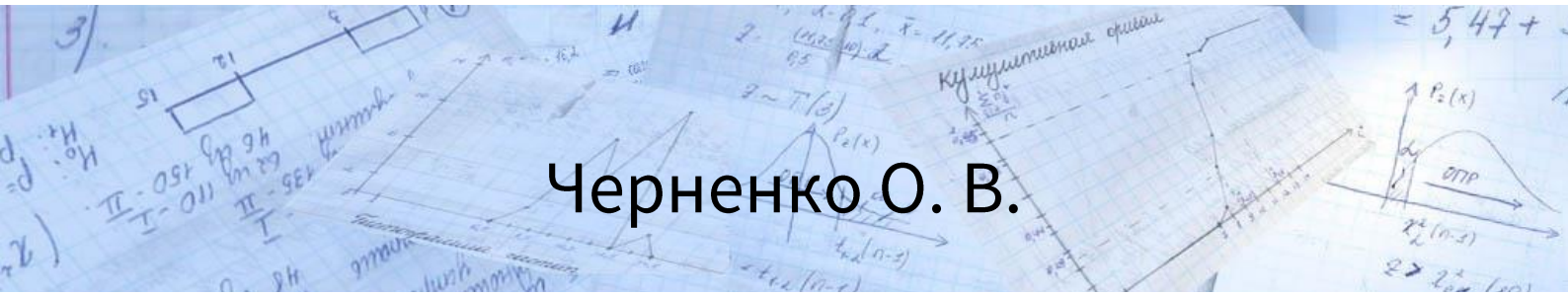
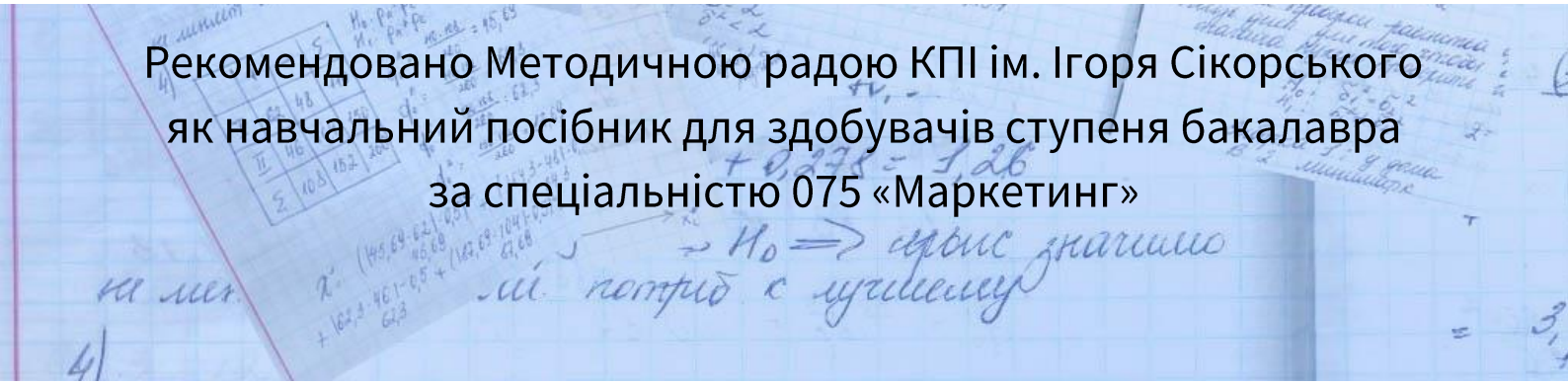
МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»



Черненко О. В.

СТАТИСТИКА:

навчально-методичний комплекс



Рекомендовано Методичною радою КПІ ім. Ігоря Сікорського
як навчальний посібник для здобувачів ступеня бакалавра
за спеціальністю 075 «Маркетинг»

MARKETING
KPI

Київ
КПІ ім. Ігоря Сікорського
2020

Статистика: Навчально-методичний комплекс [Електронний ресурс] : навч. посіб. для здоб. ступ. бакалавра за спец. 075 «Маркетинг» / уклад.: О. В. Черненко. Електронні текстові дані (1 файл: 21,5 Мбайт). Київ : КПІ ім. Ігоря Сікорського, 2020. 135 с.

Гриф надано
Методичною радою КПІ ім. Ігоря Сікорського (протокол № 5 від 14.01.2021 р.)
за поданням Вченої ради Факультету менеджменту та маркетингу
(протокол № 5 від 14.12.2020 р.)

Електронне навчальне видання

СТАТИСТИКА: НАВЧАЛЬНО-МЕТОДИЧНИЙ КОМПЛЕКС

Укладач: Черненко Оксана Володимирівна, канд. екон. наук

Відповідальний редактор: Солнцев С. О., д-р фіз.-мат. наук, проф.

Рецензент: Капустян В. О., д-р фіз.-мат. наук, проф.

Навчальний посібник містить методичні рекомендації для вивчення дисципліни «Статистика» здобувачів ступеня бакалавра за спеціальністю 075 «Маркетинг». Містить інформацію щодо змісту, методики опанування дисципліни, матеріали для роботи на практичних заняттях, написання модульної контрольної роботи, розрахункової роботи та підготовки до екзамену.

ЗМІСТ

ВСТУП

Призначення навчальної дисципліни полягає в ознайомленні студентів із методами збирання, обробки та аналізу інформації про масові соціально-економічні явища та процеси; формуванні теоретичних знань і практичних навичок для аналізу та прогнозування макроекономічних показників, проведення розрахунків окремих соціально-економічних показників, побудови конкретних статистичних моделей.

Мета дисципліни:
Мета дисципліни полягає в ознайомленні студентів із методами збирання, обробки та аналізу інформації про масові соціально-економічні явища та процеси.

Предмет дисципліни:
Розміри і кількісні співвідношення масових суспільних явищ, закономірності їх формування, розвитку та взаємозв'язку.

Навіщо це потрібно студенту?
У своїй діяльності менеджер-маркетолог стикається з великою кількістю числової інформації. Йому треба зібрати та систематизувати, коротко і інформативно відобразити у наглядних графіках та таблицях і, нарешті, використати цю інформацію для можливого прогнозу чи пошуку найкращого рішення. Вирішенню цих задач допомагають статистичні методи. Знання статистичних концепцій та методів набуває з часом все більшого значення для маркетолога.

Вивчення дисципліни дозволить сформувати у студента такі **програмні результати навчання**:
- застосовувати набуті теоретичні знання для розв’язання практичних завдань у сфері маркетингу, у тому числі на промисловому та споріднених ринках;
- збирати та аналізувати необхідну інформацію, розраховувати економічні та маркетингові показники, обґрунтовувати управлінські рішення на основі використання необхідного аналітичного й методичного інструментарію;
- використовувати цифрові інформаційні та комунікаційні технології, а також програмні продукти, необхідні для належного провадження маркетингової діяльності та практичного застосування маркетингового інструментарію, у тому числі на промисловому та споріднених ринках;
- пояснювати інформацію, ідеї, проблеми та альтернативні варіанти прийняття управлінських рішень фахівцям і нефахівцям у сфері маркетингу, представникам різних структурних підрозділів ринкового суб’єкта;
- демонструвати вміння застосовувати міждисциплінарний підхід та здійснювати маркетингові функції ринкового суб’єкта, у тому числі на промисловому та споріднених ринках;
- виявляти навички самостійної роботи, гнучкого мислення, відкритості до нових знань, бути критичним і самокритичним.

ЗМІСТ НАВЧАЛЬНОГО МАТЕРІАЛУ

Розділ 1. Статистичне спостереження, узагальнення та представлення даних
Тема 1.1. Предмет і метод статистики
Тема 1.2. Статистичне спостереження
Тема 1.3. Вибіркові спостереження
Тема 1.4. Зведення і групування статистичних даних
Тема 1.5. Абсолютні та відносні величини
Тема 1.6. Середні величини
Тема 1.7. Індекси
Тема 1.8. Показники варіації
Розділ 2. Параметрична та непараметрична статистика
Тема 2.1. Точкове оцінювання параметрів
Тема 2.2. Інтервальне оцінювання параметрів
Тема 2.3. Параметричні критерії перевірки статистичних гіпотез
Тема 2.4. Непараметричні критерії перевірки статистичних гіпотез
Розділ 3. Прикладна статистика
Тема 3.1. Дисперсійний аналіз
Тема 3.2. Елементи кореляційного та регресійного аналізу
Тема 3.3. Ряди динаміки
Тема 3.4. Організація статистики в умовах ринкової економіки

Дані, інформація, вимірювання

Дані та їх типи

Дані – (від лат. «data») факти; сукупність відомостей, які зафіксовані на деякому носії у формі, що є придатною для їх тривалого зберігання, передачі та обробки. Перетворення та обробка даних дозволяє отримати інформацію.
Дані можуть бути представлені в текстовій, числовій, графічній, відео, аудіо формі.

Дані можуть бути:
- якісні (неметричні)
- кількісні (метричні)
    - дискретні
    - неперервні

Кількісні (метричні) дані є власне числами, які мають змістовний сенс, мають одиниці вимірювання (грн, кг, шт.)
Прикладами таких даних можуть бути обсяги збуту, рівні доходів у грошовому вимірі, вік споживача тощо.
Дискретні дані – кількісні дані, які мають зчислену або злічену кількість значень (звичайно є цілими числами).
Приклади: кількість вагітностей у жінки, кількість авто в сім'ї тощо.
Неперервні дані – кількісні дані, які приймають значення на неперервному інтервалі значень.
Приклади: температура, вага, показник гемоглобіну в крові.

Якісні (неметричні) дані не можна виразити в числах (думки, враження, колір, смак, запах).
Приклади: стать споживача, район проживання, наявність домашніх тварин тощо.
Як показує досвід, при проведенні маркетингових досліджень кількісні дані складають незначну частину від усіх даних, які отримує дослідник. Всі дані, які стосуються думок, поглядів, переваг споживачів є якісними.

Отримання інформації з даних

Перетворення та обробка даних дозволяє отримати інформацію. Дані є матеріалом для отримання інформації. Даними необхідно керувати для того, щоб вони стали інформацією.

Інформація – (від лат. «informatio») роз'яснення, викладення, повідомлення; результат перетворення даних для розв'язання конкретних завдань.

Приклад: в базі даних зберігаються дані, а за певним запитом система управління базою даних видає потрібну інформацію.

Маркетингові дані та інформація – це дані та інформація, що отримані із зовнішнього макро або мікро ринкового середовища або внутрішнього середовища підприємства, та можуть бути використані для роботи маркетолога при розробці, реалізації та контролі за реалізацією продуктово-ринкової стратегії, а також для її тактичного та оперативного коригування.

Особливості маркетингових даних та інформації:
1. Величезна кількість джерел отримання
2. Різні (будь-які) носії
3. Різна періодичність надходження
4. Неможливість однозначного відбору даних, які можуть бути потрібними у майбутньому
5. Різна природа даних: якісні та кількісні, візуальні тощо
6. Практична неможливість створення єдиного банку даних

Приклади маркетингових даних та інформації:
Зовнішні дані:
- прайс-лист конкурента
- кількість чоловіків та жінок, що проживають у Києві в поточному році
- кількість проданих автомобілів Honda в м. Києві в січні місяці поточного року

Зовнішня інформація:
- прогноз зміни курсу Євро на три місяці
- думка експертів ринку про вплив на розвиток бізнесу в країні прийняття нового податкового кодексу
- динаміка ринку за останні 5 років
- отримана в результаті дослідження частка споживачів, які позитивно сприймають запропоновану модифікацію товару

Внутрішні дані:
- витрати на рекламу в поточному році
- обсяг продажу певного товару за місяць
- рівень недозавантаженості виробничих ліній
- кількість дилерів компанії

Внутрішня інформація:
- динаміка продажів за минулий рік
- співвідношення кількості проданого товару та витрат на рекламу по місяцях минулого року
- середня закупка

Тріада генерування інформації – це три складові, які взаємопов'язані між собою, які обов'язково необхідні для отримання **Інформації**:
- **Завдання**
- **Методи**
- **Дані**

Завдання інформаційного генезису:
- Опис об'єкту дослідження, вивчення структури, вивчення сили та напряму взаємозв'язків
- Класифікація, кластеризація (ділення, дроблення об'єкту дослідження)
- Зниження вимірності простору ознак досліджуваної сукупності

Методи аналізу даних поділяються:
1. Залежно від наявності вимірювання ознак досліджуваного об'єкту:
    - Якісні методи (вимірювань немає)
    - Кількісні методи (є вимірювання)

2. Залежно від того, чи враховуються взаємозв'язки між досліджуваними ознаками:
   - Одновимірний аналіз (ознаки вивчаються без урахування впливу одна на одну)
   - Багатовимірний аналіз (вивчається спільний вплив ознак)

3. Залежно від етапу кількісного аналізу:
   - **Первинний** аналіз (описова або дескриптивна статистика)
   - **Вторинний** аналіз (аналітична статистика), що поділяється:
     - **Індуктивна** (вивідна) статистика (отримання висновків про генеральну сукупність на основі вибіркової сукупності)
       - Інтервальне оцінювання параметрів
       - Статистичні тести
     - **Моделювання** об'єкту дослідження

Якісний аналіз даних – аналіз даних за допомогою сукупності технік інтерпретацій, спрямованих на те, щоб за допомогою опису, декодування, перетворення або інших способів дійти згоди про сенс явищ, які відбуваються в соціальному житті.

Кількісний аналіз даних – аналіз даних за допомогою математичних методів та моделей, в основі кількісного аналізу лежить вимірювання ознак досліджуваного об'єкту.

Вимірювання. Шкалювання

Статистика:
- масив даних
- галузь практичної діяльності людини, яка спрямована на збір, обробку та аналіз статистичних даних
- наукова дисципліна, у якій розроблені та систематизовані поняття, прийоми, математичні методи та моделі, що призначені для організації збору, стандартного запису, систематизації та обробки статистичних даних з метою їх зручного представлення, інтерпретації та отримання наукових та практичних висновків.

Вимірювання – присвоєння чисел або інших символів характеристикам об'єктів за наперед визначеними правилами.
Для вимірювання даних необхідно провести **шкалювання** даних, тобто обрати відповідну шкалу, за якою буде відбуватися вимірювання.

Кількісні дані звичайно вимірюються за наступними шкалами:
- **інтервальна шкала**
- **відносна шкала**.

Відмінність між ними полягає в тому, що для відносних даних, на відміну від інтервальних, нуль є значущою цифрою. Таким чином, інтервальні дані можна складати і віднімати, а відносні, крім цього, ще й множити і ділити.

На відміну від кількісних даних, якісні дані не можуть вимірюватися за наявними числовим шкалами. Для того, щоб виміряти якісні дані, їх необхідно класифікувати за певними категоріями. При цьому кожне отримане при вимірюванні значення повинне належати до однієї з цих категорій, причому тільки до однієї. Задавати ці категорії дослідник може по-різному, залежно від того, який рівень деталізації інформації необхідний для вирішення завдань дослідження.

Залежно від того, чи можна встановити порядок між категоріями, якісні дані вимірюються за наступними шкалами:

- **номінальна шкала** (не існує порядку між категоріями)
- **порядкова шкала** (існує порядок між категоріями).

Статистики для даних, які вимірюються за різними типами шкал

Тип даних	Якісні (неметричні)		Кількісні (метричні_	
Тип шкали	Номінальна	Порядкова	Інтервальна	Відносна
Дозволені операції	= ≠	≠ = > <	= ≠ > < + –	= ≠ > < + – * /
Приклади	Колір авто, стать, місце проживання	Оцінювання важливості, відповідності, згоди, рівень освіти	Температура за Цельсія, прибутковість/ збитковість, координата точки	Частка ринку, обсяги продажів, ширина/глибина асортименту, кількість конкурентів
Статистики одновимірні	Графік частот, відсотки, мода	Графік частот, графік накопичених частот, відсотки, мода, процентілі, медіана	Відсотки, мода, процентілі, медіана, середнє, розмах, графік розподілу	Відсотки, мода, процентілі, медіана, середнє, розмах, графік розподілу дисперсія, коефіцієнт варіації, асиметрія, ексцес
Статистики багатовимірні	Таблиця сполучення ознак, емпіричні частоти, коефіцієнт Крамера, коефіцієнт подібності	Таблиця сполучення ознак, емпіричні частоти, рангові коефіцієнти кореляції Спірмена, Кендалла	Поле частот, графікі спільних розподілів, лінійна відстань	Поле частот, графікі спільних розподілів, коефіцієнт кореляції Пірсона, Евклідова відстань

Категорії порядкових даних можна розмістити в порядку значущості, а категорії номінальних даних є рівнозначними. Формально, можливість або неможливість такого упорядкування категорій якісних даних є введенням над цими даними різних операцій.

Якщо для номінальних даних може бути тільки визначена приналежність до певної категорії, що рівносильно введенню операцій «=» и «≠», то для порядкових даних встановлення певного порядку категорій відповідає введенню, крім «=» и «≠», операцій «>», «<». Проте для будь-яких якісних даних операції «+», «-», «*» і «/» визначити неможливо.

Статистичне спостереження

Аналіз маркетингової інформації проводиться з використанням різних способів і прийомів статистичної методології. Правильність й достовірність як теоретичних, так і практичних висновків неможлива без наявності вичерпної і достовірної інформації про досліджуваний об'єкт. Для вивчення виділяють одиниці сукупності.

Одиниця сукупності – межа дроблення об'єкта дослідження, при якому зберігаються всі властивості досліджуваного об'єкта.
Генеральна сукупність – сукупність всіх умовно можливих одиниць сукупності даного типу, які підлягають вивченню, і з яких існує можливість зняти спостереження, що виконані у даному, реальному комплексі умов.

Одиниця сукупності має ознаки, що підлягають вивченню. Значення ознаки для довільної одиниці сукупності є випадковою величиною. Вимірюючи значення кожної ознаки для окремих одиниць сукупності, отримуємо масив даних. Аналізуючи дані, отримуємо інформацію про генеральну сукупність.

Статистичне спостереження

Статистичне спостереження – етап статистичного дослідження, у процесі якого збирають статистичні дані.
Статистичне спостереження може бути:

- **суцільне** (вимірюванню підлягають ознаки всіх одиниць сукупності)
- **несуцільне** (вимірюванню підлягають ознаки обраних певним чином одиниць з генеральної сукупності).

Несуцільне спостереження може бути проведене в одній з трьох можливих форм:
1. **Монографічне** спостереження (обстеження) – детальне (глибинне) вивчення однієї або декількох одиниць сукупності

Приклади: глибинне інтерв'ю, фокус-групи, детальне вивчення одного підприємства як типового представника даної галузі

Монографічне дослідження за своєю суттю є якісним. Кількісний аналіз не проводиться

2. **Метод основного масиву** – вибір для вивчення таких одиниць сукупності, які вносять максимальний внесок у явище, яке досліджується

Приклад: 95% обсягів ринку сухих будівельних сумішей припадає на 11 великих виробників (при цьому дрібні підприємства – більше 200 вітчизняних плюс імпорт – складають усього 5%). Досліджуючи структуру асортименту та його динаміку тільки по цим 11 підприємствам, маємо практично достатню точність інформації про ринок

При дослідженні методом основного масиву будують описову статистику, узагальнюють значення ознак, які вимірювали у досліджуваних одиниць сукупності. За цими ж даними можуть бути побудовані моделі.
А перенесення висновків на генеральну сукупність виконується вже не кількісно, а якісно. Індуктивна статистика не може бути використана, тому що одиниці сукупності не є однорідними та обираються за певними параметрами, які впливають на значення досліджуваних ознак.

3. **Вибіркове** спостереження – вивчення спеціальним чином відібраної з генеральної сукупності одиниць сукупності, ознаки яких підлягають вимірюванню, а отримані результати переносяться (апроксимуються) на всю генеральну сукупність.

При використанні випадкового відбору при формуванні вибірки можливе використання як описової, так й індуктивної статистики. При цьому із заданим рівнем похибки отримують значення параметрів генеральної сукупності.

Репрезентативність вибірки – правильне представлення генеральної сукупності, при якому вибіркові характеристики є близькими до характеристик генеральної сукупності. Репрезентативність вибірки залежить, передусім, від процедури відбору елементів.

Залежно від того, чи є вибір кожного з елементів вибірки випадковим чи ні, виділяють два основних типу відбору:

- **випадковий** (кожний елемент генеральної сукупності має однакову ймовірність потрапити у вибірку) відбір
- **детермінований** (або невипадковий) відбір

Основні вибіркові характеристики та їх властивості

До основних вибіркових характеристик належать (визначення та приклади побудови функцій будуть розглянуті на практичному занятті):

– вибіркова (емпірична) функція розподілу $\hat{F}(x)$ (кумулятивна крива, гістограма відносних накопичених частот);

– вибіркова (емпірична) функція щільності розподілу $\hat{f}(x)$ (полігон частот, гістограма відносних частот);

– вибіркові (емпіричні) відносні частоти $\hat{p}_i$ появи і-го можливого значення x_i дискретної випадкової величини;

– порядкові статистики $\hat{x}_p$;

– вибіркові моменти випадкової величини.

Вибіркові моменти випадкової величини

Представимо вибірку у табличному вигляді (по аналогії з дискретною випадковою величиною):

x_i	x_1	x_2	…	x_n
pi	1/n	1/n	…	1/n

Вибірковий початковий момент k-го порядку:

$$\hat{m}_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

Вибірковий центральний момент k-го порядку:

$$\hat{m}_k^0 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^k ,$$

де $\overline{X}$ – вибірковий початковий момент першого порядку:

$$\overline{X} = \hat{m}_i = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

Вибіркові показники звичайно поділяють на такі групи:

– показники центру групування (центральної тенденції)

– показники розсіювання (мінливості)

– показники форми розподілу

– квантилі (процентилі, децилі, квінтилі, квартилі)

Показники центру групування

1. **Вибіркове середнє** або середнє арифметичне:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

Можна показати, що вибіркове середнє є незсуненою, конзистентною (рос. «состоятельной»), ефективною оцінкою математичного сподівання.

2. **5% зрізане** (рос. «усеченное») **середнє** (використовується для виключення спостережень, що різко виділяються): спочатку з вибірки виключаються елементи менші за 5-у процентіль та більші за 95-у, а потім розраховується вибіркове середнє.

3. **Середнє геометричне**:

$$\hat{G} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} .$$

Використовується при розрахунках темпів змін величин (чисельність населення, динаміка ринку, індекси цін тощо)

Приклад: у 2013 році рівень доходу зріс на 10%, у порівнянні з 2012, а в 2014 році – на 25%. На скільки в середньому зріс рівень доходу за два роки?

«В середньому» означає, що при застосуванні середнього значення до кожного року, отримаємо той самий результат, що й при використанні множини вихідних значень.
Усього дохід зріс на 1,1*1,25=1,375, тобто 37,5% за 2 роки.

Є декілька варіантів для осереднення:
1) середнє арифметичне – 17,5%. Перевірка: 1,175*1,175=1,38 (різниця в 0,5%)

2) складний відсоток: 37,5% за 2 роки і 18,75% за один рік. Перевірка: 1,1875*1,1875=1,41 (різниця у 3,5%)
3) середнє геометричне: – $\sqrt{1{,}1 \cdot 1{,}25} = 1{,}1726$, тобто середній зріст доходу 17,26%. Перевірка: 1,1726*1,1726 = 1,375 !!!

4. **Середнє гармонійне**:

$$\hat{H} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i}} .$$

Використовується для аналізу середніх норм часу.

Приклад: на касах обслуговуються клієнти, за годину на першій касі розрахували 20 клієнтів, на другій – 80, на третій – 50. Визначити середню швидкість обслуговування.

Середнє арифметичне дорівнює 50 клієнтів за годину.

Середнє гармонійне:

$$\hat{H} = \frac{1}{\frac{1}{3}\left(\frac{1}{20} + \frac{1}{80} + \frac{1}{50}\right)} = 36{,}36$$ клієнтів за годину.

(Для того, щоб обслугувати 400 клієнтів на першій касі потрібно 20 годин, на другій – 5, на третій – 8. Таким чином, загальний час обслуговування дорівнює 33 години. Якщо швидкість обслуговування на всіх касах буде однаковою, то кожна каса може працювати рівно 11 годин. Якщо 400 клієнтів поділити на 11 годин, отримаємо саме 36,36 клієнтів за годину – середню швидкість обслуговування).

Для конкретної вибірки $x_1, x_2, \ldots, x_n$ завжди справедлива нерівність:

$$\hat{H}_n < \hat{G}_n < \overline{X}_n$$

До показників центру групування відносять також так звані структурні середні: моду та медіану.

Модою $\hat{d}_x$ унімодального розподілу (існує значення випадкової величини, для якого відповідна ймовірність є більшою за інші; або щільність розподілу випадкової величини має один максимум) є елемент масиву даних, який має максимальну частоту.
Медіаною називається середина варіаційного ряду (спосіб запису масиву даних, при якому елементи впорядковані від найменшого до найбільшого).
Якщо n непарне:

$$n = 2k + 1, \qquad \hat{h} = x^{(k+1)}$$

якщо парне:

$$n = 2k, \qquad \hat{h} = \frac{x^{(k)} + x^{(k+1)}}{2}$$

Медіана = 50-а процентиль = 5-а дециль = 2-а квартиль.

Показники розсіювання (мінливості)

1. Розмах (в SPSS «діапазон») масиву даних:

$$R = x_{max} - x_{min}$$

2. Дисперсія:

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

3. Вибіркова дисперсія:

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

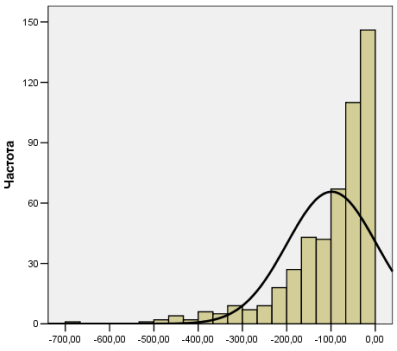
4. Середньоквадратичне відхилення – S – корінь з дисперсії.

5. Коефіцієнт варіації:

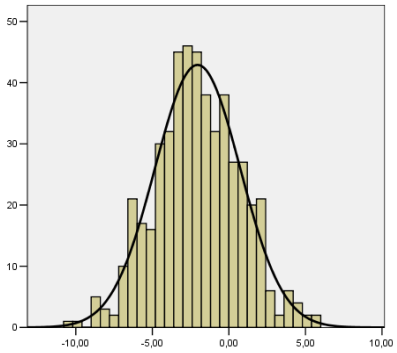
$$V = \frac{S}{\bar{X}} \cdot 100\%$$

Показники форми розподілу

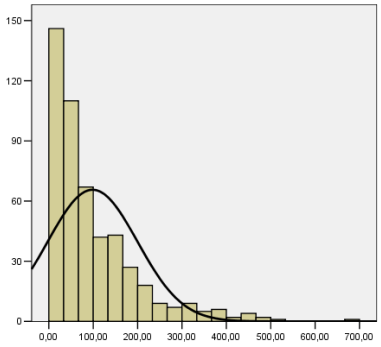
1. Асиметрія – коефіцієнт скошеності:



від'ємна асиметрія

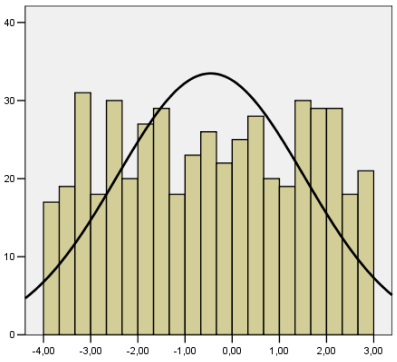


асиметрія близька до нуля

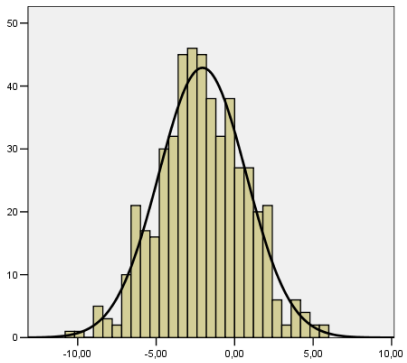


додатна асиметрія

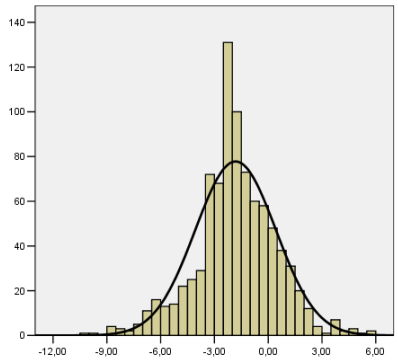
Ексцес – коефіцієнт гостровершинності (при порівнянні з нормальним розподілом):



від'ємний ексцес



ексцес близький до нуля



додатній ексцес

Близькість до нуля асиметрії та ексцесу одночасно свідчить на користь вибору нормального закону розподілу досліджуваної сукупності.

Порядкові статистики

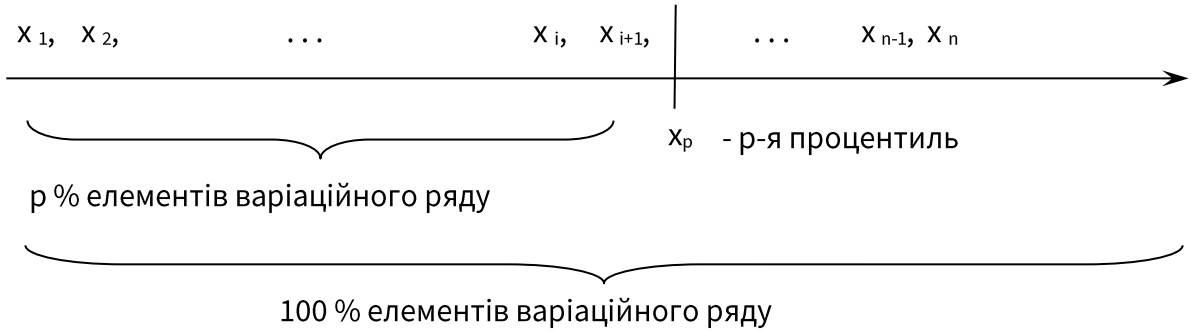
Квантилі: процентилі, децилі, квінтилі, квартилі.

Процентилі – статистики, які показують який відсоток значень всього масиву є меншим за дане значення.

0-а процентиль дорівнює мінімальному значенню в масиві

100-а процентиль дорівнює максимальному значенню в масиві

38-я процентиль дорівнює значенню, меньше або дорівнюють якому 38% усіх елементів масиву



Для того, щоб побудувати к-у процентиль необхідно відділити к % мінімальних значень варіаційного ряду. Наближеною оцінкою буде максимальний елемент у відділеній частині мінімальних елементів масиву.

Для більш точного розрахунку можна використовувати формулу:

$$\hat{x}_p = x_i + (x_{i+1} - x_i)\cdot(p\cdot(n+1) - i),$$

де $i = [(n+1)\cdot p]$ (квадратні дужки – операція відділення цілого числа),
p – потрібний порядок процентилі,
n – об'єм масиву,
x_i и x_{i+1} – значення елементів варіаційного ряду, між якими знаходиться потрібна процентиль,
i и i+1 – номера відповідних елементів в варіаційному ряді

Децилі – процентилі з кроком в 10%
1-а дециль = 10-а процентиль
2-а дециль = 20-а процентиль
…
9-а дециль = 90-а процентиль

Квартили – процентилі с кроком 25%
1-а квартиль = 25-а процентиль
2-а квартиль = 50-а процентиль = медіана
3-а квартиль = 75-а процентиль

Статистики, статистичні оцінки та їх основні властивості

Визначення. Будь-яка функція $y(x_1, x_2, ..., x_n)$ від результатів спостережень $x_1, x_2, \dots, x_n$ випадкової величини X, називається **статистикою**.

Визначення. Статистика $\hat{\theta}$, яка використовується як наближене значення невідомого параметра θ, називається **статистичною оцінкою**.

Наприклад, $\overline{X}$ є статистичною оцінкою математичного сподівання, а S^2 статистична оцінка дисперсії генеральної сукупності.

Усі статистики та статистичні оцінки є випадковими величинами. Тому отримані на різних вибірках, навіть з одної генеральної сукупності, конкретні значення статистичних оцінок будуть мати деякий неконтрольований розкид.

Для отримання **надійності** (в деякому сенсі) статистичних оцінок, до них висуваються вимоги, які формулюються за допомогою трьох **властивостей**:

1. Незсуненість (рос. «несмещенность», в сучасних підручниках також зустрічається термін «незміщеність», який є калькованим перекладом з російської)
2. Конзистентність (рос. «состоятельность», переклад: «спроможність», який є в деяких підручниках, вважаємо не досить вдалим, тому залишено англійський варіант)
3. Ефективність

Визначення. Оцінка $\hat{\theta} = \hat{\theta}(x_1, x_2, ..., x_n)$ невідомого параметра θ називається **незсуненою**, якщо при будь-якому об'ємі вибірки n результат її осереднення по всім можливим вибіркам даного об'єму приводить до точного істинного значення оцінюваного параметра:

$$E\hat{\theta}_n = \theta$$

Визначення. Оцінка $\hat{\theta} = \hat{\theta}(x_1, x_2, ..., x_n)$ невідомого параметра θ називається **конзистентною**, якщо по мірі збільшення кількості спостережень n, вона прямує по ймовірності до оцінюваного значення θ:

$$\forall \varepsilon > 0 \qquad P\left(\left|\hat{\theta} - \theta\right| > \varepsilon\right) \underset{n\to\infty}{\to} 0$$

Визначення. Оцінка $\hat{\theta} = \hat{\theta}(x_1, x_2, ..., x_n)$ невідомого параметра θ називається **ефективною**, якщо серед усіх оцінок того самого параметра вона має найменшу міру випадкового розкиду відносно істинного значення оцінюваного параметра θ:

$$D\hat{\theta}_1 < D\hat{\theta}_2, \quad \Rightarrow \quad \hat{\theta}_1 \text{ ефективніше за } \hat{\theta}_2.$$

Приклад. Перевіримо властивості вибіркового середнього як статистичної оцінки математичного сподівання.

Нехай $X_1, X_2, \dots, X_n$ – послідовність незалежних однаково розподілених випадкових величин, $x_1, x_2, \dots, x_n$ – конкретна вибірка.

Зазначені випадкові величини мають однакове математичне сподівання:

$$EX_1 = EX_2 = ... = EX_n = \theta$$

Візьмемо за його оцінку вибіркове середнє:

$$\hat{\theta} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Перевіримо незсуненість:

$$E\overline{X} = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} Ex_i = \frac{1}{n}\sum_{i=1}^{n}\theta = \frac{1}{n}\cdot n\theta = \theta$$

Таким чином, вибіркове середнє є незсуненою оцінкою.

За законом великих чисел вибіркове середнє є конзистентною оцінкою математичного сподівання:

$$\overline{X} = \frac{x_1 + x_2 + \ldots + x_n}{n} \xrightarrow[n\to\infty]{P} \theta$$

Якщо генеральна сукупність має нормальний розподіл, то вибіркове середнє є ефективною оцінкою генерального середнього в класі лінійних незсунених оцінок.

Приклад. Перевірити незсуненість вибіркової дисперсії:

$$S^2 = \hat{m}_2^0 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2$$

Нехай X_1, X_2, … , X_n – послідовність незалежних однаково розподілених випадкових величин з математичним сподіванням m та дисперсією σ^2, x_1, x_2, … , x_n – конкретна вибірка.

$$ES^2 = E\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\overline{X} + \overline{X}^2\right)\right] = \frac{1}{n}\sum_{i=1}^{n}\left(Ex_i^2 - 2E\left(x_i\overline{X}\right) + E\overline{X}^2\right)$$

Розглянемо окремо кожний з доданків.

Для дисперсії має місце тотожність: $DX = EX^2 - (EX)^2$, звідки:

$Ex_i^2 = Dx_i + (Ex_i)^2 = \sigma^2 + m^2$.

Для другої частини використовуємо властивість математичного сподівання: $E(XY) = EX\cdot EY$, якщо X та Y – незалежні випадкові величини, а також скористаємося результатом, отриманим для першого доданка:

$$E\left(x_i\overline{X}\right) = E\left(x_i\cdot\frac{1}{n}\sum_{j=1}^{n} x_j\right) = \frac{1}{n}\left(Ex_i^2 + \sum_{j=1, j\neq i}^{n} Ex_iEx_j\right) = \frac{1}{n}\left(\sigma^2 + m^2 + (n-1)m^2\right) = \frac{\sigma^2}{n} + m^2$$

І, нарешті, третя частина:

$$E\overline{X}^2 = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\cdot\frac{1}{n}\sum_{j=1}^{n} x_j^2\right) = \frac{1}{n^2}\left(\sum_{i=1}^{n} Ex_i^2 + \sum_{i,j=1, i\neq j}^{n} Ex_i\cdot Ex_j\right) =$$

$$\frac{1}{n^2}\left(n\left(\sigma^2 + m^2\right) + \left(n^2 - n\right)m^2\right) = \frac{\sigma^2}{n} + m^2$$

Підставимо отримані значення:

$$ES^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\sigma^2 + m^2 - 2\left(\frac{\sigma^2}{n} + m^2\right) + \frac{\sigma^2}{n} + m^2\right) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2.$$

Таким чином, $ES^2 \neq \sigma^2$ і оцінка дисперсії є зсуненою.

Для невеликих об’ємів вибірки (n<100) використовується підправлена оцінка дисперсії, яка є незсуненою:

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

Оцінювання параметрів генеральної сукупності

В статистиці існує два підходи до оцінювання невідомих параметрів розподілу генеральної сукупності за результатами спостережень:

1. **Точковий** підхід (в рамках описової статистики) дозволяє вказати точку, біля якої знаходиться оцінюваний параметр.
2. **Інтервальний** підхід (в рамках індуктивної або вивідної статистики) передбачає знаходження інтервалу, який з деякою достатньо великою ймовірністю накриває невідоме значення параметра.

Побудова точкових оцінок параметрів. Метод моментів

Нехай закон розподілу випадкової величини X є відомим з точністю до числових значень його параметрів $\theta_1,\theta_2,...,\theta_k$. Це означає, що для неперервної випадкової величини є відомим вид функції щільності розподілу $p(x,\theta_1,\theta_2,...,\theta_k)$, а для дискретної випадкової величини є відомим ряд розподілу з відповідними ймовірностями $P(X=x_i,\theta_1,\theta_2,...,\theta_k)$.

Нехай k початкових моментів існує. Складемо систему з k рівнянь, у кожному з яких прирівняємо відповідні теоретичні та емпіричні моменти:

$$\begin{cases} m_1(\theta_1,\theta_2,...,\theta_k)=\hat{m}_1 \\ \\ m_k(\theta_1,\theta_2,...,\theta_k)=\hat{m}_k \end{cases},$$

де теоретичний момент для дискретної випадкової величини дорівнює:

$$m_i(\theta_1,...,\theta_k)=\sum_{j=1}^{n} x_{ij}\cdot P(X=x_j,\theta_1,\theta_2,...,\theta_k),$$

теоретичний момент для неперервної випадкової величини:

$$m_i(\theta_1,...,\theta_k)=\int_{-\infty}^{\infty} x^i\cdot p(x,\theta_1,\theta_2,...,\theta_k)dx.$$

Розв'язуючи систему рівнянь відносно $\theta_1,\theta_2,...,\theta_k$ отримуємо оцінки параметрів за методом моментів.

Приклад. Розглянемо випадкову величину, що має дихотомічний розподіл $X\sim D(p)$. Вона має ряд розподілу:

xi	0	1
pi	1– p	p

Отримана вибірка: 1, 0, 1, 1, 0. Необхідно оцінити невідомий параметр p за допомогою методу моментів.

Невідомий параметр один: $\theta\equiv p$.

Теоретичний перший початковий момент:

$$m_1=EX=\sum_{i=1}^{n} x_i\cdot p_i=0\cdot(1-p)+1\cdot p=p\ .$$

Емпіричний перший початковий момент:

$$\hat{m}_1=\overline{X}=\frac{1+0+1+1+0}{5}=\frac{3}{5}=0{,}6\ .$$

Система рівнянь складається з одного:

$m_1(p)=\hat{m}_1$, підставимо отримані значення:

$\hat{p} = 0{,}6$.

Приклад. Розглянемо випадкову величину, що має нормальний розподіл $X \sim N(m, \sigma^2)$. Отримана вибірка $x_1, x_2, \ldots, x_n$. Необхідно оцінити параметри m та σ^2 за методом моментів.

Невідомих параметрів – два, тому складаємо систему з двох рівнянь:

$$\begin{cases} m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i \\ m_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \end{cases}.$$

Відомо, що: $m_1 = m$, $\sigma^2 = m_2 - (m_1)^2$, звідки:

$$\begin{cases} m_1 = m = \frac{1}{n}\sum_{i=1}^{n} x_i \\ m_2 = \sigma^2 + (m_1)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \end{cases},$$

тобто $\hat{m} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$, а $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{X})^2$.

Формулу для оцінки дисперсії можна спростити, для цього додамо та віднімемо $(\overline{X})^2$:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{X})^2 \pm (\overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2(\overline{X})^2 + (\overline{X})^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\cdot\frac{1}{n}\sum_{i=1}^{n} x_i \cdot \overline{X} + \frac{1}{n}\cdot n\cdot(\overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\cdot\overline{X} + (\overline{X})^2\right) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2$$

В результаті отримані оцінки параметрів нормального розподілу:

$$\hat{m} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{та} \quad \hat{\sigma}^2 = S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2 \ .$$

Побудова точкових оцінок параметрів. Метод максимальної вірогідності

В основі методу лежить побудова функції вірогідності (рос. «правдоподобия»). Нехай $X_1, X_2, \ldots, X_n$ – послідовність незалежних однаково розподілених випадкових величин, $x_1, x_2, \ldots, x_n$ – конкретна вибірка, закон розподілу відомий з точністю до параметрів $\theta_1, \theta_2, \ldots, \theta_k$, тобто для неперервної випадкової величини є відомим вид функції щільності розподілу $p(x, \theta_1, \theta_2, \ldots, \theta_k)$, а для дискретної випадкової величини є відомим ряд розподілу з відповідними ймовірностями $P(X = x_i, \theta_1, \theta_2, \ldots, \theta_k)$.

Функцією вірогідності називається для дискретної випадкової величини:

$$L(\theta_1, \theta_2, \ldots, \theta_k) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) =$$
$$= P(X_1 = x_1)\cdot P(X_2 = x_2)\cdot \ldots \cdot P(X_n = x_n)$$

або для неперервної випадкової величини:

$$L(\theta_1, \theta_2, \ldots, \theta_k) = p(x_1, x_2, \ldots, x_n) = p(x_1)\cdot p(x_2)\cdot \ldots \cdot p(x_n).$$

Суть методу максимальної вірогідності полягає в знаходженні невідомих параметрів $\theta_1, \theta_2, \ldots, \theta_k$ таким чином, щоб дана конкретна вибірка могла з'явитися з максимальною ймовірністю, тобто була «максимально вірогідна».

Таким чином, оцінки $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ отримують при знаходженні максимуму функції вірогідності:

$$L\left(\hat{\theta}_1,\hat{\theta}_2,...,\hat{\theta}_k\right)=\max L(\theta_1,\theta_2,...,\theta_k)$$

В деяких випадках для зручності використовують логарифмічну функцію вірогідності:

$$\ln L\left(\hat{\theta}_1,\hat{\theta}_2,...,\hat{\theta}_k\right)=\max \ln L(\theta_1,\theta_2,...,\theta_k)$$

Для знаходження оцінок $\hat{\theta}_1,\hat{\theta}_2,...,\hat{\theta}_k$ необхідно:

1. Розв'язати систему рівнянь:

$\frac{\partial L}{\partial \theta_j}=0 \quad j=\overline{1,k}$, або $\frac{\partial \ln L}{\partial \theta_j}=0 \quad j=\overline{1,k}$.

2. Серед розв'язків виділити точки максимуму.

3. Якщо система не визначена, не розв'язувана, або максимум відсутній всередині області припустимих значень для $\theta_1,\theta_2,...,\theta_k$, необхідно шукати точку максимуму на границі припустимих значень.

Приклад. Необхідно оцінити параметр p дихотомічно розподіленої випадкової величини $X\sim D(p)$ за допомогою методу максимальної вірогідності. Ряд розподілу має вигляд:

xi	0	1
pi	1– p	p

Отримана вибірка: 1, 0, 1, 1, 0, 1, 1, 1, 0, 1.

Випадкова величина є дискретною, побудуємо функцію вірогідності для конкретної вибірки:

$L(p)=P(X_1=x_1)\cdot P(X_2=x_2)\cdot ....\cdot P(X_n=x_n)=$

$=p\cdot(1-p)\cdot p\cdot p\cdot(1-p)\cdot p\cdot p\cdot p\cdot(1-p)\cdot p=p^7(1-p)^3$

Знайдемо похідну функції L(p) та прирівняємо її до 0:

$L'(p)=7p^6\cdot(1-p)^3-3p^7\cdot(1-p)^2=0$

$\hat{p}=\frac{7}{10}$

Приклад. Отримати оцінки параметрів нормального розподілу методом максимальної вірогідності.

Випадкова величина $X\sim N(m,\sigma^2)$, необхідно оцінити параметри $m\equiv\theta_1$ та $\sigma^2\equiv\theta_2$ за $x_1, x_2, \ldots, x_n$ – конкретною вибіркою.

Щільність розподілу нормально розподіленої випадкової величини:

$$p(x)=\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(x-m)^2}{2\sigma^2}}$$

Побудуємо функцію вірогідності:

$$L(\theta_1,\theta_2)=\frac{1}{\sqrt{2\pi\theta_2}}e^{-\frac{(x_1-\theta_1)^2}{2\theta_2}}\cdot...\cdot\frac{1}{\sqrt{2\pi\theta_2}}e^{-\frac{(x_n-\theta_1)^2}{2\theta_2}}=\left(\frac{1}{\sqrt{2\pi\theta_2}}\right)^n\cdot e^{-\frac{\sum_{i=1}^{n}(x_i-\theta_1)^2}{2\theta_2}}.$$

Більш простою для обчислень є логарифмічна функція вірогідності:

$$\ln L(\theta_1,\theta_2)=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln\theta_2-\sum_{i=1}^{n}\frac{(x_i-\theta_1)^2}{2\theta_2}.$$

Знайдемо похідні:

$$\begin{cases} \dfrac{\partial \ln L}{\partial \theta_1} = \sum\limits_{i=1}^{n} \dfrac{(x_i - \theta_1)}{\theta_2} \\ \dfrac{\partial \ln L}{\partial \theta_2} = -\dfrac{n}{2\theta_2} + \sum\limits_{i=1}^{n} \dfrac{(x_i - \theta_1)^2}{2\theta_2^{\,2}} \end{cases}.$$

Розв'яжемо систему рівнянь:

$$\begin{cases} \sum\limits_{i=1}^{n} (x_i - \theta_1) = 0 \\ \sum\limits_{i=1}^{n} \dfrac{(x_i - \theta_1)}{\theta_2} - n = 0 \end{cases} \Rightarrow \begin{cases} \hat{\theta}_1 = \dfrac{1}{n} \sum\limits_{i=1}^{n} x_i \\ \hat{\theta}_2 = \dfrac{1}{n} \sum\limits_{i=1}^{n} (x_i - \hat{\theta}_1)^2 \end{cases}.$$

Якщо перевірити достатні умови, можна показати, що $\hat{\theta}_1$ та $\hat{\theta}_2$ є максимумом функції вірогідності L.

Таким чином, отримані методом максимальної вірогідності оцінки параметрів нормального розподілу:

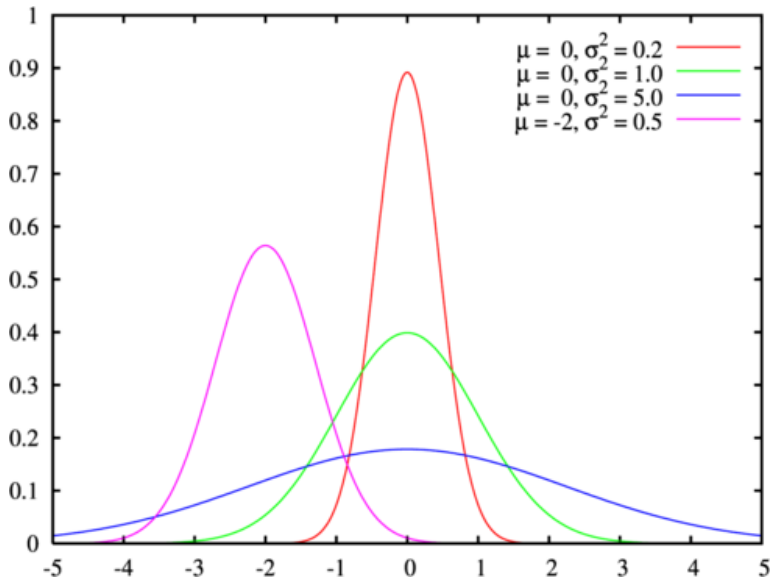
$$\hat{\theta}_1 = \hat{m} = \overline{X} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{та} \quad \hat{\theta}_2 = \hat{\sigma}^2 = S^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{X})^2 \ .$$

Розподіли, які використовуються в техніках статистичних обчислень

1. Стандартний нормальний розподіл

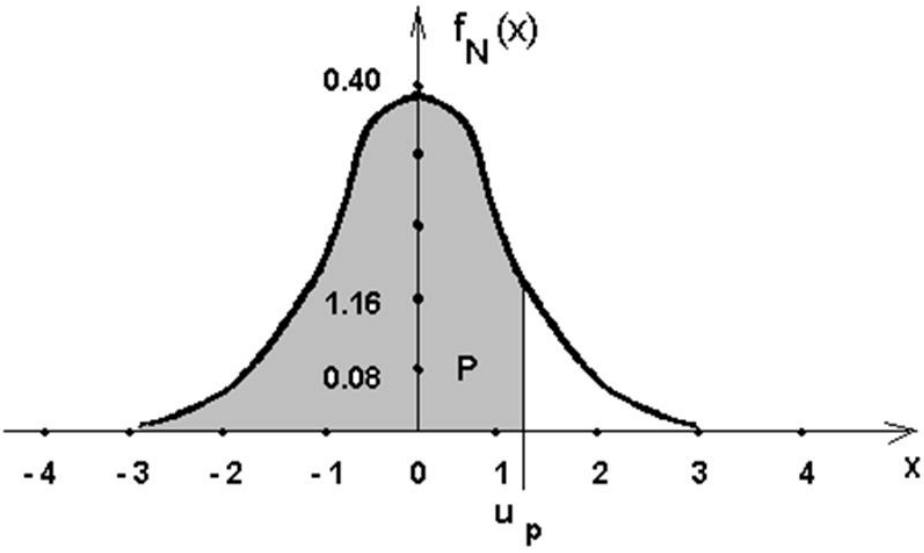
Нормальний розподіл $N(m,\sigma^2)$ має параметри m – математичне сподівання та σ^2 – дисперсія.
Щільність нормального розподілу:

$$\rho_N(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}}$$



Стандартний нормальний розподіл N(0,1) – це нормальний розподіл, в якого m=0, $\sigma^2=1$

Квантилі стандартного нормального розподілу N(0,1) позначають u_p , де p – порядок квантилі. Квантилі великих порядків можна знайти в таблицях.
Приклад: $u_{0,95}$ = 1,645(читається як “у порядку 0,95”)



Нормальний розподіл є симетричним відносно осі ординат, тому має місце наступна рівність:

$$u_p = -u_{1-p}$$

Це співвідношення використовується при знаходженні квантилів малих порядків.
Приклад: $u_{0,05} = -u_{0,95} = -1,645$

MS Excel
Для знаходження квантилів стандартного нормального розподілу використовується функція
НОРМСТОБР(вероятность)
Або у більш нових версіях:
НОРМ.СТ.ОБР(вероятность)
Наприклад, для того, щоб отримати значення квантилі $u_{0,05}$ потрібно задати в комірці:
=НОРМСТОБР(0,05)
Або:
=НОРМ.СТ.ОБР(0,05)
І натиснути Enter!
☺

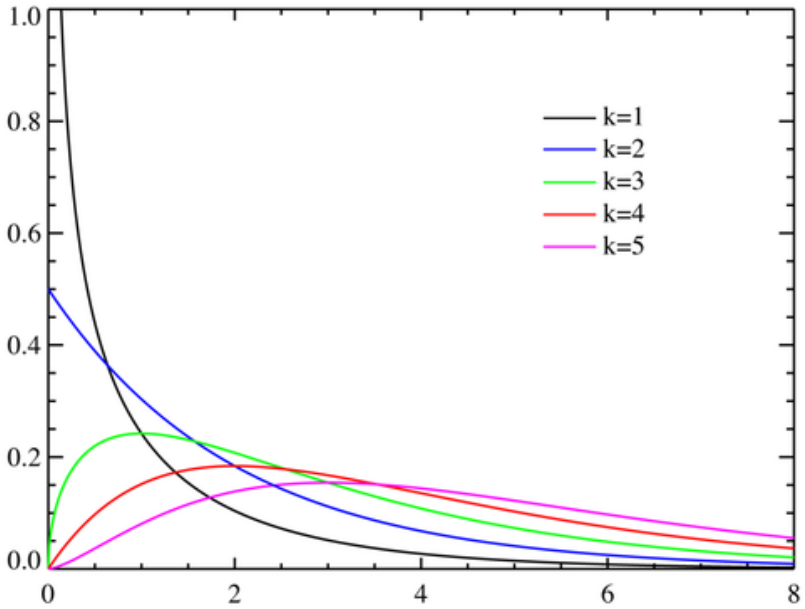
2. Хі-квадрат розподіл з k ступенями свободи

Якщо $X_1, X_2, \dots, X_k$ – незалежні випадкові величини, які мають стандартний нормальний розподіл N(0,1), тоді величина:

$$\chi^2(k) = X_1^2 + X_2^2 + \dots + X_k^2$$

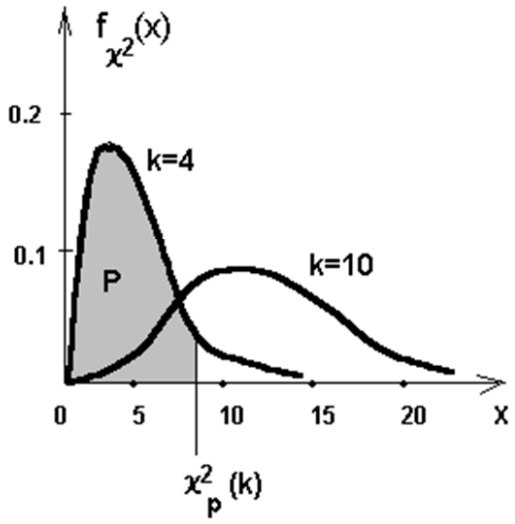
має χ^2–розподіл з k ступенями свободи.

Щільність розподілу Хі-квадрат значно залежить від кількості ступенів свободи:



Квантилі розподілу Хі-квадрат позначаються як $\chi_p^2(k)$. Їх значення можна знайти в Lesson4_table2 (у рядках k – кількість ступенів свободи, у стовпцях – p – порядок квантилі).

Приклад: $\chi^2_{0,95}(5)$ = 11,07 (читається як «квантиль хі квадрат розподілу порядку 0,95 з п'ятьма ступенями свободи)



MS Excel

Для знаходження квантилів Хі-квадрат розподілу використовується функція

ХИ2ОБР(вероятность; степени_свободы)

Або у більш нових версіях:

ХИ2.ОБР.ПХ(вероятность; степени_свободы) чи

ХИ2.ОБР(вероятность; степени_свободы)

Наприклад, для того, щоб отримати значення квантилі $\chi^2_{0,95}(5)$ потрібно задати в комірці:

= ХИ2ОБР (1-0,95;5)

Або:

=ХИ2.ОБР.ПХ(1-0,95;5) чи

=ХИ2.ОБР(0,95;5)

Будьте уважні з використанням функцій! Читайте довідки та перевіряйте значення за таблицями (один раз – просто, щоб зрозуміти, яку саме ймовірність необхідно використовувати в даній формулі).

2. Розподіл Стьюдента з k ступенями свободи

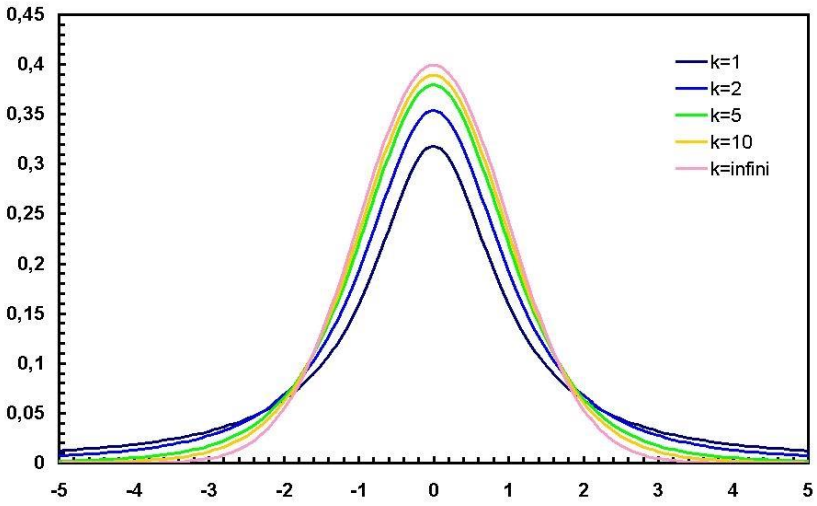
Якщо X_0, X_1, X_2, … , X_k – незалежні випадкові величини, які мають нормальний розподіл $N(0, \sigma^2)$ з нульовим математичним сподіванням та однаковою дисперсією σ^2, тоді величина:

$$T(k) = \frac{X_0}{\sqrt{\frac{X_1^2 + \cdots + X_k^2}{k}}}$$

має розподіл Стьюдента з k ступенями свободи.

При збільшенні кількості ступенів свободи до ∞ розподіл стає стандартним нормальним розподілом!

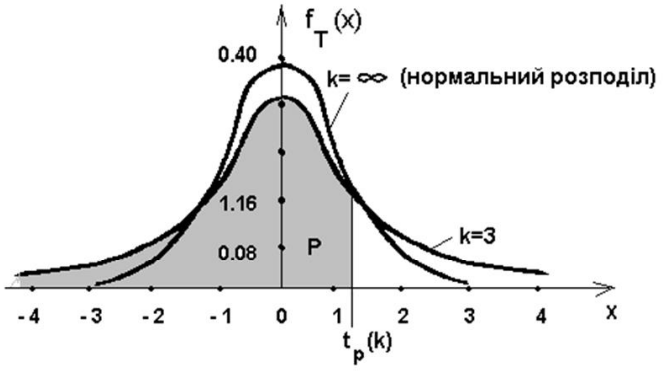
Так виглядає щільність розподілу Стьюдента для різних ступенів свободи:



Квантилі розподілу Стьюдента позначаються як $T_p(k)$. Їх значення для великих порядків можна знайти в Lesson4_table3 (у рядках k – кількість ступенів свободи, у стовпцях – p – порядок квантилі).
Приклад: $T_{0,9}(12)$ = 1,356 (читається як «квантиль розподілу Стьюдента порядку 0,9 з дванадцятьма ступенями свободи»)

Розподіл Стьюдента (як і стандартний нормальний розподіл) є симетричним відносно осі ординат, тому має місце наступна рівність:

$$t_p(k) = -t_{1-p}(k)$$



Приклад: $t_{0,05}(5) = -t_{0,95}(5) = -2,015$.

MS Excel
Для знаходження квантилів розподілу Стьюдента використовується функція
СТЬЮДРАСПОБР(вероятность; степени_свободы)
Наприклад, для того, щоб отримати значення квантилі $t_{0.995}(75)$ потрібно набрати в комірці:
= СТЬЮДРАСПОБР (2*(1-0,995);75)
І натиснути Enter.

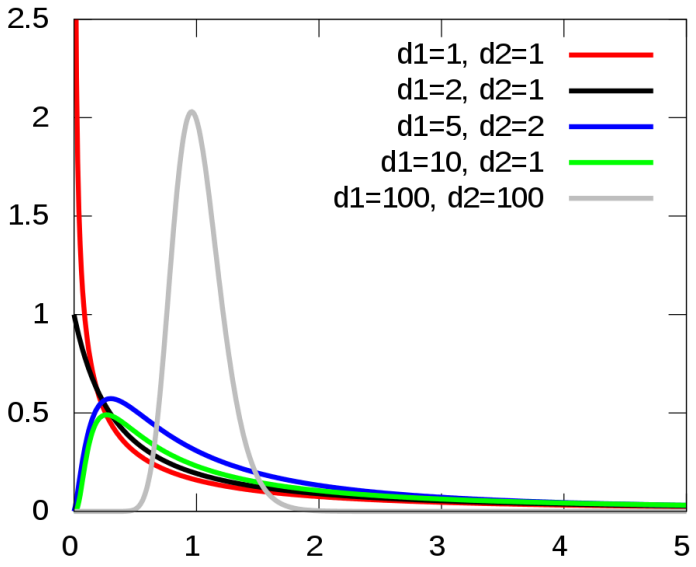
В нових версіях є 2 функції:
=СТЬЮДЕНТ.ОБР.2Х(2*(1-0,995);75)
Ця функція відповідає старій. Або більш простий варіант:
=СТЬЮДЕНТ.ОБР(0,995;75)
Зверніть увагу на знаходження квантилів малих порядків:
Наприклад, значення квантилі $t_{0.005}(75)$
У старому варіанті це буде:
= - СТЬЮДРАСПОБР (2*(0,005);75)
Або відповідне у новому:
=- СТЬЮДЕНТ.ОБР.2Х(2*(0,005);75)
І простий варіант:
=СТЬЮДЕНТ.ОБР(0,005;75)

2. Розподіл Фішера з k та n ступенями свободи

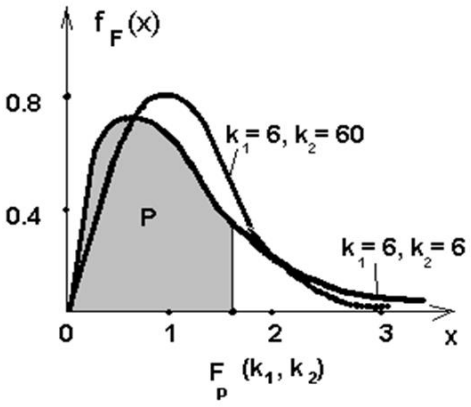
Якщо $X_1, X_2, \dots, X_k, X_{k+1}, \dots, X_{k+n}$ – незалежні випадкові величини, які мають нормальний розподіл $N(0, \sigma^2)$ з нульовим математичним сподіванням та однаковою дисперсією σ^2, тоді величина:

$$F(k,n) = \frac{\frac{X_1^2 + \cdots + X_k^2}{k}}{\frac{X_{k+1}^2 + \cdots + X_{k+n}^2}{n}}$$

має розподіл Фішера з k та n ступенями свободи.



Квантилі розподілу Фішера позначаються як $F_p(k,n)$. Їх значення для великих порядків можна знайти в Lesson4_table4 (у стовпцях k – значення першого ступеню свободи, у рядках – n – значення другого, таблиця окрема для кожного порядку квантилі p=0,9; 0,95; 0,975; 0,99).
Приклад: $F_{0,9}(5,8) = 2,73$ (читається як «квантиль розподілу Фішера порядку 0,9 з п'ятьма та вісьмома ступенями свободи).



Для розрахунку квантилів малих порядків використовується наступне співвідношення:

$$F_p(k, n) = \frac{1}{F_{1-p}(n, k)}$$

Приклад: $F_{0,01}(20,4) = 1/F_{0,99}(4,20) = 1/4,43 = 0,226$.

Для знаходження квантилів розподілу Фішера використовується функція
FРАСПОБР(вероятность; степени_свободы1; степени_свободы2)
Наприклад, для того, щоб отримати значення квантилі $F_{0,01}(20,4)$ потрібно набрати в комірці:
= FРАСПОБР (1-0,01;20;4)
Нова версія формули:
=F.ОБР.ПХ(1-0,01;20;4) – ця відповідає старій, або без ускладнень:
=F.ОБР(0,01;20;4)
І Enter.
Будьте уважні з версіями та формулами. Обирайте правильні варіанти, перевіряйте за допомогою таблиць!

Інтервальне оцінювання

Внаслідок випадковості результатів спостережень неможливо встановити достатньо вузькі межі, за які помилка точкової оцінки (тобто відхилення оцінки від істинного значення параметру, що оцінюється) не виходила б з повною гарантією. Чим меншим є об'єм вибірки, тим більше точкова оцінка може відрізнятися від істинного значення параметру.

Тому крім точкового оцінювання у статистиці застосовується **інтервальне оцінювання параметрів**: на основі вибірки отримання інтервалу, у який попадає істинне значення параметру з даною ймовірністю.

Завдання: ми не знаємо наскільки $\hat{\theta}$ відрізняється від істинного значення параметру θ. Необхідно знайти таку величину Δ («дельта»), яка з «практичною достовірністю» (тобто із заздалегідь заданою ймовірністю, яка є близькою до 1) гарантувала б, що виконується нерівність:

$$|\hat{\theta} - \theta| < \Delta$$

Δ – називають **похибкою вибірки**.

Довірчий інтервал для параметра θ – інтервал $(\hat{\theta}_1, \hat{\theta}_2)$, який містить у собі істинне значення параметру θ із заданою ймовірністю $P = 1 - \alpha$:

$$P(\hat{\theta}_1 < \theta < \hat{\theta}_2) = 1 - \alpha.$$

$1 - \alpha$ – довірча ймовірність, значення довірчої ймовірності залежить від конкретних умов, воно найчастіше дорівнює 0,90; 0,95; 0,99.
α – рівень значущості.
$\hat{\theta}_1$ та $\hat{\theta}_2$ визначаються по вибірці і є випадковими величинами.

Довірчі інтервали можуть бути двосторонніми та односторонніми:
$P(\hat{\theta}_1 < \theta < \hat{\theta}_2) = 1 - \alpha$ – двосторонній довірчий інтервал,
$P(\theta < \hat{\theta}_2) = 1 - \alpha$ – лівосторонній довірчий інтервал,
$P(\hat{\theta}_1 < \theta) = 1 - \alpha$ – правосторонній довірчий інтервал.

Для побудови довірчого інтервалу необхідно отримати статистику $Z = Z(x_1, ..., x_n, \theta)$ – функцію від елементів вибірки, – розподіл якої не залежить від θ та інших невідомих параметрів. Тоді з розподілу Z, можна визначити такі z_1 та z_2, що: $P(z_1 < Z < z_2) = 1 - \alpha$.
Тоді розв'язуючи нерівність $z_1 < Z(x_1, ..., x_n, \theta) < z_2$ відносно θ, отримуємо границі довірчого інтервалу.

Приклад.
Випадкова величина X має нормальний розподіл $N(m, \sigma^2)$. Знайти довірчий інтервал для m за результатами n спостережень, якщо σ^2 відома, довірча ймовірність дорівнює $1 - \alpha$.

Нам вже відомо, що ефективною, конзистентною, незсуненою оцінкою математичного сподівання є вибіркове середнє.

Можна показати, що вибіркове середнє має нормальний розподіл з параметрами m та σ^2/n:

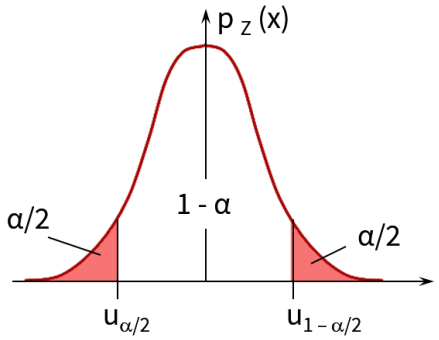
$$\bar{X} \sim N(m, \sigma^2/n).$$

Розглянемо статистику Z:

$$Z = \frac{\bar{X} - m}{\sigma/\sqrt{n}}$$

Ця статистика має стандартний нормальний розподіл: Z ~N(0,1).

Ймовірність того, що значення статистики Z знаходиться між квантиллю порядку α/2 та квантиллю порядку (1 – α/2) дорівнює 1– α, як видно з графіку щільності розподілу Z:



$$P\left(u_{\alpha/2} < Z < u_{1-\alpha/2}\right) = 1 - \alpha$$

Розглянемо нерівність:

$$u_{\alpha/2} < Z < u_{1-\alpha/2}$$

Або:

$$u_{\alpha/2} < \frac{\bar{X} - m}{\sigma/\sqrt{n}} < u_{1-\alpha/2}$$

Розв’язуючи нерівність відносно m, отримуємо довірчий інтервал для m:

$$\bar{X} - u_{1-\alpha/2} \cdot \sigma/\sqrt{n} < m < \bar{X} + u_{1-\alpha/2} \cdot \sigma/\sqrt{n}$$

Для побудови довірчих інтервалів за вибірками малих об’ємів нормальний розподіл генеральної сукупності є необхідним. При n > 50 формулою довірчого інтервалу можна користуватись і за відсутності нормального розподілу генеральної сукупності, тому що діє Центральна гранична теорема.

Перевірка статистичних гіпотез

Статистична гіпотеза – деяке припущення про генеральну сукупність, яке перевіряється за допомогою вибіркової сукупності.

Перевірка статистичних гіпотез – процедура обґрунтованого співставлення сформульованої гіпотези з отриманими в результаті експерименту вибірковими даними $x_1, x_2, \dots, x_n$. Процедура супроводжується кількісною оцінкою ступеню достовірності отриманого висновку та здійснюється за допомогою відповідного статистичного критерію.

Результат співставлення може бути або **негативним** (дані спостереження протирічать гіпотезі, що перевіряється, тому від неї слід відмовитися), або **не негативним** (дані спостереження не суперечать гіпотезі, що перевіряється, відповідно її можна прийняти як одне з природних і допустимих рішень).

Приклади статистичних гіпотез та їх застосування у маркетингу

1. Гіпотеза про числове значення параметра генеральної сукупності.
Приклад: Чи можна вважати, що середня ціна на певний продукт в точках конкурентів вища, ніж ціна у нашому магазині?

2. Гіпотеза про рівність параметрів у двох чи більше генеральних сукупностях. Приклад: Нехай респонденти оцінюють деякі характеристики товару за 5-бальною шкалою. Чи можна стверджувати, що одна з характеристик є важливішою за іншу для споживачів у генеральній сукупності (медіанний критерій). Чи можна впорядкувати характеристики за ступенем важливості для споживачів у генеральній сукупності?

3. Гіпотеза про однорідність двох чи більше вибірок (про рівність розподілу генеральної сукупності).
Приклад: Чи можна вважати, що частки споживачів, що купують різні моделі Honda, відрізняються для чотирьох вікових груп в генеральній сукупності?

4. Гіпотеза про тип розподілу генеральної сукупності.
Приклад: Чи можна вважати, що розподіл цін на куряче філе в роздрібній торгівлі є нормальним?

5. Гіпотеза про відсутність взаємозв'язку між ознаками, що аналізуються (про рівність нулю коефіцієнта кореляції).
Приклад 1: Чи залежить сума чеку від часу, який споживачі проводять у супермаркеті? (метричні дані)
Приклад 2: Чи залежить думка споживачів про якість соку від його ціни? (неметричні дані)

6. Гіпотеза про загальний вид моделей, що описують статистичну залежність між ознаками.
Приклад: Чи можна вважати залежність між часом, проведеним у супермаркеті, і сумою чеку лінійною на інтервалі від 10 до 40 хвилин.

Які бувають гіпотези?

Якщо розподіл генеральної сукупності є відомим, і за допомогою вибірки необхідно перевірити припущення про значення параметра цього розподілу, – це **параметрична гіпотеза**.

Якщо гіпотеза стосується не значень певних параметрів, які визначають розподіл, а визначає тип ймовірнісних розподілів, які має генеральна сукупність, що розглядається, – це **непараметрична гіпотеза** (наприклад, гіпотеза про те, що два невідомих розподіли є однаковими).

Якщо гіпотеза однозначно визначає розподіл випадкової величини – це **проста** гіпотеза (H_0: $\theta = \theta_0$), інакше – **складна** (H_1:$\theta \neq \theta_0$, H_2: $\theta > \theta_0$).

Гіпотезу, яку перевіряють, називають **нульовою** або **основною** і позначають $\mathbf{H_0}$.
Одночасно з гіпотезою H_0 розглядається одна з **альтернативних** гіпотез H_1.
Наприклад, якщо H_0: $\theta = \theta_0$, можна розглядати такі альтернативні гіпотези:
1. $H_1^{(1)}$:$\theta \neq \theta_0$
2. $H_1^{(2)}$:$\theta < \theta_0$
3. $H_1^{(3)}$:$\theta > \theta_0$
4. $H_1^{(4)}$:$\theta = \theta_1$ $(\theta_1 \neq \theta_0)$.

Вибір типу альтернативної гіпотези залежить від конкретної задачі.
За нульову гіпотезу частіше всього приймають таку, яку бажано відхилити.

Правило, за яким гіпотезу H_0 приймають або відхиляють (відповідно відхиляють або приймають альтернативну гіпотезу H_1), називають **критерієм перевірки гіпотез.**

Перевірка гіпотези здійснюється на основі вибірки $x_1, ..., x_n$ – реалізацій n незалежних випадкових величин за допомогою статистики Z, яка побудована за вибіркою Z ($x_1, ..., x_n$). Статистика – це функція:

$$Z : R^n \to V,$$

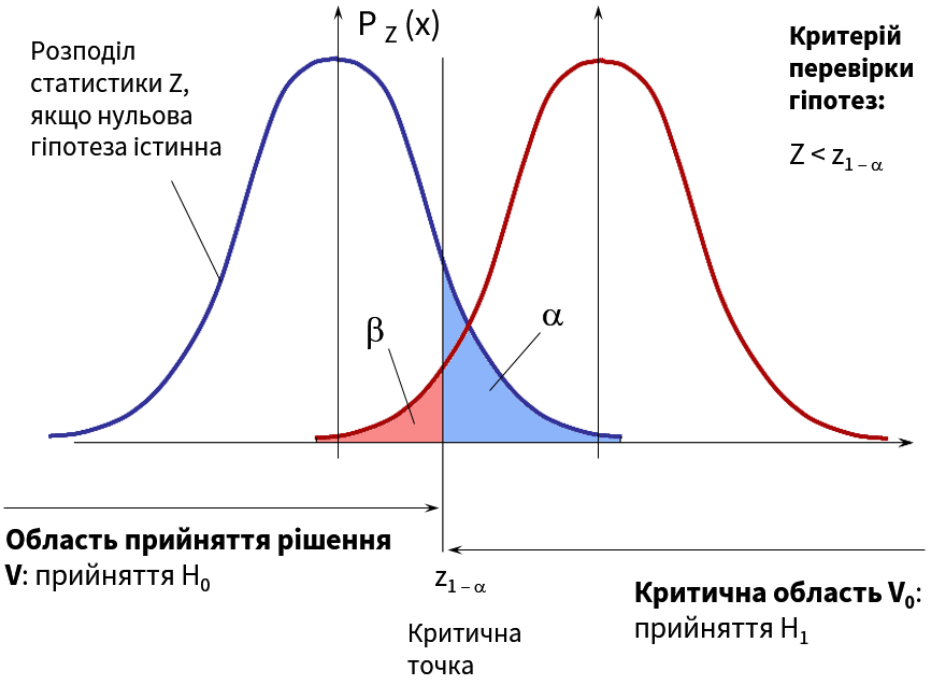
де – V область значень статистики Z.
Ця область поділяється на область прийняття основної гіпотези – **область прийняття рішення** H_0: V_0 та область відхилення основної гіпотези V_1 .

$$V = V_0 \cup V_1$$

Область відхилення V_1 – множина тих значень статистики, які ведуть до відхилення нульової гіпотези, – називається **критичною областю.**
Це така область, в яку значення статистики Z (за умови, що гіпотеза H_0 є істинною), потрапляє з малою ймовірністю α:

$$P [Z \in V_1] = \alpha$$

Ймовірність α фіксують перед здійсненням аналізу і називають **рівнем значущості**. Цю ймовірність обирають залежно від вартості помилок. Найчастіше рівень значущості приймають 0,05 (за певних умов він може бути збільшений до 0,1 або зменшений до 0,02; 0,01).



Помилка I-го роду: відхилення гіпотези H_0, за умови, що вона є істинною. Ймовірність такої помилки – α.
Помилка II-го роду: прийняття гіпотези H_0, якщо вона є хибною. Ймовірність помилки II-го роду позначають β.

Гіпотеза H_0	**Умовна ймовірність того, что**	
	Відхилена H_0 (прийнята H_1)	**Прийнята H_0 (відхилена H_1)**
Є істинною	$\alpha = P(H_1 / H_0)$ Помилка I-го роду	$1 - \beta = P(H_0 / H_0)$ Правильне рішення
Є хибною (Істинною є H_1)	$1 - \alpha = P(H_1 / H_1)$ Правильне рішення	$\beta = P(H_0 / H_1)$ Помилка II-го роду

Таким чином, якщо приймають альтернативну гіпотезу, то вона є справедливою з ймовірністю не менше, ніж $(1 - \alpha)$. Ця величина є відомою, її визначає дослідник. Ось чому звичайно намагаються прийняти альтернативну гіпотезу (відхилити нульову гіпотезу).

Зрозуміло, що бажано зменшити помилки і I-го, і II-го роду, але із зменшенням однієї, інша збільшується. Тому значення α обирають залежно від того, як співвідноситься ризик помилок I-го та II-го роду. (Наприклад: нульова гіпотеза – новий міст витримує вагу p, альтернативна – міст витримає вагу меншу, ніж p. Помилка I-го роду – це прийняти цю вагу, хоча насправді, міст її не витримає. Вочевидь, що вартість такої помилки набагато більша, ніж помилки II-го роду – прийняття меншої ваги, хоча насправді міст витримає і вагу p).

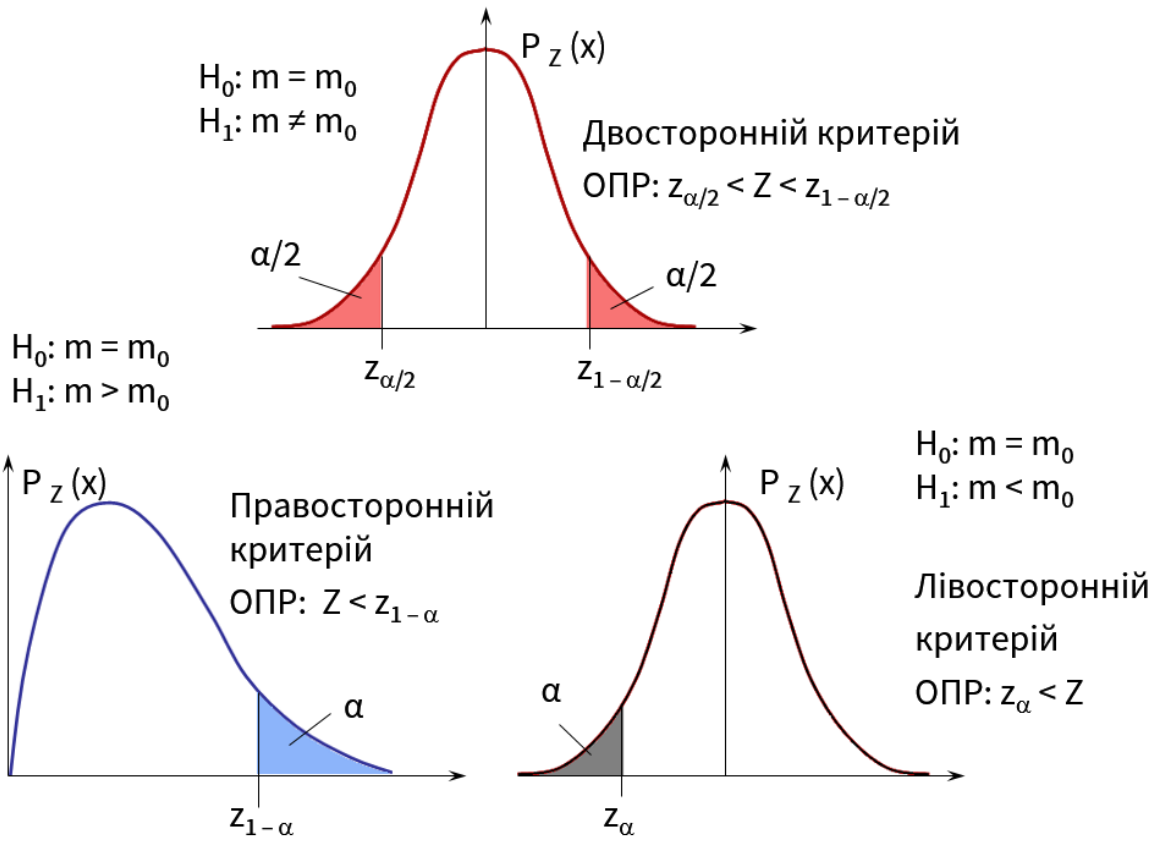
Ймовірність $(1-\beta)$ називається **потужністю критерія** перевірки гіпотези. Для перевірки тієї самої гіпотези можна використовувати різні критерії. Якість критерія залежить від властивостей даного критерія у випадках виконання чи невиконання гіпотези, яку перевіряють. Чим меншою є ймовірність β при заданому рівні значущості α, тим більш потужним є критерій.

Критерій перевірки гіпотез формулюють таким чином: для заданого рівня значущості α прийняти гіпотезу H_0, якщо $Z \in V_0$, або відхилити гіпотезу H_0 та прийняти H_1, якщо $Z \notin V_0$. Критерій перевірки гіпотез для заданого рівня значущості називають **критерієм значущості.**

Розташування критичної області V_1 на множині значень V статистики Z залежить від альтернативної гіпотези і обмежується відповідними квантилями z_α розподілу статистики Z (ці точки називають **критичними точками**), за умови, що $H_0: \theta = \theta_0$ є істинною.
Критичні точки належать до критичної області!

Критерії значущості для перевірки гіпотез:

Альтернативна гіпотеза H_1	Область V_0 прийняття гіпотези H_0	Критерій	Розташування критичної області V
$\theta \neq \theta_0$	$z_{\alpha/2} < Z < z_{1-\alpha/2}$	двосторонній	$(-\infty; z_{\alpha/2}] \cup [z_{1-\alpha/2}; \infty)$
$\theta > \theta_0$	$Z < z_{1-\alpha}$	правосторонній	$[z_{1-\alpha}; \infty)$
$\theta < \theta_0$	$z_\alpha < Z$	лівосторонній	$(-\infty; z_\alpha]$



Процедура перевірки статистичних гіпотез передбачає такі етапи:

1. Вибір нульової (основної) гіпотези H_0 та альтернативної гіпотези H_1.
2. Формулювання припущень, які включають умови на тип розподілу величин, які аналізують.
3. Вибір рівня значущості α.
4. Побудова за вибіркою статистики Z критерія перевірки гіпотези H_0.
5. Визначення критичної області.
6. Прийняття статистичного рішення:
   - прийняття H_0, якщо статистика $Z \in V_0$,
   - відхилення гіпотези H_0 та прийняття гіпотези H_1, якщо $Z \notin V_0$, тобто потрапляє в критичну область.

Для прийняття статистичного рішення перевіряють нерівність, яка визначає область прийняття рішення (область прийняття гіпотези H_0).

Якщо **нерівність виконується**, тобто статистика Z потрапляє в область прийняття рішення, то приймаємо H_0, яка є справедливою з певною невідомою ймовірністю $(1 - \beta)$. І робимо висновок, що наші емпіричні дані узгоджуються з висунутим припущенням (**інакше**, не дозволяють нам відкинути нульову гіпотезу, **інакше**, недостатньо статистично значущої різниці для того, щоб відхилити H_0). Тому якщо приймаємо H_0, то **не говоримо про ймовірність** того, що вона є справедливою!

Якщо **нерівність не виконується**, тобто статистика Z потрапляє в критичну область, то відхиляємо H_0 і приймаємо H_1, яка є справедливою з заданою, а отже, відомою ймовірністю $(1 - \alpha)$. Робимо висновок, що **з ймовірністю $(1 - \alpha)$** існує статистично значуща різниця між емпіричними даними та висунутою гіпотезою.

Статистичні тести

В сучасних програмних продуктах статистичної обробки даних реалізована не процедура перевірки статистичних гіпотез, а перевірка статистичних тестів. Розглянемо, в чому полягає спільне та відмінності між ними.

Спільне в процедурах перевірки статистичних гіпотез та перевірки статистичних тестів: задаються **гіпотези**, які перевіряють **на вибіркових даних** за допомогою відповідного **критерія**.

Різниця:

- в процесі перевірки статистичних гіпотез на вході задають **рівень значущості** α, а на виході отримують статистичне **рішення**: вибір однієї з двох альтернативних гіпотез
- в процесі перевірки тестів рівень похибки не задають, а в результаті тесту отримують число s – **значущість тесту**, за значенням якого дослідник повинен прийняти рішення.

Значущість тесту (рівень значущості, що спостерігається або досягається в експерименті) s – допустима мінімальна ймовірність помилки I-роду, тобто ймовірність відхилення гіпотези H_0, якщо вона істинна.

Рівень значущості α та значущість тесту s пов'язані між собою наступним чином:

$$s > \alpha \Rightarrow H_0$$

$$s \leq \alpha \Rightarrow H_1$$

Елементи дисперсійного аналізу

Дисперсійний аналіз – статистичний метод, призначений для оцінки впливу якісних факторів на результат експерименту, що вимірюється кількісно.

Сутність методу полягає у тому, що загальна варіація результуючого показника ділиться на частини, що відповідають окремому і спільному впливу різних якісних факторів, і остаточну варіацію, що виникає у наслідок впливу всіх неврахованих факторів.

По числу факторів, вплив яких досліджується, розрізняють:
- Однофакторний дисперсійний аналіз (ОДА);
- Двохфакторний дисперсійний аналіз;
- Багатофакторний дисперсійний аналіз.

Однофакторний дисперсійний аналіз

В основі ОДА лежить теоретико-ймовірнісна модель:

$$X_{ij} = m_i + \varepsilon_{ij},$$

де:

i = 1, …, r – кількість значень якісного фактору,

j = 1, … , n_i – кількість спостережень при і-му значенні якісного фактору, причому:

$$\sum_{i=1}^{r} n_i = n$$

n – загальний об’єм вибірки (загальна кількість спостережень при всіх значеннях якісного фактору),

X_{ij} – випадкові величини, що представляють результуючу ознаку,

m_i – математичне сподівання результуючої ознаки при і – му значенні якісного фактору.

ε_{ij} – випадкове нормально розподілене відхилення результуючої ознаки від середніх (збурення, викликане впливом неконтрольованих факторів).

$E(\varepsilon_{ij})=0$; $D(\varepsilon_{ij})=\sigma^2$ (ці випадкові похибки мають нульове математичне сподівання та однакову дисперсію).

У результаті проведення вибіркового експерименту отримуємо r-груп вибіркових значень результуючої ознаки:

$$\begin{cases} 1: & x_{11} & x_{12} & \dots & x_{1n_1} \\ 2: & x_{21} & x_{22} & \dots & x_{2n_2} \\ & & \dots & & \\ r: & x_{r1} & x_{r2} & \dots & x_{rn_r} \end{cases}$$

Гіпотези, які необхідно перевірити:

H_0: $m_1 = m_2 = \dots = m_r$ – якісний фактор не впливає на результуючу ознаку

H_1: ∃ ≠ (існує хоча б одне середнє, яке не дорівнює іншим) – якісний фактор впливає.

Введемо позначення для загального вибіркового середнього:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{ij}$$

та групових вибіркових середніх:

$$\overline{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$$

Розглянемо суму квадратів відхилень результуючої ознаки від загального середнього.

$$Q^2 = \sum_{i=1}^{r}\sum_{j=1}^{n}(x_{ij} - \bar{X})^2$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{n}\left((x_{ij} - \bar{X}_i) + (\bar{X}_i - \bar{X})\right)^2$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \bar{X}_i)^2 + \sum_{i=1}^{r}\sum_{j=1}^{n}(\bar{X}_i - \bar{X})^2 + 2\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \bar{X})(\bar{X}_i - \bar{X})$$

А тепер розглянемо окремо третій доданок і доведемо, що він дорівнює 0:

$$\sum_{i=1}^{r}\sum_{j=1}^{n}(x_{ij} - \bar{X}_i)(\bar{X}_i - \bar{X}) = \sum_{i=1}^{r}(\bar{X}_i - \bar{X}) * \sum_{j=1}^{n_i}(x_{ij} - \bar{X}_i)$$

де:

$$\sum_{j=1}^{n_i}(x_{ij} - \bar{X}_i) = \sum_{j=1}^{n_i} x_{ij} - n_i\,\bar{X}_i = \sum_{j=1}^{n_i} x_{ij} - n_i * \frac{1}{n}\sum_{j=1}^{n_i} x_{ij} = 0$$

Отже, третій доданок дорівнює нулю, тобто:

$$Q^2 = \sum_{i=1}^{r}\sum_{j=1}^{n}(x_{ij} - \bar{X})^2 + \sum_{i=1}^{r} n_i(\bar{X}_i - \bar{X})^2 = Q_R^2 + Q_A^2$$

Або коротко:

$$Q^2 = Q_R^2 + Q_A^2$$

І ще раз окремо:

Q_A^2 – сума квадратів відхилення між групами або варіація, обумовлена впливом якісного фактору:

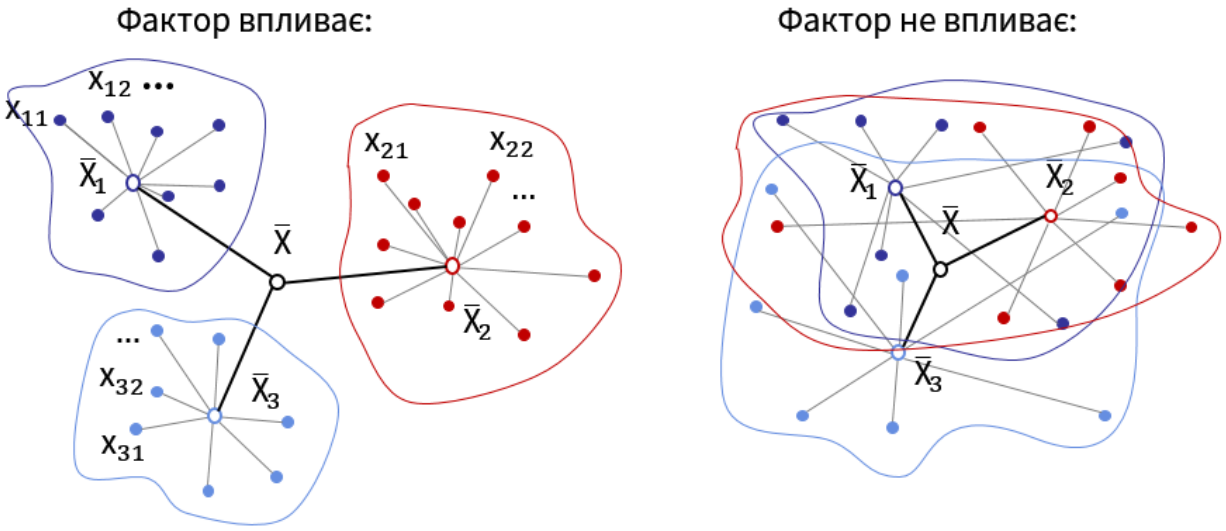
$$Q_A^2 = \sum_{i=1}^{r} n_i(\bar{X}_i - \bar{X})^2$$

Q_R^2 – сума квадратів відхилень всередині груп, залишкова варіація, обумовлена випадковими відхиленнями від групових середніх.

$$Q_R^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \bar{X}_i)^2$$

Сутність дисперсійного аналізу полягає у порівнянні варіацій між групами спостережень, кожна із яких відповідає одному значенню якісного фактору, та сумі випадкових варіацій всередині груп. Якщо перша варіація значно перевищує другу, це означає, що якісний фактор впливає на результуючу (кількісну) змінну. Якщо перша варіація незначно відрізняється від другої, то фактор не впливає.

Графічна інтерпретація сутності дисперсійного аналізу зображена на рисунку.



$x_{11}, x_{12}, \ldots$ – значення показника для значення фактору 1

$\bar{X}_1$ – середнє значення показника для значення фактору 1

$\bar{X}$ – середнє значення результуючого показника для всіх значень фактору

Критерій перевірки гіпотез:

H_0: $m_1=m_2=\ldots=m_r$

H_1: $\exists\neq$

Статистика критерію:

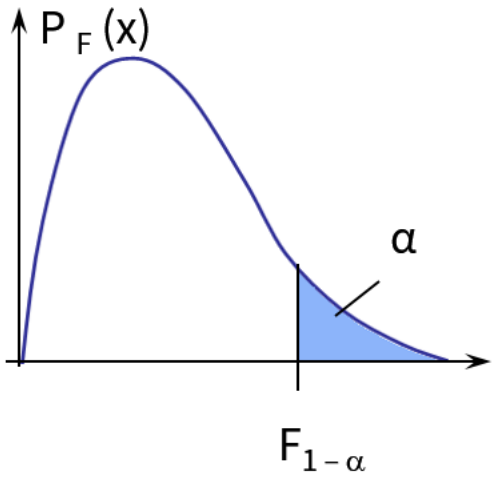
$$F = \frac{Q_A^2/(r-1)}{Q_R^2/(n-r)}$$

Статистика має розподіл Фішера з (r–1) та (n–r) ступенями свободи:

$$F \sim F(r-1;\ n-r)$$

Область прийняття рішення визначається наступним чином:

$$F < F_{1-\alpha}(r-1;\ n-r).$$



Якщо H_0 вірна, то $\bar{X}_i$ являються незсуненими та конзистентними оцінками одного й того ж математичного сподівання m і відповідно близькі між собою, тому Q_A^2 мала у порівнянні із Q_R^2.

Приклад

У трьох містах тестуються три рекламні концепції. Фіксуються зміни у відсотках об'ємів продажів у випадковим чином обраних торгових точках. Чи можна вважати всі три рекламні концепції однаково успішними.

Місто 1: 12,5% 14,3% 15,6% 12,7% 10,2% 12,4%

Місто 2: 10,1% 8,4% 12,7% 12,0% 11,6%

Місто 3: 4,2% 9,4% 14,6% 5,6% 7,8% 9,4% 11,2%

Розв’язання:

Кількість значень фактору: r=3

Кількість спостережень по кожному значенню фактору:

$n_1=6 \quad n_2=5 \quad n_3=7$

Загальна кількість спостережень: n=18

$$\bar{X}_1 = \frac{1}{6}(12{,}5 + \cdots + 12{,}4) = 12{,}95$$

$$\bar{X}_2 = \frac{1}{5}(10{,}1 + \cdots + 11{,}6) = 10{,}96$$

$$\bar{X}_3 = \frac{1}{7}(4{,}2 + \cdots + 11{,}2) = 8{,}89$$

$$\bar{X} = \frac{1}{18}(12{,}5 + \cdots + 11{,}2) = 10{,}82$$

$$Q_A^2 = 6(12{,}95 - 10{,}82)^2 + 5(10{,}96 - 10{,}82)^2 + 7(8{,}89 - 10{,}82)^2 = 53{,}51$$

$$Q_R^2 = (12{,}5 - 12{,}95)^2 + (14{,}3 - 12{,}95)^2 + \cdots + (10{,}1 - 10{,}96)^2 + (8{,}4 - 10{,}96)^2 + \cdots + (4{,}2 - 8{,}89)^2 + (9{,}4 - 8{,}89)^2 \ldots = 101{,}26$$

Зведемо все це у таблицю (аналогічну таблицю ви побачите у вікні виводу в SPSS):

Компоненти варіації	Сума квадратів	Число ступенів свободи	Середній квадрат	F
Міжгрупова	53,51	2	26,765	3,963
Внутрішньогрупова	101,26	15	6,75	
Загальна	154,77	17		

Якщо брати рівень значущості α = 0,05, тоді: $F_{0,95}(2, 15) = 3{,}68$

Перевіряючи нерівність (ОПР), отримуємо:

$F < F_{0,95}(2, 15)$: 3,963 > 3,68.

Приймаємо альтернативну гіпотезу, тобто з ймовірністю 0,95 можемо стверджувати, що середні зміни обсягів продажів відрізняються залежно від міста (або фактор – варіант концепції – впливає на зміну обсягів продажів).

Непараметричні критерії

Усі статистичні критерії (як ми вже відзначали) поділяють на критерії:
– **параметричні** (перевіряють рівність певних параметрів генеральної сукупності заданим значенням)
– **непараметричні** (критерії про вид розподілу в цілому).

Непараметричні критерії, в свою чергу, поділяють на критерії згоди та критерії однорідності.

Критерії згоди – це одновибіркові критерії, за допомогою яких перевіряють, чи дійсно вибірка взята з генеральної сукупності, яка має певний конкретний розподіл (у деяких випадках, з точністю до параметрів $\theta_1, \dots, \theta_l$ розподілу генеральної сукупності). Перевірка полягає у відповіді на питання: чи **узгоджуються** емпіричні дані (вибірка) з припущенням, сформульованим в основній гіпотезі (або протирічать їй).

Критерії однорідності – критерії, за допомогою яких перевіряють, чи були дві або більше вибірок відібрані з **однорідних** у певному сенсі генеральних сукупностей. Однорідність можуть розуміти при цьому у різних сенсах: від збігу значень деяких конкретних параметрів (наприклад, медіанний критерій – рівність медіан) до повного збігу функцій розподілу (критерій Колмогорова-Смірнова).

Звичайно, критеріїв існує дуже велика кількість для різних розподілів, різних вибірок тощо. Розглянемо як приклади критеріїв згоди та критеріїв однорідності, критерій Колмогорова-Смірнова (він є найпотужнішим в своєму класі критеріїв) та Хі-квадрат критерій, які є найбільш універсальним інструментом, тому що може бути застосований для великої кількості зовсім різних завдань.

Критерій згоди Хі-квадрат Пірсона

Цей критерій дозволяє перевірити гіпотези про вид розподілу як дискретної, так і неперервної випадкової величини, у випадках, якщо параметри $\theta_1, \dots, \theta_l$ генеральної сукупності є відомими або невідомими.

$$H_0: F(x) = F_{мод}(x)$$
$$H_1: F(x) \neq F_{мод}(x)$$

де $F(x)$ – функція розподілу випадкової величини, яку ми перевіряємо (значення вибірки, яку ми тестуємо, є реалізаціями даної випадкової величини),

$F_{мод}(x)$ – модельна функція розподілу випадкової величини, з якою ми порівнюємо випадкову величину, яку перевіряємо.

Обмеження на використання критерію:

$$\begin{cases} n \geq 30 \\ n_i^* \geq 5 \end{cases}, \qquad i = \overline{1,k}$$

де n – об'єм вибірки,
n_i^* – емпірична частота і-го інтервалу,
k – кількість інтервалів.

Статистика критерію:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i^* - n \cdot p_i)^2}{n \cdot p_i}$$

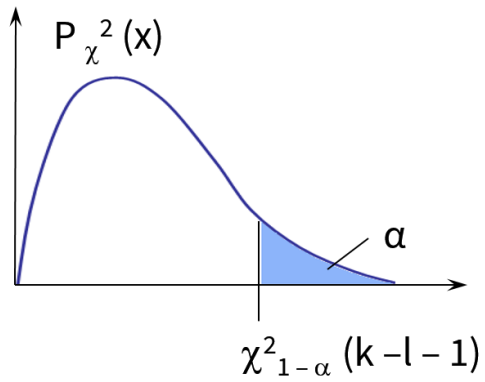
Розподіл статистики:

$$\chi^2 \sim \chi^2(k - l - 1)$$

де k – кількість інтервалів,
l – кількість невідомих параметрів генеральної сукупності,
n_i^* - емпірична частота і-го інтервалу,
$n \cdot p_i$ – теоретична (очікувана) частота і-го інтервалу.

Критична область завжди розміщена справа (не плутати з параметричними критеріями!).
Область прийняття рішення:

$$\chi^2 < \chi^2_{1-\alpha}(k - l - 1)$$



Процедура перевірки складається з наступних етапів:
1. Будують групований статистичний ряд

$$(x_i^*, n_i^*)$$

так, щоб $n_i^* \geq 5$ для всіх інтервалів (якщо є інтервали з недостатньою кількістю елементів, їх об'єднують між собою)

2. Вважаючи істинною гіпотезу H_0 , розраховують p_i – теоретичні ймовірності того, що випадкова величина знаходиться в і-му інтервалі

3. Розраховують очікувані частоти $n \cdot p_i$ для всіх інтервалів

4. Розраховують статистику χ^2

5. Перевіряють нерівність:

$$\chi^2 < \chi^2_{1-\alpha}(k - l - 1)$$

6. Приймають статистичне рішення:
- якщо нерівність виконується, приймаємо H_0 (відхиляємо H_1), отже, не можемо довести, що емпіричні дані протирічать висунутій гіпотезі – генеральна сукупність, з якої отримано вибірку, може мати даний розподіл
- якщо нерівність не виконується, приймаємо H_1 (відхиляємо H_0), отже, з ймовірністю (1 – α) можемо стверджувати, що генеральна сукупність, якої отримано вибірку, має інший розподіл.

Приклад

Візьмемо підвибірку цін на куряче філе:

73,9	82,25	77,95	78,49	85,31	77,9	77	69,95	68	69,8	65,59	76,99	80,45	71,93	73
	74,99	76	68	63,99	64,95	72,9	70,99	62,99	75	80,45	79,98	80,45	74,89	
77,99	71	76,89	75,3	80,45	77,99	74,99	76,3	75,4	82,5					

(Ми вже працювали з цією підвибіркою у темі описова статистика)
Об'єм вибірки: n = 38
Вибіркове середнє: $\overline{X}$=75 (заокруглене до цілого значення)
Оцінка середньоквадратичного відхилення: S = 5,4
Перевіримо гіпотезу про те, що ціна на куряче філе в генеральній сукупності має нормальний розподіл з математичним сподіванням 75 та дисперсією 29 (α = 0,05):

H_0: X ~ N(75, 29)
H_1: X $\nsim$ N(75, 29)

Побудуємо груповий статистичний ряд:

№ інтервалу	Границі інтервалу	Середина інтервалу x_i^*	Частота n_i^*	Теоретична ймовірність p_i	Очікувана частота $n \cdot p_i$
1	(60 – 70]	65	8	0,177	6,6
2	(70 – 75]	72,5	10	0,323	12,3
3	(75 – 80]	77,5	13	0,323	12,3
4	(80 – 90]	85	7	0,177	6,6

Розрахунок теоретичної ймовірності
Для першого інтервалу необхідно знайти ймовірність того, що значення випадкової величини, що має нормальний розподіл з середнім 75 та дисперсією 29 знаходиться в інтервалі (60 – 70].

Згадайте, як ми це робили в теорії ймовірностей:

$$P(60 < X < 70) = F (70) – F (60)$$

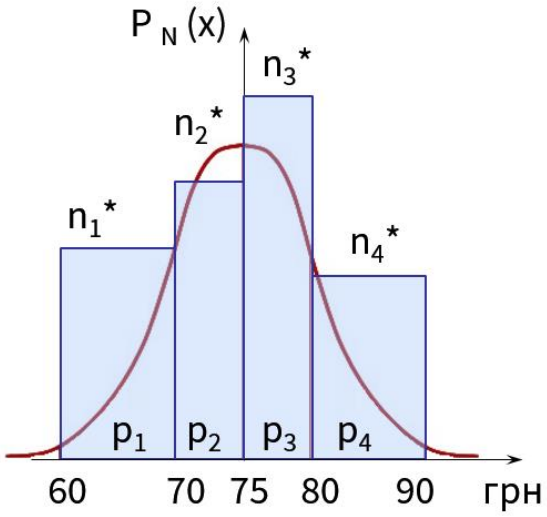
Отже, необхідно знайти значення відповідної теоретичної функції розподілу в точках 70 та 60. Для цього в MS Excel є відповідна функція:

НОРМ.РАСП(x;среднее;стандартное_откл;интегральная)
(Функція повертає функцію розподілу, якщо «интегральная» = «true», або щільність розподілу, якщо «интегральная» = «false»)

Для першого інтервалу нам необхідно розрахувати: = НОРМ.РАСП(70;75;5,4;1) – НОРМ.РАСП(60;75;5,4;1) = 0,1772 – 0,0027 = 0,175

Тоді для другого інтервалу буде: = НОРМ.РАСП(75;75;5,4;1) – НОРМ.РАСП(70;75;5,4;1) = 0,5 – 0,1772 = 0,323

Графічна інтерпретація розв'язання задачі:



Далі розраховуємо статистику Хі-квадрат:

$$\chi^2 = \frac{(8-6{,}6)^2}{6{,}6} + \frac{(10-12{,}3)^2}{12{,}3} + \frac{(13-12{,}3)^2}{12{,}3} + \frac{(7-6{,}6)^2}{6{,}6} = 0{,}79$$

Невідомих параметрів немає (математичне сподівання та дисперсію ми задали конкретними значеннями). Статистика має розподіл Хі-квадрат з трьома ступенями свободи.
Перевіряємо нерівність:

$$\chi^2 < \chi^2_{0{,}95}(4-1) = \chi^2_{0{,}95}(3) = 7{,}81$$

Або:

$$0{,}79 < 7{,}81$$

Висновок: нерівність виконується, отже приймаємо нульову гіпотезу. Наші емпіричні дані не дають можливості спростувати припущення про те, що розподіл ціни на куряче філе в генеральній сукупності може бути нормальний з математичним сподіванням 75 грн та дисперсією 29 грн^2.

Приклади використання критерію Хі-квадрат для неметричних даних у відео:
Хі-квадрат критерій згоди (двозначна вибірка)
Хі-квадрат критерій згоди (багатозначна вибірка, частки однакові)
Хі-квадрат критерій згоди (багатозначна вибірка, частки різні)

Критерій згоди Колмогорова-Смірнова

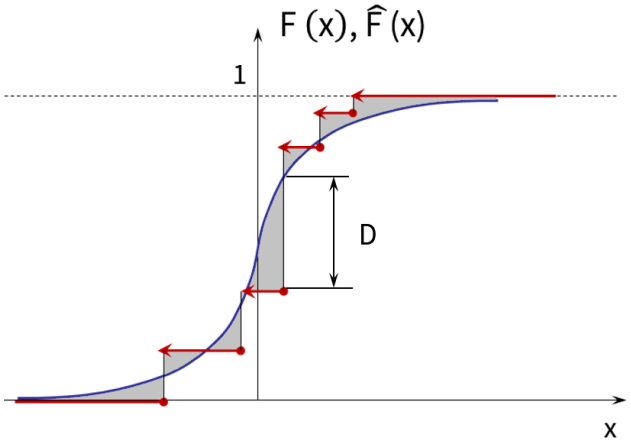
Критерій дозволяє перевірити гіпотезу про те, що генеральна сукупність, з якої отримана вибірка, має певний **неперервний** розподіл з відомими параметрами.

$$H_0: F(x) = F_{мод}(x)$$
$$H_1: F(x) \neq F_{мод}(x)$$

Для побудови статистики критерію по вибірці будують емпіричну функцію розподілу та порівнюють з теоретичною функцією.

Статистику Колмогорова розраховують, як максимальну відстань між емпіричною та теоретичною функцією розподілу по всім значенням x:

$$D = \max_x \left| \hat{F}(x) - F_{мод}(x) \right|$$



Якщо об'єм вибірки є малим (n до 35) використовують таблицю критичних значень.
Якщо n більше 35 користуються наступною таблицею відповідності критичних значень рівням значущості α:

α	0,1	0,05	0,01
$D_{кр}$	$\frac{1,22}{\sqrt{n}}$	$\frac{1,36}{\sqrt{n}}$	$\frac{1,63}{\sqrt{n}}$

Область прийняття рішення (прийняття нульової гіпотези) визначається наступним чином:

$$D < D_{кр}$$

Якщо нерівність виконується, приймають основну гіпотезу і роблять висновок про те, що генеральна сукупність, яка перевіряється (з якої була отримана вибірка) **може мати даний розподіл.**
Якщо нерівність не виконується, приймають альтернативну гіпотезу: **з ймовірністю (1 - α) генеральна сукупність має інший розподіл.**

NB! Необхідно пам'ятати, що ми перевіряємо конкретний розподіл з конкретними заданими параметрами. Тож, якщо ми довели що генеральна сукупність не має нормальний розподіл з

параметрами 8 та 24, це не означає, що вона не може бути нормально розподіленою, але з іншими параметрами!

Приклад

Перевіримо, що генеральна сукупність, з якої отримана вибірка цін на куряче філе, має рівномірний розподіл з параметрами 60 та 90 (параметрами рівномірного розподілу є межі інтервалу).

Розв'язання задачі дивіться у відео:

Критерій згоди Колмогорова-Смірнова

Критерій однорідності Колмогорова-Смірнова

Критерій дозволяє перевірити гіпотезу про те, що дві генеральні сукупності, з яких було отримано вибірки, мають однаковий розподіл (а це означає, що їх функції розподілу співпадають):

$$H_0: F_1(x) = F_2(x)$$
$$H_1: F_1(x) \neq F_2(x)$$

Для перевірки критерію будують дві емпіричні функції розподілу та знаходять найбільшу відстань між ними на всій числовій осі. Далі розраховують статистику:

$$D = \max_x \left| \hat{F}_1(x) - \hat{F}_2(x) \right|$$

Область прийняття рішення (прийняття нульової гіпотези) визначається наступним чином:

$$D < D_{кр}$$

Критичні значення для малих об'ємів вибірок беруть з таблиці критичних значень. Для великих об'ємів вибірок користуються наступною таблицею відповідності критичних значень рівням значущості α:

α	0,1	0,05	0,01
$D_{кр}$	$1{,}22\sqrt{\frac{n_1+n_2}{n_1 n_2}}$	$1{,}36\sqrt{\frac{n_1+n_2}{n_1 n_2}}$	$1{,}63\sqrt{\frac{n_1+n_2}{n_1 n_2}}$

n_1 та n_2 – об'єми першої та другої вибірок.

Якщо нерівність виконується, приймають основну гіпотезу і роблять висновок про те, що вибірки були взяті з генеральних сукупностей, які **можуть мати однаковий розподіл.**

Якщо нерівність не виконується, приймають альтернативну гіпотезу: **з ймовірністю (1 - α) вибірки отримані з генеральних сукупностей, які мають різний розподіл.**

Приклад

Компанія випустила на ринок новий продукт, плануючи омолодити цільову аудиторію. Але середній вік покупців не змінився. Перевіримо, чи можна вважати, що розподіл споживачів за віком, які купують нову та стару модель продукту компанії, відрізняється (рівень значущості 0,1).

Розв'язання задачі дивіться у відео:

Критерій однорідності Колмогорова-Смірнова

Критерій однорідності Хі-квадрат Пірсона

За допомогою даного критерію перевіряють гіпотезу про рівність законів розподілу двох або декількох генеральних сукупностей. Використовують для неперервних, дискретних та, навіть, неметричних даних. Цей критерій можна використовувати навіть для даних, що вимірюються за номінальною шкалою.
Перевіряємо гіпотези:

$$H_0: \quad F_1(x) = F_2(x) = \ldots = F_k(x)$$
$$H_1: \quad \exists \neq$$

де F_i (x) – функція розподілу і-ої випадкової величини, k – кількість генеральних сукупностей (випадкових величин) і, відповідно, кількість вибірок.

Розв'язання задачі починається з описової, у даному випадку вже **двомірної**, статистики. Отримуємо вибірку з кожної генеральної сукупності і рахуємо кількість елементів у кожній вибірці, які мають значення від 1 до m. Отримані значення емпіричних частот (рос., «наблюдаемые») заносять в таблицю, яку називають **таблицею сполучення ознак**:

	Значення 1	…	Значення m	Сума частот по всім значенням
Вибірка 1	x_{11}	…	x_{1m}	$\sum_{j=1}^{m} x_{1j}$
Вибірка 2	x_{21}		x_{2m}	
…	…		…	
Вибірка k	x_{k1}	…	x_{km}	
Сума частот по всім вибіркам	$\sum_{i=1}^{k} x_{i1}$			Загальна сума частот $\sum_{i=1}^{k}\sum_{j=1}^{m} x_{ij}$

Далі для кожної комірки таблиці сполучення ознак розраховують очікувані частоти (яку кількість елементів максимально ймовірно було б отримано по кожному значенню у кожній вибірці, якщо частки всіх значень по всім вибіркам співпадають).

Для розрахунку **очікуваної частоти x_{rv}^{o}** значения **v** по выборке **r** необхідно суму по стовпцю **v** помножити на суму по рядку **r** і поділити результат на загальну суму емпіричних частот (загальний об'єм вибірки):

$$x_{rv}^{o} = \left(\sum_{i=1}^{k} x_{iv} \cdot \sum_{j=1}^{m} x_{rj}\right) \Big/ \sum_{i=1}^{k}\sum_{j=1}^{m} x_{ij}$$

Обмеження на використання критерію: Загальний об'єм вибірки повинен бути більше 50, всі очікувані частоти не менше 5 (якщо є категорії з меншими очікуваними частотами, їх видаляють або об'єднуть з іншими).

Після того, як розраховані усі очікувані частоти, можна розрахувати статистику:

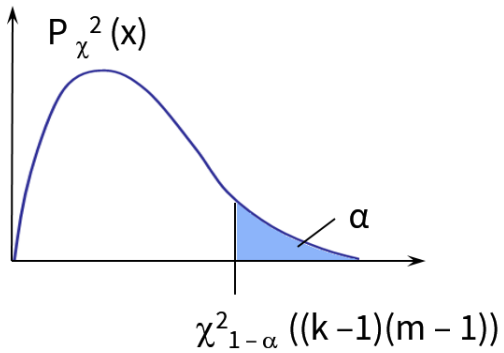
$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\left(x_{ij}^o - x_{ij}\right)^2}{x_{ij}^o}$$

Статистика має розподіл Хі-квадрат. Кількість ступенів свободи розраховують, як кількість вибірок **k мінус 1** помножити на кількість значень **m мінус 1**:

$$\chi^2 \sim \chi^2((k-1)*(m-1))$$

Область прийняття рішення (прийняття нульової гіпотези) визначають таким чином:

$$\chi^2 < \chi^2_{1-\alpha}((k-1)*(m-1))$$



NB! Якщо кількість ступенів свободи дорівнює 1 (дві двозначні вибірки) формулу для розрахунку статистики необхідно використовувати з поправкою на неперервність (поправка Йейтса):

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\left(\left|x_{ij}^o - x_{ij}\right| - 0{,}5\right)^2}{x_{ij}^o}$$

Область прийняття рішення:

$$\chi^2 < \chi^2_{1-\alpha}(1).$$

Якщо нерівність **виконується**, робимо висновок про те, що емпіричні дані не протирічать висунутій гіпотезі: вибірки, які тестуємо можуть бути отримані з генеральних сукупностей, які **мають однаковий розподіл**.
Якщо нерівність **не виконується**, отже, **з ймовірністю 1 – α існує хоча б одне значення по одній вибірці, частка якого відрізняється від інших** (не всі розподіли усіх генеральних сукупностей співпадають).

Приклади розв'язання задач дивіться у відео:
Хі-квадрат критерій однорідності (дві двозначні вибірки)
Хі-квадрат критерій однорідності (декілька багатозначних вибірок)

Кореляційний та регресійний аналіз

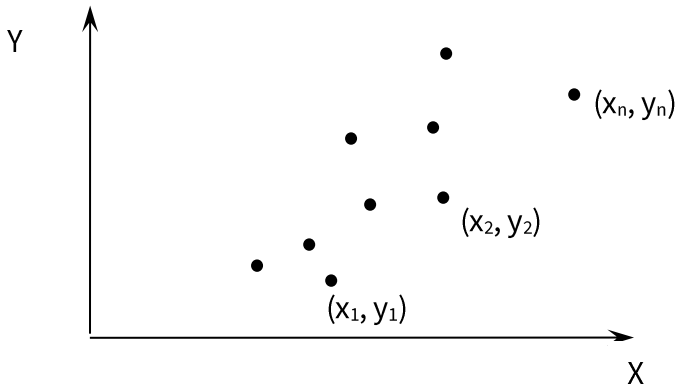
Одним з трьох основних завдань генерування інформації є вивчення взаємозв'язку процесів та явищ, їх сили та напряму. Поняття кореляції та регресії з'явилося в середині XIX ст. у працях англійських статистиків Ф. Гальтона і К. Пірсона. Термін «**кореляція**» від лат. «correlation» – співвідношення, взаємозв'язок, термін «**регресія**» від лат. «regressio» – рух назад.

Кореляційний аналіз – сукупність статистичних методів виявлення статистичних зв'язків між досліджуваними ознаками та оцінювання ступеня тісноти цих зв'язків.

У природничих науках часто йдеться про **функціональну** залежність (зв'язок), коли кожному значенню однієї змінної відповідає певне (єдине) значення іншої змінної (залежність відстані від часу, сили від маси тощо). В економіці в більшості випадків між змінними величинами існують залежності, при яких кожному значенню однієї змінної відповідає не визначене значення, а безліч можливих значень (тобто, існує певний умовний розподіл) іншої змінної. Така залежність отримала назву **статистичної** залежності. Виникнення цього поняття обумовлено тим, що залежна змінна схильна до впливу ряду неконтрольованих або не врахованих факторів, а також тим, що вимірювання значень змінних супроводжується деякими випадковими помилками.

Кореляційний аналіз. Метричні дані
Парна кореляція

Парна кореляція - взаємозв'язок між двома змінними. Нехай отримано n пар спостережень за змінними X і Y. Зображення даних точок на координатній площині називається кореляційним полем. За допомогою даного зображення можна візуально проаналізувати існування залежності між змінними.



Показником тісноти лінійного зв'язку між двома змінними є **парний коефіцієнт кореляції** (Пірсона):

$$\hat{\rho}(X,Y)=\frac{\hat{cov}(X,Y)}{\sqrt{\hat{\sigma}_X^2\cdot\hat{\sigma}_Y^2}}=\frac{\sum_{i=1}^{n}(x_i-\bar{X})(y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{X})^2(y_i-\bar{Y})^2}},\ \text{де}$$

$\hat{cov}(X,Y)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{X})(y_i-\bar{Y})$ – вибіркова коваріація змінних X та Y,

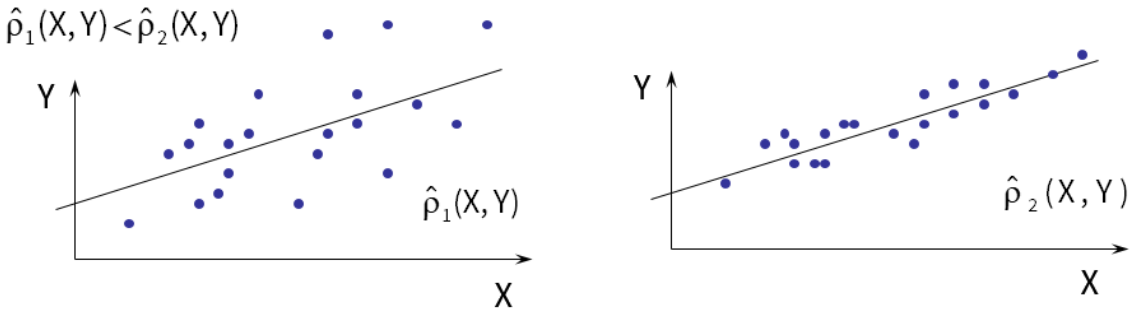
$\hat{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2$, $\hat{\sigma}_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{Y})^2$ – вибіркові дисперсії X и Y,

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n}x_i$, $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n}y_i$ – вибіркові середні X и Y.

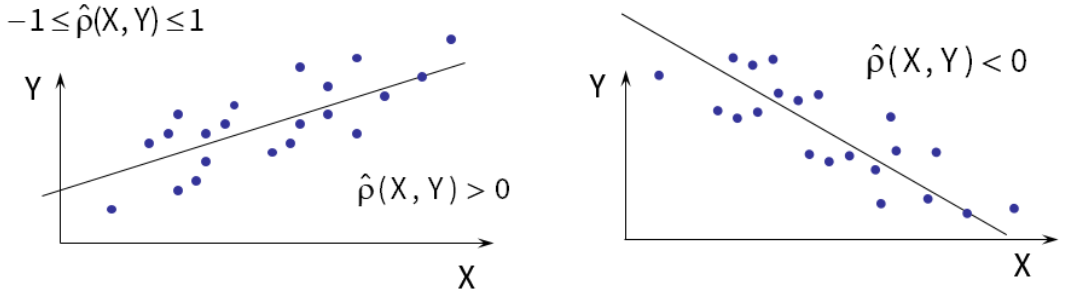
Властивості вибіркового парного коефіцієнта кореляції

1. $-1 \le \hat{\rho}(X,Y) \le 1$

Чим тісніше зв'язок між X і Y, тим більшим є модуль коефіцієнта кореляції і тим ближче точки кореляційного поля до прямої.



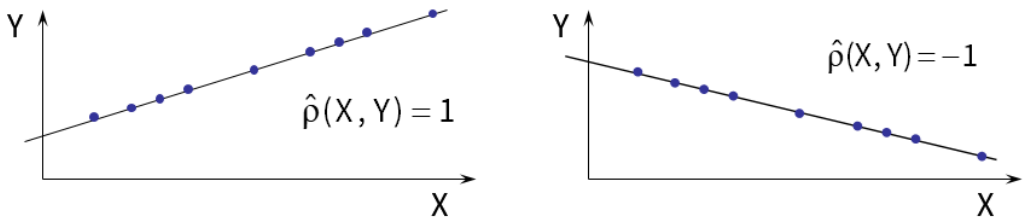
2. Якщо $\hat{\rho}(X,Y)$ значимо більше 0, між змінними X і Y існує позитивний парний зв'язок, якщо $\hat{\rho}(X,Y)$ значимо менше 0 – негативний парний зв'язок.



Позитивний статистичний зв'язок

Негативний статистичний зв'язок

3. При $\hat{\rho}(X,Y) = \pm 1$, кореляційний зв'язок являє собою лінійну функціональну залежність.



Функціональний лінійний зв'язок

4. Коефіцієнт кореляції – симетрична характеристика: $\hat{\rho}(X,Y) = \hat{\rho}(Y,X)$

5. Якщо X и Y – незалежні змінні, то $\hat{\rho}(X, Y) = 0$.
6. Якщо $\hat{\rho}(X, Y) = 0$, це означає, що між X та Y відсутній лінійний кореляційний зв'язок (але зв'язок іншої форми може існувати).

Для перевірки значущості парного коефіцієнту кореляції використовують статистичний критерій:

$H_0: \rho = 0$
$H_1: \rho \neq 0$

Нульова гіпотеза – лінійний зв'язок між змінними відсутній, альтернативна гіпотеза – існує значущий лінійний зв'язок (будьте уважні!).

Статистика критерія:

$$t = \frac{\hat{\rho} \cdot \sqrt{n-2}}{\sqrt{1-\rho^2}}, \quad t \sim T(n-2)$$

Критерій є двостороннім, область прийняття рішення:

$$|t| < t_{1-\frac{\alpha}{2}}(n-2)$$

Приклад. Метричні дані. Парний коефіцієнт кореляції Пірсона

Множинна кореляція

Нехай існує сукупність змінних $X_1, X_2, \ldots, X_k$, які мають спільний нормальний розподіл.
Побудуємо матрицю парних коефіцієнтів кореляції:

$$P = \begin{vmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1k} \\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2k} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{k1} & \rho_{k2} & \rho_{k3} & \cdots & 1 \end{vmatrix}$$

де ρ_{ij} – парний коефіцієнт кореляції між X_i та X_j.

Завдання багатовимірного кореляційного аналізу:
1. Оцінка матиці P по вибірці, тобто побудова матриці $\hat{P}$ вибіркових парних коефіцієнтів кореляції та перевірка їх значущості
2. Визначення тісноти зв'язку однієї змінної із сукупністю інших (k – 1) змінних
3. Визначення тісноти зв'язку між змінними при фіксуванні чи виключенні впливу інших змінних

Вибірковий множинний (сукупний) коефіцієнт кореляції $\hat{R}$ оцінює тісноту взаємозв'язку однієї змінної із сукупністю інших (k – 1) змінних.

У випадку k = 3 множинний коефіцієнт кореляції:

$$\hat{R}_{i.jk} = \sqrt{\frac{\hat{\rho}_{ij}^2 + \hat{\rho}_{ik}^2 - 2 \cdot \hat{\rho}_{ij} \cdot \hat{\rho}_{ik} \cdot \hat{\rho}_{jk}}{1 - \hat{\rho}_{jk}^2}}$$

Властивості множинного коефіцієнту кореляції:
1. $0 \leq \hat{R}_{i.jk} \leq 1$
2. Модуль парних коефіцієнтів кореляції між будь-якими двома змінними завжди менше відповідного множинного коефіцієнта кореляції:

$$|\hat{\rho}\ (Y, X_i)| \leq \hat{R}\ (Y, X_1, \cdots, X_k)$$

R^2 – вибірковий множинний (сукупний) **коефіцієнт детермінації**.
Він показує, яку частку варіації досліджуваної змінної Y пояснює варіація інших k змінних $X_1, X_2, \ldots, X_k$.

Для парної моделі:

$$\hat{R}^2(X, Y) = \hat{\rho}^2(X, Y)$$

Для перевірки значущості коефіцієнту детермінації використовується критерій:

$H_0: R^2 = 0$
$H_1: R^2 \neq 0$

Нульова гіпотеза – коефіцієнт детермінації є незначущим, альтернативна гіпотеза – коефіцієнт детермінації є значущим (будьте уважні!).

Статистика критерія:

$$F = \frac{\widehat{R^2} \cdot (n-k)}{(1-\widehat{R^2})(k-1)}, \quad F \sim F(k-1, n-k)$$

Критична область розміщується справа. Область прийняття рішення:

$$F < F_{1-\alpha}(k-1, n-k)$$

Вибірковий частинний (рос., частный) **коефіцієнт кореляції** між змінними X_i та X_j при фіксованих значеннях інших (k – 2) змінних дозволяє оцінити тісноту лінійного зв'язку між змінними X_i та X_j при виключенні впливу інших змінних.

У випадку k = 3 частинний коефіцієнт кореляції:

$$\hat{r}_{ij.k} = \frac{\hat{\rho}_{ij} - \hat{\rho}_{ik} \cdot \hat{\rho}_{jk}}{\sqrt{(1-\hat{\rho}_{ik}^2)(1-\hat{\rho}_{jk}^2)}}$$

$$-1 \leq \hat{r}_{ij.k} \leq 1$$

Для перевірки значущості частинного коефіцієнту кореляції використовується критерій:

$H_0: r = 0$
$H_1: r \neq 0$

Нульова гіпотеза – частинний коефіцієнт кореляції не є значущим, альтернативна гіпотеза – частинний коефіцієнт кореляції є значущим (будьте уважні!).

Статистика критерія:

$$t = \frac{\hat{r} \cdot \sqrt{n-2}}{\sqrt{1-r^2}}, \quad t \sim T(n-k+2)$$

Критерій є двостороннім, область прийняття рішення:

$$|t| < t_{1-\frac{\alpha}{2}}(n-k+2)$$

Приклад. Метричні дані. Множинна кореляція. Парний коефіцієнт кореляції
Приклад. Метричні дані. Множинна кореляція. Частинний коефіцієнт кореляції
Приклад. Метричні дані. Множинна кореляція. Множинний коефіцієнт кореляції

Кореляційний аналіз. Неметричні дані
Номінальна шкала

Для перевірки наявності зв'язку між двома або більше змінними, що вимірюються за номінальною шкалою, може бути використаний критерій Хі-квадрат (детальніше у лекції 9).

Гіпотези, що перевіряють:

H_0: Значущий зв'язок між змінними відсутній
H_1: Між змінними є значущий зв'язок

Статистика:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\left(x_{ij}^o - x_{ij}\right)^2}{x_{ij}^o}$$

Область прийняття рішення:

$$\chi^2 < \chi^2_{1-\alpha}((k-1)*(m-1))$$

де k – кількість змінних, зв'язок між якими перевіряють (за допомогою вибірок, отриманих з них), m – кількість значень кожної змінної.
Якщо нерівність не виконується, між змінними існує зв'язок з ймовірністю (1 – α).

Значення статистики Хі-квадрат може приймати значення від 0 до нескінченності:

$$0 < \chi^2 < \infty$$

Тому, за значенням статистики складно визначити, наскільки сильним є зв'язок. Тому, замість χ^2 часто використовують коефіцієнт Крамера:

$$C = \sqrt{\frac{\chi^2}{n \cdot \min(k-1, m-1)}}$$

Коефіцієнт Крамера може приймати значення від 0 до 1:

$$0 \leq C \leq 1$$

Якщо C = 0, це означає, що змінні, що аналізують є незалежними, якщо C = 1, то за значенням однієї змінної можна однозначно відновити значення іншої змінної.

Порядкова шкала. Рангові коефіцієнти кореляції

Для перевірки парного статистичного зв'язку між змінними, що вимірюються за порядковою шкалою, використовуються рангові коефіцієнти кореляції Спірмена і Кендалла. Розглянемо та побудуємо один з них.

Для побудови даних коефіцієнтів замість самих порядкових значень у кожній з двох вибірок використовують ранги.

Правила побудови рангів:

1. Вибірку необхідно представити у виді варіаційного ряду
2. Якщо елемент зустрічається у вибірці один раз, то його ранг дорівнює його порядковому номеру у варіаційному ряді
3. Якщо у вибірці є декілька однакових елементів, то всі вони мають однаковий ранг, який дорівнює середньому арифметичному їх порядкових номерів у варіаційному ряді. Такі ранги називають пов'язаними рангами.

Ранговий коефіцієнт кореляції Спірмена

Вибірковий ранговий коефіцієнт кореляції Спірмена (за відсутності пов'язаних рангів) розраховується за формулою:

$$\hat{\rho}_S = 1 - \left(6 \cdot \sum_{i=1}^{n} \left(R_i^x - R_i^y\right)^2\right) \Big/ \left(n \cdot (n^2 - 1)\right)$$

де n – об'єм вибірки,

R_i^x і R_i^y – ранги і-го елементу вибірки по змінним X та Y.

Якщо наявні пов'язані ранги, використовують поправки:

$$T_x = \frac{1}{12} \cdot \sum_{i=1}^{m_x} (t_{xi}^3 - t_{xi}) \qquad T_y = \frac{1}{12} \cdot \sum_{i=1}^{m_y} (t_{yi}^3 - t_{yi})$$

де m_x і m_y – число груп однакових рангів у змінних X та Y,
t_{xi} і t_{yi} – число самих однакових рангів, які входять в і-ту групу.

Тоді формула розрахунку рангового коефіцієнту кореляції Спірмена буде виглядати таким чином:

$$\hat{\rho}_S = 1 - \left(\sum_{i=1}^{n} \left(R_i^x - R_i^y \right)^2 \right) \Big/ \left(\frac{1}{6} \cdot n \cdot (n^2 - 1) - (T_x + T_y) \right)$$

Для перевірки значущості рангового коефіцієнту кореляції Спірмена використовують статистичний критерій (обмеження використання критерію n > 10):

H_0: $\rho_s = 0$
H_1: $\rho_s \neq 0$
Нульова гіпотеза – зв'язок між змінними відсутній, альтернативна гіпотеза – існує значущий зв'язок (будьте уважні!).

Статистика критерія:

$$t = \frac{\hat{\rho_s} \cdot \sqrt{n-2}}{\sqrt{1 - \hat{\rho}_s^2}}, \quad t \sim T(n-2)$$

Критерій є двостороннім, область прийняття рішення:

$$|t| < t_{1-\frac{\alpha}{2}}(n-2)$$

Приклад. Порядкові дані. Ранговий коефіцієнт кореляції Спірмена. Частина 1
Приклад. Порядкові дані. Ранговий коефіцієнт кореляції Спірмена. Частина 2

Порядкова шкала. Множинна кореляція. Коефіцієнт конкордації (узгодженості)

Для вимірювання ступеню статистичного зв'язку між декількома (m > 2) порядковими змінними, використовують **коефіцієнт конкордації** (узгодженості).

Для випадку, коли змінні не мають зв'язаних рангів, для розрахунку користуються формулою:

$$\widehat{W} = \frac{12}{m^2(n^3 - n)} \sum_{i=1}^{n} \left(\sum_{j=1}^{m} R_{ij} - \frac{m(n+1)}{2} \right)^2,$$

де m – число змінних, зв’язок між якими аналізують, n – число об'єктів, які досліджують, R_{ij} – ранг і-го об’єкту по j-й змінній.

Якщо є пов’язані ранги (ранги, з однаковими значеннями), використовують поправки:

$$T_j = \frac{1}{12}\sum_{k=1}^{m_j}(t_{jk}^3 - t_{jk}),$$

де m_j – число груп пов’язаних рангів по j-й змінній, t_{jk} – число однакових рангів, що входять в k-ту групу по по j-й змінній.

Тоді формула для розрахунку коефіцієнту конкордації виглядає наступним чином:

$$\widehat{W} = \sum_{i=1}^{n}\left(\sum_{j=1}^{m} R_{ij} - \frac{m(n+1)}{2}\right)^2 \Bigg/ \left(\frac{1}{12}m^2(n^3-n) - m\sum_{j=1}^{m} T_j\right).$$

Для перевірки значущості коефіцієнта конкордації використовують наступний критерій перевірки статистичних гіпотез:

Обмеження на використання: n > 7 (об’єм вибірки або кількість об’єктів, по яких вимірюють ознаки)!

Гіпотези, що перевіряються:

$H_0: W = 0$ (ознаки неузгоджені, зв′язку немає)
$H_1: W \neq 0$ (ознаки узгоджені, зв′язок значущий)

Статистика має розподіл Хі-квадрат з (n–1) ступенями свободи:

$$\chi^2 = \widehat{W} \cdot m \cdot (n-1) \quad \chi^2 \sim \chi^2(n-1)$$

Приклад. Чотири експерти оцінювали 8 брендів (від 1 – найкращий, до 8 – найгірший). Були отримані такі результати:

Бренд	1	2	3	4	5	6	7	8
Експерт1	2	7	1	3	4	6	8	5
Експерт2	1	4	3	2	5	6	7	8
Експерт3	1	4	2	5	3	7	8	6
Експерт4	2	3	4	1	6	5	7	8

Чи можна вважати, що оцінки експертів є узгодженими між собою?
Увага! У цьому прикладі немає оцінок, що повторюються, тому їх ранги співпадають з їх значеннями. Якщо значення у вибірці повторюються, необхідно спочатку розрахувати ранги і використовувати саме їх!!!

В задачі m = 4, n = 8.

Розрахуємо суми рангів R_{ij}:

Бренд	1	2	3	4	5	6	7	8
Експерт1	2	7	1	3	4	6	8	5
Експерт2	1	4	3	2	5	6	7	8
Експерт3	1	4	2	5	3	7	8	6
Експерт4	2	3	4	1	6	5	7	8
$\sum_{j=1}^{m} R_{ij}$	6	18	10	11	18	24	30	27

$$\frac{m(n+1)}{2} = 18,$$

$$\widehat{W} = \frac{12}{4^2(8^3-8)}((6-18)^2 + (18-18)^2 + (10-18)^2 + (11-18)^2 + (18-18)^2 + (24-18)^2 + (30-18)^2 + (27-18)^2) = 0{,}77$$

Перевірка значущості:
$H_0: W = 0$
$H_1: W \neq 0$

Розраховуємо статистику:
$\chi^2 = 0{,}77 \cdot 4 \cdot (8-1) = 21{,}5$

Область прийняття рішення:
$\chi^2 < \chi^2_{1-\alpha}(n-1)$

$\chi^2_{1-\alpha}(8-1) = \chi^2_{0{,}95}(7) = 14{,}1$

Висновок: нерівність не виконується, статистика потрапляє в критичну область, приймаємо альтернативну гіпотезу. Отже, коефіцієнт конкордації є значущим з імовірністю 95 %.

Регресійні моделі. Метричні дані

Для опису, аналізу та прогнозування явищ та процесів в економіці застосовують математичні моделі у формі рівнянь або функцій. Модель економічного об’єкту (виробничого процесу тощо), відображаючи основні його властивості та абстрагуючись від другорядних, дозволяє судити про його поведінку в певних умовах.

У разі застосування регресійних моделей результат дії економічної системи або об'єкта у вигляді одного або декількох вихідних показників представляється як функція від факторів, які на нього впливають. Деякі з цих факторів істотно впливають на результат, інші - несуттєво. Як правило, істотних факторів небагато, в той час як несуттєвих досить велика кількість, тому останніми повністю нехтувати не можна.

Завдання регресійного аналізу:
1) встановлення форми залежності між змінними;
2) оцінка функції регресії;
3) оцінка невідомих значень (прогноз) залежною змінною.

Маємо наступну теоретико-ймовірнісну модель:

$$Y = f(X_1,\dots,X_k,\varepsilon_1,\dots,\varepsilon_n).$$

Результуючий показник є функцією істотних ($X_1,\dots,X_k$) та не суттєвих ($\varepsilon_1,\dots,\varepsilon_n$) факторів.
Змінну Y називають також: функцією відгуку, вихідною, результуючою, ендогенною змінною.
Змінні $X_1, \dots, X_k$ – вхідні, предикторні, екзогенні змінні; фактори; регресори.
Змінні $\varepsilon_1,\dots,\varepsilon_n$ – латентні фактори; збурення; залишки.

Залежно від виду функції f регресійні моделі можна підрозділяти на лінійні та нелінійні моделі. Залежно від кількості факторів регресію розрізняють парну (розглядається один фактор) і множинну (кілька факторів). Розглянемо найпростіший випадок.

Парна лінійна регресія

У випадку парної лінійної регресії існує лише один детермінований фактор x, модель записується наступним чином:

$$Y = \alpha_0 + \alpha_1 X + \varepsilon,$$

де Y – результуюча (залежна) змінна,
X – фактор, ε – випадкова компонента ($E\varepsilon = 0, \quad D\varepsilon = \sigma^2$).

Оцінка параметрів регресії α_0, α_1 здійснюється за вибіркою:

(x_i, y_i), i=1,…,n.

Статистичні оцінки параметрів регресії вибираються таким чином, щоб емпіричні значення детермінованої складової:

$$\hat{y}_i = \hat{\alpha}_0 + \hat{\alpha}_1 x_i$$

якомога ближче знаходилися до фактичних значень результуючого ознаки. В якості міру відхилень зазвичай вибирають суму квадратів. Якщо випадкова складова має нормальний розподіл, то оцінки, що мають найкращу якість, можуть бути отримані за допомогою методу найменших квадратів.

Складемо суму квадратів відхилень як функцію параметрів α_0, α_1 :

$$Q(\alpha_0, \alpha_1) = \sum_{i=1}^{n} (y_i - \alpha_0 - \alpha_1 x_i)^2$$

Для того, щоб знайти мінімальне Q, знаходимо частинні похідні по α_0, α_1 :

$$\begin{cases} \left.\dfrac{\partial Q}{\partial \alpha_0}\right|_{\alpha_0=\hat{\alpha}_0} = -2\sum_{i=1}^{n} (y_i - \hat{\alpha}_0 - \hat{\alpha}_1 x_i) = 0 \\ \left.\dfrac{\partial Q}{\partial \alpha_1}\right|_{\alpha_1=\hat{\alpha}_1} = -2\sum_{i=1}^{n} x_i (y_i - \hat{\alpha}_0 - \hat{\alpha}_1 x_i) = 0 \end{cases}.$$

Отримаємо систему нормальних рівнянь:

$$\begin{cases} \sum_{i=1}^{n} y_i = n\hat{\alpha}_0 + \sum_{i=1}^{n} x_i \hat{\alpha}_1 \\ \sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} x_i \hat{\alpha}_0 + \sum_{i=1}^{n} x_i^2 \hat{\alpha}_1 \end{cases}.$$

З першого рівняння отримуємо:

$$\hat{\alpha}_0 = \overline{Y} - \hat{\alpha}_1 \overline{X}.$$

Підставимо отримане у друге рівняння:

$$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} x_i (\overline{Y} - \hat{\alpha}_1 \overline{X}) + \sum_{i=1}^{n} x_i^2 \hat{\alpha}_1 = \sum_{i=1}^{n} x_i \overline{Y} + \left(-\overline{X} \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} x_i^2 \right) \hat{\alpha}_1 ,$$

звідки:

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \overline{Y} = \left(\sum_{i=1}^{n} x_i^2 - \overline{X} \sum_{i=1}^{n} x_i \right) \hat{\alpha}_1 .$$

Додаємо до лівої частини рівності добуток, рівний нулю, і перегрупуємо:

$$\sum_{i=1}^{n} x_i (y_i - \overline{Y}) + \overline{X} \sum_{i=1}^{n} (y_i - \overline{Y}) = \sum_{i=1}^{n} (x_i - \overline{X})(y_i - \overline{Y})$$

Тепер розглянемо праву частину. Очевидно, що:

$$\overline{X} \sum_{i=1}^{n} x_i = n\overline{X}^2$$

Тоді справедливо:

$$\sum_{i=1}^{n} x_i^2 - \overline{X}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i^2 - 2\overline{X}\sum_{i=1}^{n} x_i + n\overline{X}^2 = \sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2$$

Таким чином, отримали оцінки параметрів регресії:

$$\begin{cases} \hat{\alpha}_1 = \dfrac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - Y\right)}{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2} = \hat{\rho}\cdot\dfrac{\hat{\sigma}_Y}{\hat{\sigma}_X} \\ \hat{\alpha}_0 = \overline{Y} - \hat{\alpha}_1\overline{X} = \overline{Y} - \hat{\rho}\cdot\dfrac{\hat{\sigma}_Y}{\hat{\sigma}_X}\overline{X} \end{cases}$$

Отримані за методом найменших квадратів оцінки параметрів регресії є незсуненими та конзистентними, а також ефективними у класі лінійних незсунених оцінок.

Приклад. Парна лінійна регресія
Приклад. Використання побудованої парної регресійної моделі

Що показують коефіцієнти регресії

Далі по тексту використовуються позначення b_0 и b_1 для коефіцієнтів регресії:

$$\hat{b}_0 = \overline{Y} - \hat{b}_1\overline{X},$$
$$\hat{b}_1 = \hat{\rho}(X,Y)\cdot\frac{\hat{\sigma}_Y}{\hat{\sigma}_X}.$$

Коефіцієнт b_0 показує, у якій точці регресійна пряма перетинає вісь Oy. Тобто, чому буде дорівнювати передбачене значення y, якщо значення x дорівнює 0.
Коефіцієнт b_1 показує, наскільки зростає змінна y при збільшенні значення змінної x на 1. Інакше, даний коефіцієнт є тангенсом кута нахилу регресійної прямої та демонструє силу лінійного зв'язку між X та Y.
Стандартизований коефіцієнт показує, на скільки стандартних відхилень збільшується залежна змінна при збільшенні незалежної змінної на одне стандартне відхилення.

Таким чином, вибіркове значення y_i являє собою суму передбаченого моделлю значення $\hat{y}_i$ та випадкового збурення ε_i:

$$y_i = \hat{y}_i + \varepsilon_i = b_0 + b_1 x_i + \varepsilon_i$$

Ступінь розбіжності вибіркового значення y_i і значення $\hat{y}_i$, передбаченого моделлю, називають в регресійному аналізі **залишками**:

$$y_i - \hat{y}_i = \varepsilon_i$$

Вплив неврахованих випадкових чинників і помилок спостережень визначається за допомогою залишкової дисперсії, оцінкою якої є вибіркова залишкова дисперсія:

$$S_\varepsilon{}^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n-2} = \frac{\sum_{i=1}^{n}\varepsilon_i^2}{n-2}$$

(у знаменнику виразу кількість ступенів свободи, вона дорівнює (n–2), а не n, тому що 2 ступеня свободи втрачаються при визначенні параметрів b_0 та b_1).

Велика кількість залишків, а відповідно і велика величина залишкової дисперсії, свідчать про слабкий зв'язок між X та Y (коефіцієнт кореляції Пірсона при цьому буде малий).

Для функції регресії (тобто для умовного математичного сподівання значення Y), а також індивідуальних значень залежної змінної можуть бути розраховані довірчі інтервали з заданим рівнем надійності (довірчої ймовірності). Величина довірчого інтервалу залежить від значення пояснюючої змінної X: при значеннях близьких до середнього значення – вона є мінімальною, а з віддаленням від середнього – величина інтервалу збільшується.

Перевірка значущості коефіцієнтів регресії

Регресія по даній залежною змінній має сенс в тому випадку, якщо відповідний коефіцієнт регресії значуще відрізняється від нуля. За абсолютною величиною коефіцієнтів регресії визначити їх статистичну близькість до нуля неможливо, тому що їх величина залежить від того, яким є масштаб вимірювання змінних X і Y. Тому для визначення значущості коефіцієнтів регресії будують оцінки їх середньоквадратичних відхилень:

$$S_{b_0} = \sqrt{\frac{S_\varepsilon^2}{n}\left(1 + \frac{\overline{X}^2}{S_X^2}\right)},$$

$$S_{b_1} = \sqrt{\frac{S_\varepsilon^2}{nS_X^2}}.$$

Перевірка значущості параметрів регресії – це перевірка наступних гіпотез:

$H_0: \quad b = 0$

$H_1: \quad b \neq 0$

Для перевірки гіпотез використовують статистику:

$$t = \frac{b}{S_b}, \qquad t \sim T(n-2).$$

Особливо важливою є перевірка на рівність 0 коефіцієнта b_1. Дійсно, якщо b_1 дорівнює 0, це означає, що регресійна пряма проходить паралельно осі абсцис і, отже, Y не залежить від X. Таким чином, якщо ми не можемо з високою ймовірністю відхилити гіпотезу про рівність b_1 нулю, це означає, що ми не можемо прийняти гіпотезу про існування лінійного зв’язку між X та Y.

Перевірка значущості рівняння регресії

Перевірити значущість рівняння регресії – означає встановити, чи відповідає математична модель, що виражає залежність між змінними, експериментальним даним і чи достатньо включених в рівняння пояснюючих змінних (однієї або декількох) для опису залежної змінної. Перевірка значущості рівняння регресії здійснюється на основі дисперсійного аналізу.

Згідно з основною ідеєю дисперсійного аналізу:

$$\sum_{i=1}^{n}(y_i-\bar{Y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{Y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

або:

$$Q=Q_R+Q_\varepsilon$$

де Q – загальна сума квадратів відхилень залежної змінної від середнього, Q_R і Q_ε – відповідно, сума квадратів, обумовлена регресією, та сума квадратів залишків, яка характеризує вплив неврахованих факторів.

Компоненти варіації:

Регресія: $Q_R=\sum_{i=1}^{n}(\hat{y}_i-\bar{Y})^2$,

Залишкова: $Q_\varepsilon=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$,

Загальна: $Q=\sum_{i=1}^{n}(y_i-\bar{Y})^2$.

Для перевірки значущості рівняння регресії:
H_o: рівняння регресії не є значущим
H_1: рівняння регресії є значущим,

– будують статистику F, яка має розподіл Фішера з (m–1) та (n–m) ступенями свободи:

$$F=\frac{Q_R(n-m)}{Q_\varepsilon(m-1)} \qquad F\sim F(m-1,n-m),$$

де m – число параметрів рівняння регресії, які оцінюються (у випадку парної регресії m = 2), n – число наблюдений.

Значення статистики F демонструє відношення варіації, обумовленої регресією, до варіації обумовленої впливом випадкових збурень.

Коефіцієнт детермінації, виражений через варіації, має вигляд:

$$R^2=\frac{Q_R}{Q}=1-\frac{Q_\varepsilon}{Q}.$$

Коефіцієнт детермінації можна розглядати як міру якості рівняння регресії (характеристику прогностичної сили аналізованої регресійній моделі): чим ближче R^2 до одиниці, тим краще регресія описує залежність між пояснюючими змінними і залежною змінною. Якщо R^2 дорівнює одиниці, це означає що модель повністю описує Y і залишки відсутні.

Недолік R^2 полягає в тому, що його значення не зменшується із зростанням кількості пояснюючих змінних. В цьому сенсі кращим є **скоригований коефіцієнт детермінації**:

$$\hat{R}^2=1-\frac{n-1}{n-m-1}\left(1-R^2\right),$$

де m – кількість пояснюючих змінних.

Цей коефіцієнт зменшується при введенні в регресійну модель змінних, що не чинять істотного впливу на залежну змінну.

Дотримання умови незалежності залишків

Дотримання умови незалежності залишків перевіряється за допомогою критерію Дарбіна-Уотсона. В ідеальній ситуації він дорівнює 2,0. Допустимі значення - від 1 до 3. Якщо цей критерій має значення менше 1 або більше 3, це означає, що умова незалежності залишків не дотримується, а значить, прогнозування за допомогою цього методу буде не зовсім коректним.

Дотримання умови нормальності розподілу залишків

Побудова регресійній моделі здійснюється в умовах припущення нормального розподілу залишків. Зокрема, на даному припущенні, відбувається оцінювання значущості коефіцієнтів регресії.
У разі відсутності нормальності розподілу залишків, не можна користуватися формулами довірчих інтервалів для коефіцієнтів регресії, а значить перенесення результатів, отриманих за вибіркою, на генеральну сукупність неможливо. Тому отриманої рівняння регресії буде представляти цінність тільки для даної конкретної вибірки.
Нормальність розподілу залишків можна подивитися на гістограмі стандартизованих залишків. Квантільная діаграма стандартизованих залишків ще краще візуалізує дотримання (недотримання) умови нормальності.
Збережені при застосуванні процедури лінійної регресії залишки можна перевірити на нормальність, наприклад, за допомогою критерію Колмогорова-Смірнова.

Дотримання умови гомоскедастичності

Обов'язковою умовою для побудови регресійної моделі є вимога однакового розкиду спостережень навколо лінії регресії для всіх значень X. Ця вимога називається вимогою гомоскедастичності, що означає однаковий розкид.
Порушення гомоскедастичності, тобто гетероскедастичність, говорить про те, що для різних значень X необхідно будувати різні регресійні моделі.
Можна перевірити умову гомоскедастичності шляхом побудови графіка залежності стандартизованих залишків (*ZRESID) від стандартизованих передбачених значень (*ZPRED). Розкид значень залишків повинен бути однаковий для всіх передбачених значень.
Графік залежності стандартизованих залишків від стандартизованих передбачених значень може мати форму трикутника, трапеції, являти собою криволінійну залежність тощо. У всіх цих випадках говорять про гетероскедастичність, що не дозволяє застосовувати лінійний регресійний аналіз.

Діагностика викидів

Точки, сильно віддалені від регресійної прямої, в регресійному аналізі називаються **викидами**. Спостереження класифікується як викид в залежності від величини залишку. За замовчуванням в SPSS викидами вважаються випадки, коли значення залишку виходить за кордон трьох

стандартних відхилень залишків (правило трьох сігм: 99% всіх значень нормально розподіленої величини не відхиляються від середнього більше, ніж на три середньоквадратичних відхилення). Наявність викидів погіршує нормальність розподілу залишків, збільшує їх дисперсію, що тягне до збільшення стандартних помилок регресійних коефіцієнтів і зменшення коефіцієнта детермінації. Зі змістовної точки зору все ще гірше. Виникає підозра, що дані неоднорідні. У них є частина спостережень, для яких характерний один вид залежності Y від X, а є інша частина, які мають іншу залежність.
Отже, викиди необхідно ретельно аналізувати, і бажано по кожному з них робити висновок з точки зору предметної області. А далі вже приймати рішення про можливість використання всієї побудованої моделі.

Повторимо ще раз **необхідні умови**, які повинні виконуватися **для використання лінійної моделі в регресійному аналізі**:
- Залежна змінна повинна бути кількісною
- Незалежна змінна повинна бути кількісною
- Спостереження (і залишки) повинні бути незалежні один від одного (перевіряється за допомогою критерію Дарбіна-Уотсона)
- Залежність між змінними повинна бути лінійною (перевіряється графічно шляхом побудови кореляційного поля і за допомогою коефіцієнта кореляції Пірсона)
- Рівняння регресії має бути значущим
- Коефіцієнти регресії повинні значуще відрізнятися від нуля
- Залишки повинні мати нормальний розподіл (перевіряється графічно за допомогою гістограм, квантільних діаграм, а також за допомогою критерію Колмогорова-Смирнова)
- Залишки повинні мати однаковий розкид на всьому інтервалі передбачених значень (або незалежної змінної)
- Вибірка повинна бути репрезентативною

Кластерний аналіз

За допомогою кластерного аналізу вирішуються **завдання багатовимірної класифікації**. Статистичні угрупування і класифікація переслідує цілі виділення якісно однорідних сукупностей, вивчення структури сукупності, дослідження існуючих залежностей.

Приклади маркетингових завдань, які вирішуються за допомогою кластерного аналізу:

- сегментування - ідентифікація стійких груп (людей, домогосподарств, підприємств, організацій, ринків), кожна з яких об'єднує об'єкти зі схожими характеристиками;

- аналіз поведінки споживачів (побудова типових моделей, визначення драйверів поведінки)

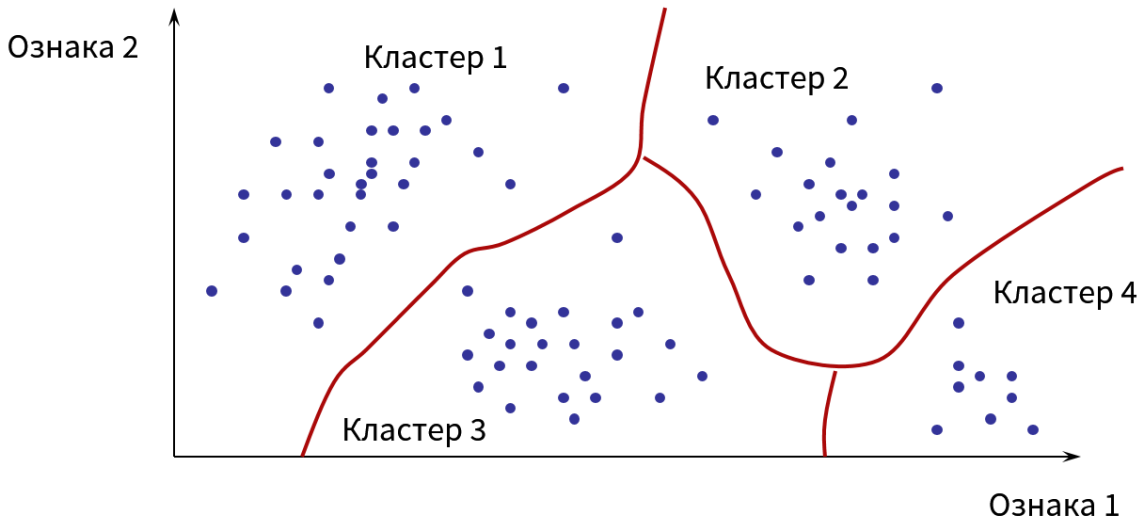
- позиціонування (побудова карти, на основі якої можна визначити рівень і характер конкуренції в різних сегментах, визначення характеристик товару, що відповідає цільовому сегменту)

- вибір цільових ринків (групування країн, ринків, магазинів, продуктів в групи по релевантним характеристикам для подальшого дослідження)

- класифікація постачальників за ступенем привабливості тощо.

Сутність методів багатовимірної класифікації полягає в наступному. Є сукупність з n об'єктів. З усіх ознак, які мають об'єкти, відбирається m значущих. По кожному об'єкту вимірюються значення кожної з ознак. Потрібно розбити сукупність на однорідні в деякому сенсі групи. Отримані в результати розбиття групи називаються кластерами (від англ. cluster - група, пучок, кущ).

Класифікація об'єктів при кластерному аналізі проводиться не послідовно за окремими ознаками, а одночасно за декількома ознаками. Цей фіксований набір ознак утворює так званий простір ознак, а кожній ознаці надається сенс координати. Якщо задано m істотних ознак сукупності, то будь-який об'єкт розглядається як точка в m-вимірному просторі ознак, і завдання класифікації зводиться до виділення згущень об'єктів в цьому просторі.



Для цього використовуються різні алгоритми, але групи (типи, класи) завжди формуються на підставі близькості об'єктів за комплексом ознак.

Етапи кластерного аналізу:

1. Визначення цілей класифікації
2. Виділення комплексу ознак
3. Визначення міри схожості об'єктів
4. Вибір алгоритму й програми класифікації
5. Розрахунок варіантів, оцінка достовірності кластеризації, прийняття рішення про кількість кластерів
6. Профілювання кластерів і змістовна інтерпретація результатів

Визначення цілей класифікації й початковий **вибір набору ознак** - завдання предметної області. Їх рішення не пов'язане математичними методами, проте саме вони впливають на результати, які будуть отримані. Тому дослідник повинен чітко розуміти, для чого необхідне застосування кластерного аналізу, в ідеалі - мати готові гіпотези того, які результати він планує отримати, і представляти, як вони будуть використані.

Після початкового вибору комплексу ознак, проводяться необхідні вимірювання. Початкові дані для завдання багатовимірної класифікації зазвичай представляють у вигляді матриці «об'єкт-ознака». Рядками її є значення ознак, що характеризують відповідний об'єкт, а стовпцями – значення кожної ознаки для даної сукупності об'єктів.

Далі бажано здійснити перевірку на відсутність ефекту мультиколінеарності ознак. **Мультиколінеарність** – наявність дуже сильної кореляції між ознаками (звичайно мультиколінеарність визначають на рівні від 0,8, іноді, 0,9). Якщо деякі ознаки серед всього набору сильно корелюють між собою, вони вносять однаковий (посилений один одним) внесок. В результаті вийде, що класифікація буде проведена саме за цими, пов'язаних один з одним ознаками, а вплив інших на результат розбиття буде мінімальним.
При виявленні дуже сильного зв'язку між деякими двома ознаками, необхідно прийняти рішення про виключення одного з них із загального набору. Вибір цей здійснюється, знову-таки, виходячи з предметної області: вибирається та змінна, яка є визначальною, основною для вирішення поставленого завдання.

На наступному етапі необхідно вибрати **міру подібності об'єктів**. Міри подібності можна розділити на міри близькості і міри відстані. Міра близькості визначає, наскільки близько один до одного знаходяться об'єкти: чим більше її величина, тим ближче об'єкти однин до одного. Міра відстані, навпаки, показує, наскільки далеко один від одного об'єкти: чим більше величина, тим далі один від одного об'єкти. Міри обирають залежно від шкал, за якими вимірюються ознаки.

Шкала даних, за якою вимірюється ознака	**Міри близькості**	**Міри відстані**
Номінальна	Коефіцієнт подібності	Коефіцієнти різниці, Хі-квадрат, Фі-квадрат
Порядкова	Рангові коефіцієнти кореляції Спірмена, Кендалла	-
Відносна	Коефіцієнт кореляції Пірсона	Евклідова відстань, Лінійна відстань тощо

Коефіцієнти подібності

Для вимірювання ступеня близькості між парами об'єктів (і та j), кожна з ознак яких вимірюється за номінальною шкалою, використовуються коефіцієнти подібності. Найбільш простий коефіцієнт подібності для двох об’єктів розраховується за формулою:

$$S_{ij} = \frac{P_{ij}}{m},$$

де P_{ij} – число співпадаючих ознак у об'єктів і та j;
m – загальне число ознак, за якими здійснюється порівняння.
Таким чином, коефіцієнт подібності може приймати значення від 0 (максимально відрізняються об'єкти) до 1 (максимально подібні об'єкти). Коефіцієнти відмінності, як міра відстані між об'єктами, розраховуються аналогічно, тільки підсумовується кількість незбіжних ознак.

Часто в якості мір подібності використовуються **коефіцієнти кореляції**. Якщо ознаки не піддаються точній кількісній оцінці, але вимірюються за порядковими шкалами, то мірами їх зв'язку служать коефіцієнти рангової кореляції.

Міри відстані. У багатьох випадках роль міри схожості відіграє функція відстані. Якщо ознаки мають різні одиниці виміру, слід звернути увагу на масштаб вимірів. Відмінності в масштабах можуть вплинути на отримані кластерні рішення. Якщо масштаб змінних сильно відрізняється (наприклад, одна змінна виміряна в тис. доларах, а інша - в частках відсотків), перед проведенням кластеризації необхідно виконати їх стандартизацію. Загальновизнаним способом стандартизації вважається заміна первинних значень ознак їх відхиленнями від середнього рівня:

$$x^{*}_{ij} = \frac{(x_{ij} - \bar{x}_j)}{\sigma_j}$$

де x_{ij} - початкове значення j-ї ознаки й і-го об'єкта,
x^{*}_{ij} - стандартизоване значення j-ї ознаки у і-го об'єкта,
$\bar{x}_j$ - середнє значення j-ї ознаки по всіх об'єктах,
σ_j – середньоквадратичне відхилення j-ї ознаки.

Існує велика кількість різних **мір відстані** між об'єктами. Найчастіше використовують наступні:

1) Лінійна відстань (відстань міських кварталів, блок, відстань Манхеттен):

$$d_{ij} = \sum_{k=1}^{m} \left| x_{ik} - x_{jk} \right|;$$

2) Евклідова відстань - для кількісних ознак:

$$d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2},$$

де x_{ik} - значення k-го ознаки в об'єкта і, а x_{jk} - у об'єкта j.

Коефіцієнти асоціативності

Велика кількість мір подібності розроблено для бінарних ознак (які приймають два значення 1 і 0).
Для розгляду цих коефіцієнтів вводять таблицю асоціативності:

	Ознака 1	
Ознака 2	1	0
1	a	b
0	c	d

Приклади коефіцієнтів асоціативності:

Евклідова відстань - корінь з числа ознак, значення за якими для двох об'єктів не збігаються:

$$d_{ij} = \sqrt{b+c}$$

Різниця довжин:

$$d_{ij} = \frac{(b-c)^2}{(a+b+c+d)^2}$$

Різниця структур:

$$d_{ij} = \frac{bc}{(a+b+c+d)^2}$$

Просте узгодження - частка співпадаючих ознак

$$d_{ij} = \frac{a+d}{a+b+c+d}$$

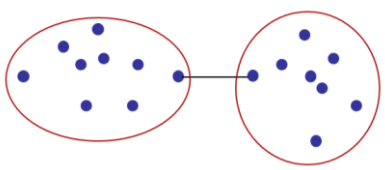
І багато інших…

Методи кластеризації, реалізовані в SPSS:

1. **Ієрархічна кластеризація**. При використанні ієрархічного методу кластеризація починається з виділення пари найближчих об'єктів (спостережень або змінних) і об'єднання їх в кластер. На кожному кроці об'єднуються або пара об'єктів, або пара кластерів, або об'єкт і кластер.

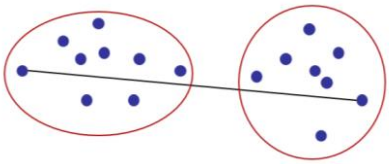
Розглядаються наступні способи об'єднання об'єктів (кластерів):

Метод ближнього сусіда
Відстанню між двома кластерами вважається відстань між двома найближчими точками з різних кластерів.



Метод далекого сусіда

Відстанню між двома кластерами вважається відстань між двома найбільш віддаленими точками з різних кластерів.



Метод міжгрупових зв'язків

У цьому методі відстань між кластерами розраховується шляхом усереднення всіх відстаней від об'єкта одного кластера до об'єкта іншого.

Метод внутрішньогрупових зв'язків

У цьому методі відстань між кластерами обчислюється як середня відстань між усіма можливими парами об'єктів, що належать обом кластерам, в тому числі об'єктів, розташованих усередині одного і того ж кластера.

Центроїдний метод

При злитті двох кластерів центроїд нового кластера розраховується як зважене по числу об'єктів в кожному кластері середнє значення центроїдів початкових кластерів.

Медіанний метод

При злитті двох кластерів центроїд нового кластера розраховується шляхом усереднення координат центроїдів двох кластерів, які об’єднують. Число об'єктів, що входять в ці кластери, до уваги не береться.

Метод Варда

Цей метод відрізняється від всіх інших методів, оскільки він використовує методи дисперсійного аналізу для оцінки відстаней між кластерами. Метод мінімізує суму квадратів (SS) для будь-яких двох (гіпотетичних) кластерів, які можуть бути сформовані на кожному кроці.

Процес об'єднання об'єктів і кластерів в **ієрархічній кластеризації** триває до тих пір, поки всі дані не потраплять в один кластер. Відмінною особливістю методу є те, що після об'єднання двох об'єктів або кластерів вони залишаються разом до останнього кроку. Таким чином, кластер, сформований на останньому кроці, містить кластери попереднього кроку, які, в свою чергу, містять кластери з більш ранніх кроків.

2. **Метод k-середніх**. Процедура починається з використання значень перших k спостережень файлу даних в якості попередніх оцінок k-кластерних середніх, де k - число кластерів, що задається користувачем. Початкові кластерні центри формуються наступним чином: кожне спостереження призначається в кластер з найближчим центром, а потім значення центру переобчислюють. Далі використовується ітеративний процес. На кожному кроці спостереження групуються в кластер з найближчим центром, і кластерні центри переобчислюють. Цей процес триває до тих пір, поки центри кластерів не перестають змінюватися або поки кількість ітерацій не досягне заданого максимуму. Існує можливість задавати центри кластерів. Процедура кластеризації k-середніми більше підходить для великої кількості спостережень, ніж ієрархічна. Однак для використання даного методу необхідно, щоб всі ознаки вимірювалися за відносними шкалами.

3. **Двоетапний кластерний аналіз**. При необхідності роботи як з кількісними (наприклад, дохід), так і з категоріальним (наприклад, сімейний стан) змінними, а також якщо обсяг даних досить

великий, використовується метод Двоетапного кластерного аналізу, який являє собою масштабну процедуру кластерного аналізу, що дозволяє працювати з даними різних типів. Для цього на першому етапі роботи записи попередньо кластеризуються в велику кількість субкластерів. На другому етапі субкластери групуються в необхідну кількість результуючих кластерів.

ПРАКТИЧНІ ЗАНЯТТЯ

Дескриптивна статистика

Увага! Перед тим, як почати роботу, перевірте, будь ласка, що у вас завантажений Аналіз даних!

> Кнопка **Анализ данных** знаходиться в групі **Анализ** на вкладці **Данные.** Якщо команда **Анализ данных** недоступна, необхідно завантажити загрузить додатковий компонент «Пакет анализа».
> Активація **Пакет анализа**
> 1. Відкрийте вкладку **Файл**, натисніть кнопку **Параметры**, оберіть категорію **Надстройки**.
> (Якщо ви використовуєте Excel 2007, натисніть кнопку **Microsoft Office**, а потім – кнопку **Параметры Excel**).
> 2. У списку, що розкриється, **Управление** оберіть пункт **Надстройки Excel** и натисніть кнопку **Перейти**.
> 3. Якщо ви використовуєте Excel для Mac, в рядку **Меню** відкрийте вкладку **Средства** та в списку, що розкривається, оберіть пункт **Надстройки для Excel**.
> 4. У діалоговому вікні **Надстройки** встановіть прапорець **Пакет анализа**, а потім натисніть кнопку **ОК**.

1. Графічне представлення розподілу досліджуваної змінної

Оберемо три різні змінні з точки зору типу шкали, за якою вони вимірюються:
1. Тип_магазина (номінальна шкала)
2. Оценка_ассортимента (порядкова шкала)
3. Цена_куриного_филе_1_кг (відносна шкала)
Візуалізуємо емпіричний розподіл.

1.1. Графічне представлення розподілу змінної **Тип_магазина**

1.1.1. Створимо заготовку для зведеної таблиці
Встановіть курсор на комірці А2 листа **'Результаты наблюдения'.**
На вкладці **Вставка** в групі **Таблицы** обрати **Сводные таблицы**
В діалоговому вікні, що відкриється, в полі **Выбрать таблицу или диапазон** задати діапазон: **'Результаты наблюдения'!$A$1:$AL$225** (якщо курсор у вас стояв правильно, то даний діапазон з’явиться автоматично).
Задати розміщення звіту зведеної таблиці: **'Описательная статистика'!$A$1**.
При натисканні **Ок** буде створений новий лист із заготовкою зведеної таблиці.

1.1.2. Згрупуємо дані за змінною **Тип_магазина.**
Для цього з вікна з переліком полів перетягніть поле **Тип_магазина** у вікно **Строки.** В результаті в першому стовпці зведеної таблиці з’явилася назва змінної та список її значень.
Тепер перетягніть поле **Тип_магазина** у вікно **Значения.** З’явився стовпчик групованих даних. Автоматично може рахуватися **Сумма** по полю **Тип_магазина** (а нам потрібна Кількість).
Натисніть на прапорець, який розкриває список у вікні **Значения** на кнопці **Сумма…** У діалоговому вікні, що відкрилося, оберіть **Параметры полей значений.** Задайте ім’я: Частота.
Операцію **Сумма** змініть на операцію **Количество.** В результаті ви отримали таблицю зі значеннями статистичного ряду для змінної **Тип_магазина.**

1.1.3. Побудуємо відносні частоти (відсотки значень від загальної кількості).
Для цього ще раз перетягніть з вікна з переліком полів поле **Тип_магазина** у вікно **Значения**. Ще раз отримали **Сумма…** Натисніть на прапорець, що розкриває список, оберіть **Параметры полей значений.** Задайте ім'я: Відсоток. Операцію **Сумма** змініть на операцію **Количество.** З вкладки **Операция** перейдіть на вкладку **Дополнительные вычисления.** Заміть **Без вычислений** в списку **Дополнительные вычисления** оберіть **% от суммы по столбцу.**

1.1.4. Візуалізувати отримані результати можна за допомогою, наприклад, секторної діаграми по часткам.
Виділіть дані в стовпці Частота (або Відсоток) і на вкладці Вставка оберіть найпростішу кругову діаграму.
Натисніть на **+** , приберіть прапорці **Название** и **Легенда**, додайте прапорець **Подписи данных** та зайдіть в додаткові параметри. В **Параметрах подписей** додайте до **Значений Доли**.

1.2. Графічне представлення розподілу змінної **Оценка_ассортимента**

Виконаємо таку саму послідовність дій.
Розташуйте зведену таблицю нижче на тому ж листі.
При побудові графіку оберемо стовпчикову діаграму.

Зверніть увагу на те, що стовпці впорядковані в алфавітному порядку, який нам не підходить. Натисніть на кнопку поряд з назвою змінної **Оценка_ассортимента** в таблиці. В діалоговому вікні, що видкриється, оберіть Додаткові параметри сортування. Перевірте, що встановлений прапорець **Вручную**. Після цього, натискаючи на кожне значення (поле підсвічується зеленою рамкою) та праву кнопку миші, у діалоговому вікні натискаємо **переместить.** И пересуваємо наші стовпчики так, як хочемо.

Тепер, я сподіваюся, картинки у нас однакові!
Існує можливість групування категорій: наприклад, можна згрупувати «отвратительно» і «плохо» в «Плохо», а «хорошо» і «идеально» в «Хорошо».
Далі можна експериментувати і покращувати скільки бажаєте…

1.3. Графічне представлення розподілу змінної **Цена_куриного_филе_1_кг**

Для даної змінної скористуємося функцією Аналізу даних **Гистограмма.**
Для роботи цієї функції необхідно задати інтервали групування (кишені).

На тому ж листі **'Описательная статистика'** нижче нашого аналізу категоріальних змінних розмістимо вертикально (в стовпці А, наприклад) наступні значення: 60, 65, 70, 75, 80, 85, 90. У мене це комірки А45:А50.

На вкладці Данные оберіть **Анализ данных**. У діалоговому вікні, що відкриється, з інструментів аналізу оберіть **Гистограмму.**

Задайте параметри.
Вхідний інтервал: **'Результаты наблюдения'!$H$2:$H$225**,
Інтервал кишень: **'Описательная статистика'!$A$45:$A$50**,
Вихідний інтервал: **'Описательная статистика'!$B$45**
и поставте прапорці **Интегральный процент** і **Вывод графика**.

Змініть параметри ряду Частота: перекриття рядів 100%, бічний зазор 0%. Додайте підписи даних. І далі удосконалюйте за смаком…

2. Описові статистики

2.1. Міри центральної тенденції: **мода, медіана, середнє**.
Приклад 1. Отримана вибірка: кількість марок-конкурентів в 10 магазинах:

12, 3, 2, 0, 6, 19, 2, 5, 2, 9

Перепишіть елементи у порядку зростання:

_____
Дане представлення вибірки називається **варіаційний ряд**.
Знайдіть елемент, що зустрічається найчастіше. Це **мода**.
d = _____

Знайдіть середину варіаційного ряду (якщо кількість елементів – парне число, знаходимо середнє арифметичне між двома серединними елементами, якщо непарне – сам серединний елемент). Це **медіана**.
h = _____

Знайдіть **середнє** арифметичне між елементами (це оцінка математичного сподівання, фізичний сенс – центр мас). Для цього необхідно додати всі елементи і поділити на їх кількість.
$\overline{X}$ = _____

Як бачите, мода, медіана та середнє відрізняються між собою.

Пример 2. Нехай дані отримані відразу у вигляді групованого ряду:
Ціни на куряче філе:
від 65,00 до 70,00 грн. – 40 магазинів,
від 70,00 до 75,00 грн. – 30 магазинів,
від 75,00 до 80,00 грн. – 10 магазинів,
від 80,00 до 85,00 грн. – 20 магазинів.
Як отримати оцінки у цьому випадку?
Знайдіть середину інтервалу з максимальною частотою. Це **мода**:
d = _____

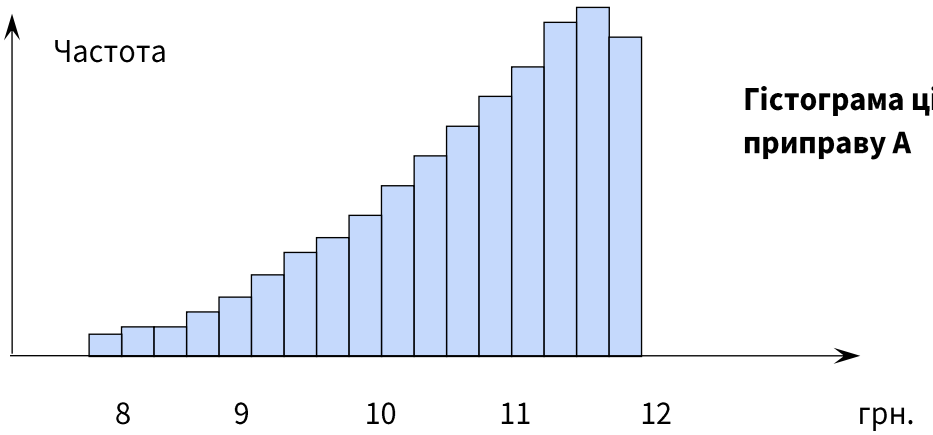
Знайдіть середину інтервалу, в якому знаходиться середина варіаційного ряду. Це **медіана**:
h = _____

Знайдіть **середнє** арифметичне за формулою: $\overline{X}^* = \frac{1}{n}\sum_{i=1}^{k} x_i^* n_i^*$,

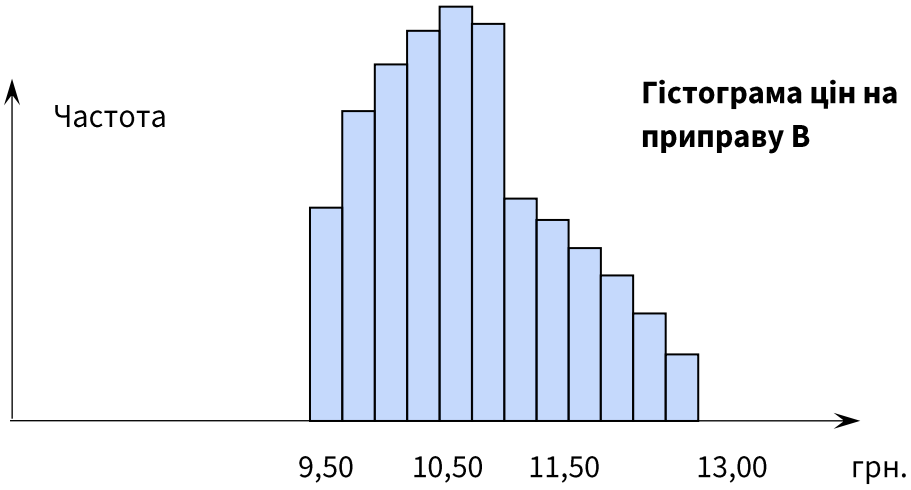
где n – об'єм вибірки,
k – кількість інтервалів,
x_i^* – середина і-го інтервалу,
n_i^* – частота і-го інтервалу.

$\overline{X}^* =$ _____

Спробуємо оцінити міри центральної тенденції за графіками.
Приклад 3. Нехай отримані два вибіркових розподілу цін (приправа А і приправа В).



Гістограма цін на приправу А



Гістограма цін на приправу В

Намалюйте на графіках та порівняйте (приблизно, звичайно) моду, медіану та середнє.

$d_A =$ _____ < = > $d_B =$ _____

$h_A =$ _____ < = > $h_B =$ _____

$\overline{X}_A =$ _____ < = > $\overline{X}_B =$ _____

2.2. Міри мінливості: **розмах, дисперсія, середньоквадратичне відхилення, коефіцієнт варіації**.

Приклад 1. Оцінимо мінливість кількості марок-конкурентів
Поверніться до варіаційного ряду. Від максимального елементу віднімiть мінімальний.
Це **розмах**. R = _____
Розрахуємо **дисперсію**. Зверніть увагу, що формула дисперсії по вибірці (це особливо важливо, якщо об'єм вибірки малий) є підправленою (ділити на n-1, а не на n):

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

S^2 = _____

Знайдіть корінь з дисперсії. Це **середньоквадратичне відхилення**.

S = _____

Розділіть середньоквадратичне відхилення на середнє. Це **коефіцієнт варіації**.

V = _____

Приклад 2. Якщо дані були отримані відразу у вигляді групованого ряду, можна розрахувати дисперсію за формулою:

$$S^{2^*} = \frac{1}{n-1}\sum_{i=1}^{k}\left(x_i^* - \bar{X}^*\right)^2 n_i^*$$

Розрахуємо дисперсію:

S^{2^*} = _____

Середньоквадратичне відхилення:

S^* = _____

Коефіцієнт варіації:

V^* = _____

Які дані мають більшу мінливість: з прикладу 1 чи з прикладу 2?

2.3. Показники форми розподілу: **асиметрія** та **ексцес**

Приклад 3. Розгляньте графіки та оцініть (приблизно!) асиметрію та ексцес.

Асиметрія цін на приправу А < = > 0

Асиметрія цін на приправу В < = > 0

Ексцес цін на приправу А < = > 0

Ексцес цін на приправу В < = > 0

Тепер знову працюємо в MS Excel, **PripravkaKPI2**.

2.4. Отримаємо описові статистики для змінної **Цена_кур_филе**.

На вкладці Дані в Аналізі даних оберіть **Описательная статистика**.
Задайте Вхідний інтервал: **'Результаты наблюдения'!$H$2:$H$225**,
Вихідний інтервал: **'Описательная статистика'!$B$68**
И поставте прапорець **Итоговая статистика**.

Аналогічно отримайте описові статистики для змінної **Количество_мест_с_приправами.**

Розрахуйте коефіцієнти варіації та порівняйте їх для двох змінних. Зробіть висновки.

Нехай необхідно розрахувати 40-у, 80-у и 90-у процентилі.
Можемо використати функцію (Увага! Назва функції для різних версій MS Excel може відрізнятися. Запитуйте!)
=ПРОЦЕНТИЛЬ.ВКЛ(Массив;Требуемая вероятность)
Зверніть увагу, що в функцію підставляють не 40, а 0,4 – порядок потрібної квантилі.

В MS Excel існує ще багато інших способів розрахувати статистики та побудувати графіки розподілів. Головне під час продедення розраунків та побудови графіків враховувати шкалу змінної, яку ви аналізуєте.

А способи розрахунків і елементи візуалізації кожен вибирає на свій власний смак ☺

Квантилі розподілів, які використовуються в техніках статистичних обчислень

Завдання за варіантами – знайти квантилі – варіант 1, 2, 3, 4:

$\chi^2_{0.8}(10)$, $\chi^2_{0.01}(16)$, $\chi^2_{0.9}(20)$, $\chi^2_{0.05}(15)$

$\chi^2_{0.999}(130)$, $\chi^2_{0.99}(150)$, $\chi^2_{0.995}(110)$, $\chi^2_{0.99}(95)$

$\chi^2_{0.001}(130)$, $\chi^2_{0.01}(150)$, $\chi^2_{0.05}(110)$, $\chi^2_{0.01}(95)$

$t_{0.01}(25)$, $t_{0.005}(17)$, $t_{0.001}(9)$, $t_{0.05}(4)$

$t_{0.01}(114)$, $t_{0.005}(127)$, $t_{0.001}(158)$, $t_{0.05}(145)$

$F_{0.9}(5,4)$, $F_{0.95}(9,10)$, $F_{0.99}(12,24)$, $F_{0.9}(30,40)$;

$F_{0.1}(9,20)$, $F_{0.05}(30,18)$, $F_{0.01}(2,12)$, $F_{0.1}(30,30)$.

Завдання для індивідуальної роботи: $\mathbf{u_{p1}}$, $\mathbf{\chi 2_{p2}(k1)}$, $\mathbf{t_{p3}(k2)}$, $\mathbf{F_{p4}(k3,k4)}$.

Варіант	p1	p2	k1	p3	k2	p4	k3	k4
1	0,92	0,021	1	0,002	30	0,992	30	15
2	0,93	0,022	2	0,003	29	0,993	29	14
3	0,94	0,023	3	0,004	28	0,994	28	13
4	0,96	0,024	4	0,006	27	0,996	27	12
5	0,97	0,026	5	0,007	26	0,997	26	11
6	0,98	0,027	6	0,008	25	0,998	25	10
7	0,02	0,992	7	0,021	24	0,02	24	9
8	0,03	0,993	8	0,022	23	0,03	23	8
9	0,04	0,994	9	0,023	22	0,04	22	7
10	0,06	0,996	10	0,024	21	0,06	21	6
11	0,07	0,997	11	0,026	20	0,07	20	5
12	0,08	0,998	12	0,027	19	0,08	19	4
13	0,992	0,92	13	0,02	18	0,002	18	3
14	0,993	0,93	14	0,03	17	0,003	17	2
15	0,994	0,94	15	0,04	16	0,004	16	1
16	0,996	0,96	16	0,06	1	0,006	15	30
17	0,997	0,97	17	0,07	2	0,007	14	29
18	0,998	0,98	18	0,08	3	0,008	13	28
19	0,002	0,02	19	0,92	4	0,992	12	27
20	0,003	0,03	20	0,93	5	0,993	11	26
21	0,004	0,04	21	0,94	6	0,994	10	25
22	0,006	0,06	22	0,96	7	0,996	9	24
23	0,007	0,07	23	0,97	8	0,997	8	23
24	0,008	0,08	24	0,98	9	0,998	7	22
25	0,021	0,002	25	0,992	10	0,92	6	21
26	0,022	0,003	26	0,993	11	0,93	5	20
27	0,023	0,004	27	0,994	12	0,94	4	19
28	0,024	0,006	28	0,996	13	0,96	3	18
29	0,026	0,007	29	0,997	14	0,97	2	17
30	0,027	0,008	30	0,998	15	0,98	1	16

Інтервальне оцінювання

Побудова довірчих інтервалів для параметрів нормального розподілу

1. Побудувати довірчий інтервал для середньої ціни на куряче філе в районі (дані вибираємо з бази PripravkaKPI, вважаємо генеральну сукупність нормально розподіленою):

Варіант	Район	Рівень значущості
1	Голосіївський	0,01
2	Печерський	0,05
3	Святошинський	0,1
4	Солом’янський	0,01

1.1. Вважаємо дисперсію відомою та рівною 40 грн^2.

1.2. Дисперсія є невідомою та оцінюється по вибірці.

2. Побудувати довірчий інтервал для різниці середніх цін на куряче філе у районах:

Варіант	Район 1	Район 2	Рівень значущості
1	Голосіївський	Святошинський	0,05
2	Печерський	Солом’янський	0,1
3	Святошинський	Шевченківський	0,01
4	Солом’янський	Дарницький	0,05

3. Побудувати довірчий інтервал для дисперсії ціни на куряче філе:

Варіант	Район	Рівень значущості
1	Дарницький	0,1
2	Шевченківський	0,01
5	Печерський	0,01
6	Голосіївський	0,05

Перевірка статистичних гіпотез

Перевірка гіпотез про параметри нормального розподілу

В декількох торгових точках були заміряні ціни на шоколадки:

Корона	26,5	24,6	26,7	28,5	21,4	24,5	26,9	
Мілка	28,2	29,4	30,1	34,5	32,8	37,2	31,4	32,5
Світоч	24,4	24,4	24,6	24,1	24,6	25,5		
Рошен	26,5	26,2	28,4	27,2	26,4	25,2		

Перевірити гіпотези:

1. Чи дорівнює середня ціна на шоколадку Корона в генеральній сукупності 30 грн. (рівень значущості 0,05)

2. Чи можна вважати, що середня ціна на шоколадку Мілка перевищує 30 грн. (рівень значущості 0,01) в генеральній сукупності

3. Чи можна вважати, що середня ціна на шоколадку Світоч менше 30 грн. (рівень значущості 0,1), якщо **дисперсія генеральної сукупності відома та дорівнює 1**

4. Чи можна вважати, що дисперсія ціни на шоколадку Рошен більше 1 (рівень значущості 0,05)

5. Чи можна вважати, що дисперсія ціни на шоколадку Світоч менше 4 (рівень значущості 0,05)

6. Чи можна вважати, що ціни на шоколадки Корона та Мілка рівні в генеральній сукупності (рівень значущості 0,05)

7. Чи можна вважати, що ціни на шоколадки Корона та Світоч рівні в генеральній сукупності (рівень значущості 0,05)

Перевірка гіпотез про параметр p біноміального розподілу

Задача 1

Для перевірки ефективності рекламної компанії було проведене опитування. Всього у ньому взяли участь 500 респондентів. З них 270 бачили рекламний ролик, 230 – не бачили. Серед тих, хто бачив ролик, 200 респондентів позитивно ставляться до продукту компанії, решта – негативно. Серед тих, хто не бачив ролик, 160 чоловік позитивно налаштовані, інші – негативно. Чи можна стверджувати, що ролик змінює ставлення споживачів на краще (перевірити, чи збільшується частка позитивно налаштованих у генеральній сукупності, $\alpha = 0,05$)

Задача 2

Під час проведення апробації анкети було проведене 58 інтерв'ю. На останнє відкрите питання відмовилися дати відповідь 49 респондентів. Чи можна вважати, що частка відмов у всій генеральній сукупності перевищить 75% ($\alpha = 0,01$).

Задача 3

В результаті опитування були отримані наступні результати: Серед 85 жінок 47 надають перевагу чорному шоколаду, серед 50 чоловіків 35 надають перевагу чорному шоколаду.

3.1. Чи можна вважати, що частки жінок та чоловіків, які надають перевагу чорному шоколаду, в генеральній сукупності співпадають ($\alpha = 0,05$).

3.2. Чи можна вважати, що частка жінок, які надають перевагу чорному шоколаду, в генеральній сукупності перевищує 50 % ($\alpha = 0,1$).

Однофакторний дисперсійний аналіз

В декількох торгових точках були заміряні ціни на шоколадки:

Корона	26,5	24,6	26,7	28,5	21,4	24,5	26,9	
Мілка	28,2	29,4	30,1	34,5	32,8	37,2	31,4	32,5
Світоч	24,4	24,4	24,6	24,1	24,6	25,5		
Рошен	26,5	26,2	28,4	27,2	26,4	25,2		

Перевірити гіпотези:

1. Чи можна вважати, що ціни на всі шоколадки рівні в генеральній сукупності (рівень значущості 0,1)
2. Чи можна вважати, що ціни на шоколадки Корона, Світоч та Рошен рівні в генеральній сукупності (рівень значущості 0,01)

Непараметричні критерії перевірки статистичних гіпотез

1. Чи можна вважати, що частка споживачів, які позитивно ставляться до продукту, у генеральній сукупності перебільшує p_0, якщо з n опитаних у ході досліджень x споживачам подобається продукт, а (n – x) – не подобається?
1) $p_0 = 0.6$, $n = 120$, $x = 85$, $\alpha = 0.01$,
2) $p_0 = 0.75$, $n = 250$, $x = 190$, $\alpha = 0.05$,
3) $p_0 = 0.8$, $n = 300$, $x = 255$, $\alpha = 0.01$,
4) $p_0 = 0.9$, $n = 450$, $x = 409$, $\alpha = 0.05$.

2. При проведенні маркетингових досліджень було опитано n споживачів. Питання передбачало вибір одного з k варіантів відповідей. I-ий варіант обрали ni споживачів . Чи можна вважати з заданим рівнем значущості, що частки споживачів у генеральній сукупності відрізняються?
1) $n = 200$, $k = 3$, $n_1 = 62$, $n_2 = 58$, $n_3 = 80$, $\alpha = 0.05$,
2) $n = 200$, $k = 4$, $n_1 = 38$, $n_2 = 43$, $n_3 = 54$, $n_4 = 65$, $\alpha = 0.10$,
3) $n = 300$, $k = 4$, $n_1 = 82$, $n_2 = 74$, $n_3 = 79$, $n_4 = 65$, $\alpha = 0.01$,
4) $n = 200$, $k = 5$, $n_1 = 32$, $n_2 = 41$, $n_3 = 38$, $n_4 = 59$, $n_5 = 30$, $\alpha = 0.025$.

3. Споживачам з двох різних сегментів було запропоновано висловити своє ставлення до різних характеристик продуктів:
1) Упаковка: подобається 45 з 60 респондентів з першого сегменту, та 58 з 80 респондентів з другого;
2) Технічні характеристики: 31 з 70 – перший, 45 з 95 – другий;
3) Ергономічність продукту: 28 з 49 – перший, 42 з 60 – другий;
4) Сервіс: 42 з 120 – перший, 69 з 135 – другий.
Значущість 0,05. Перевірити різницю між сегментами для генеральної сукупності.

4. У ході опитування відвідувачів ТРЦ задавалися питання щодо частоти відвідування та наявності авто. В результаті первинного аналізу даних була отримана така перехресна таблиця:

Частота відвідування	Частіше, ніж раз на тиждень	3-4 рази на місяць	1-2 рази на місяць	Рідше, ніж раз на місяць
Мають авто	22	47	85	24
Не мають авто	15	48	109	50

Необхідно перевірити, чи є статистично значуща різниця у частоті відвідування ТРЦ між покупцями, які мають та які не мають авто.

5. За результатами попередніх досліджень частки споживачів в генеральній сукупності за сегментами становлять: 0,25; 0,40; 0,20; 0,10 та 0,05. В ході досліджень, в яких було опитано 259 респондентів, було отримано наступний розподіл по сегментам: 72, 97, 56, 19 та 15. Чи можна вважати гіпотезу про частки сегментів вірною?

6. Підприємство має традиційний магазин та інтернет-магазин. Керівництво поставило перед аналітичним відділом завдання дослідити, чи відрізняються продажі через ці два канали. Були заміряні суми чеків за один тиждень (серійна вибірка). Дані були згруповані:

Традиційний магазин	
Сума чеку, грн	Кількість чеків
Від 0 до 50	45
Від 50 до 150	68
Від 150 до 300	95
Від 300 до 450	42
Від 450 до 600	22
Від 600 до 1000	10

Інтернет-магазин	
Сума чеку, грн	Кількість чеків
Від 0 до 200	35
Від 200 до 400	28
Від 400 до 600	54
Від 600 до 800	25
Від 800 до 1000	18

Чи можна вважати що розподіли сум чеків є однаковими у двох типах магазинів?
(для розв’язання завдання скористатися критерієм Колмогорова-Смірнова)

7. У ході досліджування необхідно було побудувати портрет покупця. Серед інших параметрів вивчали вік та час користування послугою. Отримали такий групований ряд:

Вік	Кількість покупців
Від 12 до 18	5
Від 18 до 25	26
Від 25 до 35	52
Від 35 до 55	15
Від 55 до 70	4

Час користування послугою, місяців	Кількість покупців
Від 0 до 3	25
Від 3 до 6	32
Від 6 до 12	28
Від 12 до 24	16
Від 24 до 48	10

Чи можна вважати, що вік покупців має нормальний розподіл, а час користування послугою розподілений рівномірно від 0 до 48 місяців?

Кореляційний аналіз

1. У деяких випадково обраних магазинах фірма провела кампанію з просування, яка включала дегустацію продукту та поширення рекламних матеріалів про продукт. Витрати та зміна у відсотках обсягів продажів наводяться в таблиці:

Номер магазину	Відсоток підвищення обсягів продажів, %	Витрати на дегустації, тис. грн	Витрати на рекламні матеріали, тис. грн
	X	Y	Z
1	11,75	4,2	2,8
2	6,00	3,7	1,4
3	12,50	9,2	6,2
4	6,25	2,4	3,1
5	12,50	6,8	5,4
6	10,25	5,2	3,1
7	11,25	6,2	2,4
8	12,25	6,8	4,5
9	9,75	2,4	2,1
10	11,00	6,5	5,2
11	5,00	0	1,2
12	9,75	3,1	0,9
13	9,75	2,1	0,7
14	12,50	10,4	5,2
15	9,75	3,4	2,4

Маркетингова задача:
Чи впливають витрати на просування на зміну обсягів продажів? Який з інструментів є більш ефективним? Чи можна залишити лише один з двох інструментів?

Задача в термінах статистики:
Розрахувати парні та частинні коефіцієнти кореляції, перевірити значущість, зробити висновки. Розрахувати множинний коефіцієнт кореляції (та коефіцієнт детермінації) впливу на варіацію змінної X змінних Y,Z. Перевірити значущість зробити висновки.

Дати маркетологу обґрунтовані статистично рекомендації!

2. В ході дослідження клієнти оцінювали 15 різних магазинів за кількома характеристиками (від 10 – найвища оцінка до 1 – найнижча оцінка), а також давали загальну оцінку всьому магазину. Після узагальнення даних були отримані наступні результати:

№	Характеристика	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	Якість асортименту	8	5	4	2	10	5	7	4	10	9	9	9	6	10	7
2	Якість обслуговування	6	8	3	7	9	6	4	5	9	9	8	3	5	9	6
3	Акуратність викладки	6	6	5	4	8	7	5	6	9	10	10	6	7	9	8
4	Рівень цін	5	6	5	4	8	4	5	6	9	8	10	7	3	7	9
	Загальна оцінка	6	6	4	3	8	6	6	4	9	7	9	8	6	8	6

Маркетингова задача:

Чи зв’язані між собою оцінки по різним характеристикам? Які є зв’язаними найбільш щільно?

Чи зв’язані між собою оцінки по всім характеристикам (клієнт має певний портрет магазину)

Проранжувати характеристики по ступеню важливості для клієнтів (по щільності зв’язку з загальною оцінкою).

Задача в термінах статистики:

По варіантам для зменшення навантаження:

1. Розрахувати коефіцієнт кореляції Спірмена між змінними X 1 та X 2, між змінними X 3 та загальною оцінкою. Перевірити значущість

3. Розрахувати коефіцієнт конкордації між змінними 1-4. Перевірити значущість (Не забутьте про пов’язані ранги!!!)

Варіант	X1	X2	X3
1	1	2	3
2	2	3	4
3	1	4	2
4	3	4	1

Всі інші коефіцієнти отримати за допомогою SPSS, звести всі парні коефіцієнти та їх значущість в таблицю (4x4).

Зробити один загальний висновок і передати результати маркетологу для впровадження! ☺

Регресійний аналіз

Відкрийте базу даних **RegressionSushiya.sav**

1. Кореляція

Для того, щоб визначитися із залежними змінними, перевіримо, чи існує зв'язок між часткою відвідувачів з дітьми та характеристиками мікрорайону, у якому він знаходиться.

Оберіть процедуру **Анализ-Корреляции-Парные**. Будемо розраховувати коефіцієнт кореляції Пірсона. Усі метричні змінні, які є у переліку, обираємо.
Запускаємо аналіз.
Робимо висновок, з якими змінними кореляція Частки відвідувачів з дітьми є значущою:

	Коэффициент корреляции	Значимость
Бізнес_центри_тис_м2		
Торгові_центри_тис_м2		
Дитячі_магазини_тис_м2		
Парки_га		
Розмір_мікрорайону_тис_кв		
Кінотеатри_тис_місць		

Крім того, можемо відразу перевірити наявність/відсутність мультиколінеарності меж змінними, за якими будемо будувати регресію.

Висновок: _____

2. Кореляційне поле

Обираємо змінну з максимальним коефіцієнтом кореляції як незалежну.

Це змінна X:_____
Для того, щоб подивитися, як виглядає залежність між змінними, можна побудувати кореляційне поле.

Оберіть **Графика-Устаревшие окна-Рассеяния/точки**. У вікні, що відкриється, оберіть: Простая диаграмма рассеяния. Нажмите **Задать**.
По Осі X оберіть незалежну змінну.
По Осі Y оберіть змінну **Частка відвідувачів з дітьми.**
Запустіть побудову графіку.

Відкрийте редактор діаграм (два рази клікнути мишкою на графіку).
Оберіть **Элементы-Линия аппроксимации для итога**. Проставьте метод аппроксимации **Линейная регрессия** и Доверительные интервалы для среднего.
Зайдіть ще раз і додайте **Доверительные интервалы для отдельных значений.**
Так можна побачити лінію регресії та її інтервальні оцінки для генеральної сукупності.

3. Парна лінійна регресія

Оберіть процедуру **Анализ-Регрессия-Линейная.**
Оберіть залежну змінну: **Частка відвідувачів з дітьми**
Оберіть незалежну змінну: **Кінотеатри_тис_місць**
Натисніть кнопку **Статистики** та оберіть:
Оценки для коефіцієнтів регресії,
Статистики остатков: Дарбина-Уотсона и Диагностику по наблюдениям, замініть значення 3 на значення 2.
Інші параметри для парної регресії не потрібні.
Натисніть кнопку **Графики** та оберіть:
Диаграмму рассеяния: графік залежності стандартизованих залишків по Y (*ZRESID) від стандартизованих передбачених значень по X (*ZPRED);
Графики стандартизованных остатков: Гистограмму и Нормальный вероятностный график.
Натисніть кнопку **Сохранить** і оберіть:
Предсказанные значения: Нестандартизированные
Остатки: Нестандартизированные
Интервалы предсказания: Отдельное значение
Запустіть виконання процедури.

Розглянемо отримані результати:
1. Введенные или удаленные переменные – не потрібна для парної регресії

2. Сводка для моделі
Значення **коэффициента детерминации** і статистика **Дарбина-Уотсона**

Висновки:_____

_____

_____

3. Дисперсійний аналіз
Перевірка значущості моделі

Висновки: _____

_____

4. Таблиця коефіцієнтів регресії, їх середньоквадратичних відхилень та перевірка їх значущості

Висновки: _____

_____

5. Статистики залишків
Можемо побачити такі статистики: мінімальне, максимальне та середнє передбачене значення, залишки та інші.

6. Гістограма та ймовірнісний графік (квантильна діаграма) стандартизированих залишків.
Діаграма розсіювання
Візуально визначаємо нормальність залишків та гомоскедастичність. Для цього потрібний достатньо великий об'єм вибірки та досвіт регресійного аналізу.

Перевіримо нормальність залишків за допомогою критерію Колмогорова-Смірнова.
Оберіть процедуру **Анализ-Непараметрические критерии-Устаревшие окна-Одновыборочный критерий Колмогорова-Смирнова**
Оберіть у список змінних, що перевіряють: **RES_1.**
Перевіряємо на нормальний розподіл. Запустіть процедуру.
Подивіться результат.

Висновок:_____

3. Множинна лінійна регресія

Перевіримо, чи можна покращити прогноз, якщо застосувати не парну, а множинну регресію.

Оберіть прцедуру **Анализ-Регрессия-Линейная.**
Оберіть залежну змінну: **Частка відвідувачів з дітьми**
Оберіть незалежні змінні: **Бізнес_центри_тис_м2, Торгові_центри_тис_м2, Дитячі_магазини_тис_м2, Кінотеатри_тис_місць**
Оберіть **Метод** шагового відбору (він дозволить використати оптимальну кількість незалежних змінних).
Натисніть кнопку **Статистики** і оберіть:
Оценки и Доверительные интервалы для коефіцієнтів регресії,
Статистики остатков: Дарбина-Уотсона і Диагностику по наблюдениям, замініть значення 3 на значення 2.
Согласие модели і **Изменение R-квадрат**
Натисність **Графики** і оберіть:
Диаграмму рассеяния: графік залежності стандартизованих залишків по Y (*ZRESID) від стандартизованих передбачених значень по X (*ZPRED);
Графики стандартизованных остатков: Гистограмму и Нормальный вероятностный график.
Натисніть кнопку **Сохранить** і оберіть:
Предсказанные значения: Нестандартизированные **Остатки:** Нестандартизированные
Интервалы предсказания: Отдельное значение
Запустіть виконання процедури.
Висновки: _____

_____

_____

Кластерний аналіз

1. Міри схожості об’єктів

1.1. Побудуємо матрицю **відстані** між об’єктами:

$$\begin{pmatrix}0\\2\\3\\5\end{pmatrix}, \begin{pmatrix}1\\5\\4\\3\end{pmatrix}, \begin{pmatrix}4\\2\\3\\1\end{pmatrix}$$

Побудуємо матрицю Евклідових відстаней:

І матрицю відстаней Манхеттена:

Висновок:_____

_____

1.2. Разраховуємо **коефіцієнти подібності** для об’єктів:

$$\begin{pmatrix}\text{є}\\\text{супермаркет}\\\text{Солом'янський}\\\text{ні}\\\text{відмінно}\\\text{добре}\end{pmatrix}, \begin{pmatrix}\text{є}\\\text{мінімаркет}\\\text{Солом'янський}\\\text{є}\\\text{добре}\\\text{добре}\end{pmatrix}, \begin{pmatrix}\text{ні}\\\text{ринок}\\\text{Шевченківський}\\\text{ні}\\\text{погано}\\\text{погано}\end{pmatrix}, \begin{pmatrix}\text{ні}\\\text{супермаркет}\\\text{Солом'янский}\\\text{є}\\\text{відмінно}\\\text{добре}\end{pmatrix}, \begin{pmatrix}\text{є}\\\text{гипермаркет}\\\text{Святошинський}\\\text{ні}\\\text{відмінно}\\\text{відмінно}\end{pmatrix}$$

Побудуємо матрицю коефіцієнтів подібності:

Висновки: _____

_____

1.3. Розрахуємо **коефіцієнти асоціативності** для двух бінарних змінних:

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
Спостереження1	1	0	1	0	1	1	0	1	1	0	0	1	1	0	1
Спостереження2	1	1	0	0	1	0	0	0	1	0	1	1	1	0	1

Побудуємо таблицю асоціативності:

	Спостереженння 1	
Спостереження 2	1	0
1	a	b
0	c	d

Евклідова відстань:

$$d = \sqrt{b+c} = \_\_\_\_\_$$

Різниця довжин:

$$d = \frac{(b-c)^2}{(a+b+c+d)^2} = \_\_\_\_\_$$

Різниця структур:

$$d = \frac{bc}{(a+b+c+d)^2} = \_\_\_\_\_$$

Просте узгодження :

$$d = \frac{a+d}{a+b+c+d} = \_\_\_\_\_$$

2. Ієрархічний кластерний аналіз

Відкрийте базу данних **CompetitiveAnalysis.sav.**

Нехай необхідно класифікувати 12 підприємств (конкурентів компанії Аврора) за такими ознаками:

- частка ринку
- ширина асортименту
- рівень цін
- кількість торгових точок
- обсяги продажів.

2.1. Перевіримо ознаки на відсутність мультиколінеарності: для цього запустимо процедуру **Аналіз-Кореляції-Парні**, виберемо 5 потрібних змінних і розрахуємо **коефіцієнт кореляції Пірсона** (всі змінні метричні). Проаналізуємо результати:

_____

_____

2.2. Стандартизація змінних необхідна, так як наші змінні мають різний масштаб.

Для стандартизації змінних є дві можливості:

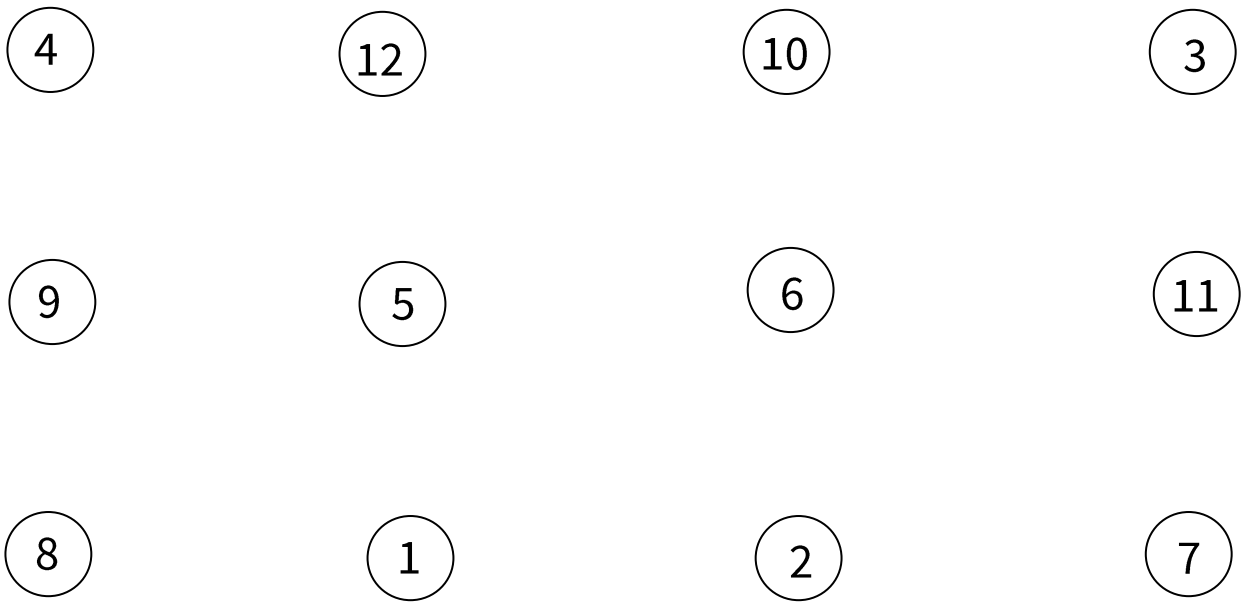
- є безпосередньо в процедурі **Ієрархічний кластерний аналіз** або
- можна виконати її самостійно, запустивши процедуру **Аналіз-Описові статистики-Описові**, вибравши необхідні змінні і поставивши прапорець **Зберегти стандартизовані значення.**

2.3. Необхідно вибрати міру подібності об'єктів
Виберемо **Евклідова відстань**.
Його можна розрахувати в процедурі **Аналіз-Кореляції-Відстані**. Виберіть потрібні змінні (не забудьте, що вибирати необхідно вже стандартизовані змінні) і задайте міру відмінності.
Повинна вийти така таблиця:

Евклидово расстояние

	1	2	3	4	5	6	7	8	9	10	11	12
1	0,00	3,20	2,92	2,96	2,58	3,19	2,92	1,83	0,65	2,70	3,36	2,54
2	3,20	0,00	2,25	3,73	1,09	0,49	0,67	3,98	3,38	2,03	2,33	1,97
3	2,92	2,25	0,00	4,79	1,29	1,97	2,03	4,05	3,17	0,42	0,89	1,66
4	2,96	3,73	4,79	0,00	3,82	4,02	3,64	3,81	2,58	4,52	5,22	4,01
5	2,58	1,09	1,29	3,82	0,00	0,88	1,06	3,68	2,78	1,00	1,70	1,14
6	3,19	0,49	1,97	4,02	0,88	0,00	0,96	4,00	3,41	1,73	2,10	1,62
7	2,92	0,67	2,03	3,64	1,06	0,96	0,00	3,74	3,12	1,90	2,04	2,14
8	1,83	3,98	4,05	3,81	3,68	4,00	3,74	0,00	2,33	3,88	4,18	3,73
9	0,65	3,38	3,17	2,58	2,78	3,41	3,12	2,33	0,00	2,94	3,70	2,70
10	2,70	2,03	0,42	4,52	1,00	1,73	1,90	3,88	2,94	0,00	1,20	1,25
11	3,36	2,33	0,89	5,22	1,70	2,10	2,04	4,18	3,70	1,20	0,00	2,35
12	2,54	1,97	1,66	4,01	1,14	1,62	2,14	3,73	2,70	1,25	2,35	0,00

2.4. Приступимо до агломерації. Для того, щоб не потрібно було перераховувати відстані, виберемо **Метод ближнього сусіда** для об'єднання кластерів. Перебираючи послідовно всі відстані в таблиці від мінімального до максимального, будемо об'єднувати об'єкти, записуючи номер кроку і відстань:



2.5. Виділення кластерів.
В результаті об'єднання вийшов один великий кластер. Тепер, для того, щоб отримати 2, 3 і т.д. кластерів, необхідно виробляти зворотні дії: розривати послідовно зв'язки, починаючи з найслабшою (останньої). Так, щоб отримати 3 кластера, потрібно розірвати 2 останніх зв'язки: видалити 10-й і 11-й кроки. Зробимо це і подивимося, що вийшло. Запишіть приналежність об'єктів до кластерів:

Кластер 1:_____ Кластер 2:_____ Кластер 3:_____

2.6. А тепер зробимо те ж саме в SPSS. Виберіть процедуру **Аналіз-Класифікація-ієрархічна кластеризація**. Виберіть змінні.
Натисніть кнопку **Графіки** і поставте галочку **Дендрограма**.
Натисніть кнопку **Метод** і виберіть **Метод ближнього сусіда**, виберіть интервальную міру **Евклідова відстань**, внизу є можливість задати стандартизацію змінних.
Натисніть кнопку **Зберегти** і виберіть зберегти приналежність до кластерів - одне рішення - 3 кластера.
Запустіть процедуру. Розглянемо результати.

3. Метод К-середніх
3.1. Розглянемо роботу алгоритму. При використанні методу К-середніх задається кількість кластерів. Можуть бути задані також початкові центри кластерів (якщо є припущення про їх розташуванні). Інакше алгоритм сам вибирає початкові центри, таким чином, щоб між ними була максимальна відстань.
У нашому випадку обрані 8-й, 4-й і 11-й елементи. Далі для кожного з об'єктів (не центрів кластерів) знаходиться найближчий центр, і об'єкт приєднується до цього кластеру.

Евклідова відстань

	1	2	3	4	5	6	7	8	9	10	11	12
8	1,83	3,98	4,05	3,81	3,68	4,00	3,74	0,00	2,33	3,88	4,18	3,73
4	2,96	3,73	4,79	0,00	3,82	4,02	3,64	3,81	2,58	4,52	5,22	4,01
11	3,36	2,33	0,89	5,22	1,70	2,10	2,04	4,18	3,70	1,20	0,00	2,35

Запишіть приналежність об’єктів до кластерів:

Кластер 1:_____ Кластер 2:_____ Кластер 3:_____
Далі для отриманих кластерів перераховуються центри, і знову перевіряється приналежність об'єктів до кластерів. У нашому випадку на другому кроці алгоритм зупиняється.

3.2. А тепер зробимо те ж саме в SPSS. Виберіть процедуру **Аналіз-Класифікація-Кластеризація К-середніми**. Виберіть змінні (зверніть увагу, в цій процедурі стандартизації немає, тому необхідно вибирати тільки стандартизовані змінні). Задайте число кластерів 3.
Натисніть кнопку **Зберегти** і поставте прапорець **Приналежність до кластера**.
Натисніть кнопку **Параметри** і поставте прапорці **Таблиця дисперсійного аналізу** і **Кінцевий кластер** для спостережень. Запустіть процедуру. Розгляньте результати.

А далі можна пробувати і тренуватися з алгоритмами, заходами, відстанями і кількостями кластерів і таке інше хоч все життя ...☺

САМОСТІЙНІ РОБОТИ

Самостійна робота 1. Аналітичне та графічне представлення вибірки

Мета роботи:
Навчитися працювати з вибірками різного об'єму, будувати описові статистики, отримувати різні графічні представлення вибірки за допомогою можливостей MS Excel

Завдання:
З використанням MS Excel для вибірок об'ємом 50, 200 і 1000 елементів:
1. Побудувати групований варіаційний ряд
2. Побудувати гістограму частот, полігон відносних частот, кумулятивну криву
3. Отримати розмах вибірки, оцінки математичного сподівання, дисперсії, моди, медіани, квантилів порядку 0,25 та 0,9 з використанням повної вибірки та групованого ряду, порівняти отримані результати.
4. Сформувати гіпотезу про розподіл генеральної сукупності.

Для цього завдання вашим варіантом є номер за списком. Оберіть дані з відповідної змінної з файлу **TaskData_3.xlsx**

Хід роботи:
1. Відібрати перші 50 елементів вибірки та скопіювати їх на новий лист. Наприклад, можете дати назву листу: 'Дані 50'
2. За допомогою команди **Сортировка** в меню **Данные** побудувати варіаційний ряд.
Розрахувати розмах вибірки:

$$R = x_{max} - x_{min}$$

3. На окремому листі створити таблицю, яка має наступні стовпці:

№ інтервалу	Ліва межа інтервалу	Права межа інтервалу	Середина інтервалу	Частота інтервалу	Відносна частота інтервалу	Відносна накопичена частота інтервалу
1						
…						
20						

Наприклад, шапка таблиці може бути розміщена в області A5:G5, а вся таблиця – в області A5:G25.

4. Для того, щоб мати змогу змінювати межі інтервалів, необхідно передбачити додаткові комірки (до чи після таблиці): «Ліва межа першого інтервалу» та «Довжина інтервалу». Наприклад, це можуть бути комірки $F$2 і $F$3 відповідно.

5. Заповнення таблиці:
5.1. Для першого інтервалу ліва межа задається рівною спеціально призначеною коміркою (=F2), для всіх наступних інтервалів ліва межа дорівнює правій межі попереднього інтервалу.

5.2. Права межа визначається як ліва межа плюс довжина інтервалу.

5.3. Середина інтервалу – середнє арифметичне лівої та правої межі.
Наприклад, для другого інтервалу, яких буде розміщено в рядку 7, в комірці, яка визначає праву межу інтервалу, буде наступна формула: =B7+$F$3, а в комірці, яка визначає середину інтервалу: =(B7+C7)/2.

5.4. Визначення частоти інтервалу передбачає підрахунок кількості таких елементів вибірки, які більше лівої межі інтервалу, але менше правої.
Можна запропонувати багато різних способів, якими можна заповнити стовпець частот. Один з них – використати стандартну функцію СЧЁТЕСЛИ(диапазон;условие).
Наприклад, якщо дані розміщені в області $A$1:$A$50 на листі ‘Дані 50’, частота другого інтервалу може бути розрахована наступним чином:
=СЧЁТЕСЛИ('Дані 50'!$A$1:$A$50;"<"&C7)-СУММ($E$6:E6).

5.5. Відносна частота інтервалу розраховується як частота інтервалу поділена на кількість елементів вибірки, а відносна накопичена частота для довільного інтервалу – сума частот з першого інтервалу до даного включно.

6. Для побудови гістограми частот в **Мастере диаграмм** обрати **Гистограмма**, в **Диапазон данных** задати область частот інтервалів, в закладці **Ряд** визначити **Подписи оси X**, наприклад, вказуючи область середин інтервалів. Додати назву в закладці **Заголовки.** Після отримання попереднього графіка, клацнути мишкою на стовпчиках і увійти в **Формат рядов данных.** Зменшити **Ширину зазора** до 0. Змінюючи інші параметри, можна значно змінити зовнішній вигляд графіка.

7. Для побудови полігона відносних частот чи кумулятивної кривої в **Мастере диаграмм** обрати **График** з маркерами, які відмічають точки даних, в **Диапазон данных** задати область відносних частот інтервалів чи накопичених відносних частот.

8. Після побудови графіків, змінюючи ліву межу першого інтервалу та довжину інтервалу, необхідно обрати оптимальне представлення вибірки – такий групований ряд, для якого графіки є найбільш гладкими, при цьому всі елементи повинні потрапити в побудовані інтервали.

9. За виглядом гістограми та полігону відносних частот висунути гіпотезу про вид розподілу генеральної сукупності, з якої було отримано вибірку.

10. Для знаходження значень вибіркового середнього, вибіркової дисперсії, оцінок моди, медіани та квантилів можна скористатись наступними стандартними функціями MS Excel (будьте уважні, читайте довідки по функціях – в різних версіях функції можуть відрізнятися!):

СРЗНАЧ(Массив) – розраховує вибіркове середнє; використовуючи функцію необхідно пам’ятати, що пусті комірки за замовчуванням рахуються такими, що мають нульове значення;

ДИСП(Массив) або ДИСП.В(Массив) – розраховує незсунену оцінку дисперсії; якщо в масиві є логічні значення чи текст, вони ігноруються;

МОДА(Массив) або МОДА.ОДН(Массив) – повертає значення в масиві, яке зустрічається найчастіше; якщо вибірка не має значень, які повторюються, функція повертає значення #Н/Д – помилка;

МЕДИАНА(Массив) – повертає число, яке є серединою множини чисел: половина чисел менше медіани, половина – більше; якщо у множині парна кількість чисел, то функція МЕДИАНА() обчислює середнє двох чисел, які знаходяться в середині множини;

ПЕРСЕНТИЛЬ(Массив;k) – повертає k-у перцентиль (вибіркову квантиль порядку k) для значень з інтервалу; k приймає значення від 0 до 1.

11. Використовуючи групований статистичний ряд, отримати оцінку математичного сподівання та вибіркову дисперсію:

$$\bar{X}^* = \frac{1}{n}\sum_{i=1}^{k} x_i^* n_i^* \qquad S^{2^*} = \frac{1}{n-1}\sum_{i=1}^{k}\left(x_i^* - \bar{X}^*\right)^2 n_i^*$$

де n – об'єм вибірки,
k – кількість інтервалів,
x_i^* – середина i-го інтервалу,
n_i^* – частота i-го інтервалу.

Оцінка моди за групованим рядом:

$$\hat{d}^* = a_d + b \cdot \left(\frac{n_d^* - n_{d-1}^*}{2 \cdot n_d^* - n_{d-1}^* - n_{d+1}^*}\right),$$

де a_d – нижня межа інтервалу, у якому найбільша кількість елементів,
b – довжина інтервалу,
n_d^* – частота інтервалу, у якому найбільша кількість елементів,
n^*_{d-1} та n^*_{d+1} – частота відповідно попереднього та наступного інтервалів.

Оцінка медіани за групованим рядом:

$$\hat{h}^* = a_h + b \cdot \left(\frac{\frac{n}{2} - \left(n_1^* + \ldots + n_{h-1}^*\right)}{n_h^*}\right),$$

де a_h – ліва межа інтервалу, у якому знаходиться медіана,
b – довжина інтервалу,
n_h^* – частота інтервалу, у якому знаходиться медіана,
$(n^*_1 + \ldots + n^*_{h-1})$ – кількість елементів у попередніх інтервалах.

Оцінки квантилів можна отримати по кумулятивній кривій: вибіркова квантиль порядку p – абсциса точки на кумулятивній кривій, ордината якої дорівнює p.

12. Порівняти оцінки, отримані по повній та груповаій вибірці, пояснити чому вони відрізняються.

13. Побудувати графіки та розрахувати числові характеристики по вибіркам об'ємом 200 елементів (відібрати перші 200 з 1000 елементів) та об'ємом 1000 елементів.
14. Сформувати гіпотезу про розподіл генеральної сукупності та оцінити його параметри.

Контрольні теоретичні питання:

1. Статистика
2. Статистична оцінка
3. Незсуненість статистичної оцінки
4. Конзистентність статистичної оцінки
5. Ефективність статистичної оцінки
6. Гістограма частот
7. Полігон частот
8. Емпірична функція розподілу
9. Кумулятивна крива
10. Варіаційний ряд
11. Статистичний ряд
12. Групований статистичний ряд
13. Порядкові статистики
14. Міри центральної тенденції
15. Міри мінливості
16. Характеристики форми розподілу

Контрольні практичні завдання:

1. Вміти розрахувати числові характеристики та побудувати графіки за запропонованою невеликою вибіркою
2. Вміти розрахувати числові характеристики та побудувати графіки за запропонованим групованим статистичним рядом

Самостійна робота 2. Інтервальне оцінювання

Інтервальне оцінювання параметрів генеральної сукупності. Побудова довірчих інтервалів для параметра p біноміально розподіленої генеральної сукупності

Мета роботи:
Навчитися будувати довірчі інтервали на великих та малих об'ємах за допомогою можливостей MS Excel. Звернути увагу на те, як довжина довірчого інтервалу (похибка вибірки) залежить від об'єму вибірки, довірчої ймовірності та дисперсії генеральної сукупності (значення точкової оцінки h)

Завдання:
1. Було опитано n споживачів з метою визначення переваг щодо наповнювачів йогуртів. Отримані наступні результати:

Наповнювач	Кількість респондентів, які люблять даний йогурт
1. Шоколадний	x1
2. Ананасовий	x2
3. Полуничний	x3
4. Смородиновий	x4
5. Киви	x5
6. Малиновий	x6
7. Вишневий	x7
8. Абрикосовий	x8
9. Ванільний	x9
10. Персиковий	x10

Визначити, яким наповнювачам віддадуть перевагу найбільша кількість споживачів в генеральній сукупності.

2. В процесі лабораторного експерименту споживачам, які віддають перевагу йогуртам різним виробників, було запропоновано спробувати новий полуничний йогурт та порівняти його смак з тим йогуртом, який вони звичайно купують. Були отримані наступні результати:

ТМ йогурту, який подобається цим респондентам	Кількість респондентів	Кількість респондентів, яким сподобався йогурт, що тестується
Баланс	n1	k1
Данон	n2	k2
Чудо	n3	k3
Фанни	n4	k4
Президент	n5	k5

Споживачам яких ТМ йогуртів (в генеральній сукупності) більше за інших сподобається новинка?

Хід роботи:

1. Розрахувати точкову оцінку частки споживачів **p**:

$$\mathbf{h} = \frac{\mathbf{x}}{\mathbf{n}}$$

2. Розрахувати ліві та праві межі довірчих інтервалів, користуючись формулою:

$$h - u_{1-\frac{\alpha}{2}}\sqrt{\frac{h\cdot(1-h)}{n}} < p < h + u_{1-\frac{\alpha}{2}}\sqrt{\frac{h\cdot(1-h)}{n}}$$

NB! Для того, щоб мати можливість подивитися, як залежить похибка вибірки від довірчої ймовірності, зробіть окрему комірку для рівня значущості **α** та потім у формулі посилайтеся на цю комірку!!

3. Побудувати графік: обираємо тип графіку - **Гистограмма с группировкой**, в **Диапазон данных** задати область, в якій знаходяться ліві та праві межі довірчих інтервалів. Додати **Подписи оси X** – вказати область значень назв наповнювачів. Клацнути мишкою на одному зі стовпчиків та увійти в **Формат рядов данных.** В закладці **Параметры ряда** збільшити **Перекрытие** до 100%. Далі натиснути на область побудови графіка та в діалоговому вікні обрати **Выбрать данные**. В **Элементы легенды** ряди даних поміняти місцями (натиснути на трикутничок у правому верхньому куті). В результаті стануть видимі менші стовпчики. Подальше удосконалення за бажанням автора.
4. За графіком та стовпцями меж інтервалів обрати інтервал з максимальною лівою межею.
5. Обрати всі наповнювачі, довірчі інтервали для яких перетинаються з інтервалом, який має максимальну ліву межу. Це й будуть наповнювачі, які найбільше подобаються споживачам в генеральній сукупності.
6. Змінити рівень значущості та перевірити, як змінюється при цьому довжина довірчих інтервалів.
7. Повторити ці дії для розв'язання задачі 2. Враховувати, що невиконання будь-якої з умов:

$$\begin{cases} n > 50 \\ x > 5 \\ (n-x) > 5 \end{cases}$$

вимагає застосування для побудови довірчих інтервалів іншої формули:

$$\frac{x\cdot F_{\frac{\alpha}{2}}\left(2\cdot x, 2\cdot(n-x+1)\right)}{n-x+1+x\cdot F_{\frac{\alpha}{2}}\left(2\cdot x, 2\cdot(n-x+1)\right)} < p <$$

$$< \frac{(x+1)\cdot F_{1-\frac{\alpha}{2}}\left(2\cdot(x+1), 2\cdot(n-x)\right)}{n-x+(x+1)\cdot F_{1-\frac{\alpha}{2}}\left(2\cdot(x+1), 2\cdot(n-x)\right)}$$

Контрольні питання:

3. Що таке довірчий інтервал?
4. Що таке рівень значущості та довірча ймовірність?

5. Що таке похибка вибірки?
6. Як змінюється довірчий інтервал при зміні рівня значущості?
7. Як змінюється похибка вибірки із збільшенням об’єму вибірки?
8. Від чого ще залежить довжина довірчого інтервалу?
9. Якими є переваги та недоліки інтервальних оцінок у порівнянні з точковими?
10. Які існують умови використання формул довірчого інтервалу для параметру p біноміально розподіленої генеральної сукупності?
11. Яким є розподіл генеральної сукупності в задачах про йогурти?
12. Для двох певних наповнювачів зробити висновок про існування статистично значущої різниці для частки споживачів в генеральній сукупності.

Самостійна робота 3. Перевірка статистичних гіпотез

T-тести в SPSS

Мета роботи:
Навчитися перевіряти параметричні статистичні гіпотези в SPSS.
Ознайомитися з Т-тестом для однієї вибірки та Т-тестом для двох незалежних вибірок
Використовуємо базу даних **PripravkaKPI.sav**

Завдання:
1. Перевірити, чи дорівнює середня ціна на куряче філе (змінна в масиві **Цена_кур_филе**) в районі К міста Києва m0 грн

Варіант	Район К	Значення ціни m0
1	Шевченківський	73,00
2	Дарницький	77,00
3	Дніпровський	75,00
4	Деснянський	71,00

2. Перевірити, чи дорівнює середня ціна на Приправку Ексклюзивну для курки (змінна в масиві **Цена_ПриправкаЭкс_для_кур**) в районі К міста Києва m0 грн

Варіант	Район К	Значення ціни m0
1	Дніпровський	16,00
2	Деснянський	18,00
3	Шевченківський	19,00
4	Дарницький	17,00

3. Перевірити, чи дорівнює середня ціна на куряче філе (змінна в масиві **Цена_кур_филе**) в двох типах магазинів міста Києва

Варіант	Тип магазину 1	Тип магазину 2
1	Супермаркети	Ринки
2	Мінімаркети	Ринки
3	Магазини біля дому	Ринки
4	Ринки	Гіпермаркети

4. Перевірити, чи дорівнює середня ціна на Приправку Ексклюзивну для курки (змінна в масиві **Цена_ПриправкаЭкс_для_кур**) в двох типах магазинів міста Києва

Варіант	Тип магазину 1	Тип магазину 2
1	Магазини біля дому	Мінімаркети
2	Магазини біля дому	Гіпермаркети
3	Супермаркети	Мінімаркети
4	Ринки	Супермаркети

Контрольні питання:

1. Статистична гіпотеза
2. Етапи перевірки статистичних гіпотез
3. Критерій перевірки статистичних гіпотез
4. Помилка I-го роду
5. Помилка II-го роду
6. Ймовірність помилки I-го роду
7. Ймовірність помилки II-го роду
8. Потужність критерію
9. Область прийняття рішення
10. Критична область
11. Статистичне рішення
12. Статистичний тест
13. Значущість статистичного тесту

РОЗРАХУНКОВА РОБОТА

1. Мета розрахункової роботи

Розрахункова робота з дисципліни “Статистика” передбачає самостійне поглиблене вивчення одного з методів, які розглядаються на практичних заняттях та його реалізації за допомогою програми статистичної обробки даних.

Студент у ході виконання розрахункової роботи повинен:

- продемонструвати розуміння сфери застосування запропонованого йому статистичного метода,
- вміти підібрати дані з практики маркетингових досліджень чи маркетингової діяльності підприємства, які можна опрацьовувати з допомогою цього метода,
- показати, як з цих даних можна отримати інформацію і як ця інформація може бути використана в управлінні підприємством,
- навести відповідні розрахунки (у разі складних багатовимірних методик – приклади, частину розрахунків), які зроблені самостійно,
- за допомогою запропонованої програми статистичної обробки даних отримати необхідні результати,
- детально описати алгоритм використання програмного засобу для статистичної обробки даних.

2. Тематика розрахункових робіт

Кожен студент повинен самостійно або за допомогою викладача обрати собі тему розрахункової роботи.

Приблизна тематика розрахункових робіт:

1. Графіки в Microsoft Excel.
2. Статистичні функції Microsoft Excel.
3. Дескриптивна статистика в Microsoft Excel.
4. Інтервальне оцінювання в Microsoft Excel.
5. Т-тести та F-тест в Microsoft Excel.
6. Однофакторний дисперсійний аналіз в Microsoft Excel.
7. Двохфакторний дисперсійний аналіз в Microsoft Excel.
8. Регресія в Microsoft Excel.
9. Графіки в SPSS.
10. Дескриптивна статистика в SPSS.
11. Порівняння середніх в SPSS.
12. Кореляція. Коефіцієнт кореляції в SPSS.
13. Лінійна регресія в SPSS.
14. Нелінійна регресія в SPSS.
15. Однофакторний дисперсійний аналіз в SPSS.
16. Двохфакторний дисперсійний аналіз в SPSS.
17. Непараметрична кореляція в SPSS.
18. Непараметричні критерії перевірки статистичних гіпотез. Хі-квадрат критерії. SPSS.
19. Непараметричні критерії перевірки статистичних гіпотез. Критерій Колмогорова-Смірнова. SPSS.

20. Непараметричні критерії перевірки статистичних гіпотез. Критерій Манна-Уітні. SPSS.
21. Непараметричні критерії перевірки статистичних гіпотез. Критерій Краскала-Уоліса. SPSS.
22. Непараметричні критерії перевірки статистичних гіпотез. Критерій знаків. SPSS.
23. Непараметричні критерії перевірки статистичних гіпотез. Критерій Вілкоксона в SPSS.
24. Непараметричні критерії перевірки статистичних гіпотез. Критерій серій. SPSS.
25. Непараметричні критерії перевірки статистичних гіпотез. Критерій Фрідмана в SPSS.
26. Непараметричні критерії перевірки статистичних гіпотез. Біноміальний критерій. SPSS.
27. Кластерний аналіз (ієрархічна кластерізація) в SPSS.
28. Кластерний аналіз (кластерізація методом К-середніх) в SPSS.
29. Дискримінантний аналіз в SPSS.
30. Факторний аналіз в SPSS.

3. Вимоги до виконання

До виконання розрахункової роботи висуваються такі вимоги:

3.1. Розрахункова робота є індивідуальною роботою, яка виконується студентом самостійно під керівництвом викладача. Робота виконується на базі теоретичних знань та практичних навичок, отриманих студентом упродовж вивчення кредитного модулю «Статистика» і самостійної роботи. За робочим навчальним планом на підготовку студентом розрахункової роботи передбачено 15 годин самостійної роботи студента. Робота повинна мати теоретичну та практичну значущість.

3.2. Процес виконання роботи включає кілька етапів, а саме:

- вибір теми;
- самостійне поглиблене опанування статистичного методу;
- вибір маркетингової задачі, яку доцільно вирішити за допомогою даного статистичного методу;
- проведення розрахунків;
- оформлення роботи;
- захист розрахункової роботи.

3.3. Маркетингова задача, що обирається студентом самостійно, повинна відповідати таким вимогам:

- дані, обрані для аналізу, стосуються маркетингової діяльності деякого підприємства на ринку або процесу маркетингових досліджень;
- дані відповідають обраному статистичному методу (об’єм вибірки, тип даних, інші обмеження методу тощо);
- висновки, які буде отримано після проведення статистичного аналізу, можна застосувати для розробки рекомендацій з коригування продуктово-ринкової стратегії.

Для усунення помилок під час вибору і підвищення практичної значущості результатів маркетингову задачу необхідно узгодити з викладачем.

4. Структура і зміст роботи

Розрахункова робота з кредитного модуля «Статистика» повинна мати відповідну структуру:

- Титульний аркуш
- Зміст

- Вступ
- Розділ 1. Теоретичні відомості з обраного статистичного методу
- Розділ 2. Формулювання маркетингової задачі і розв'язок її за допомогою статистичного аналізу
- Розділ 3. Використання програмного забезпечення для розв'язання маркетингової задачі
- Висновки
- Список використаної літератури

Титульний аркуш виконується стандартно, на ньому вказується вид роботи, назва кредитного модуля, тема роботи, прізвище та ініціали студента, який виконував роботу, номер залікової книжки, дата здачі роботи, прізвище та ініціали викладача, який буде перевіряти роботу.

Зміст – перелік розділів роботи та інших її частин (див. структуру роботи) з відповідним номером сторінки у крайньому правому положенні в рядку.

У вступі необхідно коротко сформулювати отримане завдання з розрахункової роботи, включаючи метод статистичного аналізу, особливості маркетингової задачі та запропоноване програмне забезпечення для реалізації статистичного аналізу.

У першому розділі необхідно надати максимально стислі теоретичні відомості з обраного статистичного методу.

У другому розділі детально розглядається постановка маркетингової задачі обраного підприємства, надається опис об'єктів, які підлягають вивченню, визначаються змінні, які вимірюються, наводиться масив даних. Далі перевіряється виконання обмежень на використання статистичного методу, наводяться деталізовані розрахунки. Результат статистичного аналізу надається спочатку в термінах статистики, а потім трансформується в предметну область.

Третій розділ показує володіння студентом програмного забезпечення зі статистичного аналізу. У ньому необхідно надати покроковий алгоритм для отримання результатів, які було вручну розраховано у другому розділі, за допомогою використання спеціальної комп'ютерної програми.

У висновках необхідно стисло викласти отримані у ході статистичного аналізу результати і надати практичні рекомендації щодо їх використання.

5. Рекомендації до виконання

Для успішного виконання розрахункової роботи студенту необхідно:

5.1. Ретельно переглянути літературу з математичної статистики, з декількох джерел обрати найбільш легке та зрозуміле для сприйняття викладення статистичного методу, обраного для розрахункової роботи. Самостійно опанувати процес статистичного аналізу на рівні більш детальному, ніж пропонується в лекціях. За обраним теоретичним матеріалом скласти максимально стислий, але змістовий та грамотний конспект, з якого студенту зрозуміло, що являє собою обраний статистичний метод, для яких практичних задач він може бути застосований, які існують обмеження на використання вхідних даних, і як власне проводити розрахунки і робити

статистичні висновки. Результат цього опрацювання наводиться у першому розділі розрахункової роботи.

5.2. Самостійно придумати маркетингову задачу, яка може бути вирішена за допомогою даного метода, та підібрати дані, з урахуванням обмежень на використання метода щодо типу даних, типу шкали вимірювання, об’єму масиву даних тощо. Сформулювати задачу спочатку на змістовному рівні, а потім формалізувати її у термінах статистики. Розв’язати задачу вручну, з наведенням максимально деталізованих розрахунків. У тому випадку, якщо розрахунки є дуже великими за обсягом, можливе часткове обчислення, за умови зрозумілого ходу розрахунків. Зробити висновки спочатку в термінах статистики, а потім на змістовному рівні. Практичне застосування статистичного аналізу викладається у другому розділі.

5.3. Розв’язати задачу за допомогою запропонованого програмного продукту. При цьому слід звернути увагу на детальному викладенні всіх дій, які необхідно здійснити для отримання результатів статистичного аналізу, та їх послідовності. Дозволяється наводити рисунки вікон редактору даних, діалогові вікна, вікна виводу. При цьому значення кожного з прапорців, які можуть бути використані, повинно бути детально роз’яснене. Зміст таблиць виводу також необхідно пояснити. Розв’язання задачі за допомогою програмного продукту наводиться у третьому розділі.

5.4. У висновках викласти отримані результати статистичного аналізу. Розробити та навести рекомендації щодо використання результатів, які були отримані, в практичному маркетингу.

МОДУЛЬНА КОНТРОЛЬНА РОБОТА

Загальна кількість контрольних (модульних) робіт: одна МКР, яка поділяється дві роботи тривалістю по 1 академічній годині кожна.

Контрольна робота 1: Розділ 1. Статистичне спостереження, узагальнення та представлення даних
Ця контрольна робота передбачає:

- перевірку засвоєння студентами таких понять статистики як “статистичне спостереження”, “вибірка”, “генеральна сукупність”, “статистика”, “числова оцінка параметру”;
- перевірку вміння будувати точкові оцінки параметрів генеральної сукупності;
- демонстрацію студентами вміння вдало графічно представляти вибірку.

Контрольна робота 2: Розділ 2. Параметрична та непараметрична статистика Розділ 3. Прикладна статистика
Ця контрольна робота передбачає:
перевірку засвоєння студентами основних понять інтервального оцінювання, статистичної перевірки гіпотез, регресійного та кореляційного аналізу;
перевірку вміння будувати інтервальні оцінки параметрів генеральної сукупності;
перевірку вміння формувати та перевіряти статистичні критерії;
демонстрацію студентами вміння розраховувати коефіцієнт кореляції та будувати лінійну регресію.

МЕТОДИКА ОПАНУВАННЯ ДИСЦИПЛІНИ

В межах вивчення дисципліни протягом семестру заплановано проведення лекційних та практичних занять, написання модульної контрольної роботи.

Навчальним планом передбачено індивідуальне завдання у вигляді розрахункової роботи.

Опанування студентами дисципліни передбачає вивчення теоретичного матеріалу, який викладається на лекціях та пропонується студентам для самостійної підготовки, та здобуття практичних навичок з розв'язання практичних задач, яке відбувається на практичних заняттях. Ознайомлення з новою темою на практичному занятті передбачає короткий виклад теоретичних відомостей (нагадування студентам потрібної інформації з лекційного матеріалу, роз'яснення незрозумілих моментів тощо), детальний розгляд особливостей розв'язання задачі кожного типу, які супроводжуються поясненнями викладача. В кінці кожного заняття студенти отримують перелік номерів задач для домашнього завдання. На наступному занятті студенти мають можливість отримати від викладача відповіді на питання, що виникли при їх розв'язанні. Після опанування теми кожний студент самостійно розв'язує типові задачі, маючи при цьому можливість отримати консультацію викладача. Ці задачі оформлюються та здаються у вигляді самостійних робіт.

Під час вивчення матеріалу застосовуються інформаційно-комунікаційні технології, що забезпечують проблемно-дослідницький характер процесу навчання та активізацію самостійної роботи студентів (електронні презентації для лекційних занять, використання аудіо-, відео-підтримки навчальних занять), доповнення традиційних навчальних занять засобами взаємодії на основі мережевих комунікаційних можливостей (інтернет-лекції, інтернет-семінари під час дистанційного навчання).

ПОЛІТИКА ДИСЦИПЛІНИ

Порушення термінів виконання завдань та заохочувальні бали:

Ключовими заходами при викладанні кредитного модуля є ті, які формують семестровий рейтинг студента. Тому студенти мають своєчасно виконувати завдання на практичних заняттях, писати модульну контрольну роботу у середині викладення курсу.

Штрафні бали з кредитного модуля передбачено за порушення термінів здачі розрахункової роботи (-5 штрафних бали за запізнення).

Заохочувальні бали студент може отримати за поглиблене вивчення окремих тем курсу, що може бути представлене у вигляді наукових тез, наукової статті, есе, презентації тощо, а також за активну участь у дискусіях на практичних та лекційних заняттях.

Відвідування занять та поведінка на заняттях:

Відвідування занять є вільним, бали за присутність на лекції не додаються, і штрафні бали за пропуски занять не передбачено. Втім, вагома частина рейтингу студента формується через активну участь у заходах на практичних заняттях, а саме у вирішенні завдань, груповій та індивідуальній роботі. Тому пропуск практичного заняття не дає можливість отримати студенту бали у семестровий рейтинг.

На заняттях студенту дозволяється користуватись інтерактивними засобами навчання, в т. ч. виходити в інтернет із метою пошуку навчальної або довідкової інформації, якщо це передбачено тематикою завдання. Активність студента на парах, його готовність до дискусій та участь в обговоренні навчальних питань може бути оцінена заохочувальними балами на розсуд викладача.

Захист індивідуального семестрового завдання передбачено у вигляді стислої доповіді за виконаним завданням, та відповідей на запитання. За форс-мажорних обставин, що зумовили нестачу часу, індивідуальне семестрове завдання зараховується за результатами представленої готової роботи та із урахуванням відповідей на запитання викладача щодо виконаної роботи, поставлених в індивідуальному порядку.

Пропущені контрольні заходи:

Якщо контрольні заходи пропущені з поважних причин (хвороба або вагомі життєві обставини), студенту надається можливість додатково скласти контрольне завдання протягом найближчого тижня. В разі порушення термінів і невиконання завдання з неповажних причин, студент не допускається до складання екзамену в основну сесію.

Політика щодо академічної доброчесності докладно описано у Кодексі Честі КПІ ім. Ігоря Сікорського. Це передбачає, що студент бере повну відповідальність за те, що всі виконані ним завдання відповідають принципам академічної доброчесності.

РЕЙТИНГОВА СИСТЕМА ОЦІНЮВАННЯ РЕЗУЛЬТАТІВ НАВЧАННЯ (РСО)

Оцінювання ґрунтується на застосуванні рейтингової системи, яка передбачає систематичну роботу студентів протягом семестру і складається з наступних заходів:

1. Рейтинг студента з кредитного модуля розраховується з 100 балів, з них 50 балів складає стартова шкала і 50 балів студент отримує за екзаменаційну контрольну роботу. Стартовий рейтинг (протягом семестру) складається з балів, що студент отримує за:
- роботу на практичних заняттях: три письмових самостійних роботи, які виконуються на 18 практичних заняттях;
- дві контрольні роботи (одна МКР поділяється на дві контрольні роботи тривалістю по одній академічній годині);
- одну розрахункову роботу.
-

2. Критерії нарахування балів:

2.1. Робота на практичних заняттях

На практичних заняттях студенти самостійно розв'язують та здають типові задачі. Ваговий бал rпз = 5 балів. Усього здається три самостійні роботи. Максимальна кількість балів на практичних заняттях дорівнює 5 х 3 = 15 балів.

Критерії оцінювання:

– «відмінно», повне виконання (не менше 95% потрібної інформації), – 4,75-5 балів;
– «дуже добре», майже повне виконання (не менше 85% потрібної інформації) – 4,25-4,74 бали;
– «добре», достатньо повне виконання (не менше 75% потрібної інформації) – 3,75-4,24 бали;
– «задовільно», робота виконана частково (не менше 65% потрібної інформації) – 3,25-3,74 бали;
– «достатньо», робота задовольняє мінімальним вимогам (не менше 60% потрібної інформації) – 3-3,24 бали;
– «незадовільно», робота не задовольняє вимогам або роботи немає – 0 балів.

2.2. Модульний контроль

Ваговий бал rмкр = 10 балів. Усього проводиться дві контрольні роботи. Максимальна кількість балів дорівнює 10 х 2 = 20 балів.

Критерії оцінювання:

– «відмінно», повна відповідь (не менше 95% потрібної інформації) – 9,5-10 балів;
– «дуже добре», майже повна відповідь (не менше 85% потрібної інформації) – 8,5-9,4 бали;

– «добре», достатньо повна відповідь (не менше 75% потрібної інформації) – 7,5-8,4 балів;
– «задовільно», неповна відповідь (не менше 65% потрібної інформації) – 6,5-7,4 балів;
– «достатньо», відповідь задовольняє мінімальним вимогам (не менше 60% потрібної інформації) – 6 -6,4 балів;
– «незадовільно», відповідь не задовольняє вимогам або відповіді немає – 0 балів.

2.3. Розрахункова робота

Ваговий бал rрр = 15 балів.

Критерії оцінювання:

– «відмінно», виконані всі вимоги до роботи, студент вільно орієнтується у наведених розрахунках – 14,25-15 балів;
–«дуже добре», майже повне виконання (не менше 85% потрібної інформації) або студент недостатньо орієнтується у наведених розрахунках – 12,75-14,24 балів;

– «добре», достатньо повне виконання (не менше 75% потрібної інформації) або студент недостатньо орієнтується у наведених розрахунках – 11,25-12,74 балів;
– «задовільно», робота виконана частково (не менше 65% потрібної інформації) – 9,75-11,24 балів;
– «достатньо», робота задовольняє мінімальним вимогам (не менше 60% потрібної інформації) – 9-9,74 бали;
– «незадовільно», робота не задовольняє вимогам або роботи немає – 0 балів.
За копіювання чужої роботи студент отримує штраф до - 5 балів. Студенти, які беруть активну участь у роботі на практичних заняттях, пропонують нестандартні підходи до вирішення задач, виконують додаткові завдання отримують заохочувальні бали (до 5 балів).

Розрахунок шкали (R) рейтингу
Рейтингова шкала з дисципліни складає:
RC + RE = 50 + 50 = 100 балів.

Максимальна сума балів стартової складової дорівнює 50:
RC = rпз x 3 + rмкр x 2 + rрр = 5 x 3 + 10 x 2 + 15 = 50 балів.

3. Календарний контроль проводиться у вигляді двох атестацій. Умовою першої атестації є поточний рейтинг не менше 8 балів. Умовою другої атестації – отримання не менше 18 балів.

4. Семестровий контроль проводиться у вигляді екзамену.

5. Умовою допуску до екзамену є здача розрахункової роботи не менше, ніж на 9 балів.

6. На екзамені студенти відповідають на два теоретичних питання, кожне з яких оцінюється у 10 балів, та розв'язують 10 коротких задач, кожна оцінюється у 3 бали.
Система оцінювання теоретичних питань:
– «відмінно», повна відповідь (не менше 95% потрібної інформації) – 9,5-10 балів;
– «дуже добре», майже повна відповідь (не менше 85% потрібної інформації) – 8,5-9,4 бали;
– «добре», достатньо повна відповідь (не менше 75% потрібної інформації) – 7,5-8,4 балів;
– «задовільно», неповна відповідь (не менше 65% потрібної інформації) – 6,5-7,4 балів;
– «достатньо», відповідь задовольняє мінімальним вимогам (не менше 60% потрібної інформації) – 6 -6,4 балів;
– «незадовільно», відповідь не задовольняє вимогам або відповіді немає – 0 балів.
Система оцінювання задач:
– «відмінно», задача розв'язана вірно, наведені детальні пояснення (не менше 95% потрібної інформації) – 2,85-3 балів;
– «дуже добре», майже повне виконання (не менше 85% потрібної інформації) – 2,55-2,84 балів;
– «добре», розв'язок задачі вірний, але є незначні помилки у розрахунках та/або поясненнях (не менше 75% потрібної інформації) – 2,25-2,54 балів;
– «задовільно», розв'язок задачі неповний, є помилки у розрахунках та/або поясненнях (не менше 65% потрібної інформації) – 1,95-2,24 балів;
– «достатньо», розв'язання задачі задовольняє мінімальним вимогам (не менше 60% потрібної інформації) – 1,8-1,94 бали;
– «незадовільно», розв'язок невірний або відсутній – 0 балів.

7. Сума стартових балів і балів за екзамен переводиться до екзаменаційної оцінки згідно з таблицею:

Бали: практичні заняття + МКР + РР + + екзаменаційна контрольна робота	Оцінка
100…95	Відмінно
94…85	Дуже добре
84…75	Добре
74…65	Задовільно
64…60	Достатньо
Менше 60	Незадовільно
Не зарахована розрахункова робота	Не допущено

КОНТРОЛЬНІ ПИТАННЯ

1. Дані, інформація, вимірювання. У чому різниця між даними та інформацією
2. Дані, інформація, вимірювання. Тріада генерування інформації
3. Дані, інформація, вимірювання. Якісні дані
4. Дані, інформація, вимірювання. Кількісні дані
5. Дані, інформація, вимірювання. Номінальна шкала вимірювання
6. Дані, інформація, вимірювання. Порядкова шкала вимірювання
7. Дані, інформація, вимірювання. Особливості маркетингових даних
8. Дані, інформація, вимірювання. Завдання інформаційного генезису
9. Статистичне спостереження. Суцільне спостереження. Особливості аналізу даних
10. Статистичне спостереження. Монографічне спостереження. Особливості аналізу даних
11. Статистичне спостереження. Метод основного масиву. Особливості аналізу даних
12. Статистичне спостереження. Вибіркове спостереження. Особливості аналізу даних
13. Дескриптивна статистика. Особливості точкових оцінок
14. Дескриптивна статистика. Оцінка моди
15. Дескриптивна статистика. Оцінка медіани
16. Дескриптивна статистика. Оцінка математичного сподівання
17. Дескриптивна статистика. Оцінка квантилів
18. Дескриптивна статистика. Оцінка дисперсії
19. Дескриптивна статистика. Оцінка коефіцієнту варіації
20. Дескриптивна статистика. Оцінка асиметрії
21. Дескриптивна статистика. Оцінка ексцесу
22. Дескриптивна статистика. Полігон частот
23. Дескриптивна статистика. Гістограма частот
24. Дескриптивна статистика. Емпірична функція розподілу
25. Дескриптивна статистика. Варіаційний ряд
26. Дескриптивна статистика. Статистичний ряд
27. Дескриптивна статистика. Групований статистичний ряд
28. Дескриптивна статистика. Якість оцінок: Незсуненість
29. Дескриптивна статистика. Якість оцінок: Конзистентність
30. Дескриптивна статистика. Якість оцінок: Ефективність
31. Інтервальне оцінювання. Визначення довірчого інтервалу
32. Інтервальне оцінювання. Особливості, переваги та недоліки інтервальних оцінок
33. Інтервальне оцінювання. Доцільність використання
34. Квантілі розподілів, які використовуються в індуктивній статистиці. Стандартний нормальний розподіл
35. Квантілі розподілів, які використовуються в індуктивній статистиці. Хі-квадрат розподіл
36. Квантілі розподілів, які використовуються в індуктивній статистиці. Розподіл Стьюдента
37. Квантілі розподілів, які використовуються в індуктивній статистиці. Розподіл Фішера
38. Перевірка статистичних гіпотез. Основні етапи
39. Перевірка статистичних гіпотез. Що таке статистична гіпотеза
40. Перевірка статистичних гіпотез. Які бувають статистичні гіпотези
41. Перевірка статистичних гіпотез. Критерій перевірки статистичних гіпотез

42. Перевірка статистичних гіпотез. Статистика критерію перевірки гіпотез
43. Перевірка статистичних гіпотез. Область прийняття рішення
44. Перевірка статистичних гіпотез. Критична область
45. Перевірка статистичних гіпотез. Помилка I-го роду
46. Перевірка статистичних гіпотез. Помилка II-го роду
47. Перевірка статистичних гіпотез. Ймовірність помилки I-го роду
48. Перевірка статистичних гіпотез. Ймовірність помилки II-го роду
49. Перевірка статистичних гіпотез. Потужність критерія
50. Перевірка статистичних гіпотез. Прийняття статистичного рішення
51. Перевірка статистичних гіпотез. Двосторонній, правосторонній, лівосторонній критерії
52. Перевірка статистичних гіпотез. Статистичні тести, їх відмінність від процесу перевірки гіпотез
53. Перевірка статистичних гіпотез. Значущість статистичного тесту
54. Перевірка статистичних гіпотез. Одновибірковий T-тест
55. Перевірка статистичних гіпотез. Двохвибірковий T-тест для незалежних вибірок
56. Перевірка статистичних гіпотез. Двохвибірковий F-тест
57. Перевірка статистичних гіпотез. Однофакторний дисперсійний аналіз. Сутність
58. Перевірка статистичних гіпотез. Однофакторний дисперсійний аналіз. Внутрішньогрупова варіація
59. Перевірка статистичних гіпотез. Однофакторний дисперсійний аналіз. Міжгрупова варіація
60. Перевірка статистичних гіпотез. Однофакторний дисперсійний аналіз. F-статистика
61. Перевірка статистичних гіпотез. Однофакторний дисперсійний аналіз. Прийняття рішення
62. Непараметричні критерії перевірки статистичних гіпотез. Критерії згоди. Сутність. Приклади
63. Непараметричні критерії перевірки статистичних гіпотез. Критерії однорідності. Сутність. Приклади
64. Непараметричні критерії перевірки статистичних гіпотез. Одновибірковий критерій Колмогорова-Смірнова. Сутність. Особливості побудови статистики
65. Непараметричні критерії перевірки статистичних гіпотез. Двохвибірковий критерій Колмогорова-Смірнова. Сутність. Особливості побудови статистики
66. Непараметричні критерії перевірки статистичних гіпотез. Критерій згоди Хі-квадрат. Сутність. Особливості побудови статистики
67. Непараметричні критерії перевірки статистичних гіпотез. Критерій однорідності Хі-квадрат з одним ступенем свободи. Сутність. Особливості побудови статистики
68. Непараметричні критерії перевірки статистичних гіпотез. Критерій однорідності Хі-квадрат зі ступенями свободи > 1. Сутність. Особливості побудови статистики
69. Непараметричні критерії перевірки статистичних гіпотез. Критерій однорідності Хі-квадрат зі ступенями свободи > 1. Розрахунок очікуваних частот
70. Непараметричні критерії перевірки статистичних гіпотез. Біноміальний критерій. Особливості використання
71. Непараметричні критерії перевірки статистичних гіпотез. Критерій точної ймовірності Фішера. Особливості використання
72. Кореляційний аналіз. Кореляційний аналіз для даних, що вимірюються за номінальною шкалою
73. Кореляційний аналіз. Кореляційний аналіз для даних, що вимірюються за порядковою шкалою. Парна кореляція

74. Кореляційний аналіз. Кореляційний аналіз для даних, що вимірюються за порядковою шкалою. Множинна кореляція
75. Кореляційний аналіз. Розрахунок рангів
76. Кореляційний аналіз. Коефіцієнт кореляції Пірсона. Формула
77. Кореляційний аналіз. Коефіцієнт кореляції Пірсона. Властивості
78. Кореляційний аналіз. Коефіцієнт кореляції Пірсона. Перевірка значущості
79. Кореляційний аналіз. Частинний коефіцієнт кореляції. Сутність, особливості використання
80. Кореляційний аналіз. Множинний коефіцієнт кореляції. Сутність, особливості використання
81. Кореляційний аналіз. Коефіцієнт детермінації. Сутність, особливості використання
82. Регресійний аналіз. Види моделей
83. Регресійний аналіз. Парна лінійна регресія. Особливості використання
84. Регресійний аналіз. Парна лінійна регресія. Що показують коефіцієнти регресії
85. Регресійний аналіз. Сутність перевірки значущості коефіцієнтів регресії
86. Регресійний аналіз. Сутність перевірки значущості рівняння регресії
87. Регресійний аналіз. Ефект мультиколінеарності
88. Регресійний аналіз. Сутність коефіцієнту детермінації
89. Регресійний аналіз. Сутність перевірки незалежності залишків та нормальності їх розподілу
90. Регресійний аналіз. Умова гетероскедастичності
91. Кластерний аналіз. Основні етапи
92. Кластерний аналіз. Сфери застосування в маркетингу
93. Кластерний аналіз. Міри близькості (зв'язку) об'єктів
94. Кластерний аналіз. Міри відстані
95. Кластерний аналіз. Евклідова відстань між об'єктами
96. Кластерний аналіз. Лінійна (сіті-блок) відстань між об'єктами
97. Кластерний аналіз. Коефіцієнти подібності
98. Кластерний аналіз. Особливості використання ієрархічного кластерного аналізу
99. Кластерний аналіз. Методи агломерації
100. Кластерний аналіз. Особливості використання методу К-середніх

НАВЧАЛЬНІ МАТЕРІАЛИ ТА РЕСУРСИ

Основна література:

1. Статистика: Навчально-методичний комплекс [Електронний ресурс] : навч. посіб. для здоб. ступ. бакалавра за спец. 075 «Маркетинг» / уклад.: О. В. Черненко. Електронні текстові дані (1 файл: 21,5 Мбайт). Київ : КПІ ім. Ігоря Сікорського, 2020. 135 с.
2. Кремер Н. Ш. Теория вероятностей и математическая статистика. Учебник. Юнити-Дата, 2012. // [Електронне видання] - Режим доступу : https://may.alleng.org/d/math/math328.htm.
3. Барковський В.В., Барковська Н. В., Лопатін О. К. Теорія ймовірності та математична статистика. 6-е видання. Київ : Центр учбової літератури, 2016. 424 с.
4. Пушак Я.С., Лозовий Б.Н. Теорія ймовірностей і елементи математичної статистики. Навчальний посібник. Магнолія 2006, 2018.
5. Теорія ймовірностей та математична статистика: навчальний посібник / О. І. Огірко, Н. В. Галайко. Львів: ЛьвДУВС, 2017. 292 с.

Додаткова література:

1. Маркетингове забезпечення інноваційних процесів промислових підприємств : монографія / Є. В. Гнітецький та ін. Київ : КПІ ім. Ігоря Сікорського, 2017. 166 с.
2. Солнцев С. О., Москаленко О. Д., Черненко О. В. Система моніторингу маркетингового середовища підприємства. Економічний вісник НТУУ «КПІ». 2018. №15. С. 341–354.
3. Черненко О. В. Архітектоніка маркетингової інформаційної системи в умовах інформаційно-комунікативного середовища. Бізнес-Інформ. 2016. № 11. С. 433–440.
4. Черненко О. В. Маркетингова інформація в управлінні підприємством. Економічний вісник НТУУ «КПІ». 2017. №14. С. 369–374.
5. Solntsev S., Chernenko O. The use of modern information and communication technologies by Ukrainian enterprises-producers of domestic boilers. Economic&Edication. Internation Scientific Journal. ISMA University, Riga, 2018. Vol.3, Issue 1. Pp.47–53.
6. Айвазян С. А., Мхитарян В. С. Прикладная статистика и основы эконометрики. Учебник для вузов. М.: ЮНИТИ. 1998.
7. Гихман И. И., Скороход А. В., Ядренко М. Й. Теория вероятностей и математическая статистика. К.: Выща шк. 1988. 440 с.
8. Гмурман В. С. Теория вероятностей и математическая статистика. М.: Высш.шк. 1972. 368 с.
9. Колемаев В.А., Калинина В.Н. Теория вероятностей и математическая статистика. М.: Высш. Шк.: ИНФРА-М. 1997. 302 с.
10. Сборник задач по математике для втузов. Специальные курсы. М.: Наука. Главная редакция физико-математической литературы, 1984. 608 с.

ДОДАТКИ

Квантилі стандартного нормального розподілу N (0,1)

p	0,9	0,95	0,975	0,99	0,995	0,999	0,9995
u_p	1,282	1,645	1,960	2,326	2,576	3,090	3,290

Квантилі Хі-квадрат розподілу $\chi^2(k)$

k	p										
	0,005	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99	0,995	0,999
1	0,000	0,000	0,001	0,004	0,016	2,71	3,84	5,02	6,63	7,88	10,83
2	0,01	0,02	0,05	0,10	0,21	4,61	5,99	7,38	9,21	10,60	13,82
3	0,07	0,11	0,22	0,35	0,58	6,25	7,81	9,35	11,34	12,84	16,27
4	0,21	0,30	0,48	0,71	1,06	7,78	9,49	11,14	13,28	14,86	18,47
5	0,41	0,55	0,83	1,15	1,61	9,24	11,07	12,83	15,09	16,75	20,51
6	0,68	0,87	1,24	1,64	2,20	10,64	12,59	14,45	16,81	18,55	22,46
7	0,99	1,24	1,69	2,17	2,83	12,02	14,07	16,01	18,48	20,28	24,32
8	1,34	1,65	2,18	2,73	3,49	13,36	15,51	17,53	20,09	21,95	26,12
9	1,73	2,09	2,70	3,33	4,17	14,68	16,92	19,02	21,67	23,59	27,88
10	2,16	2,56	3,25	3,94	4,87	15,99	18,31	20,48	23,21	25,19	29,59
11	2,60	3,05	3,82	4,57	5,58	17,28	19,68	21,92	24,73	26,76	31,26
12	3,07	3,57	4,40	5,23	6,30	18,55	21,03	23,34	26,22	28,30	32,91
13	3,57	4,11	5,01	5,89	7,04	19,81	22,36	24,74	27,69	29,82	34,53
14	4,07	4,66	5,63	6,57	7,79	21,06	23,68	26,12	29,14	31,32	36,12
15	4,60	5,23	6,26	7,26	8,55	22,31	25,00	27,49	30,58	32,80	37,70
16	5,14	5,81	6,91	7,96	9,31	23,54	26,30	28,85	32,00	34,27	39,25
17	5,70	6,41	7,56	8,67	10,09	24,77	27,59	30,19	33,41	35,72	40,79
18	6,26	7,01	8,23	9,39	10,86	25,99	28,87	31,53	34,81	37,16	42,31
19	6,84	7,63	8,91	10,12	11,65	27,20	30,14	32,85	36,19	38,58	43,82
20	7,43	8,26	9,59	10,85	12,44	28,41	31,41	34,17	37,57	40,00	45,31
21	8,03	8,90	10,28	11,59	13,24	29,62	32,67	35,48	38,93	41,40	46,80
22	8,64	9,54	10,98	12,34	14,04	30,81	33,92	36,78	40,29	42,80	48,27
23	9,26	10,20	11,69	13,09	14,85	32,01	35,17	38,08	41,64	44,18	49,73
24	9,89	10,86	12,40	13,85	15,66	33,20	36,42	39,36	42,98	45,56	51,18
26	11,16	12,20	13,84	15,38	17,29	35,56	38,89	41,92	45,64	48,29	54,05
28	12,46	13,56	15,31	16,93	18,94	37,92	41,34	44,46	48,28	50,99	56,89
30	13,79	14,95	16,79	18,49	20,60	40,26	43,77	46,98	50,89	53,67	59,70
35	17,19	18,51	20,57	22,47	24,80	46,06	49,80	53,20	57,34	60,27	66,62
40	20,71	22,16	24,43	26,51	29,05	51,81	55,76	59,34	63,69	66,77	73,40
45	24,31	25,90	28,37	30,61	33,35	57,51	61,66	65,41	69,96	73,17	80,08
50	27,99	29,71	32,36	34,76	37,69	63,17	67,50	71,42	76,15	79,49	86,66
75	47,21	49,48	52,94	56,05	59,79	91,06	96,22	100,8	106,4	110,3	118,6
100	67,33	70,06	74,22	77,93	82,36	118,5	124,3	129,6	135,8	140,2	149,4

Квантилі розподілу Стьюдента T(k)

k	p							
	0,75	0,9	0,95	0,975	0,99	0,995	0,995	0,999
1	1,000	3,078	6,314	12,706	31,821	63,656	63,656	318,29
2	0,816	1,886	2,920	4,303	6,965	9,925	9,925	22,328
3	0,765	1,638	2,353	3,182	4,541	5,841	5,841	10,214
4	0,741	1,533	2,132	2,776	3,747	4,604	4,604	7,173
5	0,727	1,476	2,015	2,571	3,365	4,032	4,032	5,894
6	0,718	1,440	1,943	2,447	3,143	3,707	3,707	5,208
7	0,711	1,415	1,895	2,365	2,998	3,499	3,499	4,785
8	0,706	1,397	1,860	2,306	2,896	3,355	3,355	4,501
9	0,703	1,383	1,833	2,262	2,821	3,250	3,250	4,297
10	0,700	1,372	1,812	2,228	2,764	3,169	3,169	4,144
11	0,697	1,363	1,796	2,201	2,718	3,106	3,106	4,025
12	0,695	1,356	1,782	2,179	2,681	3,055	3,055	3,930
13	0,694	1,350	1,771	2,160	2,650	3,012	3,012	3,852
14	0,692	1,345	1,761	2,145	2,624	2,977	2,977	3,787
15	0,691	1,341	1,753	2,131	2,602	2,947	2,947	3,733
16	0,690	1,337	1,746	2,120	2,583	2,921	2,921	3,686
17	0,689	1,333	1,740	2,110	2,567	2,898	2,898	3,646
18	0,688	1,330	1,734	2,101	2,552	2,878	2,878	3,610
19	0,688	1,328	1,729	2,093	2,539	2,861	2,861	3,579
20	0,687	1,325	1,725	2,086	2,528	2,845	2,845	3,552
21	0,686	1,323	1,721	2,080	2,518	2,831	2,831	3,527
22	0,686	1,321	1,717	2,074	2,508	2,819	2,819	3,505
23	0,685	1,319	1,714	2,069	2,500	2,807	2,807	3,485
24	0,685	1,318	1,711	2,064	2,492	2,797	2,797	3,467
25	0,684	1,316	1,708	2,060	2,485	2,787	2,787	3,450
26	0,684	1,315	1,706	2,056	2,479	2,779	2,779	3,435
27	0,684	1,314	1,703	2,052	2,473	2,771	2,771	3,421
28	0,683	1,313	1,701	2,048	2,467	2,763	2,763	3,408
30	0,683	1,310	1,697	2,042	2,457	2,750	2,750	3,385
40	0,681	1,303	1,684	2,021	2,423	2,704	2,704	3,307
60	0,679	1,296	1,671	2,000	2,390	2,660	2,660	3,232
120	0,677	1,289	1,658	1,980	2,358	2,617	2,617	3,160
∞	0,674	1,282	1,645	1,960	2,326	2,576	2,576	3,090

Квантилі розподілу Фішера F(k,n)

k	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120
n	p = 0,9																
1	40	50	54	56	57	58	59	59	60	60	61	61	62	62	63	63	63
2	8,5	9,0	9,2	9,2	9,3	9,3	9,3	9,4	9,4	9,4	9,4	9,4	9,4	9,5	9,5	9,5	9,5
3	5,5	5,5	5,4	5,3	5,3	5,3	5,3	5,3	5,2	5,2	5,2	5,2	5,2	5,2	5,2	5,2	5,1
4	4,5	4,3	4,2	4,1	4,1	4,0	4,0	4,0	3,9	3,9	3,9	3,9	3,8	3,8	3,8	3,8	3,8
5	4,1	3,8	3,6	3,5	3,5	3,4	3,4	3,3	3,3	3,3	3,3	3,2	3,2	3,2	3,2	3,1	3,1
6	3,8	3,5	3,3	3,2	3,1	3,1	3,0	3,0	3,0	2,9	2,9	2,9	2,8	2,8	2,8	2,8	2,7
7	3,6	3,3	3,1	3,0	2,9	2,8	2,8	2,8	2,7	2,7	2,7	2,6	2,6	2,6	2,5	2,5	2,5
8	3,5	3,1	2,9	2,8	2,7	2,7	2,6	2,6	2,6	2,5	2,5	2,5	2,4	2,4	2,4	2,3	2,3
9	3,4	3,0	2,8	2,7	2,6	2,6	2,5	2,5	2,4	2,4	2,4	2,3	2,3	2,3	2,2	2,2	2,2
10	3,3	2,9	2,7	2,6	2,5	2,5	2,4	2,4	2,3	2,3	2,3	2,2	2,2	2,2	2,1	2,1	2,1
11	3,2	2,9	2,7	2,5	2,5	2,4	2,3	2,3	2,3	2,2	2,2	2,2	2,1	2,1	2,1	2,0	2,0
12	3,2	2,8	2,6	2,5	2,4	2,3	2,3	2,2	2,2	2,2	2,1	2,1	2,1	2,0	2,0	2,0	1,9
13	3,1	2,8	2,6	2,4	2,3	2,3	2,2	2,2	2,2	2,1	2,1	2,1	2,0	2,0	1,9	1,9	1,9
14	3,1	2,7	2,5	2,4	2,3	2,2	2,2	2,2	2,1	2,1	2,1	2,0	2,0	1,9	1,9	1,9	1,8
15	3,1	2,7	2,5	2,4	2,3	2,2	2,2	2,1	2,1	2,1	2,0	2,0	1,9	1,9	1,8	1,8	1,8
16	3,0	2,7	2,5	2,3	2,2	2,2	2,1	2,1	2,1	2,0	2,0	1,9	1,9	1,8	1,8	1,8	1,8
17	3,0	2,6	2,4	2,3	2,2	2,2	2,1	2,1	2,0	2,0	2,0	1,9	1,9	1,8	1,8	1,8	1,7
18	3,0	2,6	2,4	2,3	2,2	2,1	2,1	2,0	2,0	2,0	1,9	1,9	1,8	1,8	1,8	1,7	1,7
19	3,0	2,6	2,4	2,3	2,2	2,1	2,1	2,0	2,0	2,0	1,9	1,9	1,8	1,8	1,7	1,7	1,7
20	3,0	2,6	2,4	2,2	2,2	2,1	2,0	2,0	2,0	1,9	1,9	1,8	1,8	1,7	1,7	1,7	1,6
21	3,0	2,6	2,4	2,2	2,1	2,1	2,0	2,0	1,9	1,9	1,9	1,8	1,8	1,7	1,7	1,7	1,6
22	2,9	2,6	2,4	2,2	2,1	2,1	2,0	2,0	1,9	1,9	1,9	1,8	1,8	1,7	1,7	1,6	1,6
23	2,9	2,5	2,3	2,2	2,1	2,0	2,0	2,0	1,9	1,9	1,8	1,8	1,7	1,7	1,7	1,6	1,6
24	2,9	2,5	2,3	2,2	2,1	2,0	2,0	1,9	1,9	1,9	1,8	1,8	1,7	1,7	1,6	1,6	1,6
25	2,9	2,5	2,3	2,2	2,1	2,0	2,0	1,9	1,9	1,9	1,8	1,8	1,7	1,7	1,6	1,6	1,6
26	2,9	2,5	2,3	2,2	2,1	2,0	2,0	1,9	1,9	1,9	1,8	1,8	1,7	1,6	1,6	1,6	1,5
27	2,9	2,5	2,3	2,2	2,1	2,0	2,0	1,9	1,9	1,8	1,8	1,7	1,7	1,6	1,6	1,6	1,5
28	2,9	2,5	2,3	2,2	2,1	2,0	1,9	1,9	1,9	1,8	1,8	1,7	1,7	1,6	1,6	1,6	1,5
30	2,9	2,5	2,3	2,1	2,0	2,0	1,9	1,9	1,8	1,8	1,8	1,7	1,7	1,6	1,6	1,5	1,5
40	2,8	2,4	2,2	2,1	2,0	1,9	1,9	1,8	1,8	1,8	1,7	1,7	1,6	1,5	1,5	1,5	1,4
60	2,8	2,4	2,2	2,0	1,9	1,9	1,8	1,8	1,7	1,7	1,7	1,6	1,5	1,5	1,4	1,4	1,3
120	2,7	2,3	2,1	2,0	1,9	1,8	1,8	1,7	1,7	1,7	1,6	1,5	1,5	1,4	1,4	1,3	1,3
∞	2,7	2,3	2,1	1,9	1,8	1,8	1,7	1,7	1,6	1,6	1,5	1,5	1,4	1,3	1,3	1,2	1,2

Квантилі розподілу Фішера F(k,n)

k	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120
n	p = 0,95																
1	161	199	216	225	230	234	237	239	241	242	244	246	248	250	251	252	253
2	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19
3	10	9,6	9,3	9,1	9,0	8,9	8,9	8,8	8,8	8,8	8,7	8,7	8,7	8,6	8,6	8,6	8,5
4	7,7	6,9	6,6	6,4	6,3	6,2	6,1	6,0	6,0	6,0	5,9	5,9	5,8	5,7	5,7	5,7	5,7
5	6,6	5,8	5,4	5,2	5,1	5,0	4,9	4,8	4,8	4,7	4,7	4,6	4,6	4,5	4,5	4,4	4,4
6	6,0	5,1	4,8	4,5	4,4	4,3	4,2	4,1	4,1	4,1	4,0	3,9	3,9	3,8	3,8	3,7	3,7
7	5,6	4,7	4,3	4,1	4,0	3,9	3,8	3,7	3,7	3,6	3,6	3,5	3,4	3,4	3,3	3,3	3,3
8	5,3	4,5	4,1	3,8	3,7	3,6	3,5	3,4	3,4	3,3	3,3	3,2	3,2	3,1	3,0	3,0	3,0
9	5,1	4,3	3,9	3,6	3,5	3,4	3,3	3,2	3,2	3,1	3,1	3,0	2,9	2,9	2,8	2,8	2,7
10	5,0	4,1	3,7	3,5	3,3	3,2	3,1	3,1	3,0	3,0	2,9	2,8	2,8	2,7	2,7	2,6	2,6
11	4,8	4,0	3,6	3,4	3,2	3,1	3,0	2,9	2,9	2,9	2,8	2,7	2,6	2,6	2,5	2,5	2,4
12	4,7	3,9	3,5	3,3	3,1	3,0	2,9	2,8	2,8	2,8	2,7	2,6	2,5	2,5	2,4	2,4	2,3
13	4,7	3,8	3,4	3,2	3,0	2,9	2,8	2,8	2,7	2,7	2,6	2,5	2,5	2,4	2,3	2,3	2,3
14	4,6	3,7	3,3	3,1	3,0	2,8	2,8	2,7	2,6	2,6	2,5	2,5	2,4	2,3	2,3	2,2	2,2
15	4,5	3,7	3,3	3,1	2,9	2,8	2,7	2,6	2,6	2,5	2,5	2,4	2,3	2,2	2,2	2,2	2,1
16	4,5	3,6	3,2	3,0	2,9	2,7	2,7	2,6	2,5	2,5	2,4	2,4	2,3	2,2	2,2	2,1	2,1
17	4,5	3,6	3,2	3,0	2,8	2,7	2,6	2,5	2,5	2,4	2,4	2,3	2,2	2,1	2,1	2,1	2,0
18	4,4	3,6	3,2	2,9	2,8	2,7	2,6	2,5	2,5	2,4	2,3	2,3	2,2	2,1	2,1	2,0	2,0
19	4,4	3,5	3,1	2,9	2,7	2,6	2,5	2,5	2,4	2,4	2,3	2,2	2,2	2,1	2,0	2,0	1,9
20	4,4	3,5	3,1	2,9	2,7	2,6	2,5	2,4	2,4	2,3	2,3	2,2	2,1	2,0	2,0	1,9	1,9
21	4,3	3,5	3,1	2,8	2,7	2,6	2,5	2,4	2,4	2,3	2,3	2,2	2,1	2,0	2,0	1,9	1,9
22	4,3	3,4	3,0	2,8	2,7	2,5	2,5	2,4	2,3	2,3	2,2	2,2	2,1	2,0	1,9	1,9	1,8
23	4,3	3,4	3,0	2,8	2,6	2,5	2,4	2,4	2,3	2,3	2,2	2,1	2,0	2,0	1,9	1,9	1,8
24	4,3	3,4	3,0	2,8	2,6	2,5	2,4	2,4	2,3	2,3	2,2	2,1	2,0	1,9	1,9	1,8	1,8
25	4,2	3,4	3,0	2,8	2,6	2,5	2,4	2,3	2,3	2,2	2,2	2,1	2,0	1,9	1,9	1,8	1,8
26	4,2	3,4	3,0	2,7	2,6	2,5	2,4	2,3	2,3	2,2	2,1	2,1	2,0	1,9	1,9	1,8	1,7
27	4,2	3,4	3,0	2,7	2,6	2,5	2,4	2,3	2,3	2,2	2,1	2,1	2,0	1,9	1,8	1,8	1,7
28	4,2	3,3	2,9	2,7	2,6	2,4	2,4	2,3	2,2	2,2	2,1	2,0	2,0	1,9	1,8	1,8	1,7
30	4,2	3,3	2,9	2,7	2,5	2,4	2,3	2,3	2,2	2,2	2,1	2,0	1,9	1,8	1,8	1,7	1,7
40	4,1	3,2	2,8	2,6	2,4	2,3	2,2	2,2	2,1	2,1	2,0	1,9	1,8	1,7	1,7	1,6	1,6
60	4,0	3,2	2,8	2,5	2,4	2,3	2,2	2,1	2,0	2,0	1,9	1,8	1,7	1,6	1,6	1,5	1,5
120	3,9	3,1	2,7	2,4	2,3	2,2	2,1	2,0	2,0	1,9	1,8	1,8	1,7	1,6	1,5	1,4	1,4
∞	3,8	3,0	2,6	2,4	2,2	2,1	2,0	1,9	1,9	1,8	1,8	1,7	1,6	1,5	1,4	1,3	1,2

Квантилі розподілу Фішера F(k,n)

k	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120
n	p = 0,975																
1	648	799	864	900	922	937	948	957	963	969	977	985	993	1001	1006	1010	1014
2	39	39	39	39	39	39	39	39	39	39	39	39	39	39	39	39	39
3	17	16	15	15	15	15	15	15	14	14	14	14	14	14,1	14,0	14,0	13,9
4	12	11	10	9,6	9,4	9,2	9,1	9,0	8,9	8,8	8,8	8,7	8,6	8,5	8,4	8,4	8,3
5	10	8,4	7,8	7,4	7,1	7,0	6,9	6,8	6,7	6,6	6,5	6,4	6,3	6,2	6,2	6,1	6,1
6	8,8	7,3	6,6	6,2	6,0	5,8	5,7	5,6	5,5	5,5	5,4	5,3	5,2	5,1	5,0	5,0	4,9
7	8,1	6,5	5,9	5,5	5,3	5,1	5,0	4,9	4,8	4,8	4,7	4,6	4,5	4,4	4,3	4,3	4,2
8	7,6	6,1	5,4	5,1	4,8	4,7	4,5	4,4	4,4	4,3	4,2	4,1	4,0	3,9	3,8	3,8	3,7
9	7,2	5,7	5,1	4,7	4,5	4,3	4,2	4,1	4,0	4,0	3,9	3,8	3,7	3,6	3,5	3,4	3,4
10	6,9	5,5	4,8	4,5	4,2	4,1	3,9	3,9	3,8	3,7	3,6	3,5	3,4	3,3	3,3	3,2	3,1
11	6,7	5,3	4,6	4,3	4,0	3,9	3,8	3,7	3,6	3,5	3,4	3,3	3,2	3,1	3,1	3,0	2,9
12	6,6	5,1	4,5	4,1	3,9	3,7	3,6	3,5	3,4	3,4	3,3	3,2	3,1	3,0	2,9	2,8	2,8
13	6,4	5,0	4,3	4,0	3,8	3,6	3,5	3,4	3,3	3,2	3,2	3,1	2,9	2,8	2,8	2,7	2,7
14	6,3	4,9	4,2	3,9	3,7	3,5	3,4	3,3	3,2	3,1	3,1	2,9	2,8	2,7	2,7	2,6	2,6
15	6,2	4,8	4,2	3,8	3,6	3,4	3,3	3,2	3,1	3,1	3,0	2,9	2,8	2,6	2,6	2,5	2,5
16	6,1	4,7	4,1	3,7	3,5	3,3	3,2	3,1	3,0	3,0	2,9	2,8	2,7	2,6	2,5	2,4	2,4
17	6,0	4,6	4,0	3,7	3,4	3,3	3,2	3,1	3,0	2,9	2,8	2,7	2,6	2,5	2,4	2,4	2,3
18	6,0	4,6	4,0	3,6	3,4	3,2	3,1	3,0	2,9	2,9	2,8	2,7	2,6	2,4	2,4	2,3	2,3
19	5,9	4,5	3,9	3,6	3,3	3,2	3,1	3,0	2,9	2,8	2,7	2,6	2,5	2,4	2,3	2,3	2,2
20	5,9	4,5	3,9	3,5	3,3	3,1	3,0	2,9	2,8	2,8	2,7	2,6	2,5	2,3	2,3	2,2	2,2
21	5,8	4,4	3,8	3,5	3,3	3,1	3,0	2,9	2,8	2,7	2,6	2,5	2,4	2,3	2,2	2,2	2,1
22	5,8	4,4	3,8	3,4	3,2	3,1	2,9	2,8	2,8	2,7	2,6	2,5	2,4	2,3	2,2	2,1	2,1
23	5,7	4,3	3,8	3,4	3,2	3,0	2,9	2,8	2,7	2,7	2,6	2,5	2,4	2,2	2,2	2,1	2,0
24	5,7	4,3	3,7	3,4	3,2	3,0	2,9	2,8	2,7	2,6	2,5	2,4	2,3	2,2	2,1	2,1	2,0
25	5,7	4,3	3,7	3,4	3,1	3,0	2,8	2,8	2,7	2,6	2,5	2,4	2,3	2,2	2,1	2,1	2,0
26	5,7	4,3	3,7	3,3	3,1	2,9	2,8	2,7	2,7	2,6	2,5	2,4	2,3	2,2	2,1	2,0	2,0
27	5,6	4,2	3,6	3,3	3,1	2,9	2,8	2,7	2,6	2,6	2,5	2,4	2,3	2,1	2,1	2,0	1,9
28	5,6	4,2	3,6	3,3	3,1	2,9	2,8	2,7	2,6	2,5	2,4	2,3	2,2	2,1	2,0	2,0	1,9
30	5,6	4,2	3,6	3,2	3,0	2,9	2,7	2,7	2,6	2,5	2,4	2,3	2,2	2,1	2,0	1,9	1,9
40	5,4	4,1	3,5	3,1	2,9	2,7	2,6	2,5	2,5	2,4	2,3	2,2	2,1	1,9	1,9	1,8	1,7
60	5,3	3,9	3,3	3,0	2,8	2,6	2,5	2,4	2,3	2,3	2,2	2,1	1,9	1,8	1,7	1,7	1,6
120	5,2	3,8	3,2	2,9	2,7	2,5	2,4	2,3	2,2	2,2	2,1	1,9	1,8	1,7	1,6	1,5	1,4
∞	5,0	3,7	3,1	2,8	2,6	2,4	2,3	2,2	2,1	2,0	1,9	1,8	1,7	1,6	1,5	1,4	1,3

Квантилі розподілу Фішера F(k,n)

k	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120
n	p = 0,99																
1	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056	6107	6157	6209	6260	6286	6313	6340
2	98,5	99,0	99,2	99,3	99,3	99,3	99,4	99,4	99,4	99,4	99,4	99,4	99,4	99,5	99,5	99,5	99,5
3	34	31	29	29	28	28	28	27	27	27	27	27	27	26,5	26,4	26,3	26,2
4	21	18	17	16	16	15	15	15	15	15	14	14	14	13,8	13,7	13,7	13,6
5	16	13	12	11	11	11	10	10	10	10	9,9	9,7	9,6	9,4	9,3	9,2	9,1
6	14	11	9,8	9,1	8,7	8,5	8,3	8,1	8,0	7,9	7,7	7,6	7,4	7,2	7,1	7,1	7,0
7	12	9,5	8,5	7,8	7,5	7,2	7,0	6,8	6,7	6,6	6,5	6,3	6,2	6,0	5,9	5,8	5,7
8	11	8,6	7,6	7,0	6,6	6,4	6,2	6,0	5,9	5,8	5,7	5,5	5,4	5,2	5,1	5,0	4,9
9	11	8,0	7,0	6,4	6,1	5,8	5,6	5,5	5,4	5,3	5,1	5,0	4,8	4,6	4,6	4,5	4,4
10	10	7,6	6,6	6,0	5,6	5,4	5,2	5,1	4,9	4,8	4,7	4,6	4,4	4,2	4,2	4,1	4,0
11	9,6	7,2	6,2	5,7	5,3	5,1	4,9	4,7	4,6	4,5	4,4	4,3	4,1	3,9	3,9	3,8	3,7
12	9,3	6,9	6,0	5,4	5,1	4,8	4,6	4,5	4,4	4,3	4,2	4,0	3,9	3,7	3,6	3,5	3,4
13	9,1	6,7	5,7	5,2	4,9	4,6	4,4	4,3	4,2	4,1	4,0	3,8	3,7	3,5	3,4	3,3	3,3
14	8,9	6,5	5,6	5,0	4,7	4,5	4,3	4,1	4,0	3,9	3,8	3,7	3,5	3,3	3,3	3,2	3,1
15	8,7	6,4	5,4	4,9	4,6	4,3	4,1	4,0	3,9	3,8	3,7	3,5	3,4	3,2	3,1	3,0	3,0
16	8,5	6,2	5,3	4,8	4,4	4,2	4,0	3,9	3,8	3,7	3,6	3,4	3,3	3,1	3,0	2,9	2,8
17	8,4	6,1	5,2	4,7	4,3	4,1	3,9	3,8	3,7	3,6	3,5	3,3	3,2	3,0	2,9	2,8	2,7
18	8,3	6,0	5,1	4,6	4,2	4,0	3,8	3,7	3,6	3,5	3,4	3,2	3,1	2,9	2,8	2,7	2,7
19	8,2	5,9	5,0	4,5	4,2	3,9	3,8	3,6	3,5	3,4	3,3	3,2	3,0	2,8	2,8	2,7	2,6
20	8,1	5,8	4,9	4,4	4,1	3,9	3,7	3,6	3,5	3,4	3,2	3,1	2,9	2,8	2,7	2,6	2,5
21	8,0	5,8	4,9	4,4	4,0	3,8	3,6	3,5	3,4	3,3	3,2	3,0	2,9	2,7	2,6	2,5	2,5
22	7,9	5,7	4,8	4,3	4,0	3,8	3,6	3,5	3,3	3,3	3,1	3,0	2,8	2,7	2,6	2,5	2,4
23	7,9	5,7	4,8	4,3	3,9	3,7	3,5	3,4	3,3	3,2	3,1	2,9	2,8	2,6	2,5	2,4	2,4
24	7,8	5,6	4,7	4,2	3,9	3,7	3,5	3,4	3,3	3,2	3,0	2,9	2,7	2,6	2,5	2,4	2,3
25	7,8	5,6	4,7	4,2	3,9	3,6	3,5	3,3	3,2	3,1	3,0	2,9	2,7	2,5	2,5	2,4	2,3
26	7,7	5,5	4,6	4,1	3,8	3,6	3,4	3,3	3,2	3,1	3,0	2,8	2,7	2,5	2,4	2,3	2,2
27	7,7	5,5	4,6	4,1	3,8	3,6	3,4	3,3	3,1	3,1	2,9	2,8	2,6	2,5	2,4	2,3	2,2
28	7,6	5,5	4,6	4,1	3,8	3,5	3,4	3,2	3,1	3,0	2,9	2,8	2,6	2,4	2,4	2,3	2,2
30	7,6	5,4	4,5	4,0	3,7	3,5	3,3	3,2	3,1	3,0	2,8	2,7	2,5	2,4	2,3	2,2	2,1
40	7,3	5,2	4,3	3,8	3,5	3,3	3,1	3,0	2,9	2,8	2,7	2,5	2,4	2,2	2,1	2,0	1,9
60	7,1	5,0	4,1	3,6	3,3	3,1	3,0	2,8	2,7	2,6	2,5	2,4	2,2	2,0	1,9	1,8	1,7
120	6,9	4,8	3,9	3,5	3,2	3,0	2,8	2,7	2,6	2,5	2,3	2,2	2,0	1,9	1,8	1,7	1,5
∞	6,6	4,6	3,8	3,3	3,0	2,8	2,6	2,5	2,4	2,3	2,2	2,0	1,9	1,7	1,6	1,5	1,3

Довірчі інтервали для середнього нормально розподіленої генеральної сукупності

Вихідні дані: $\mathbf{x_1, x_2, \ldots, x_n}$ – вибірка
n – об'єм вибірки

1. Побудова довірчого інтервалу при відомій дисперсії генеральної сукупності:

Параметр, що оцінюється: Математичне сподівання **m**
Припущення: **Дисперсія генеральної сукупності дорівнює σ^2**

Точкова оцінка параметра:
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Розподіл оцінки:
$$\overline{X} \sim N\left(m, \frac{\sigma}{\sqrt{n}}\right)$$

Довірчий інтервал:
$$\overline{X} - \frac{\sigma}{\sqrt{n}} \cdot u_{1-\frac{\alpha}{2}} < m < \overline{X} + \frac{\sigma}{\sqrt{n}} \cdot u_{1-\frac{\alpha}{2}}$$

2. Побудова довірчого інтервалу при невідомій дисперсії генеральної сукупності:

Параметр, що оцінюється: Математичне сподівання **m**
Припущення: **Дисперсія генеральної сукупності є невідомою**

Точкова оцінка параметра:
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Вибіркова дисперсія:
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2$$

Розподіл оцінки:
$$(\overline{X} - m) \cdot \frac{S}{\sqrt{n}} \sim T(n-1)$$

Довірчий інтервал:
$$\overline{X} - \frac{S}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}(n-1) < m < \overline{X} + \frac{S}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}(n-1)$$

Довірчі інтервали для різниці у середніх нормально розподілених генеральних сукупностей

Вихідні дані: $x_1^1, x_2^1, \ldots, x_{n_1}^1$ – вибірка з першої генеральної сукупності,

$x_1^2, x_2^2, \ldots, x_{n_2}^2$ – вибірка з другої генеральної сукупності,

n_1 и n_2 – об'єми першої та другої вибірки відповідно

1. Побудова довірчого інтервалу, якщо дисперсії генеральних сукупностей відомі:

Параметр, що оцінюється: Різниця між математичними сподіваннями $m_1 - m_2$

Припущення: Дисперсії генеральних сукупностей відомі та дорівнюють σ_1^2 і σ_2^2 відповідно

Точкова оцінка параметра: $\overline{X}_1 - \overline{X}_2$, де

$$\bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i^1$$

$$\bar{x}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_i^2$$

Розподіл оцінки:

$$\overline{X}_1 - \overline{X}_2 \sim N\left(m_1 - m_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Довірчий інтервал:

$$\left(\overline{X}_1 - \overline{X}_2\right) - u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < m_1 - m_2 <$$

$$< \left(\overline{X}_1 - \overline{X}_2\right) + u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

2. Побудова довірчого інтервалу, якщо дисперсії генеральної сукупності невідомі:

Параметр, що оцінюється: Різниця між математичними сподіваннями $m_1 - m_2$

Припущення: Дисперсії генеральних сукупностей невідомі, але відомо, що вони однакові

$\sigma_1^2 = \sigma_2^2$

Точкова оцінка параметра: $\overline{X}_1 - \overline{X}_2$

де

$$\bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i^1$$

$$\bar{x}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_i^2$$

Оцінка дисперсії:

$$S^2 = \frac{(n_1 - 1)\cdot S_1^2 + (n_2 - 1)\cdot S_2^2}{n_1 + n_2 - 2},$$

де S_1^2 та S_2^2 – вибіркові дисперсії першої та другої вибірки відповідно:

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(x_i^1 - \overline{X}_1\right)^2,$$

$$S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}\left(x_i^1 - \overline{X}_2\right)^2,$$

де $\overline{X}_1 = \sum_{i=1}^{n_1} x_i^1, \quad \overline{X}_2 = \sum_{i=1}^{n_2} x_i^2$

Розподіл оцінки:

$$\frac{\left[\left(\overline{X}_1 - \overline{X}_2\right) - \left(m_1 \_ m_2\right)\right]}{S\cdot\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim T\left(n_1 + n_2 - 2\right)$$

Довірчий інтервал:

$$\left(\overline{X}_1 - \overline{X}_2\right) - t_{1-\frac{\alpha}{2}}\left(n_1 + n_2 - 2\right)\cdot S\cdot\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < m_1 - m_2 <$$

$$< \left(\overline{X}_1 - \overline{X}_2\right) + t_{1-\frac{\alpha}{2}}\left(n_1 + n_2 - 2\right)\cdot S\cdot\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Довірчі інтервали для дисперсії нормально розподіленої генеральної сукупності

Вихідні дані: $x_1, x_2, \ldots, x_n$ – вибірка
n – об'єм вибірки

Побудова довірчого інтервалу, якщо математичне сподівання відоме:

Параметр, що оцінюється: Дисперсія σ^2

Припущення: **Математичне сподівання генеральної сукупності дорівнює m**

Точкова оцінка параметра:
$$S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$$

Розподіл оцінки:
$$\frac{n \cdot S_0^2}{\sigma^2} \sim \chi^2(n)$$

Довірчий інтервал:
$$\frac{n \cdot S_0^2}{\chi^2_{1-\frac{\alpha}{2}}(n)} < \sigma^2 < \frac{n \cdot S_0^2}{\chi^2_{\frac{\alpha}{2}}(n)}$$

Побудова довірчого інтервалу, якщо математичне сподівання невідоме:

Параметр, що оцінюється: Дисперсія σ^2

Припущення: **Математичне сподівання невідоме**

Вибіркове середнє:
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Точкова оцінка параметра:
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2$$

Розподіл оцінки:
$$\frac{n \cdot S^2}{\sigma^2} \sim \chi^2(n-1)$$

Довірчий інтервал:
$$\frac{(n-1) \cdot S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} < \sigma^2 < \frac{(n-1) \cdot S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}$$

Довірчі інтервали для параметра p біноміально розподіленої генеральної сукупності

Вихідні дані: **n** – кількість експериментів
x – кількість успіхів в експериментах

Побудова довірчого інтервалу за умови можливості нормальної апроксимації:

Параметр, що оцінюється: **p**

Умови застосування: $\begin{cases} n > 50 \\ x > 5 \\ (n - x) > 5 \end{cases}$

Точкова оцінка параметра: $h = \frac{x}{n}$

Розподіл оцінки: $\frac{h - p}{\sqrt{\frac{p \cdot (1 - p)}{n}}} \sim N(0,1)$

Довірчий інтервал: $h - u_{1-\frac{\alpha}{2}} \sqrt{\frac{h \cdot (1 - h)}{n}} < p < h + u_{1-\frac{\alpha}{2}} \sqrt{\frac{h \cdot (1 - h)}{n}}$

Побудова довірчого інтервалу у випадку неможливості нормальної апроксимації:

Параметр, що оцінюється: **p**

Умови застосування: $\begin{cases} n < 50 \\ x < 5 \\ (n - x) < 5 \end{cases}$

Довірчий інтервал: $\frac{x \cdot F_{\frac{\alpha}{2}}(2 \cdot x, 2 \cdot (n - x + 1))}{n - x + 1 + x \cdot F_{\frac{\alpha}{2}}(2 \cdot x, 2 \cdot (n - x + 1))} < p < \frac{(x + 1) \cdot F_{1-\frac{\alpha}{2}}(2 \cdot (x + 1), 2 \cdot (n - x))}{n - x + (x + 1) \cdot F_{1-\frac{\alpha}{2}}(2 \cdot (x + 1), 2 \cdot (n - x))}$

Співвідношення між квантилями: $F_p(m,k) = \frac{1}{F_{1-p}(k,m)}$

Критерії значущості для перевірки гіпотез про середнє нормально розподіленої генеральної сукупності

Вихідні дані: $x_1, x_2, \ldots, x_n$ – вибірка
n – об'єм вибірки

Якщо дисперсія генеральної сукупності відома:

Основна гіпотеза: $H_0: m = m_0$

Умова використання: **Дисперсія генеральної сукупності відома і дорівнює σ^2**

Точкова оцінка параметра:
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Статистика:
$$Z = \frac{\overline{X} - m_0}{\sigma/\sqrt{n}}$$

Розподіл статистики, якщо основна гіпотеза є істинною:
$$Z \sim N(0,1)$$

Область прийняття рішення для двостороннього критерія:
$$\frac{|\overline{X} - m_0|}{\sigma/\sqrt{n}} < u_{1-\frac{\alpha}{2}}$$

Якщо дисперсія генеральної сукупності є невідомою:

Основна гіпотеза: $H_0: m = m_0$

Умова використання: **Дисперсія генеральної сукупності є невідомою**

Точкова оцінка параметру:
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Точкова оцінка дисперсії:
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2$$

Статистика:
$$Z = \frac{\overline{X} - m_0}{S/\sqrt{n}}$$

Розподіл статистики, якщо основна гіпотеза є істинною:
$$Z \sim T(n-1)$$

Область прийняття рішення для двостороннього критерія:
$$\frac{|\overline{X} - m_0|}{S/\sqrt{n}} < t_{1-\frac{\alpha}{2}}(n-1)$$

Критерії значущості про рівність середніх нормально розподілених генеральних сукупностей

Вихідні дані: $x_1^1, x_2^1, \ldots, x_{n_1}^1$ – вибірка з першої генеральної сукупності,

$x_1^2, x_2^2, \ldots, x_{n_2}^2$ – вибірка з другої генеральної сукупності,

n_1 и n_2 – об'єми відповідно першої та другої вибірок

Якщо дисперсії генеральних сукупностей є відомими:

Основна гіпотеза: $H_0: m_1 = m_2$

Умова використання: Дисперсії генеральних сукупностей є відомими і дорівнюють відповідно σ_1^2 и σ_2^2

Точкові оцінки параметрів: $\overline{X}_1$ и $\overline{X}_2$, де

$$\overline{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i^1$$

$$\overline{X}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_i^2$$

Статистика:

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Розподіл статистики, якщо основна гіпотеза є істинною:

$$Z \sim N(0,1)$$

Область прийняття рішення для двостороннього критерія:

$$\frac{\left|\overline{X}_1 - \overline{X}_2\right|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < u_{1-\frac{\alpha}{2}}$$

Якщо дисперсії генеральних сукупностей є невідомими, і гіпотеза про їх рівність є істинною:

Основна гіпотеза: $H_0: m_1 = m_2$

Умова використання: Дисперсії генеральних сукупностей σ_1^2 и σ_2^2 є невідомими,

Гіпотеза $\mathbf{H_0 : \sigma_1^2 = \sigma_2^2}$ є істинною

Точкові оцінки параметрів: $\overline{\mathbf{X}}_1$ и $\overline{\mathbf{X}}_2$, де

$$\overline{X}_1=\frac{1}{n_1}\sum_{i=1}^{n_1} x_i^1$$

$$\overline{X}_2=\frac{1}{n_2}\sum_{i=1}^{n_2} x_i^2$$

$$\mathbf{S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(x_i^1 - \overline{X}_1\right)^2}$$

$$\mathbf{S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}\left(x_i^2 - \overline{X}_2\right)^2}$$

$$\mathbf{S = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}}$$

Статистика: $\mathbf{Z = \frac{\overline{X}_1 - \overline{X}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}$

Розподіл статистики, якщо основна гіпотеза є істинною: $\mathbf{Z \sim T(n_1 + n_2 - 2)}$

Область прийняття рішення для двостороннього критерія: $\mathbf{\frac{\left|\overline{X}_1 - \overline{X}_2\right|}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < t_{1-\frac{\alpha}{2}}(n_1 + n_2 - 2)}$

Якщо дисперсії генеральних сукупностей є невідомими, і гіпотеза про їх рівність є хибною:

Основна гіпотеза: H_0: $m_1 = m_2$

Умова використання: Дисперсії генеральних сукупностей σ_1^2 и σ_2^2 є невідомими,

Гіпотеза $\mathbf{H_0 : \sigma_1^2 = \sigma_2^2}$ є хибною

Точкові оцінки параметрів: $\overline{\mathbf{X}}_1$ и $\overline{\mathbf{X}}_2$, де

$$\overline{X}_1=\frac{1}{n_1}\sum_{i=1}^{n_1} x_i^1$$

$$\bar{X}_2=\frac{1}{n_2}\sum_{i=1}^{n_2} x_i^2$$

$$S_1^2=\frac{1}{n_1-1}\sum_{i=1}^{n_1}\left(x_i^1-\bar{X}_1\right)^2$$

$$S_2^2=\frac{1}{n_2-1}\sum_{i=1}^{n_2}\left(x_i^2-\bar{X}_2\right)^2$$

Статистика:

$$Z=\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$$

Розподіл статистики, якщо основна гіпотеза є істинною:

$$Z\sim T(k),\quad \text{где}$$

$$k=\frac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2-1}}$$

Область прийняття рішення для двостороннього критерія:

$$\frac{\left|\bar{X}_1-\bar{X}_2\right|}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}<t_{1-\frac{\alpha}{2}}(k)$$

Критерії значущості для перевірки гіпотез про дисперсії нормально розподіленої генеральної сукупності

Вихідні дані: $x_1, x_2, \ldots, x_n$ – вибірка
n – об'єм вибірки

Якщо математичне сподівання генеральної сукупності є відомим:

Основна гіпотеза: $H_0: \sigma^2 = \sigma_0^2$

Умова використання: **Математичне сподівання генеральної сукупності відоме і дорівнює m**

Точкова оцінка параметра:
$$S_0^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2$$

Статистика:
$$Z = \frac{nS^2}{\sigma_0^2}$$

Розподіл статистики, якщо основна гіпотеза є істинною:
$$Z \sim \chi^2(n)$$

Область прийняття рішення для двостороннього критерія:
$$\chi^2_{\frac{\alpha}{2}}(n) < \frac{nS_0^2}{\sigma_0^2} < \chi^2_{1-\frac{\alpha}{2}}(n)$$

Якщо математичне сподівання генеральної сукупності є невідомим:

Основна гіпотеза: $H_0: \sigma^2 = \sigma_0^2$

Умова використання: **Математичне сподівання генеральної сукупності невідоме**

Точкова оцінка параметра:
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Точкова оцінка дисперсії:
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2$$

Статистика:
$$Z = \frac{(n-1)S^2}{\sigma_0^2}$$

Розподіл статистики, якщо основна гіпотеза є істинною:
$$Z \sim \chi^2(n-1)$$

Область прийняття рішення для двостороннього критерія:
$$\chi^2_{\frac{\alpha}{2}}(n-1) < \frac{(n-1)S^2}{\sigma_0^2} < \chi^2_{1-\frac{\alpha}{2}}(n-1)$$

Критерії значущості для перевірки гіпотез про рівність дисперсій нормально розподілених генеральних сукупностей

Вихідні дані: $x_1^1, x_2^1, \ldots, x_{n_1}^1$ – вибірка з першої генеральної сукупності,

$x_1^2, x_2^2, \ldots, x_{n_2}^2$ – вибірка з другої генеральної сукупності,

n_1 и n_2 – об'єми відповідно першої та другої вибірок

Якщо математичні сподівання генеральних сукупностей відомі:

Основна гіпотеза:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Умова використання: **Математичні сподівання генеральних сукупностей відомі і дорівнюють m_1 та m_2**

Точкові оцінки параметрів:

$$S_{01}^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(x_i^1 - m_1 \right)^2$$

$$S_{02}^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} \left(x_i^2 - m_2 \right)^2$$

Статистика:

$$Z = \frac{S_{01}^2}{S_{02}^2}, \quad S_{01}^2 > S_{02}^2$$

Розподіл статистики, якщо основна гіпотеза є істинною:

$$Z \sim F(n_1, n_2)$$

Область прийняття рішення для двостороннього критерія:

$$\frac{S_{01}^2}{S_{02}^2} < F_{1-\frac{\alpha}{2}}(n_1, n_2)$$

Якщо математичні сподівання генеральних сукупностей невідомі :

Основна гіпотеза:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Умова використання: Математичні сподівання невідомі

Точкові оцінки параметрів: $\overline{X}_1$ и $\overline{X}_2$, де

$$\bar{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i^1$$

$$\bar{X}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_i^2$$

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(x_i^1 - \overline{X}_1\right)^2$$

$$S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}\left(x_i^2 - \overline{X}_2\right)^2$$

Статистика:

$$Z = \frac{S_1^2}{S_2^2}, \quad S_1^2 > S_2^2$$

Розподіл статистики, якщо основна гіпотеза є істинною:

$$Z \sim F(n_1 - 1, n_2 - 1)$$

Область прийняття рішення для двостороннього критерія:

$$\frac{S_1^2}{S_2^2} < F_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 1)$$

Критерії значущості для перевірки гіпотез про параметр p біноміально розподіленої генеральної сукупності

У випадку однієї вибірки:

Вихідні дані: **n** – кількість випробувань
x – кількість успіхів у випробуваннях

Основна гіпотеза:
$$H_0 : p = p_0$$
$$H_1 : p \neq p_0$$

Умови використання:
$$\begin{cases} n > 50 \\ x > 5 \\ (n - x) > 5 \end{cases}$$

Точкова оцінка параметра:
$$h = \frac{x}{n}$$

Статистика:
$$Z = \frac{h - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

Розподіл статистики, якщо основна гіпотеза є істинною:
$$Z \sim N(0,1)$$

Область прийняття рішення для двостороннього критерія:
$$|Z| < u_{1-\frac{\alpha}{2}}$$

У випадку двох вибірок:

Вихідні дані: n_1 и n_2 – кількість випробувань
x_1 и x_2 – кількість успіхів у випробуваннях

Основна гіпотеза:
$$H_0 : p_1 = p2$$
$$H_1 : p_1 \neq p_2$$

Умови використання:
$$\begin{cases} n_1 > 50 \\ x_1 > 5 \\ (n_1 - x_1) > 5 \end{cases} \quad \begin{cases} n_2 > 50 \\ x_2 > 5 \\ (n_2 - x_2) > 5 \end{cases}$$

Точкова оцінка параметра:
$$h_1 = \frac{x_1}{n_1} \quad h_2 = \frac{x_2}{n_2} \quad \begin{matrix} x_1 + x_2 = x \\ n_1 + n_2 = n \end{matrix} \quad h = \frac{x}{n}$$

Статистика:
$$Z = \frac{h_1 - h_2}{\sqrt{h(1 - h)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Розподіл статистики, якщо основна гіпотеза є істинною:
$$Z \sim N(0,1)$$

Область прийняття рішення для двостороннього критерія:
$$|Z| < u_{1-\frac{\alpha}{2}}$$