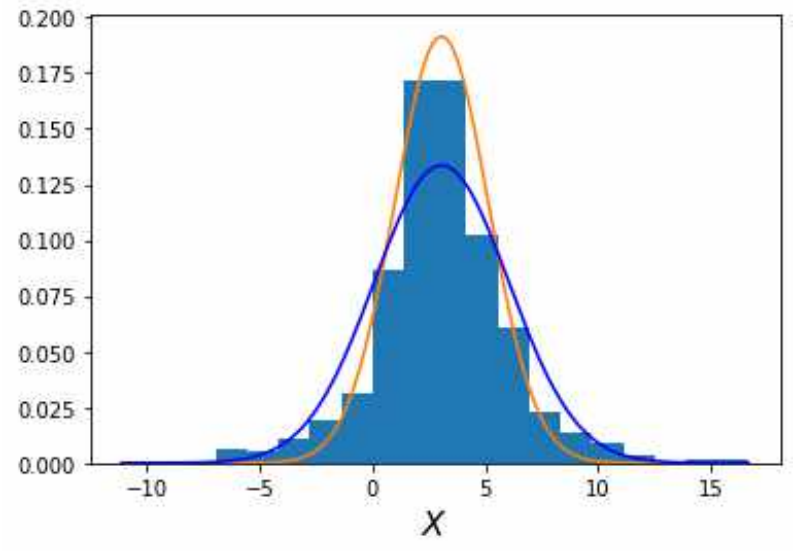


Рис. 3.19. Графіки та діаграма розподілу Гаусса

Тут начебто теоретичний розподіл Гаусса й дані добре узгоджуються – маємо пік усередині та два «хвости». Але можна побачити, що є і певні відхилення. Якщо взяти синю криву, то в центрі розподілу стовпці вищі за

криву, а справа і зліва від максимуму стовпці вже нижчі за криву. Це системне явище, чи ні? Невідомо. Звісно, навіть при розподілі Гаусса дані можуть бути згенеровані так, як показано на рис. 3.19. А може статися, що це не розподіл Гаусса, а інший, і така картина є системною й закономірною. Далі беремо до уваги кількість даних – 700 елементів. Це немало, і тому ймовірність випадкової генерації такого виду, як інтуїтивно розуміємо, є малою.

Отже, виникає задача: «Як оцінити ймовірність того, що емпіричні дані належать до певного теоретичного розподілу, параметри якого було визначено за вибіркою?» Або задачу можна сформулювати так: «Чи можна сказати, що з заданою ймовірністю (наприклад, 0,9) емпіричні дані підпорядковуються певному теоретичному розподілу?»

Зауваження. Задача може ще більш ускладнитися, наприклад: є дві вибірки, треба визначити, чи належать вибірки (з певною ймовірністю) до одного й того ж розподілу чи ні.

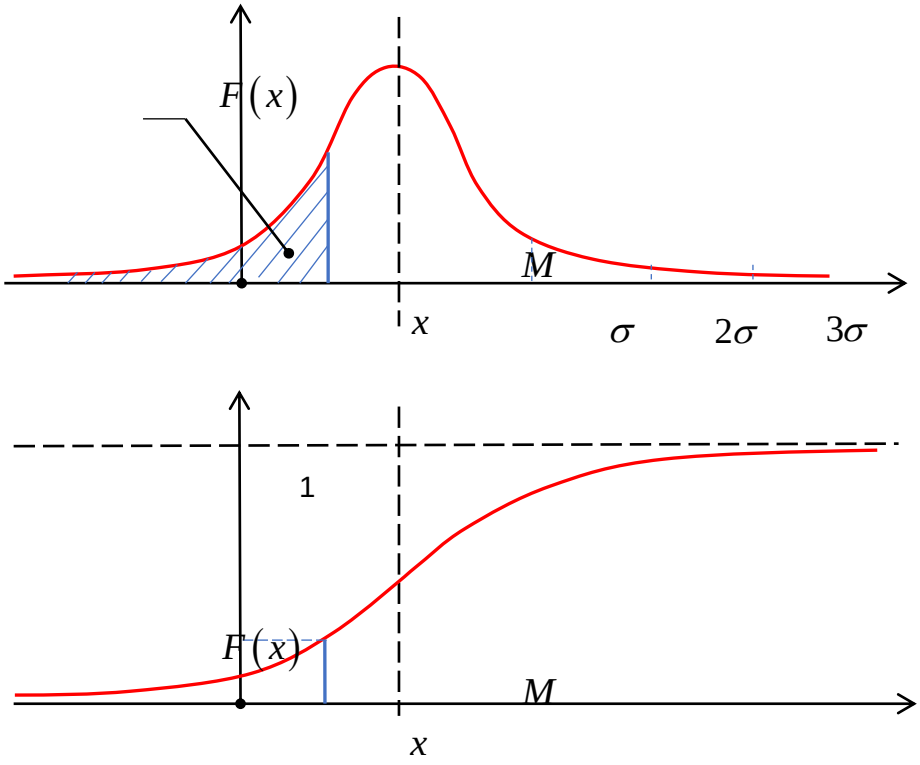


Рис. 3.20. Квантиль розподілу

Одним з елементів аналізу близькості теоретичного й фактичного розподілів даних є так званий **графік Q–Q**, або **квантильна діаграма**. Опишемо квантиль на прикладі нормального розподілу. На рис. 3. 20 зображено графіки нормального розподілу $f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-M)^2}{2\sigma^2}}$ і функції

розподілу, яка є інтегралом від щільності розподілу ймовірності $F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{(t-M)^2}{2\sigma^2}}\,dt$. Зрозуміло, що $F(-\infty)=0$, $F(M)=0{,}5$ і $F(\infty)=1$.

В інтервалі $(-3\sigma, 3\sigma)$ знаходиться більша частина вибірки. Тому зазвичай по горизонтальній осі графіка Q–Q значення змінної лежать від -3 до 3, а посередині 0, тобто дані – центровані, і відлік починається від математичного сподівання, а одиничний відрізок дорівнює σ.

У наведеному вище прикладі емпіричні дані знаходяться в інтервалі (-8,84; 18,84), тому на вертикальній осі графіка Q–Q відкладаємо значення від -8,84 до 18,84. Зрозуміло, що між значеннями змінної x і кількістю значень σ існує лінійна залежність

$$x = M + n\sigma:$$

- якщо $n=0$, то $x=M=3{,}039$;
- якщо $n=3$, то $x=M+3s=3{,}039+3\cdot 2{,}767=11{,}34$;
- якщо $n=-3$, то $x=M-3s=3{,}039+3\cdot 2{,}7673=-5{,}26$.

Як дисперсію беремо вибіркову дисперсію всієї вибірки. Графік Q–Q нормального розподілу являє собою **пряму лінію**. Аналогічним чином на цій діаграмі можна розташувати й емпіричні дані, а потім порівняти фактичний розподіл з теоретичним (рис. 3.21).

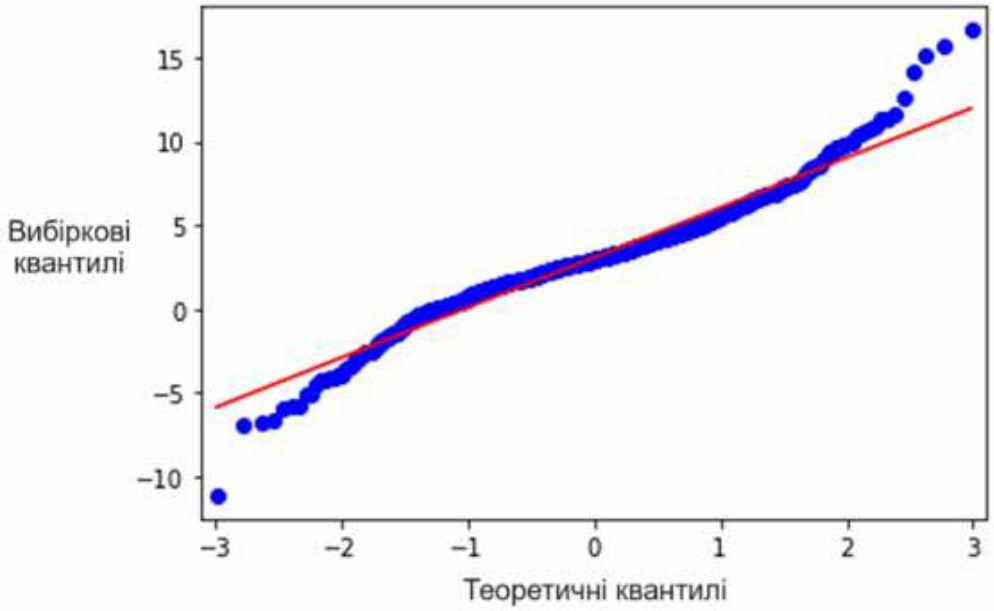


Рис. 3.21. Графік Q–Q

У мові програмування Python для побудови цієї діаграми застосовується функція **statsmodels.api.qqplot(a, line='s')**, аргументами якої є масив даних.

На графіку Q–Q дані розташовуються майже на прямій, але після

середини вони трохи відхиляються від прямої, потім перетинають її і на кінцях відхиляються ще більше. Це свідчить про те, що є якісь системні порушення нормального закону. Але як це виміряти? Вочевидь, треба ввести певну кількісну міру відхилення емпіричного розподілу даних від теоретичного, а також відповідні **критерії узгодженості** (теоретичного й емпіричного розподілів або двох емпіричних розподілів).

3.8. Критерій узгодженості Пірсона

Для порівняння теоретичного й емпіричного розподілів треба ввести якусь числову міру, яка є випадковою величиною (унаслідок того, що вибірка є випадковим відображенням популяції). Таких критеріїв розроблено декілька, і найбільш простим серед них є критерій узгодженості Пірсона.

Ідея критерію Пірсона полягає в тому, що гістограму розподілу порівнюють з теоретичним розподілом, а саме: знаходять різниці площ стовпців гістограми і площ, які відсікають ці стовпці від графіка розподілу, підносять ці різниці до квадрата й діляють на площу стовпця теоретичного розподілу, далі просумовують ці величини за всіма стовпцями (рис. 3.22). Чим більшим є відхилення, тим більшою буде сума.

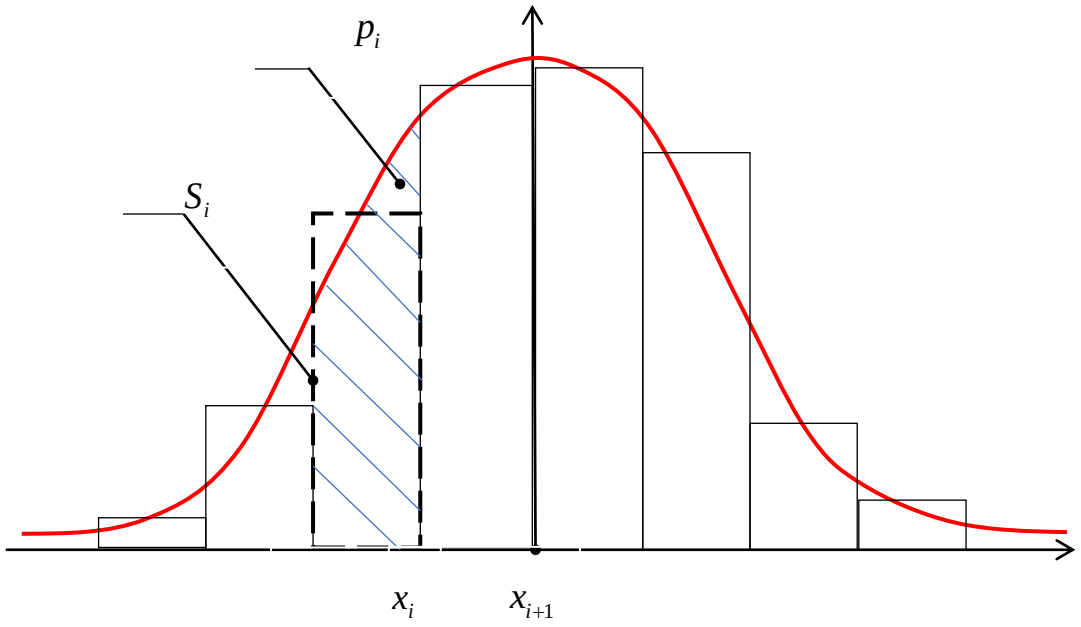


Рис. 3.22. Схема розрахунку критерію Пірсона

При кількості стовпців n критерій Пирсона визначається формулою

$$\Omega = \sum_{i=1}^{n} \frac{\left(p_i - S_i\right)^2}{p_i}.$$

Помноживши цей вираз на N – обсяг вибірки (кількість її елементів), маємо

$$P = N\Omega = \sum_{i=1}^{n} \frac{N^2\left(p_i - S_i\right)^2}{Np_i} = \sum_{i=1}^{n} \frac{\left(Np_i - N_i\right)^2}{Np_i},$$

де $N_i = NS_i$ – фактична кількість потраплянь елементів вибірки в певний інтервал; $p_i = \int_{x_i}^{x_{i+1}} f\left(x\right)dx$ (де $f\left(x\right)$ – щільність теоретичного розподілу) – імовірність потрапляння випадкової величині в інтервал $\left[x_i; x_{i+1}\right]$.

Якщо вибірка підпорядковується **нормальному закону**, то величина P має χ^2-розподіл зі ступенем свободи, що дорівнює кількості стовпців у діаграмі мінус один. Зрозуміло, що якщо величина P є великою, то ймовірність відхилення теоретичного розподілу від емпіричного – дуже мала, а це означає, що теоретичний розподіл може описати фактичні дані з малою ймовірністю, тобто, скоріше за все, цей теоретичний розподіл не відповідає дійсності.

На рис. 3.23 зображено графік розподілу χ^2 та односторонню оцінку.

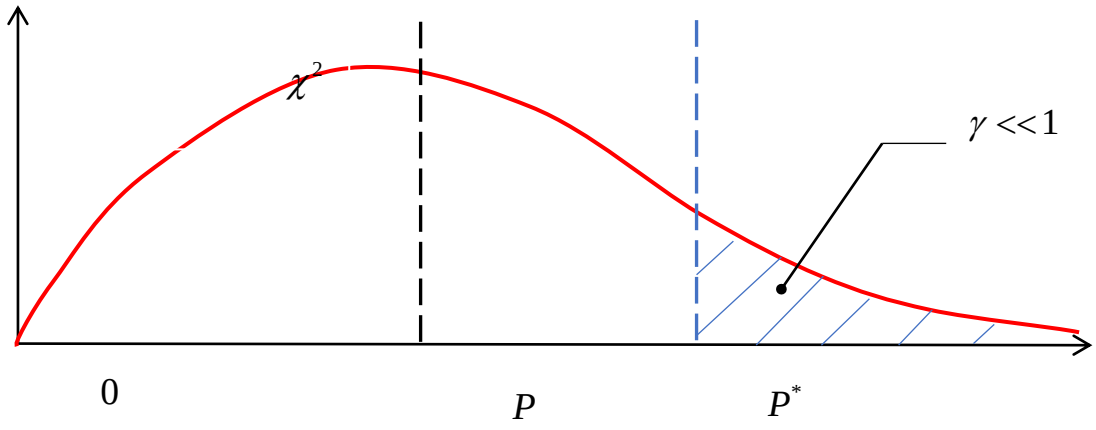


Рис. 3.23. Довірчий інтервал для P

Знаючи кількість стовпців n у діаграмі рис. 3.22, розраховуємо критичне значення P^* із застосуванням односторонньої оцінки розподілу χ^2. Для цього знаходимо значення P^*. Задаємо певну малу ймовірність γ того, що значення випадкової величини з розподілом χ^2 перевищує

значення P^*, яка зазвичай береться такою, що дорівнює 0,05 або 0,1.

Якщо розрахована величина P перевищує P^*, то це означає, що теоретичний розподіл не підходить для опису емпіричних даних, а точніше, імовірність того, що емпіричні дані підпорядковуються теоретичному розподілу, є недопустимо малою, і потрібно відкинути гіпотезу про цей вид теоретичного розподілу.

Якщо ж значення P знаходиться в межах від нуля (а воно не може бути меншим від нуля) до P^*, то з допустимою ймовірністю теоретичний розподіл описує емпіричні дані, і гіпотезу про цей вид теоретичного розподілу приймаємо.

Зауваження. Зрозуміло, що значення P^* і P залежать від кількості інтервалів n на діаграмі, зображеній на рис. 3.22. Тому слід декілька разів проводити розрахунки для різних величин n. Загалом кількість n залежить від обсягу вибірки N. Рекомендується брати n від 5 до 30 залежно від N.

3.9. Застосування методу Bootstrap для визначення узгодженості

Потужний метод Bootstrap не має обмежень щодо нормального розподілу популяції (генеральної сукупності) і може застосовуватися при будь-якому розподілі.

Основі етапи **алгоритму**:

1. Уводять гіпотезу про певний вид розподілу генеральної сукупності (нормальний, Лапласа, Стьюдента тощо) і визначають його параметри – точкову оцінку математичного сподівання й вибіркову дисперсію.

2. Розраховують критерій P або Ω для вибірки.

3. За теоретичним розподілом генерують велику кількість (1000 і більше) вибірок того ж обсягу, що й задано в умовах задачі. Для кожної з вибірок розраховують критерій узгодженості і створюють масив із цих критеріїв.

4. Від масиву критеріїв узгодженості згенерованих вибірок відсікають певну кількість (наприклад, 10 %) максимальних значень. Максимальне значення з тих, що залишилися в масиві, і є критичним значенням P^* (або відповідно Ω^*).

5. Приймають рішення стосовно теоретичного розподілу. Якщо знайдене в п. 2 значення критерію узгодженості перевищує критичне значення, то гіпотезу про вид теоретичного розподілу відкидають. Якщо значення критерію узгодженості є меншим за критичне, то теоретичний розподіл уважають допустимим.

Покажемо це на прикладі розглянутої вище вибірки з 700 значень, діаграму розподілу якої показано на рис. 3.19.

Критерій Пірсона беремо у формі $\Omega = \sum_{i=1}^{n} \frac{(p_i - S_i)^2}{p_i}$.

Як пробний беремо нормальний розподіл з параметрами $M = 3,034$ та $s = 2,026$, які визначаємо з вибірки.

Для кількості стовпців у діаграмі $n = 10$ розрахунковий критерій Пірсона $\Omega = 0,63$.

Генеруємо $NN = 1000$ вибірок за розподілом Гаусса із застосуванням функції **np.random.normal(M, s_2, NN)** і знаходимо для них критерії Пірсона, створивши масив з 1000 елементів. Відкидаємо з цього масиву 10 % найбільших значень **scipy.stats.trim1(PRS_mass_G, 0.1, tail='right')**. На рис. 3.24 зображено цей масив у вигляді діаграми та фактичне значення критерію Пірсона.

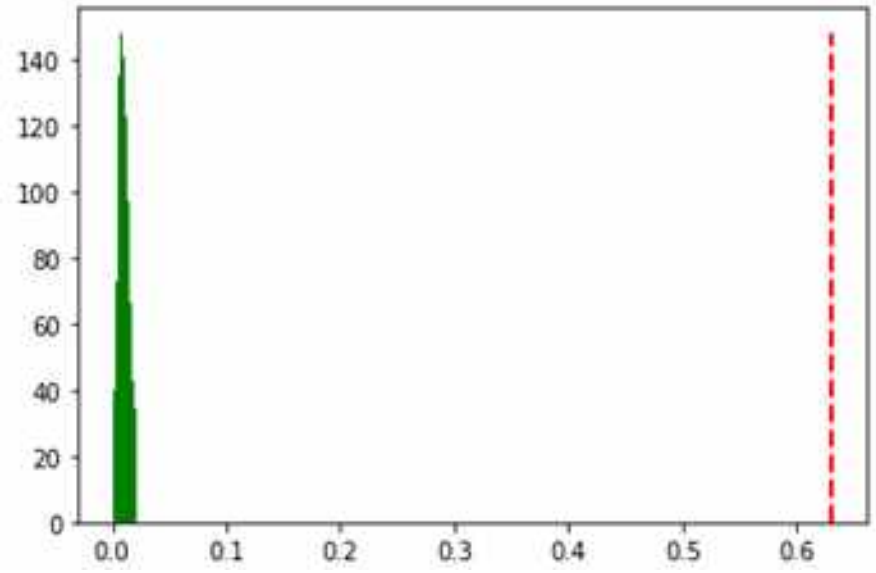


Рис. 3.24. Критерій Пірсона при гіпотезі про нормальний розподіл

Значення $\Omega = 0,63$ є занадто великим, щоб вважати його допустимим. Імовірність того, що задана в задачі вибірка була згенерована нормальним розподілом, – занадто мала. Отже, гіпотезу про нормальний розподіл популяції (генеральної сукупності) **відкидаємо!**

Далі спробуємо застосувати інший розподіл для опису вибірки, наприклад розподіл Лапласа. Цей розподіл являє собою сукупність двох експонент, має більш товсті «хвости», ніж розподіл Гаусса, і гострішу вершину. У цьому випадку критерій Пірсона $\Omega = 0,0105$, що явно значно менше за попередній розрахунок.

Для генерації вибірок при цьому розподілі застосовуємо функцію **np.random.laplace(M, s, NN)**. Для згенерованих вибірок розраховуємо критерій Пірсона, відсікаємо 10 % найбільших значень і будуємо діаграму (рис. 3.25).

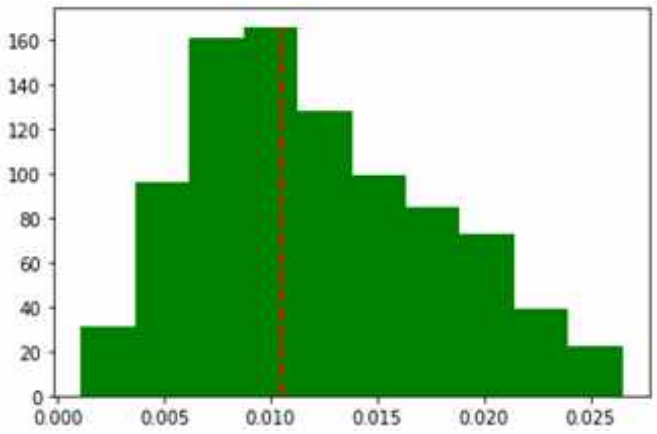


Рис. 3.25. Критерій Пірсона при гіпотезі про розподіл Лапласа

Значення критерію Пірсона $\Omega = 0,0105$ потрапляє в допустимий інтервал, отже, дані підпорядковуються закону розподілу Лапласа, тобто гіпотезу про розподіл Лапласа генеральної сукупності **приймаємо**. На рис. 3.26 зображено гістограму даних і графік відповідного розподілу Лапласа.

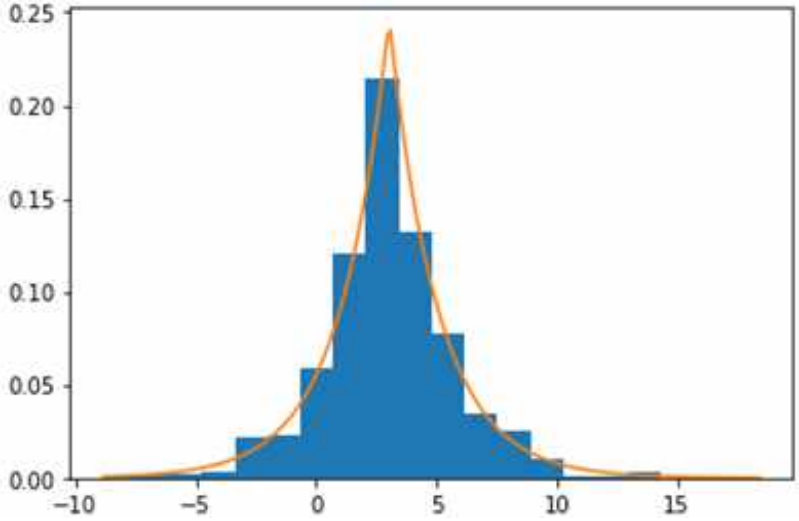


Рис. 3.26. Розподіл даних і графік розподілу Лапласа

Зауваження. Було визначено, що генеральна сукупність описується законом розподілу Лапласа з параметрами $M = 3,034$ та $s = 2,026$. Це зрозуміло, оскільки для її генерації було застосовано розподіл Лапласа з параметрами $M = 3$ та $s = 2$.

3.10. Перевірка статистичної гіпотези про належність вибірок до одного розподілу. Переставний тест

Почнемо з задачі: маємо два сайти з продажів одного й того ж продукту. Треба визначити, який сайт є кращим з огляду на ефективність

продажів. Для цього збираємо інформацію про загальну кількість відвідувань сайтів і кількість відвідувань, що закінчилися покупкою. Ці дані наведено в табл. 3.1.

Таблиця 3.1

Результат	Сайт А	Сайт Б
Конверсія	200	182
Немає конверсії	23 539	22 406

Сайт А конвертує (тобто перетворює відвідувачів на покупців) майже на 5 % краще, ніж сайт Б (0,8425 % проти 0,8057 % – різниця 0,0368 процентних пунктів); це достатньо велике значення. Тут налічується понад 45 000 точок даних, і виникає спокуса розглядати їх як великі дані, що не потребують перевірки щодо статистичної значущості (що є необхідною головним чином для враховування вибіркової варіабельності у невеликих вибірках). Однак рівень конверсій є настільки низьким (менше 1 %), що фактичні змістовні значення – конверсії – становлять лише соті долі, при цьому необхідний розмір вибірки дійсно визначається цими конверсіями. Можна перевірити, чи знаходиться різниця в конверсіях сайтів А і Б у межах діапазону випадкової варіації. Застосуємо для цього процедуру **повторного відбору**. Під випадковою варіацією розуміємо випадкову варіацію, породжену ймовірнісною моделлю нульової гіпотези про те, що між рівнями конверсій не існує різниці.

Під час наступної переставної процедури ставиться запитання: «Якщо ці сайти мають однаковий рівень конверсії, то чи зможе випадкова варіація привести до різниці аж 5 %?»

Тобто намагаємося з'ясувати, чи є спостережувана різниця в продажах сайтів випадковою, чи це – скоріше за все, певна закономірність.

Щоб відповісти на запитання, застосуємо сучасний метод Bootstrap за таким **алгоритмом:**

1. Створимо коробку з усіма вибірковими результатами, яка являтиме собою уявний спільний рівень конверсії з 382 одиниць і 45 945 нулів.

2. Перетасуємо й виймемо повторну вибірку розміром 23 739 (кількість *n* така сама, що й на сайті А), запишемо кількість одиниць.

3. Запишемо кількість одиниць в останніх 22 588 (кількість *n* така сама, що й на сайті В).

4. Запишемо різницю або у вигляді частки одиниць від загальної кількості елементів, або у вигляді кількості одиниць.

5. Повторимо кроки 2–4 багаторазово.

Розглянемо 2000 циклів. На рис. 3.27 зображено гістограму розподілу частки одиниць. Синіми лініями показано межі інтервалу, що становить 90 % значень масиву, з кожного боку відкинуто 5 % масиву. Червоною лінією показано спостережуване значення різниці в кількості одиниць у вигляді частки від загальної кількості спостережень.

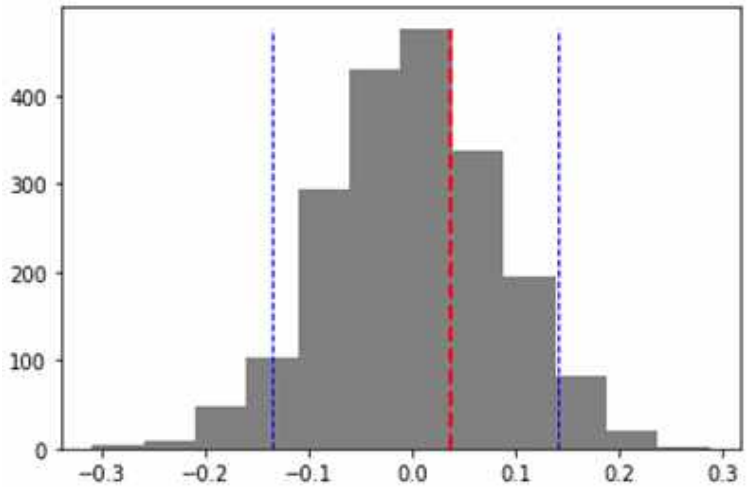


Рис. 3.27. Результати моделювання

Перестановний тест показав, що спостережувана різниця у частці одиниць є, найімовірніше, наслідком випадкових змінень і жодним чином не є наслідком різної ефективності сайтів. Тобто сайти не мають статистично значущої різниці в ефективності.

Щоб довести ефективність методу, в умові задачі змінимо кількість успіхів на сайті Б з 182 на 152. У цьому випадку, прогнавши алгоритм 2000 разів, отримаємо діаграми (рис. 3.28).

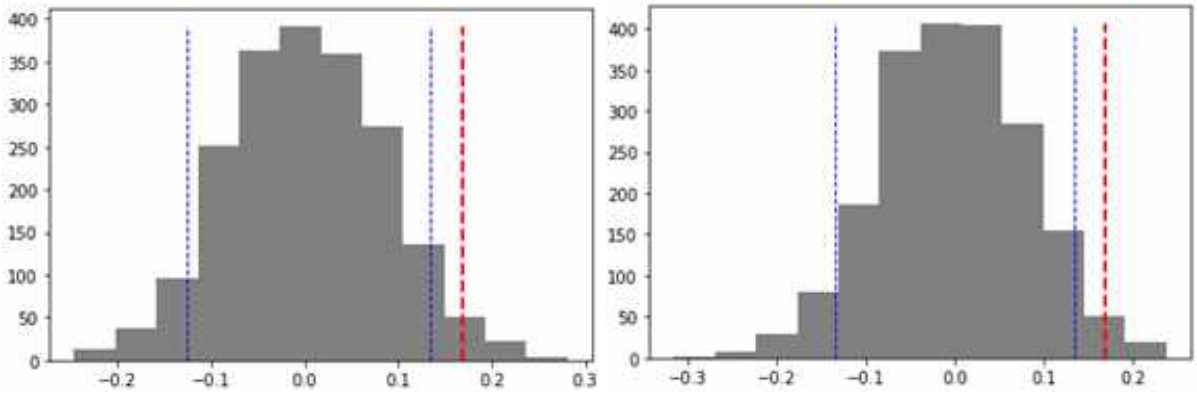


Рис. 3.28. Діаграми

Тут різниця в кількості конверсій є малоймовірною, а це означає, що це, найімовірніше, не є випадковим збігом, а є наслідком різної ефективності сайтів.

Зауваження. Перестановні тести є корисними евристичними процедурами для дослідження ролі випадкової варіації. Ці тести, що відносно легко програмуються, інтерпретуються й пояснюються, є обхідним шляхом замість формалізму й «помилкового детермінізму» статистики, що базується на формулах. Одна з переваг повторного відбору, на відміну від підходів на основі формул, полягає в тому, що він є максимально наближеним до універсального (єдиного для всіх) підходу, що застосовується для статистичного висновку. Дані можуть бути десятковими або двійковими. Розміри вибірок можуть бути однаковими або різними. Припущення про нормально розподілені дані не потребується.

Ключові ідеї для повторного відбору:

– у перестановному тесті численні вибірки об'єднуються й перемішуються;

– перетасовані значення діляться на ті, що вилучаються повторно з вибірки, а потім обчислюється цільова статистика (наприклад, різниця «успіхів» у вигляді частки, як у наведеному прикладі);

– цей процес повторюється, і повторно випробувана статистика зводиться в масив;

– порівняння спостережуваного значення статистики з повторно випробуваним розподілом дає змогу судити про те, чи може бути спостережувана різниця між вибірками випадковою.

Останніми роками велася активна полеміка стосовно використання p-значення. Один із журналів з психології пішов настільки далеко, що «заборонив» використання p-значень у поданих статтях на тій підставі, що публікаційні рішення, що базуються виключно на p-значеннях, спричиняли появу публікацій дослідних робіт поганої якості. Занадто багато дослідників лише туманно уявляють, чим насправді є p-значення, риються в даних і серед різних можливих гіпотез, що підлягають перевірці, поки не знаходять комбінацію, яка приводить до потрібного p-значення і, отже, до роботи, що підходить для публікації.

Справжня проблема полягає в тому, що люди хочуть отримувати від p-значення більший сенс, ніж воно має. Ось те, що ми хотіли б, – щоб p-значення містило **ймовірність, що результат є випадковим**.

Сподіваємося на низьке значення і тому можемо зробити висновок, що щось довели. Саме так багато редакторів журналів інтерпретують p-значення. Але ось те, що p-значення являє собою насправді, – це **ймовірність, що з урахуванням випадкової моделі можуть вийти такі самі граничні результати, що й ті, що спостерігаються.**

У березні 2016 року Американська статистична асоціація (American Statistical Association, ASA) після довгих внутрішніх дискусій

продемонструвала ступінь непорозуміння з приводу p-значень, коли випустила застереження щодо їх використання.

У заяві ASA було наведено шість принципів для дослідників і редакторів журналів:

1) p-значення можуть свідчити про те, наскільки дані є несумісними із заданою статистичною моделлю;

2) p-значення не вимірюють імовірності того, що досліджувана гіпотеза є істинною, або ймовірність того, що дані були породжені виключно в силу випадкової можливості;

3) наукові висновки й бізнес або стратегічні рішення не мають базуватися тільки на тому, чи переходить p-значення певний поріг;

4) належний висновок потребує повної звітності й прозорості;

5) p-значення, або статистична значущість, не є мірою ефекту або важливості результату;

6) p-значення як таке не забезпечує хорошої міри доведення щодо моделі або гіпотези.

Робота, проведена аналітиками даних, зазвичай не призначена для публікації в наукових журналах, тому дебати про сенс p-значення є дещо академічними. Для аналітика даних p-значення – це корисний метричний показник у ситуаціях, коли необхідно знати, чи знаходиться модельний результат, який здається цікавим і корисним, у діапазоні нормальної випадкової варіабельності. Як інструмент для прийняття рішення в експерименті p-значення слід уважати не вирішальним показником, а просто ще однією точкою інформації, яка впливає на рішення. Наприклад, p-значення іноді використовуються як проміжні входи в якісь статистичні моделі прийняття рішень або в моделі машинного навчання – ознака може бути включена до складу моделі або виключена з її складу залежно від p-значення ознаки.

Під час визначення статистичної значущості можливими є такі помилки:

- помилка 1-го роду, коли помилково вважають, що ефект є реальним, тоді як у дійсності він є суто випадковим;
- помилка 2-го роду, коли помилково вважають, що ефект не є реальним (тобто є випадковим), тоді як він є насправді реальним.

Помилка 2-го роду по суті є не стільки помилкою, скільки судженням, що розмір вибірки є занадто маленьким для того, щоб можна було виявити ефект. Коли p-значення є далеким від статистичної значущості (наприклад, перевищує 5 %), фактично можна сказати, що ефект не доведено. Може виявитися, що більша вибірка приведе до меншого p-значення.

Основна функція перевірок значущості (або так званих перевірок гіпотез) – захист від того, щоб експериментатор був обманутий випадковою можливістю. Отже, перевірки зазвичай будуються таким чином, щоб мінімізувати помилки 1-го роду.

3.11. Лінійна регресія

3.11.1. Проста лінійна регресія

Регресійний аналіз — розділ математичної статистики, у якому розглядаються методи аналізу залежності однієї величини від іншої. На відміну від кореляційного аналізу під час регресійного аналізу не визначається, чи є істотним зв'язок, а шукається модель цього зв'язку, виражена **функцією регресії**.

Регресійний аналіз використовується в тому випадку, коли зв'язки між змінними можна виразити кількісно у вигляді деякої комбінації цих змінних. Отримана комбінація використовується для прогнозування значення, якого може набувати цільова (залежна) змінна, яка обчислюється на заданому наборі значень вхідних (незалежних) змінних. У найпростішому випадку для цього використовуються стандартні статистичні методи, такі як **лінійна регресія**. На жаль, більшість реальних моделей не вкладаються в межі лінійної регресії. Більш складні моделі розглянемо пізніше.

Проста **лінійна регресія**, або лінійна парна регресія, моделює зв'язок між величиною однієї змінної і величиною іншої, наприклад: у міру збільшення X збільшується Y; у міру збільшення X зменшується Y. **Кореляція** – це ще один спосіб визначити, яким чином зв'язані дві змінні (уже розглядали коефіцієнт кореляції). Відмінність між ними полягає в тому, що кореляція вимірює зв'язок між двома змінними, тоді як регресія вказує на функціональну природу зв'язку.

Наведемо основні поняття:

– **незалежна змінна** (independent variable) – змінна, яка використовується для прогнозування відгуку; синоніми: незалежна змінна, X-змінна, предиктор, ознака, атрибут;

– **відгук** (response) – змінна, яку намагаються спрогнозувати; синоніми: залежна змінна, Y-змінна, мета, результат;

– **підігнані значення** (fitted values) $\hat{Y}$ – оцінки, отримані з кореляційної моделі; синонім – прогнозовані значення;

– **залишки** (residuals) – різниця між значеннями, що спостерігаються, і підігнаними значеннями: $Y_i - \hat{Y}_i$; синонім – похибки.

Постановка задачі: маємо масив точок (X_i, Y_i), показаний на діаграмі розсіювання (рис. 3.29).

Можна побачити, що між випадковими величинами X_i і Y_i є певний функціональний зв'язок, наприклад лінійний:

$$\hat{Y} = b_1 X + b_0 + \varepsilon,$$

де b_1 і b_0 – коефіцієнти; ε – випадкова величина, похибка моделі.

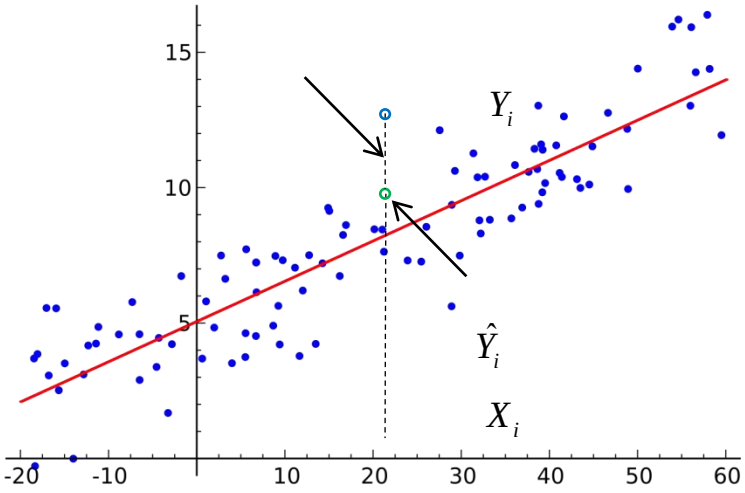


Рис. 3.29. Точки даних і регресійна пряма

Таким чином, за заданим масивом точок (X_i, Y_i) треба визначити коефіцієнти моделі b_1 і b_0, які мінімізують відхилення прогнозованого значення $\hat{Y}(X_i)$ від фактичного значення Y_i.

Як же визначити коефіцієнти? Для цього треба скласти цільову функцію, яка залежить від відхилення фактичних даних від прогнозованих, тобто від $Y_i - \hat{Y}_i$, і підібрати коефіцієнти b_1 і b_0 таким чином, щоб цільова функція набула мінімуму. Такою функцією зазвичай є сума квадратичних залишків, яку також називають залишковою сумою квадратів або RRS (residual sum of squares):

$$RRS = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - b_1 X_i - b_0\right)^2 \to \min.$$

Метод мінімізації суми квадратичних залишків називають регресією найменших квадратів або звичайним методом найменших квадратів (звичайним МНК). Цей метод часто приписують Карлу Фрідріху Гауссу, німецькому математику, але 1805 року його вперше опублікував французький математик Андре-Марі Лежандр.

Як відомо, умовами екстремуму функції декількох змінних є рівність нулю похідних функції за всіма змінними. У цьому випадку RRS є функцією

невідомих коефіцієнтів b_1 і b_0. Таким чином, умовами екстремуму є рівняння

$$\frac{\partial}{\partial b_1} RRS = 0, \quad \frac{\partial}{\partial b_0} RRS = 0,$$

звідки отримуємо систему лінійних рівнянь

$$\begin{cases} \left(\sum_{i=1}^{n} X_i^2 \right) b_1 + \left(\sum_{i=1}^{n} X_i \right) b_0 = \sum_{i=1}^{n} X_i Y_i; \\ \left(\sum_{i=1}^{n} X_i \right) b_1 + n b_0 = \sum_{i=1}^{n} Y_i. \end{cases}$$

Розв'язок цієї системи має вигляд

$$b_1 = \frac{\sum_{i=1}^{n} \left(Y_i - \bar{Y} \right) \left(X_i - \bar{X} \right)}{\sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2}, \quad b_0 = \bar{Y} - b_1 \bar{X},$$

де $\bar{Y}$ і $\bar{X}$ – вибіркові середні масивів X_i і Y_i.

Звичайно, знайдені коефіцієнти b_1 і b_0 не є точними значеннями рівняння регресії всієї популяції (генеральної сукупності), а є вибірковими значеннями, які відповідають цій вибірці, тому їх можна позначати $\hat{b}_1$ і $\hat{b}_0$. Отже, унаслідок того, що набір даних $\left(X_i, Y_i \right)$ є випадковим, знайдені коефіцієнти моделі b_1 і b_0 являють собою також випадкові величини – **оцінки** коефіцієнтів рівняння регресії.

Зауваження. Зазначимо, що лінія лінійної регресії проходить через точку з координатами $\left(\bar{x}, \bar{y} \right)$, тобто через точку вибіркового середнього змінних.

Історично так склалося, що обчислювальна зручність є однією з причин широкого застосування методу найменших квадратів у регресії. З появою великих обсягів даних обчислювальна швидкість, як і раніше, є важливим фактором. Метод найменших квадратів, як і середнє значення, є чутливим до викидів, хоча ця проблема є значущою тільки в невеликих або помірних за розміром задачах.

Коли аналітики й дослідники використовують термін «регресія» окремо, вони зазвичай посилаються на лінійну регресію; у центрі уваги – зазвичай розроблення лінійної моделі для пояснення зв'язку між предикторними змінними й числовою змінною результату (результівною змінною). З огляду на формальний статистичний сенс регресія також

охоплює нелінійні моделі, які дають функціональний зв'язок між предикторними змінними й змінною результату. У співтоваристві машинного навчання цей термін також часом використовується у вільному тлумаченні для посилання на використання будь-якої моделі прогнозування, яка дає прогнозований числовий результат (на відміну від **методів класифікації**, що прогнозують бінарний або категоріальний результат).

Раніше вже розглядалися проблеми виявлення причиново-наслідкового зв'язку змінних, що входять у модель. Регресійна модель встановлює зв'язок між X_i і Y_i , але не визначає, що є наслідком, а що – причиною. Для цього кожен має сам, виходячи з власних міркувань, визначити, які дані є незалежними змінними X_i, а які – відгуком Y_i. Звісно, якщо поміняти положення наборів даних у моделі, то можна знайти зв'язок між ними у формі $\hat{X} = a_1 Y + a_0$.

3.11.2. Діагностика моделі

Після отримання рівняння регресії необхідно визначити, наскільки добре воно описує емпіричні дані. Тут може бути кілька варіантів і кілька напрямів покращання моделі. Розглянемо рис. 3.30, де наведено приклад, автором якого є Френсіс Анскомб.

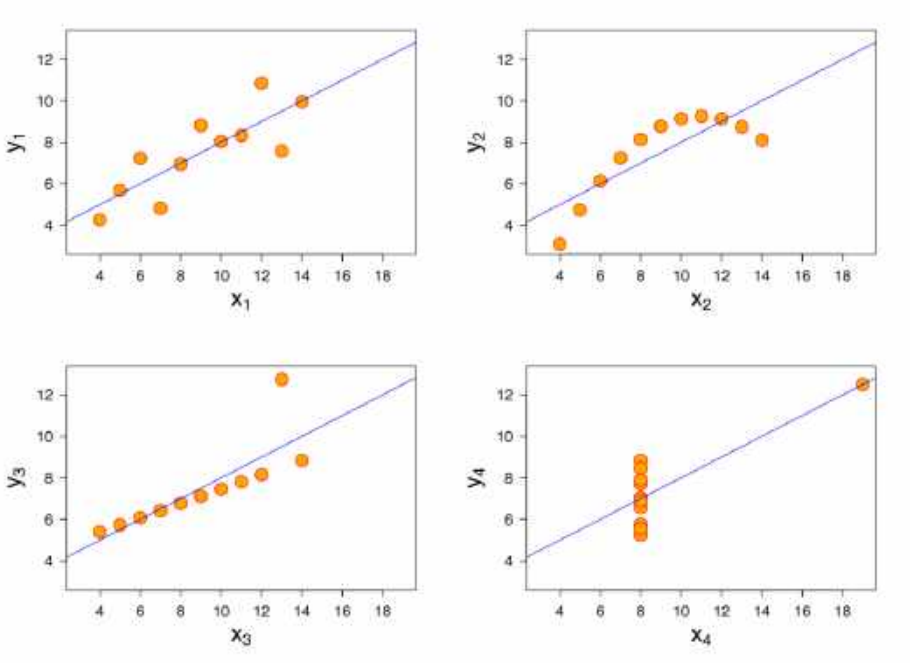


Рис. 3.30. Дані й лінія регресії

У всіх чотирьох випадках змінні y мають однакові середнє значення (7,5), дисперсію (4,12), кореляцію (0,816) і лінію регресії (y = 3 + 0,5x).

Однак, як видно на графіках, розподіл змінних є різним.

Перший розподіл (угорі ліворуч), здається, є нормальним і відповідає тому, що можна було б очікувати, розглядаючи дві корельовані змінні і зважаючи на припущенням про нормальність. Другий розподіл (угорі праворуч) не є нормальним, хоча можна спостерігати очевидний взаємозв'язок між двома змінними, він не є лінійним (а, мабуть, квадратичний). У третьому випадку (унизу ліворуч) лінійна залежність – ідеальна, за винятком одного викиду, що достатньо впливає, щоб коефіцієнт кореляції зменшився з 1 до 0,816. Нарешті, четвертий розподіл (унизу праворуч) показує, що одного викиду достатньо для отримання високого коефіцієнта кореляції, навіть якщо зв'язок між двома змінними не є лінійним.

Ці приклади показують, що коефіцієнт кореляції як зведена статистика не може замінити візуальний аналіз даних. Іноді кажуть, що приклади демонструють, що кореляція Пірсона припускає, що дані підпорядковуються нормальному розподілу, але це неправильно.

Розглянемо **числові оцінки якості наближення**:

1. Найважливішим метричним показником результативності з точки зору науки про дані є середньоквадратична помилка, або RMSE (root mean squared error). RMSE – це квадратний корінь з середньоквадратичної помилки в прогнозованих значеннях:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}}.$$

RMSE визначає загальну точність моделі і є підставою для її порівняння з іншими моделями (включаючи моделі, які підігнано з використанням спеціальних прийомів машинного навчання).

2. Аналогічною RMSE є стандартна похибка залишків, або RSE (residual standard error). У цьому випадку маємо *p* предикторів, і RSE задається такою формулою:

$$RSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n - p - 1}}.$$

Єдина відмінність полягає в тому, що знаменник у цій формулі є кількістю ступенів свободи, а в попередній формулі – кількістю записів. Чим зумовлена ця відмінність?

Зауваження. Це важливий момент. Можна покращити RMSE, узявши не лінійну модель регресії, а, наприклад, квадратичну, або ще більшого порядку. Більший порядок моделі покращує RMSE, але в цю модель входить більша кількість коефіцієнтів. Якщо кількість коефіцієнтів моделі буде дорівнювати кількості точок, то регресійна крива пройде через усі

точки й відхилення буде нульовим! Але якість такої моделі буде нульовою, тому що через високий порядок полінома крива буде осцилювати, а в проміжках між фіксованими точками емпіричних даних і поза межами інтервалу даних буде мати великі відхилення (рис. 3.31).

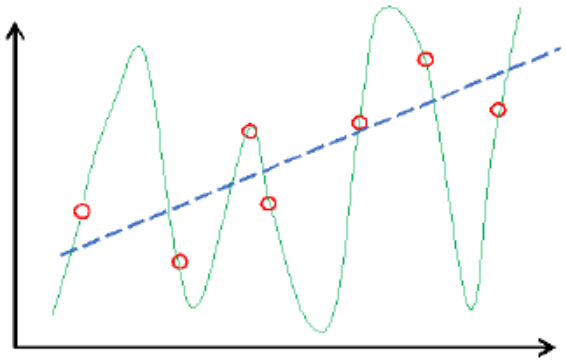


Рис. 3.31. Лінійна регресія й апроксимація

Тому, щоб компенсувати покращання точності моделі у фіксованому наборі точок (X_i, Y_i), наслідком якого є погіршення збіжності в нових точках, які можуть бути додані до моделі, у знаменник уводять кількість коефіцієнтів моделі *p*, у простій лінійній моделі *p* = *2*.

3. Ще один корисний метричний показник, який можна побачити в даних на виході з програмних систем, – це коефіцієнт детермінації, або ***R*-квадрат**, або R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}.$$

R-квадрат змінюється в межах від 0 до 1 і визначає пояснювану в моделі частку варіації в даних і є корисним головним чином під час використання регресії в пояснювальних цілях, коли потрібно проаналізувати, наскільки добре модель підігнано до даних.

Отже, дисперсія даних містить дві складові, одна з яких пояснюється моделлю, а інша є наслідком випадкової варіації (або є більш складною залежністю, яка не входить у модель). І критерій *R*-квадрат визначає частку пояснюваної дисперсії. Якщо $R^2 = 1$, то це означає, що вся дисперсія даних y_i пояснюється моделлю і випадкових варіацій немає взагалі.

3.11.3. Перехресна перевірка

Класичні статистичні метричні показники регресії (R^2, F-статистики і p-значення) є внутрішньовибірковими показниками і застосовуються до

тих же даних, які використовувалися для підгонки моделі. На інтуїтивному рівні можна побачити, що має сенс відкласти трохи вихідних даних, не використовуючи їх для підгонки моделі, і далі застосувати модель до зарезервованих (відкладених) даних, щоб визначити, чи здатна модель виконувати певну роботу. Зазвичай значна частина даних буде використовуватися для підгонки моделі, а решта – для її перевірки.

Така ідея позавибіркової перевірки не є новою, але ще не затвердилася, оскільки великі набори даних ще не набули широкого застосування; маючи невеликий набір даних, аналітики зазвичай хочуть використовувати всі наявні дані й на їх основі виконувати підгонку кращої моделі.

Однак, використання контрольної вибірки з відкладеними даними ставить дослідника в залежність від певної невизначеності, яка виникає просто через варіабельності в малій контрольній вибірці. Наскільки будуть різнитися результати діагностики моделі, якщо взяти іншу контрольну вибірку з відкладеними даними?

Перехресна перевірка розширює ідею контрольної вибірки з відкладеними даними до множинних послідовних контрольних вибірок.

Алгоритм базової k-блокової перехресної перевірки має такий вигляд:

1. Відкладають $1/k$ даних як **контрольну вибірку**.

2. Натреновують модель на дані, що залишилися (тобто знаходять $\hat{b}_1$ і $\hat{b}_0$).

3. Застосовують модель до контрольної вибірки $1/k$ (оцінюють результати) і записують необхідні метричні показники діагностики моделі.

4. Відновлюють перші $1/k$ даних і відкладають наступні $1/k$ (виключаючи будь-які записи, які було вибрано в перший раз).

5. Повторюють кроки 2 і 3.

6. Повторюють доти, доки кожен запис не буде використовуватися в процентній частці, призначеній для контрольної вибірки.

7. Усереднюють або комбінують метричні показники діагностики моделі.

Розподіл даних на тренувальну вибірку й контрольну вибірку також називають поділом на блоки (fold).

3.11.4. Приклад побудови однофакторної регресійної моделі

За посиланням http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv можна скачати файл з даними про 1599 вин, який містить 12 записів у кожному рядку – характеристики вина за 11 критеріями й середня оцінка певного вина, яку поставили

дегустатори, тобто записи (рядки) містять числові характеристики *fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, quality.*

Візьмемо два фактори – *fixed acidity* та *citric acid*. Будемо вважати, що *fixed acidity* є незалежною змінною, а *citric acid* – відгуком.

Для побудови регресійної моделі застосуємо Python. Слід зазначити, що певні бібліотеки за дефолтним варіантом будують модель регресії без члена $\hat{b}_0$, тобто модель вигляду $\hat{Y}_i = b_1 X_i$. Це зумовлене насамперед тим, що дані, які входять у модель, частіше за все подаються у центрованому й нормованому виглядах, тобто дані для моделі зазвичай попередньо обробляються, що може полягати в тому, що від кожного набору даних віднімається вибіркове середнє. У такому випадку лінія регресії проходить через точку (0; 0) і коефіцієнт $\hat{b}_0$ автоматично дорівнює нулю.

Щоб указати для процедури, що дані є нецентрованими, слід створити масив, який має більшу розмірність, ніж масив X:

X_ = sm.add_constant(X),

де **import statsmodels.api as sm.**

Далі застосуємо

lmRegModel_1 = sm.OLS(Y, X_)
result_1 = lmRegModel_1.fit()

Параметри моделі **result_1.params:**

[-0.35427, 0.07515299],

де перший елемент – це $\hat{b}_0$, а другий – це $\hat{b}_1$.

Детальнішу інформацію можна знайти за посиланням https://www.statsmodels.org/stable/generated/statsmodels.regression.linear_model.OLS.html

На рис. 3.32 зображено графік регресії та діаграму розсіювання.

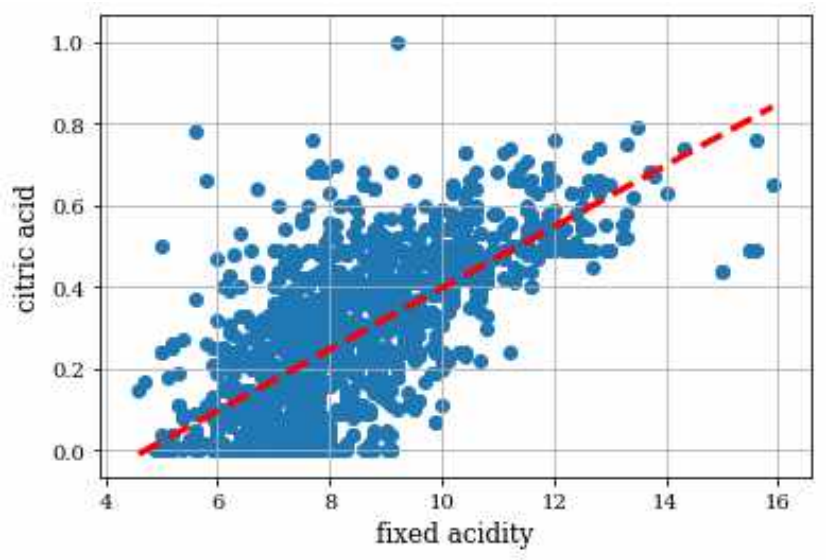


Рис. 3.32. Лінія регресії і діаграма розсіювання

Усі дані про модель отримати досить легко:
print(result_1.summary())

OLS Regression Results

Dep. Variable:	y	R-squared:	0.451
Model:	OLS	Adj. R-squared:	0.451
Method:	Least Squares	F-statistic:	1313
Date:	Sun, 21 Mar 2021	Prob (F-statistic):	2.54e-210
Time:	16:14:46	Log-Likelihood:	826.92
No. Observations:	1599	AIC:	-1650
Df Residuals:	1597	BIC:	-1639
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.3543	0.018	-20.095	0.000	-0.389	-0.320
x1	0.0752	0.002	36.234	0.000	0.071	0.079

Omnibus:	82.195	Durbin-Watson:	1.506
Prob(Omnibus):	0.000	Jarque-Bera (JB):	100.566
Skew:	0.519	Prob(JB):	1.45e-22
Kurtosis:	3.658	Cond. No.	42.1

Окремі властивості моделі:

- **result_1.rsquared** – статистика R^2, яка в цій задачі дорівнює 0,4511855;
- **result_1.bse** – стандартна похибка коефіцієнтів $\hat{b}_1$ і $\hat{b}_0$:

$$s = RSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-2}} = 0{,}144358,$$

$$S_{b_1} = \frac{s}{\sqrt{\sum_{i=1}^{n}\left(x - \overline{x}\right)^2}} = \frac{0{,}144358}{69{,}60031} = 0{,}0020740968,$$

$$S_{b_0} = s\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}} = 0{,}01762932.$$

Ці параметри потрібні для оцінювання значущості коефіцієнтів моделі $\hat{b}_1$ і $\hat{b}_0$. Ці коефіцієнти є випадковими величинами і залежать від вибірки. Знайшовши певні значення, слід визначити, чи можна вважати, що з заданою ймовірністю коефіцієнт дорівнюватиме нулю (тобто не буде входити в модель). Це можна зробити за допомогою методів математичної статистики, а можна застосувати метод Bootstrap.

Статистику

$$t_{b_1} = \frac{|b_1|}{S_{b_1}} = \frac{0{,}07515}{0{,}01762932} = 4{,}26295, \qquad t_{b_0} = \frac{|b_0|}{S_{b_1}} = \frac{0{,}35427}{0{,}0020741} = 170{,}80688$$

можна описати розподілом Стьюдента, і це дає змогу розрахувати критичні значення інтервалів цих статистик і перевірити статистичну гіпотезу про значущість коефіцієнтів моделі.

Кількість ступенів свободи (1599 – 2 = 1597) є дуже великою, тому розподіл Стьюдента можна замінити нормальним розподілом. Критичним значенням при рівні значущості $\alpha = 0{,}05$, тобто якщо площа кожного з хвостів дорівнює 0,025, є **1,96**.

Розраховані t-статистики коефіцієнтів $\hat{b}_1$ і $\hat{b}_0$ перевищують критичне значення. Це свідчить про те, що обидва коефіцієнти моделі є статистично значущими.

На рис. 3.33 показано діаграму розподілу залишків $Y_i - \hat{Y}_i$.

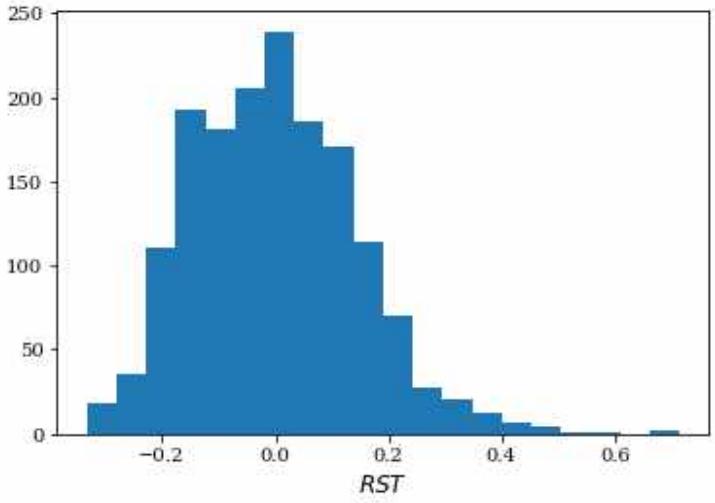


Рис. 3.33. Розподіл залишків

Як було зазначено вище, коефіцієнти $\hat{b}_1$ і $\hat{b}_0$ є випадковими величинами, тобто їх знаходять за вибіркою, а вибірка вибирається випадково з популяції (генеральної сукупності). Існують статистичні методи, які дають змогу знайти **інтервальні оцінки коефіцієнтів** $\hat{b}_1$ і $\hat{b}_0$. Але існує ефективний метод Bootstrap, з допомогою якого можна швидко

отримувати інтервальні оцінки без зайвих зусиль.

Для пошуку інтервальних оцінок можна застосувати такий алгоритм:

1. Із вибірки навмання беруть один елемент (X_i, Y_i) з поверненням і заносять його в нову вибірку. Повторюють цю операцію стільки разів, скільки елементів у вибірці (X_i, Y_i), і формують нову вибірку.
2. Для нової вибірки знаходять коефіцієнти $\hat{b}_1$ і $\hat{b}_0$ і заносять їх значення у відповідний масив.
3. Повторюють операції 1, 2 велику кількість разів, наприклад 1000.
4. Ранжують масиви $\hat{b}_1$ і $\hat{b}_0$ за збільшенням і відсікають потрібну частку найбільших і найменших елементів (наприклад, 5 % справа і 5 % зліва).
5. Крайні значення елементів, що залишились, указують на інтервальну оцінку коефіцієнтів рівняння регресії.

Типові діаграми розподілу значень коефіцієнтів у цій задачі зображено на рис. 3.34.

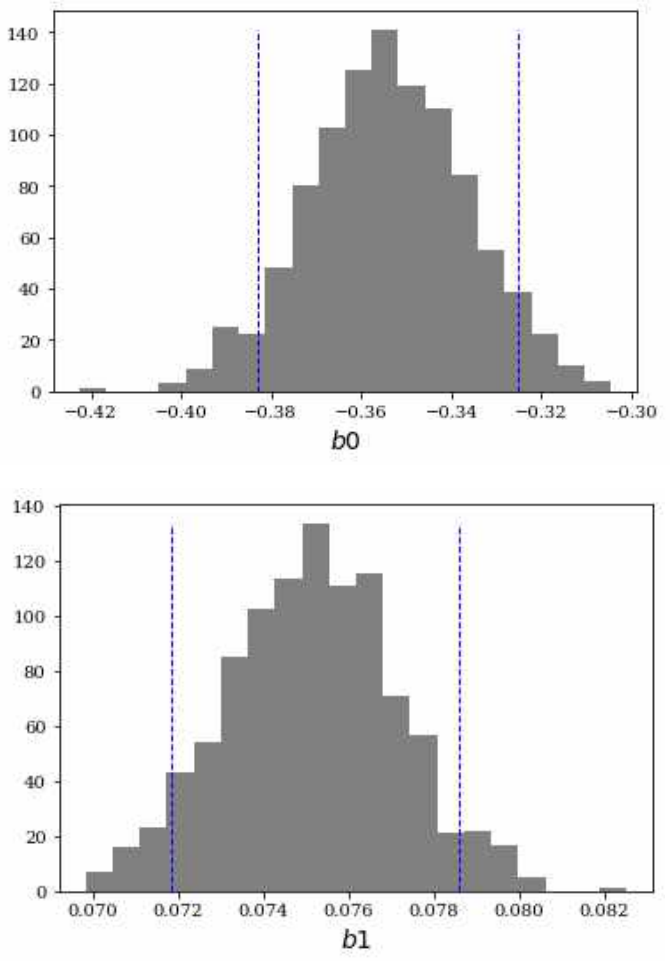


Рис. 3.34. Інтервальні оцінки вибіркових коефіцієнтів рівняння регресії $\hat{b}_1$ і $\hat{b}_0$

Маємо інтервали:

– для $\hat{b}_1$: [-0.38309673665564964, -0.3251511829537248];

– для $\hat{b}_0$: [0.07184711721191309, 0.07861336789940294].

Безумовно, в інших прогонах алгоритму ці значення будуть трохи різнитися в третьому-четвертому знаках після коми.

Які висновки можна зробити з того, що коефіцієнти $\hat{b}_1$ і $\hat{b}_0$ мають певний розподіл? З цього випливає, що рівняння регресії являє собою сім'ю прямих, а прогнозоване значення регресії $\hat{Y}$, розраховане за певним значенням аргументу X, також є випадковою величиною, яка має власний розподіл (Гаусса), математичне сподівання розподілу збігається з $\hat{Y}$, а ось середньоквадратичне відхилення збільшується при віддаленні аргументу X від вибіркового середнього $\overline{x}$. На рис. 3.35 показано діаграму розсіювання, вибіркову лінію регресії й інтервал, у який потрапляють прогнозовані значення з імовірністю 90 %.

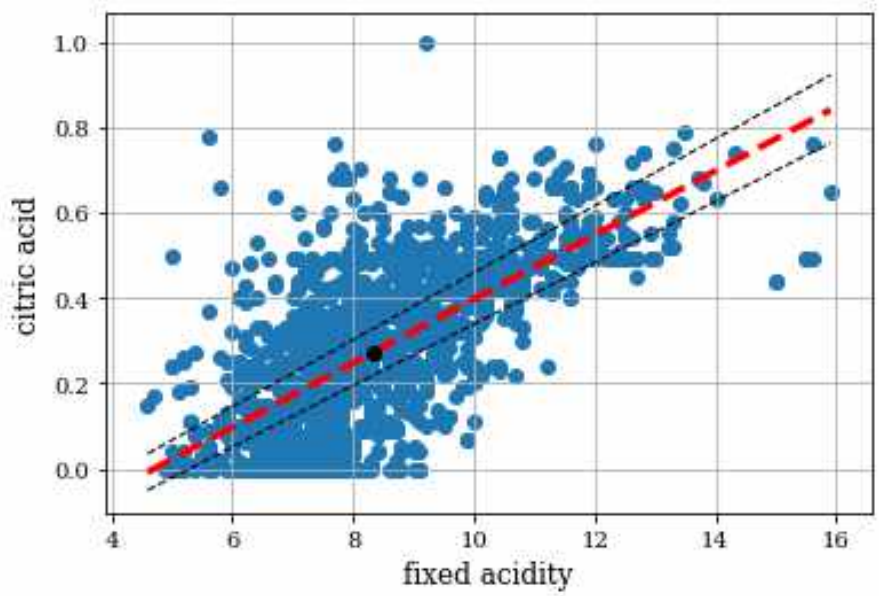


Рис. 3.35. Лінія регресії, точка $(\overline{x}, \overline{y})$ та область можливих положень лінії регресії

Як бачимо, на віддаленні від середнього значення **точність прогнозу зменшується.**

3.12. Багатофакторна регресія

3.12.1. Модель лінійної багатофакторної регресії (множинна регресія)

Якщо існує кілька предикторів, то рівняння просто розширюється для їх розміщення:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + ...b_m X_m + \varepsilon .$$

Усі інші поняття з простої лінійної регресії, як-от підгонка найменшими квадратами й визначення підігнаних значень і залишків, розширюються на множинну лінійну регресію. Наприклад, підігнані значення задаються такою формулою:

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_{1,i} + \hat{b}_2 x_{2,i} + ...\hat{b}_m x_{m,i} .$$

Тут $\hat{b}_0$, …, $\hat{b}_m$ – вибіркові коефіцієнти моделі, які визначаються за вибіркою; $\hat{y}_i$ – прогнозоване значення фактора за моделлю.

Дані $\left\{x_{k,i}\right\}$ створюють масив, до якого додається ще стовпець y_i. У такому масиві (матриці) міститься певна кількість рядків, кожен з яких утворює певний запис, або точку (вектор) у багатовимірному просторі $\left(y_i, x_{1,i}, x_{2,i}, ..., x_{m,i}\right)$.

Таким чином, дані створюють матрицю $\mathbf{X}$ і вектор $\mathbf{Y}$ з однаковою кількістю рядків:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & ... & x_{1,m} \\ x_{2,1} & x_{2,2} & ... & x_{2,m} \\ ... & ... & ... & ... \\ x_{n,1} & x_{n,2} & ... & x_{n,m} \end{pmatrix}.$$

А коефіцієнти регресії утворюють вектор $\mathbf{A} = \begin{pmatrix} b_1 \\ b_2 \\ ... \\ b_n \end{pmatrix}$.

Рівняння регресії можна записати у вигляді

$$\mathbf{Y} = \mathbf{XA} .$$

Якщо мінімізувати суму квадратів відхилення прогнозованих результатів від емпіричних, то отримаємо

$$RSS = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 \to \min ,$$

$$\frac{\partial RSS}{\partial b_0} = 0 , \ \frac{\partial RSS}{\partial b_1} = 0 , \ …. , \ \frac{\partial RSS}{\partial b_n} = 0 .$$

Розв’язок цієї системи лінійних рівнянь можна записати в матричному вигляді:

$$\mathbf{A} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}.$$

Результатом розрахунків за цією формулою буде вектор вибіркових коефіцієнтів регресії $\mathbf{A} = \begin{pmatrix} b_1 & b_2 & \dots & b_n \end{pmatrix}^T$, у якому немає коефіцієнта $\hat{b}_0$. Щоб включити до складу моделі $\hat{b}_0$, слід увести в матрицю $\mathbf{X}$ зліва стовпець одиниць:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,m} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,m} \\ & \dots & \dots & \dots & \dots \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,m} \end{pmatrix}.$$

У мовах програмування існують спеціалізовані інструменти для розв'язання задачі регресії. У мові Python існує інструмент

statsmodels.regression.linear_model.OLS(),

параметрами якого є:

- **Endog** – одновимірна змінна відклику $\mathbf{Y}$;
- **Exog** – масив $\mathbf{X}$.

Іще кілька параметрів, атрибути, методи можна знайти за посиланням: https://www.statsmodels.org/stable/generated/statsmodels.regression.linear_model.OLS.html

Факторні змінні в моделі регресії, які називають також **категоріальними змінними**, набувають граничної кількості дискретних значень. Наприклад, метою позики можуть бути «консолідація заборгованості», «весілля», «автомобіль» тощо. Бінарна (так/ні) змінна, яку називають також індикаторною змінною, є особливим випадком факторної змінної. Для регресії потребуються на вході числові дані, тому факторні змінні потрібно перекодувати, щоб їх можна було використовувати в моделі. Загальноприйнятий підхід полягає в конвертації змінної в набір двійкових фіктивних змінних.

Фіктивні змінні (dummy variables) – бінарні змінні у форматі 0–1, отримані шляхом перекодування факторних даних для використання в регресії та інших моделях.

Деякі факторні змінні відображають рівні фактора; і їх називають порядковими факторними змінними або порядковими категоріальними змінними. Наприклад, рівень позики може бути A, B, C і т. д. – кожен рівень несе більший ризик, ніж попередній. Порядкові факторні змінні зазвичай конвертуються в числові значення й використовуються як є.

У множинній регресії змінні часто корелюють одна з одною. Таке явище називають **мультиколінеарністю**.

Граничний випадок корельованих змінних приводить до

мультиколінеарності – до умови, коли існує надлишок предикторних змінних. Ідеальна мультиколінеарність трапляється, коли одна предикторна змінна може бути виражена як лінійна комбінація інших.

Мультиколінеарність спостерігається, якщо:

- змінна включається до складу моделі багаторазово помилково;
- із факторної змінної створюються P фіктивних змінних, замість $P-1$;
- дві змінні майже ідеально корелюють одна з одною.

Мультиколінеарність у регресії необхідно усувати – змінні слід прибирати доти, доки мультиколінеарність не зникне. Регресія не має чітко визначеного розв'язку за наявності ідеальної мультиколінеарності. Багатопрограмні пакети, у тому числі R, обробляють певні типи мультиколінеарності автоматично.

3.12.2. Парадокс Сімпсона

Парадокс Сімпсона (ефект Юла – Сімпсона, парадокс об'єднання) – це парадокс у статистиці, коли при об'єднанні двох груп даних, у кожній з яких спостерігається однаково спрямована залежність, ця залежність або зникає, або її напрямок змінюється на протилежний.

Наведемо приклад. Уявімо, що можна ототожнити всіх членів мережі дослідників даних зі сходу України та із заходу України. Необхідно з'ясувати, дослідники даних якого регіону є більш товариськими (табл. 3.2).

Таблиця 3.2

Регіон	Кількість користувачів	Середня кількість друзів
Західний	101	8,2
Східний	103	6,5

Очевидно, що дослідники даних із заходу України є більш товариськими. Можна зробити різні припущення, чому так відбувається: можливо, справа в Карпатах, або в каві, або в органічних продуктах, або у впливі близької Європи.

Унаслідок вивчення даних виявляється щось дуже незвичайне. Якщо враховувати тільки членів з науковим ступенем, то виявляється, що дослідники зі сходу України в середньому мають більше друзів (табл. 3.3), а якщо розглядати тільки членів без наукового ступеня, то знову виявляється що дослідники зі сходу України в середньому мають більше друзів!

Таблиця 3.3

Регіон	Науковий ступінь	Кількість користувачів	Середня кількість друзів
Західний	Є	35	3,1
Східний	Є	70	3,2
Західний	Немає	66	10,9
Східний	Немає	33	13,4

Як тільки починається враховування наукових ступенів, кореляція спрямовується в протилежний бік! Під час групування даних за ознакою «схід – захід» приховується той факт, що серед дослідників зі Східної України є сильна асиметрія в бік наукових ступенів.

Подібний феномен виникає в реальному світі з певною регулярністю. Головним моментом є те, що кореляція вимірює зв'язок між двома змінними за інших рівних умов. Якщо дані розбивати на класи випадковим чином, як і має бути в добре поставленому експерименті, то припущення «за інших рівних умов» може бути й непоганим. Єдиний реальний спосіб уникнути таких неприємностей – це знати свої дані й забезпечувати перевірку всіх можливих сплутаних факторів. Очевидно, це не завжди можливо. Якби не було даних про освіту 204 дослідників даних, то можна було б вирішити, що є щось, властиве людям із Західної України, що робить їх більш комунікабельними.

3.12.3. Статистична значущість коефіцієнтів регресії

Статистичну значущість коефіцієнтів регресії розглянемо на прикладі даних, наведених у підрозд. 3.11, – результатів хімічного аналізу червоних вин та їх тестування дегустаторами. Результати тестування подано за шкалою від 3 до 8. Хімічний аналіз проведено за 10 критеріями. Таким чином, модель регресії має вигляд

$$Y = b_0 + b_1 X_1 + b_2 X_2 + ... + b_{11} X_{11} + \varepsilon .$$

Отже, модель містить 12 коефіцієнтів $b_0,...,b_{11}$. Матрицю даних $\mathbf{X}$ треба доповнити зліва стовцем з одиниць, який відповідає коефіцієнту b_0. Матриця $\mathbf{X}$ містить 12 стовпців і 1599 рядків (табл. 3.4).

Таблиця 3.4

Y
y_1
y_2
…
y_i
…
y_{1599}

	X_1	X_2	…	X_{11}
1	$x_{1,1}$	$x_{1,2}$	…	$x_{1,11}$
1	$x_{2,1}$	$x_{2,2}$	…	$x_{2,11}$
…	…	…	…	…
1	$x_{i,1}$	$x_{i,2}$	…	$x_{i,11}$
…	…	…	…	…
1	$x_{1599,1}$	$x_{1599,2}$	…	$x_{1599,11}$

Параметри моделі регресії визначаємо за допомогою таких операцій:

lmRegModel_1 = sm.OLS(Y, X_)
result_1 = lmRegModel_1.fit()
BM = result_1.params # коефіцієнти моделі
SE = result_1.bse # стандартна похибка коефіцієнтів

Після обчислень отримуємо (**result_1.summary( )**):

OLS Regression Results

==========

Dep. Variable: y R-squared: 0.361
Model: OLS Adj. R-squared: 0.356
Method: Least Squares F-statistic: 81.35
Date: Tue, 30 Mar 2021 Prob (F-statistic): 1.79e-145
Time: 19:27:49 Log-Likelihood: -1569.1
No. Observations: 1599 AIC: 3162
Df Residuals: 1587 BIC: 3227
Df Model: 11
Covariance Type: nonrobust

==========

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	21.9652	21.195	1.036	0.300	-19.607	63.538
x1	0.0250	0.026	0.963	0.336	-0.026	0.076

x2	-1.0836	0.121	-8.948	0.000	-1.321	-0.846
x3	-0.1826	0.147	-1.240	0.215	-0.471	0.106
x4	0.0163	0.015	1.089	0.276	-0.013	0.046
x5	-1.8742	0.419	-4.470	0.000	-2.697	-1.052
x6	0.0044	0.002	2.009	0.045	0.000	0.009
x7	-0.0033	0.001	-4.480	0.000	-0.005	-0.002
x8	-17.8812	21.633	-0.827	0.409	-60.314	24.551
x9	-0.4137	0.192	-2.159	0.031	-0.789	-0.038
x10	0.9163	0.114	8.014	0.000	0.692	1.141
x11	0.2762	0.026	10.429	0.000	0.224	0.328

Omnibus:	27.376	Durbin-Watson:	1.757
Prob(Omnibus):	0.000	Jarque-Bera (JB):	40.965
Skew:	-0.168	Prob(JB):	1.27e-09
Kurtosis:	3.708	Cond. No.	1.13e+05

Коефіцієнти рівняння регресії подано у стовпці **coef**, рядки відповідають змінним. Як бачимо, критерій R^2 дорівнює 0,361 (R-squared).

Метод **result_1.predict()** дає масив результатів розрахунку $\hat{y}_i$ за моделлю. На рис. 3.36 показано гістограму розподілу залишків $y_i - \hat{y}_i$.

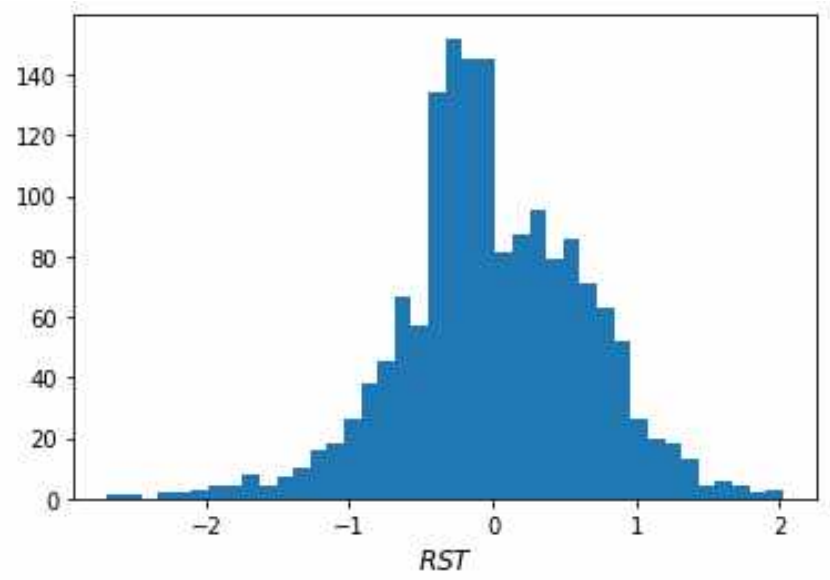


Рис. 3.36. Гістограма залишків RST

Важливо, щоб розподіл залишків був нормальним, тобто підпорядковувався розподілу Гаусса. Для оцінювання гіпотези про вид розподілу можна застосувати діаграму Q–Q та критерій Пірсона. Із того, що розподіл залишків є нормальним, випливає можливість застосувати критерій Стьюдента для оцінювання значущості коефіцієнтів моделі регресії.

Слід розуміти, що отримані коефіцієнти регресії є випадковими величинами. Можна сказати, що за вибіркою знайдено точкові оцінки їх значень. Але зрозуміло, що діючи формально, за формулами можна знайти ці коефіцієнти у будь-якому випадку, навіть якщо немає явної залежності між $\mathbf{X}$ та $\mathbf{Y}$. Модель не буде працювати. Та може статися так, що деякі фактори впливають на $\mathbf{Y}$, а деякі – ні, хоча й входять до складу моделі. Як же відсортувати фактори за значущістю? Як відокремити важливі фактори від випадкових, що не впливають на $\mathbf{Y}$ і потрапили до задачі помилково?

У наведеній вище таблиці бачимо стандартну похибку коефіцієнтів (std err), яка являє собою середньоквадратичне відхилення відповідного коефіцієнта й розраховується виходячи з гіпотези про нормальний розподіл залишків. Якщо поділити коефіцієнт на його середньоквадратичне відхилення, то отримаємо випадкову величину, яка має розподіл Стьюдента

$$t_i = \frac{|b_i|}{se_i}.$$

Ці дані можна розрахувати самостійно, використавши знайдені вище параметри **BM[i]/SE[i]**, а можна взяти з таблиці.

Якщо

$$t_i = \frac{|b_i|}{se_i} > t_{\alpha,n-m-1},$$

то коефіцієнт є значущим, і гіпотезу про його рівність нулю можна відкинути.

Якщо ж

$$t_i = \frac{|b_i|}{se_i} < t_{\alpha,n-m-1},$$

то гіпотеза про те, що b_i з імовірністю α дорівнює нулю, є правильною.

Тобто у цьому випадку відповідний фактор можна виключити з моделі регресії і зменшити кількість факторів. Тут $t_{\alpha,n-m-1}$ – граничне значення розподілу Стьюдента, яке забезпечує ймовірність α; n – кількість значень у моделі (1599); m – кількість параметрів у моделі (тут 12). Це значення можна розрахувати самостійно:

_ ,ttest = stats.t.interval(0.95, NN - 12 - 1, loc = 0, scale = 1)

Отримуємо ttest = 1.9614608646662874. Узагалі, ці ж результати бачимо в таблиці, у стовпці «**P>|t|**». Рядки, у яких елементи позначено червоним (0.000), відповідають змінним і певним коефіцієнтам моделі регресії, які слід залишити в моделі і які впливають на $\mathbf{Y}$. Інші змінні не

мають статистично визначеного впливу на прогноз $\mathbf{Y}$.

Зауваження. Під час виключення з моделі зайвих змінних не слід відкидати разом усі змінні, а виключати їх по одній, починаючи з тих, у яких t_i є найбільшим, і перераховувати показники моделі на кожному кроці. Може статися, що після виключення декількох (або одного) факторів у моделі всі фактори стануть значущими.

Метод Bootstrap – ефективний інструмент, який не має обмежень гіпотези про нормальний розподіл залишків регресійної моделі. За допомогою цього методу можна знайти розподіл коефіцієнтів моделі $b_0,...,b_{11}$.

Алгоритм:

1. Із масиву $\mathbf{X}$ та вектора $\mathbf{Y}$ беруть випадковим чином з поверненням елементи з однаковими індексами (рядок із випадковим індексом i у табл. 3.4 позначено темним фоном) і створюють із них нову вибірку $\mathbf{X}^*$ та $\mathbf{Y}^*$ тієї ж розмірності, що й задано в задачі.
2. Для нової вибірки розраховують коефіцієнти моделі регресії $b_0,...,b_{11}$ і заносять їх у відповідні масиви.
3. Пункти 1 і 2 виконують велику кількість разів (декілька тисяч).
4. Отримавши статистичні ряди коефіцієнтів $b_0,...,b_{11}$, розраховують для цих вибірок відповідні середньоквадратичні відхилення й будують гістограми розподілу.
5. Відкинувши з масивів значень коефіцієнтів $b_0,...,b_{11}$ певну кількість найменших і найбільших коефіцієнтів (наприклад, 2,5 % зліва та 2,5 % справа, залишивши 95 % значень), будують гістограми. Якщо у знайдені інтервали потрапив нуль, то відповідний коефіцієнт можна прирівняти до нуля й виключити з моделі.

На рис. 3.37 показано результати розрахунків моделювання за методом Bootstrap. Виконано 5000 випробувань.

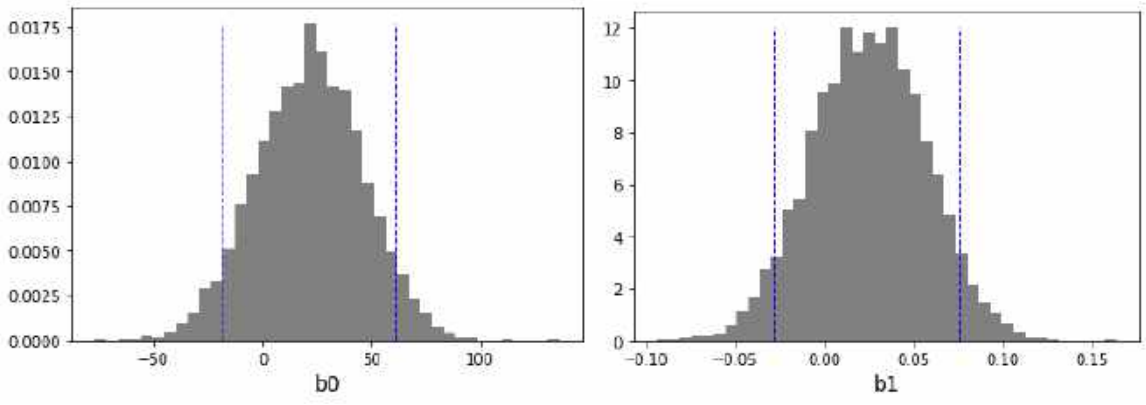


Рис. 3.37. Розподіл значень коефіцієнтів $b_0,...,b_{10}$

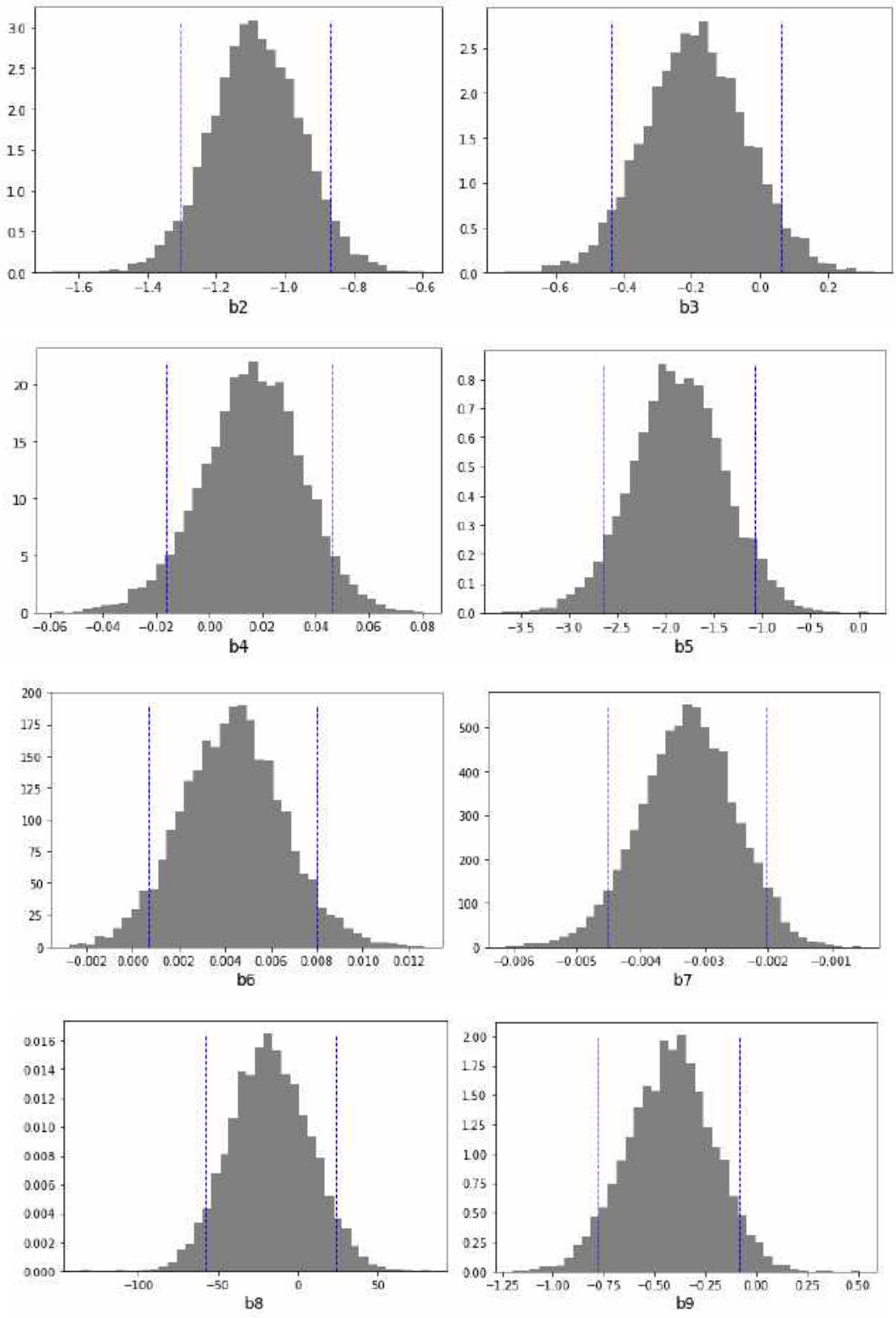


Рис. 3.37. Продовження

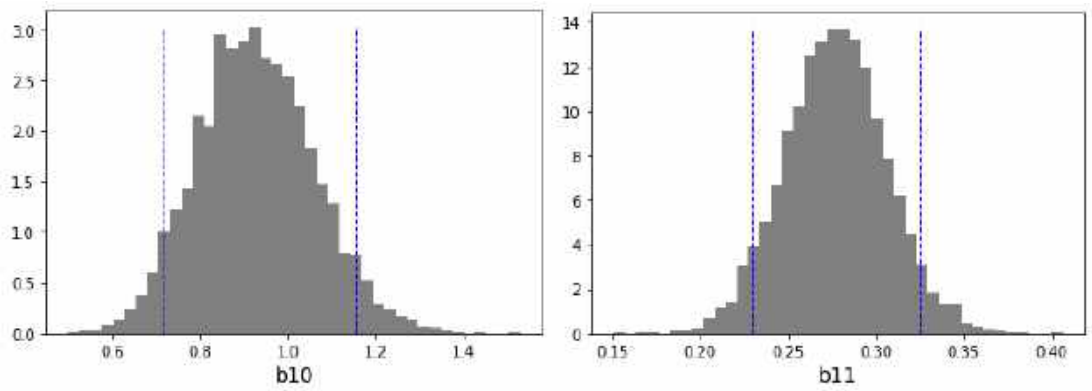


Рис. 3.37. Закінчення

Як бачимо, нуль потрапив у центральний проміжок між коефіцієнтами 0, 1, 3, 4 та 8, що цілком збігається з даними таблиці результатів роботи процедури **OLS** (Regression Results). Але цей підхід не базується на гіпотезі про нормальний розподіл залишків, що може бути корисним.

Якщо розрахувати середньоквадратичне відхилення наведених розподілів (за допомогою, наприклад, **BB[i].std(ddof = 1)**, де **BB[i]** – варіаційні ряди відповідного коефіцієнта (**i=0, …, 11**), які містять у цьому випадку 5000 елементів), то можна помітити, що вони з похибкою 1...10 % збігаються зі знайденими вище результатами **SE = result_1.bse** та стовпцем **std err**.

Для покращання та спрощення моделі з неї слід виключити змінні, коефіцієнти при яких не є значущими. Змінними є величини з індексами 0 (вільний член), 1, 3, 4 та 8.

Діяти треба послідовно, виключаючи найбільш незначущі фактори й відповідні коефіцієнти з моделі, і здійснювати перерахунок моделі на кожному кроці. Найменші значення t-критерію мають змінні з індексами 8 та 1. Виключивши їх, бачимо, що треба ще виключити змінні з індексами 3, а потім і 4. Таким чином, у моделі залишається вільний член b_0, хоча перший розрахунок показав, що цей коефіцієнт не є значущим. Але після відкидання декількох інших факторів модель змінилася і цей коефіцієнт став значущим. Виконавши три ітерації, отримуємо модель, у якій усі коефіцієнти є значущими:

OLS Regression Results

Dep. Variable:	y	R-squared:	0.359
Model:	OLS	Adj. R-squared:	0.357
Method:	Least Squares	F-statistic:	127.6
Date:	Wed, 31 Mar 2021	Prob (F-statistic):	5.32e-149
Time:	16:01:48	Log-Likelihood:	-1570.5
No. Observations:	1599	AIC:	3157
Df Residuals:	1591	BIC:	3200

Df Model: 7
Covariance Type: nonrobust

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.4301	0.403	10.995	0.000	3.640	5.220
x1	-1.0128	0.101	-10.043	0.000	-1.211	-0.815
x2	-2.0178	0.398	-5.076	0.000	-2.798	-1.238
x3	0.0051	0.002	2.389	0.017	0.001	0.009
x4	-0.0035	0.001	-5.070	0.000	-0.005	-0.002
x5	-0.4827	0.118	-4.106	0.000	-0.713	-0.252
x6	0.8827	0.110	8.031	0.000	0.667	1.098
x7	0.2893	0.017	17.225	0.000	0.256	0.322

Omnibus:	24.204	Durbin-Watson:	1.750
Prob(Omnibus):	0.000	Jarque-Bera (JB):	35.245
Skew:	-0.156	Prob(JB):	2.22e-08
Kurtosis:	3.657	Cond. No.	1.71e+03

Таким чином, в отриманій моделі регресії критерій R^2 зменшився зовсім мало – з 0,361 до 0,359. Але з моделі виключено фактори X_8 (*density*), X_1 (*fixed acidity*), X_3 (*citric acid*) та X_4 (*residual sugar*), що не мають статистичного впливу на оцінку вина, яку дали дегустатори. Під час відновлення назв факторів слід бути уважними, тому що початкова їх нумерація змінилася через уведення константи.

Фактори, які залишилися в моделі:

- базова оцінка, константа ($b_0 = 4.4301$);
- *volatile acidity* ($b_1 = -1.0128$);
- *chlorides* ($b_2 = -2.0178$);
- *free sulfur dioxide* ($b_3 = 0.0051$);
- *total sulfur dioxide* ($b_4 = -0.0035$);
- *pH* ($b_5 = -0.4827$);
- *sulphates* ($b_6 = 0.8827$);
- *alcohol* ($b_7 = 0.2893$).

Знак і величина коефіцієнта вказують на напрямок і величину впливу відповідного фактора на оцінку вина Y. Як бачимо, алкоголь та сульфати (не плутати з сульфітами та діоксидом сірки!!!) позитивно впливають на якість вина, а хлориди та pH – негативно.

3.13. Парна кореляція

Складання рівняння множинної регресії починається з вирішення питання про специфікацію моделі, що містить два кола питань: відбір факторів і вибір вигляду рівняння регресії. Уведення в рівняння множинної регресії того чи іншого набору факторів пов'язане насамперед з уявленням дослідника про природу взаємозв'язку між показником Y та іншими факторами. Фактори, що включаються в множинну регресію, мають відповідати таким вимогам:

1. Фактори мають бути кількісно вимірними. Якщо необхідно включити в модель якісний фактор, який не має кількісного виміру, то йому потрібно надати кількісну визначеність.
2. Фактори не мають бути інтеркорельованими і, тим більше, не можуть бути в точному функціональному зв'язку.

Уведення в модель факторів з високою кореляцією може призвести до небажаних наслідків – система нормальних рівнянь може виявитися погано обумовленою й спричинити нестійкість і ненадійність оцінок коефіцієнтів регресії. Якщо між факторами існує висока кореляція, то не можна визначити їх ізольований вплив на результівний показник, і параметри рівняння регресії неможливо інтерпретувати.

Для оцінювання статистичного зв’язку між двома вибірками X_i і X_j застосовується вибірковий коефіцієнт кореляції

$$r_{i,j} = \frac{\sum\left(X - \bar{X}\right)\left(Y - \bar{Y}\right)}{\sqrt{\sum\left(X - \bar{X}\right)^2}\sqrt{\sum\left(Y - \bar{Y}\right)^2}}.$$

Якщо розрахувати ці коефіцієнти для всіх пар факторів, то можна отримати симетричну матрицю, на головній діагоналі якої стоять одиниці. Це зумовлено тим, що $r_{i,j} = r_{j,i}$ і $r_{i,i} = 1$.

У мові програмування Python розрахунок матриці здійснюється за допомогою процедури

CORR = np.corrcoef(XC,XC).

Детальніше – за посиланням: https://numpy.org/doc/stable/reference/generated/numpy.corrcoef.html

Якщо розрахувати матрицю для масиву даних winequality-red.csv, який містить дані про 1599 вин за 11 факторами, то можна отримати матрицю 11×11 (рис. 3.38):

plt.pcolormesh(CORR, edgecolors="k", shading='flat')

Матрицю також можна розрахувати за допомогою бібліотеки **Seaborn, heatmap**, де **data** – це дані у форматі DataFrame (рис. 3.39).

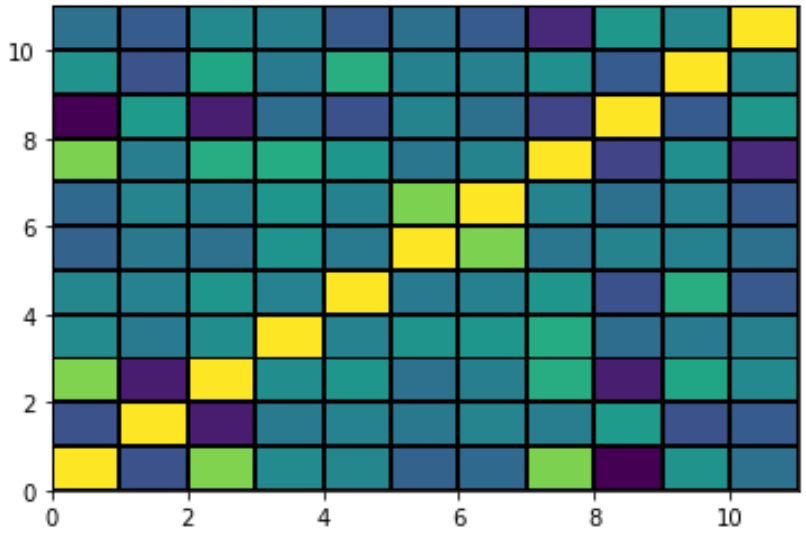


Рис. 3.38. Візуалізація матриці парних кореляцій

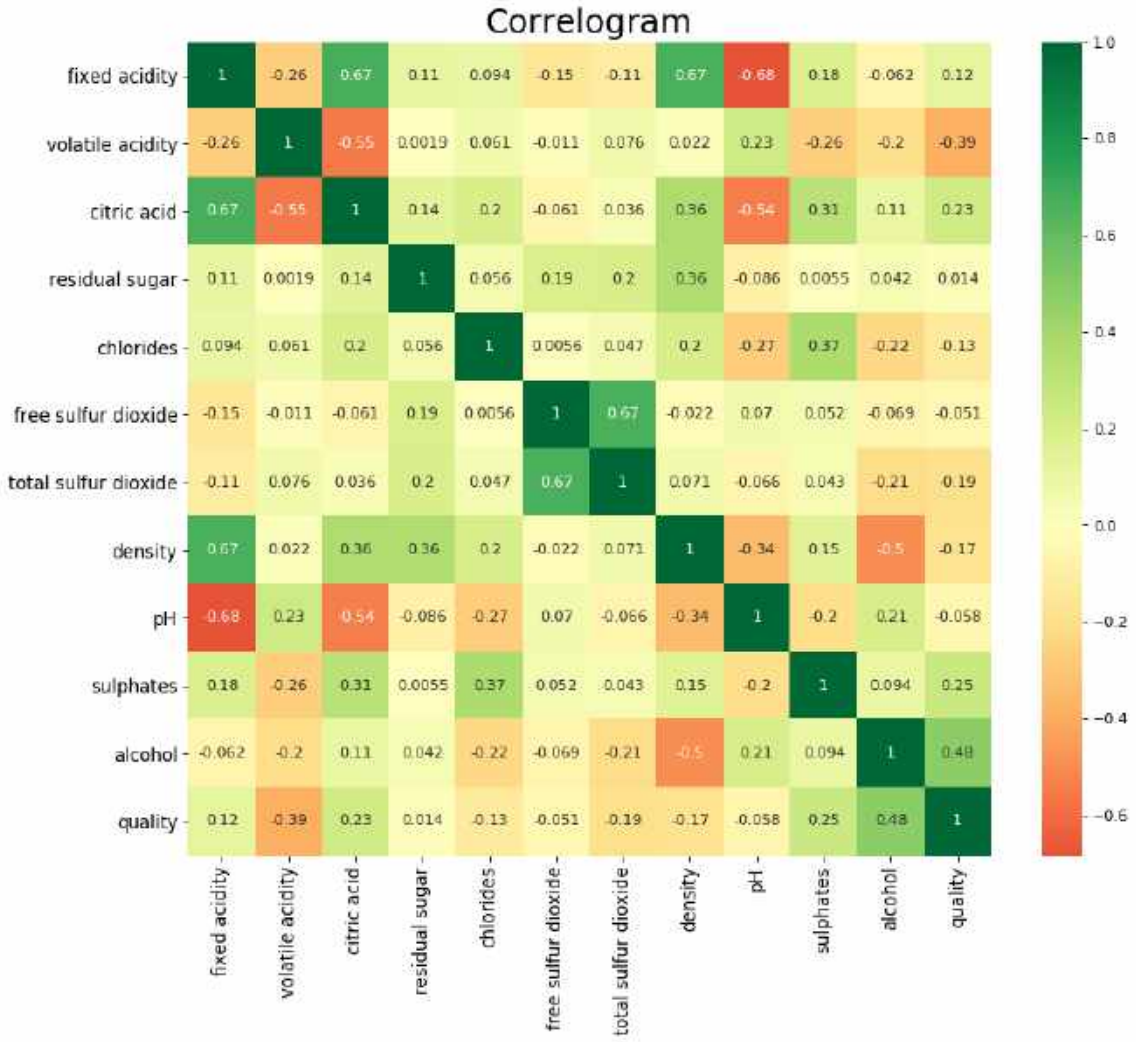


Рис. 3.39. Візуалізація матриці парних кореляцій за допомогою бібліотеки **Seaborn, heatmap**

```
plt.figure(figsize=(12,10), dpi= 80)
sns.heatmap(data.corr(), xticklabels=data.corr().columns,
        yticklabels=data.corr().columns, cmap='RdYlGn', center=0,
annot=True)
# Decorations
plt.title('Correlogram', fontsize=22)
plt.xticks(fontsize=12)
plt.yticks(fontsize=12)
plt.show()
```

Як бачимо, існує висока кореляція між факторами *free sulfur dioxide* та *total sulfur dioxide*, що є цілком зрозумілим, також між *pH* та *fixed acidity*, що також є зрозумілим і може бути прогнозованим.

3.14. Багатовимірний нормальний розподіл

Багатовимірний нормальний розподіл (або багатовимірний гавсів розподіл) у теорії ймовірностей – це узагальнення одновимірного нормального розподілу для випадку із багатьма вимірами. Відповідно до одного із означень вектор випадкових величин має k-варіативний нормальний розподіл, якщо кожна лінійна комбінація його k компонент має одновимірний нормальний розподіл. В основному його важливість випливає із узагальнення центральної граничної теореми для багатьох вимірів. Багатовимірний нормальний розподіл часто використовують, щоб описати, принаймні наближено, будь-яку множину (можливо) корельованих випадкових величин із дійсними значеннями, кожна з яких скупчується навколо середнього значення.

У випадку двовимірного розподілу цей розподіл можна подати як деякий еліпс розсіювання (рис. 3.40).

У випадку двовимірного розподілу щільність описується залежністю

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\mathrm{e}^{-\frac{1}{2\left(1-\rho^2\right)}\left[\left(\frac{x-m_x}{\sigma_x}\right)^2-2\rho\left(\frac{x-m_x}{\sigma_x}\right)\left(\frac{y-m_y}{\sigma_y}\right)+\left(\frac{y-m_y}{\sigma_y}\right)^2\right]},$$

де ρ – коефіцієнт кореляції між X та Y; m_x, m_y – координати математичного сподівання розподілу, або математичні сподівання маргінальних розподілів; σ_x, σ_y – дисперсія факторів X та Y ($\sigma_x > 0, \sigma_y > 0$).

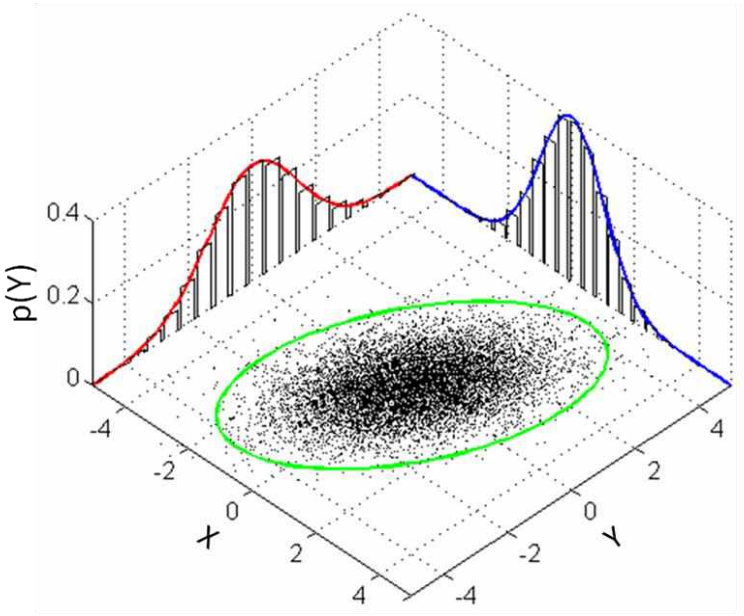


Рис. 3.40. Приклад двовимірного нормального розподілу

У матричній формі, яку можна узагальнити на довільну кількість вимірів, нормальний розподіл має вигляд

$$f\left(x_1, x_2, ..., x_k\right) = \frac{1}{\sqrt{\left(2\pi\right)^k \det\left(\boldsymbol{\Sigma}\right)}} \mathrm{e}^{-\frac{1}{2}\left[\left(\mathbf{x}-\mathbf{m}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\mathbf{m}\right)\right]},$$

де $\boldsymbol{\Sigma}$ – матриця коваріацій (коваріаційна матриця), $\boldsymbol{\Sigma} = \mathrm{cov}\left(X_i, X_j\right)$, де $1 \le i, j \le k$; $\det\left(\boldsymbol{\Sigma}\right)$ – визначник матриці $\boldsymbol{\Sigma}$; $\mathbf{x} = \left(x_1, x_2, ..., x_k\right)^T$ – вектор змінних; $\mathbf{m} = \left(m_1, m_2, ..., m_k\right)^T$ – вектор середніх значень змінних.

Для двовимірного розподілу

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \mathbf{m} = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}.$$

Вплив дисперсій на вигляд діаграми розсіювання двовимірного розподілу з нульовими математичними сподіваннями та фіксованим коефіцієнтом кореляції показано на рис. 3.41.

Кут нахилу лінії регресії залежить не тільки від коефіцієнта кореляції, а й від відношення $\frac{\sigma_y}{\sigma_x}$.

На рис. 3.42 зображено діаграми розсіювання (у кореляційному аналізі їх називають кореляційними полями) двовимірних нормальних розподілів при різних значеннях коефіцієнта кореляції та фіксованих дисперсіях $\sigma_x = \sigma_y = 1$.

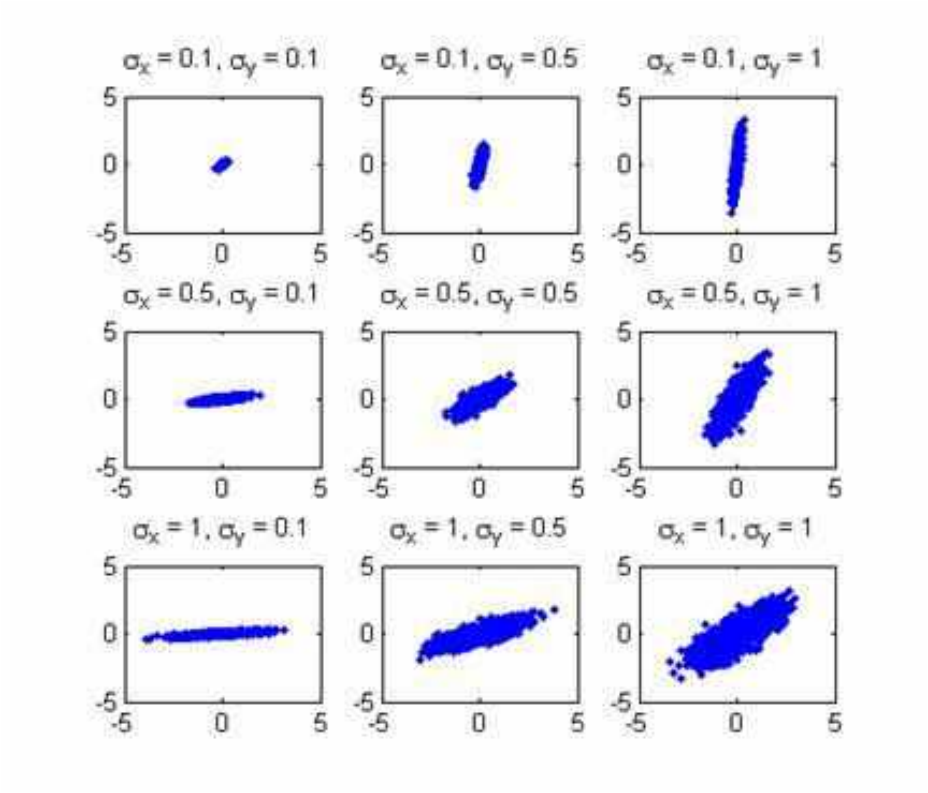


Рис. 3.41. Діаграми розсіювання

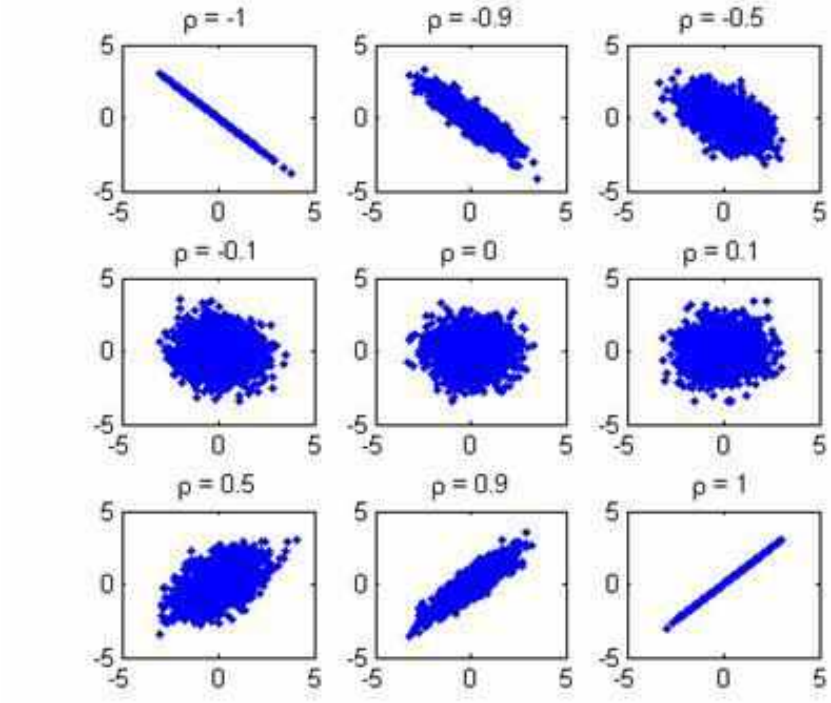


Рис. 3.42. Діаграми розсіювання двовимірного розподілу з нульовими математичними сподіваннями та фіксованим коефіцієнтом кореляції

На рис. 3.43 показано коваріаційну матрицю для даних щодо червоних вин.

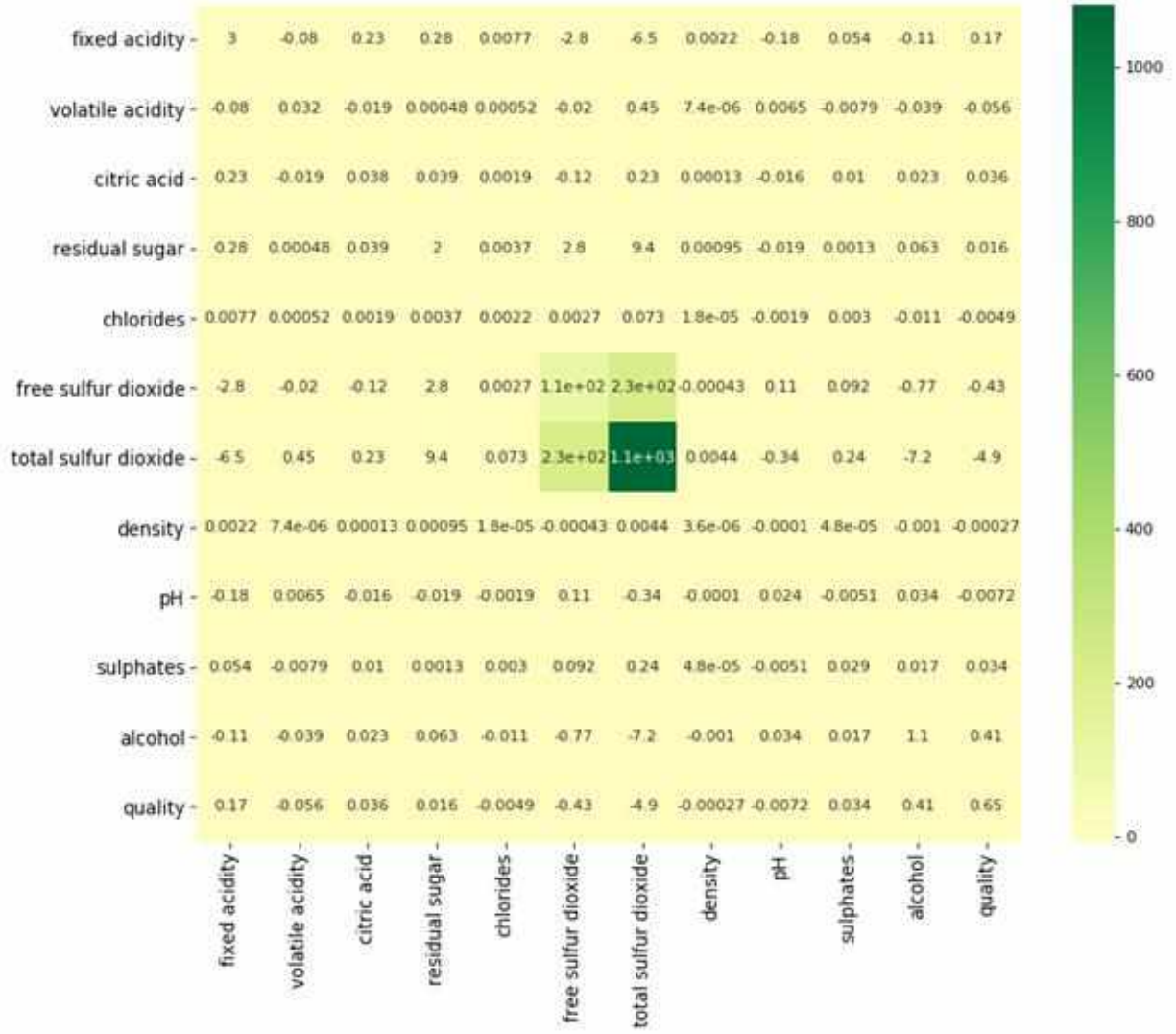


Рис. 3.43. Коваріаційна матриця

Вектор середніх значень $\mathbf{m}$:

fixed acidity	8.319637
volatile acidity	0.527821
citric acid	0.270976
residual sugar	2.538806
chlorides	0.087467
free sulfur dioxide	15.874922
total sulfur dioxide	46.467792
density	0.996747
pH	3.311113
sulphates	0.658149
alcohol	10.422983
quality	5.636023

3.15. Метод головних компонент

Метод головних компонент (МГК) — один із найпоширеніших методів факторного аналізу. Серед інших подібних методів, що дають змогу узагальнювати значення елементарних ознак, МГК вирізняється простою логічною конструкцією, і, разом з тим, на його прикладі стають зрозумілими загальна ідея й мета числових методів факторного аналізу.

Метод головних компонент дає можливість за кількістю початкових ознак *m* виокремити *r* головних компонент, або узагальнених ознак. Простір головних компонент – ортогональний. Математична модель методу головних компонент базується на логічному припущенні, що значення множини взаємозалежних ознак породжують деякий загальний результат.

Розв'язування задачі методом головних компонент зводиться до поетапного перетворення матриці початкових даних $\mathbf{X}$.

Задача аналізу головних компонент має щонайменше чотири базові версії:

- апроксимація даних лінійними множинами меншої розмірності;
- визначення підпросторів меншої розмірності, в ортогональній проєкції на які розкид даних (тобто середньоквадратичне відхилення від середнього значення) є максимальним;
- знаходження підпросторів меншої розмірності, в ортогональній проєкції на які середньоквадратична відстань між точками є максимальною;
- побудова для багатовимірної випадкової величини такого ортогонального перетворення координат, унаслідок якого кореляції між окремими координатами перетворюються на нуль.

Перші три версії оперують скінченними множинами даних, є еквівалентними і не використовують жодної гіпотези про статистичне породження даних. Четверта версія оперує випадковими величинами. Скінченні множини виникають тут як вибірки з певного розподілу, а розв'язки трьох перших задач — як наближення до розкладу за теоремою Кархунена — Лоєва (істинного перетворення Кархунена – Лоєва). При цьому виникає додаткове і не цілком тривіальне питання про точність цього наближення.

Можна сформулювати постановку задачі так: необхідно знайти лінійне перетворення даних таким чином, щоб у нових координатах дані були некорельованими, тобто матриця парних кореляцій була діагональною.

Розглянемо геометричне трактування методу головних компонент. На рис. 3.44 зображено поле розсіювання вхідних двовимірних даних.

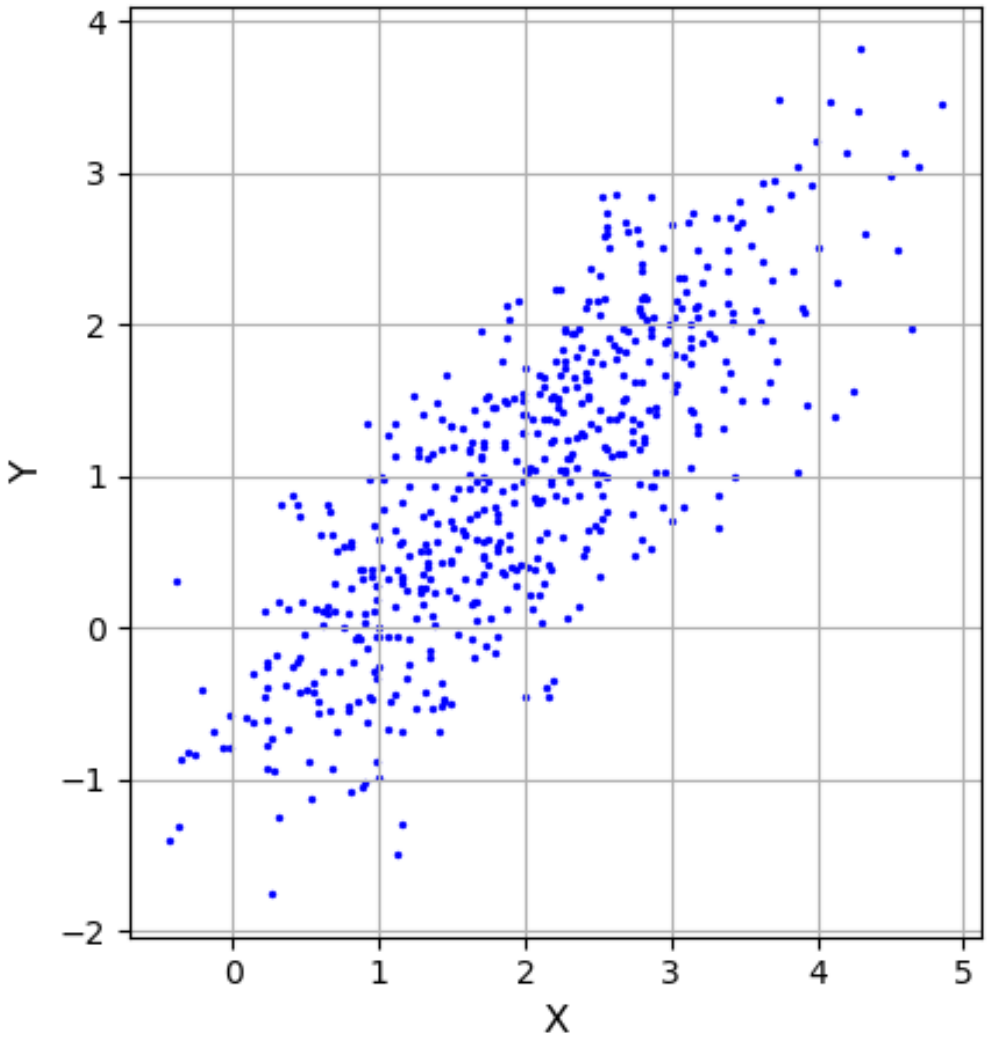


Рис. 3.44. Вхідні двовимірні дані

На першому етапі (рис. 3.45) робимо паралельний перенос системи координат таким чином, щоб вибіркові середні за обома координатами дорівнювали нулеві:

$$x_i^{(1)} = x_i - \overline{x};$$

$$y_i^{(1)} = y_i - \overline{y},$$

де $\overline{x}$ та $\overline{y}$ – вибіркові середні; $x_i^{(1)}$ і $y_i^{(1)}$ – нові змінні.

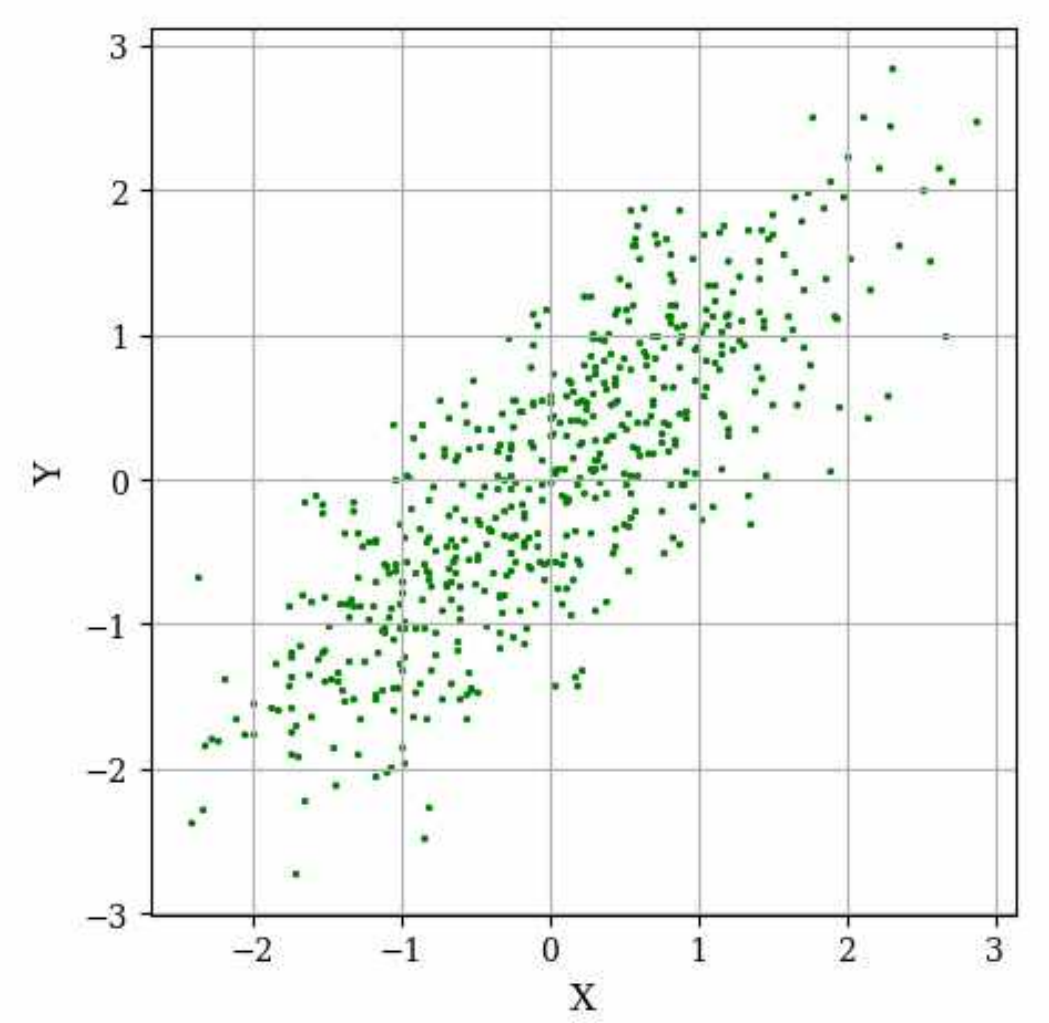


Рис. 3.45. Дані після центрування

На другому етапі проводимо поворот системи координат. Лінійне перетворення координат у матричній формі має вигляд

$$\mathbf{X}_i^{(2)} = \mathbf{H}^{-1}\,\mathbf{X}_i^{(1)},$$

де $\mathbf{H}$ – матриця ортогонального перетворення; $\mathbf{X}_i^{(1)} = \begin{pmatrix} x_i^{(1)} \\ y_i^{(1)} \end{pmatrix}$, $\mathbf{X}_i^{(2)} = \begin{pmatrix} x_i^{(2)} \\ y_i^{(2)} \end{pmatrix}$ – координати однієї й тієї ж точки у двох системах координат.

Матриця ортогонального перетворення координат має бути такою, щоб у змінних у новій системі координат не було кореляції, тобто матриця коваріацій має бути діагональною.

Елементи матриці коваріацій розраховують за формулою

$$c_{i,j} = \frac{1}{n}\sum_{k=1}^{n}\left(X_k^{(i)} - \bar{X}^{(i)}\right)\left(X_k^{(j)} - \bar{X}^{(j)}\right).$$

У Python можна застосовувати **MCov = np.cov(X1,Y1).**

З лінійної алгебри добре відомо, що матриця буде мати діагональний вигляд, якщо матриця лінійного перетворення системи координат буде складатися із власних векторів матриці.

Для пошуку власної матриці можна застосувати команду

HS = np.linalg.eig(MCov).

Тут **HS[0]** – вектор власних значень матриці **MCov**; **HS[1]** – матриця, що складається із відповідних власних векторів.

Зауваження. Матриця ортогонального перетворення визначається неоднозначно, з точністю до знака власних векторів і порядку їх розташування в матриці. Це відповідає напрямкам осей координат і найменуванню осей координат при перетворенні. Кожна вісь може мати два протилежні напрямки (найменування X2 або Y2 відповідно):

```
H = HS[1]
H_inv = np.linalg.inv(H)   # обернена матриця
X2 = H_inv[0,0]*X1 + H_inv[0,1]*Y1
Y2 = H_inv[1,0]*X1 + H_inv[1,1]*Y1
```

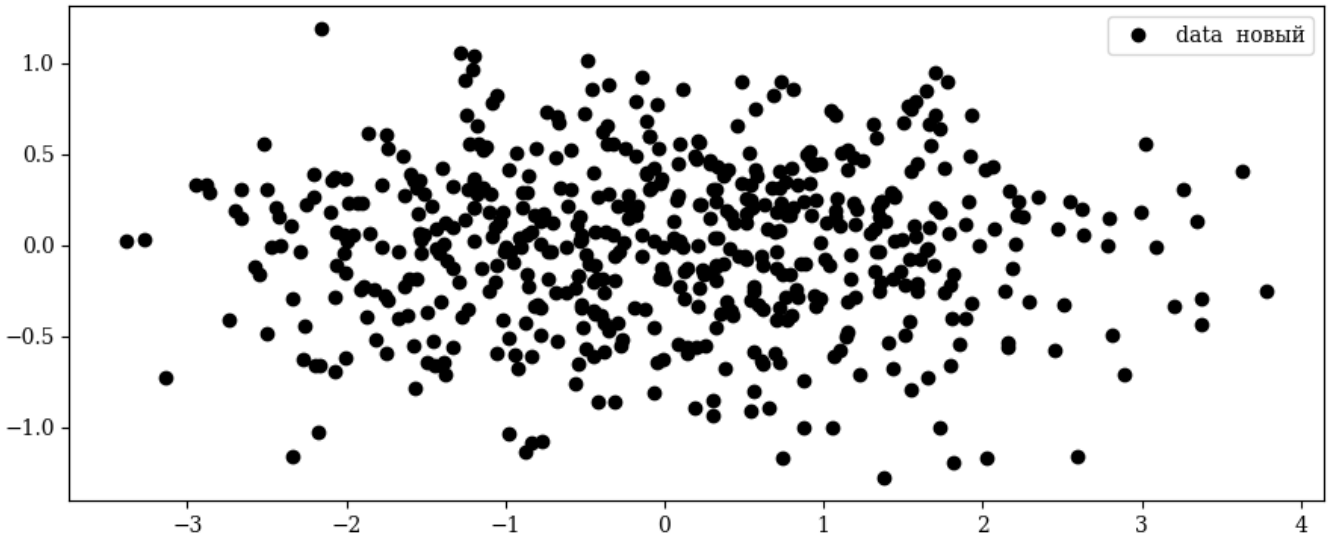


Рис. 3.46. Перетворені дані

У новій, повернутій, системі координат змінні X2, Y2 не є корельованими (рис. 3.46).

БІБЛІОГРАФІЧНИЙ СПИСОК

Bruce, P. Practical Statistics for Data Scientists / P. Bruce, A. Bruce, P. Gedeck. – Sebastopol : O'Reilly Media, Inc., 2020. – 363 p.

Grus, J. Data Science from Scratch / J. Grus. – Sebastopol : O'Reilly Media, Inc., 2015. – 656 p.

McKinney, W. Python for Data Analysis / W. McKinney – Sebastopol : O'Reilly Media, Inc., 2013. – 470 p.

Müller, Andreas C. Introduction to Machine Learning with Python A Guide for Data Scientists / Andreas C. Müller and Sarah Guido. – Published by O'Reilly Media, Inc., 2017. – 378 p.

Publishing, AI. Statistics for beginners in data science theory and applications of essential statistics concepts using python/ AI Publishing, 2020. –197 p.

Wolf, A. The Machine Learning Simplified: A Gentle Introduction to Supervised Learning Kindle Edition / A. Wolf. 2022. – 199 p.

Бідюк, П. І. Прикладна статистика / П. І. Бідюк, О. М. Терентьєв, Т. І. Просянкіна-Жарова. – Вінниця : ПП «ТД Едельвейс і К», 2013. – 304 с.

Практичний курс вищої математики : навч. посіб. У 4 кн. Кн. 1. Лінійна алгебра та аналітична геометрія. Диференціальне числення функцій однієї та декількох змінних / І. В. Брисіна [та ін.]. – Харків : ХАІ, 2004. – 355 с.

Практичний курс вищої математики : навч. посіб. У 4 кн. Кн. 3. Ряди. Інтеграл Фур'є. Функції комплексної змінної та операційне числення. Теорія ймовірностей і математична статистика / І. В. Брисіна [та ін.]. – Харків : ХАІ, 2004. – 228 с.

Навчальне видання

Курєннов Сергій Сергійович
Барахов Костянтин Петрович

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ

Редактор Т. О. Іващенко

Зв. план, 2024
Підписано до видання 23.07.2024
Ум. друк. арк. 7,3. Обл.-вид. арк. 8,25. Електронний ресурс

Видавець і виготовлювач
Національний аерокосмічний університет ім. М. Є. Жуковського
«Харківський авіаційний інститут»
61070, Харків-70, вул. Чкалова, 17
http://www.khai.edu
Видавничий центр «ХАІ»
61070, Харків-70, вул. Чкалова, 17
izdat@khai.edu

Свідоцтво про внесення суб'єкта видавничої справи
до Державного реєстру видавців, виготовлювачів і розповсюджувачів
видавничої продукції сер. ДК № 391 від 30.03.2001

С. С. Курєннов, К. П. Барахов

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ

2024