МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ТАРАСА ШЕВЧЕНКА

Є. О. Лебєдєв
Г. В. Лівінська
І. В. Розора
М. М. Шарапов

МАТЕМАТИЧНА СТАТИСТИКА

Начальний посібник

КИЇВСЬКИЙ УНІВЕРСИТЕТ 2016

ПЕРЕДМОВА

Мета пропонованого посібника – викласти в доступній для початкового вивчення формі елементи основних напрямів сучасної статистичної теорії. При цьому акцент робиться на дослідженні питань оптимальності відповідних статистичних процедур та їх практичної реалізації. У посібнику широко використовуються методи теорії ймовірностей, значна увага приділяється питанням прикладної інтерпретації матеріалу та отриманих результатів.

Зазвичай у ВНЗ викладанню математичної статистики передує обов'язковий курс теорії ймовірностей, тому автори даного посібника відмовились від традиційного підходу, коли значне місце на початку курсу відводилось переліченню основних фактів і положень зазначеної теорії, на яких базується викладання статистичної теорії. Припускатимемо, що необхідний понятійний мінімум теорії ймовірностей студентам уже відомий. Деякі її додаткові положення наведені у відповідних місцях посібника.

Викладення матеріалу ведеться на рівні, доступному студентам факультету кібернетики і прикладної математики класичних університетів, а також технічних ВНЗ. Автори спираються тільки на знання студентами основ математичного аналізу та лінійної алгебри, що викладаються на першому курсі. Матеріал не містить громіздких математичних викладок і доведень, проте підкреслюється статистична суть проблеми, що розглядається.

У першому розділі аналізується ситуація, що відповідає моделі повторних незалежних спостережень над деякою скалярною випадковою величиною ξ. Вводяться основні поняття вибіркової теорії, яка вивчає стохастичні властивості випадкової вибірки; наводяться фундаментальні теореми математичної статистики; досліджуються в точній і асимптотичній (при збільшенні розміру вибірки) постановках властивості деяких харак-

теристик випадкової вибірки; розглядаються розподіли, що відіграють важливу роль у статистиці.

Другий і третій розділи присвячено викладенню основ теорії оцінок невідомих параметрів і функцій від них у межах довільної параметричної моделі. Розглянуто два традиційні підходи до розв'язання цих задач: точкове та інтегральне оцінювання. При викладенні теорії точкового оцінювання переважно розглядаються незсунені оцінки й за міру точності різних оцінок береться величина їх дисперсії. Основну увагу приділено методам побудови оптимальних оцінок. У третьому розділі розглянуто різні підходи до побудови надійних інтервалів.

Четвертий розділ фактично є вступом до теорії перевірки статистичних гіпотез. Тут наведені основні поняття зазначеної теорії (статистичної гіпотези, статистичного критерію, критичної області тощо). Сформульовані типові й найпоширеніші на практиці статистичні гіпотези. На прикладах розв'язання задач показані загальні принципи побудови й дослідження критеріїв згоди.

У п'ятому розділі розглядаються гіпотези про істинне значення невідомого параметра, який задає множину розподілів. Прикладами проілюстровані принципи побудови оптимальних або асимптотично оптимальних критеріїв перевірки таких гіпотез, в основі яких лежить запропонований Є. Нейманом і Е. Пірсоном метод відношення вірогідності.

У шостому розділі викладено класичний метод найменших квадратів для оцінювання параметрів моделі та його оптимальні властивості, розглянуто питання застосування зазначеного методу на практиці.

У кожному розділі наведено задачі для самостійного розв'язання. Цей важливий матеріал допомагає засвоїти теорію до того рівня, який дозволяє використовувати її на практиці. У кожному теоретичному блоці містяться модельні задачі з їх розв'язанням.

Розділ 1
ЕЛЕМЕНТИ ВИБІРКОВОЇ ТЕОРІЇ

1.1. Задачі математичної статистики

На відміну від теорії ймовірностей, де модель явища вважалася заданою та проводилися розрахунки ймовірностей можливих змін, у математичній статистиці ми виходимо з відомих реалізацій випадкових подій (статистичних даних) і розробляємо методи, які дозволяють за цими даними підібрати відповідну теоретико-ймовірнісну модель.

Основна математична модель випадкового явища базується на понятті ймовірнісного простору (Ω, U, P). При вивченні конкретного експерименту ймовірність $P(\cdot)$ рідко буває відомою повністю. Часто апріорі можна стверджувати лише те, що $P(\cdot)$ є елементом деякої сім'ї ймовірностей $\mathbb{P}$.

Клас $\mathbb{P}$ може містити всі ймовірності, які можна задати на U – ситуація повної невизначеності. В інших випадках є деякою вужчою сім'єю ймовірностей, заданою в якій-небудь формі. Якщо фіксовано клас $\mathbb{P}$, то кажуть, що задано статистичну (імовірнісно-статистичну) модель, і розуміють під цим набір $(\Omega, U, \mathbb{P})$.

Приклад 1.1. Розглянемо стохастичний експеримент, який проводимо згідно зі схемою Бернуллі: експеримент складається з n незалежних випробувань, у кожному з яких спостерігається або 1 – "успіх", або 0 – "невдача" з імовірностями p і $q = 1 - p$, відповідно. Результат експерименту можна зобразити n-вимірним вектором $\omega = (\varepsilon_1, ..., \varepsilon_n)$, де $\varepsilon_i = 0,1$ при $i = 1, 2, ..., n$. Отже, $\Omega = \{\omega : \omega = (\varepsilon_1, ..., \varepsilon_n), \varepsilon_i = 0 \text{ або } 1, \quad i = 1, 2, ..., n\}$, U – су-

купність усіх підмножин ... Якщо ймовірність успіху p відома, то маємо модель теорії ймовірностей (Ω, U, P), де P визначається ймовірностями появи елементарних подій

$$P(\omega) = p^{\sum_{i=1}^{n}\varepsilon_i} \cdot q^{n-\sum_{i=1}^{n}\varepsilon_i}.$$

У межах цієї моделі можна розв'язувати ймовірнісні задачі, наприклад знайти ймовірність появи рівно k успіхів.

Припустимо тепер, що ймовірність успіху заздалегідь невідома. Позначимо її через θ. Про неї відомо тільки, що $\theta \in \Theta = [0,1]$. Таким чином, у даному випадку маємо статистичну модель $(\Omega, U, \mathbb{P})$, де

$$\mathbb{P} = \left\{ P_\theta, \theta \in \Theta = [0,1] : P_\theta(\omega) = \theta^{\sum_{i=1}^{n}\varepsilon_i} \cdot (1-\theta)^{n-\sum_{i=1}^{n}\varepsilon_i}, \omega \in \Omega \right\}.$$

Статистична модель описує ситуацію, коли в імовірнісній моделі експерименту, що розглядається, є невизначеність у виборі P. Задача математичної статистики полягає в тому, щоб зменшити цю невизначеність, використовуючи інформацію, що подається у вигляді результатів експерименту (статистичних даних). У певному розумінні математична статистика розв'язує задачі, обернені до задач теорії ймовірностей: вона уточнює (виявляє) структуру стохастичних моделей за результатами спостережень, що проводяться.

1.2. Основна статистична модель експерименту

Часто результат експерименту характеризується скінченною сукупністю випадкових величин $\xi' = (\xi_1, ..., \xi_n)$. У цьому випадку казатимемо, що експеримент складається з n випробувань, у яких результат i-го випробування описується випадковою величиною ξ_i, $i = 1, 2, ..., n$.

Вибіркою називатимемо сукупність випадкових величин $\xi' = (\xi_1,...,\xi_n)$, що спостерігаються в експерименті. Величини ξ_i, $i = 1,2,...,n$, – це елементи вибірки, n – розмір вибірки. Реалізацію вибірки ξ позначатимемо через $x' = (x_1,...,x_n)$.

Вибірковий простір – це множина всіх можливих значень вибірки $\xi : X = \{x\}$. Вибірковий простір може бути або всім n-вимірним евклідовим простором R^n, або його частиною, наприклад складатися зі скінченної або зліченної кількості точок з R^n, якщо ξ має дискретний розподіл. У кожному конкретному випадку задається σ-алгебра U на вибірковому просторі X. Під статистичною моделлю експерименту будемо розуміти набір $(X,U,\mathbb{P})$, де $\mathbb{P}$ – клас усіх допустимих розподілів випадкового вектора ξ.

Розподіл імовірностей довільного випадкового вектора однозначно визначається його функцією розподілу. Отже, статистична модель експерименту визначається вибірковим простором і сім'єю функцій розподілу $\mathbb{F}$, якій належить невідома функція розподілу

$$F_\xi(z_1,...,z_n) = P\{\xi_1 \le z_1,...,\xi_n \le z_n\}$$

вибірки $\xi' = (\xi_1,...,\xi_n)$. Статистичну модель маємо у вигляді $(X,U,\mathbb{F})$.

Часто виникають ситуації, коли компоненти $\xi_1,...,\xi_n$ незалежні й розподілені так, як деяка випадкова величина ξ_0. Це відповідає експерименту, у якому проводяться повторні незалежні спостереження над випадковою величиною ξ_0. Таку модель можна задавати в термінах функції розподілу $F_{\xi_0}(\cdot)$, оскільки $F_\xi(z) = F_{\xi_1}(z_1)\cdot...\cdot F_{\xi_n}(z_n)$ і $F_{\xi_i}(\cdot) = F_{\xi_0}(\cdot)$ для $i = 1,2,...,n$. У цьому випадку говорять, що $\xi' = (\xi_1,...,\xi_n)$ – вибірка з генеральної сукупності з розподілом $F_{\xi_0}(\cdot)$. Статистичну модель для повторних незалежних спостережень коротко позначатимемо

$\left(X, U, \left\{F_{\xi_0}\right\}\right)$, де $\left\{F_{\xi_0}(.)\right\}$ – клас допустимих функцій розподілу випадкової величини ξ_0.

Параметрична модель – це така модель, у якій функції розподілу з класу $\mathbb{F}$ задані з точністю до значень деякого параметра θ, який набуває значення з множини Θ

$$\mathbb{F} = \left\{F(z,\theta), \theta \in \Theta\right\}.$$

Говорять, що в цьому випадку відомий тип розподілу випадкової величини, що спостерігається. Невідомим є тільки параметр, від якого залежить розподіл.

Приклад 1.2. 1) Нехай відомо, що $F_{\xi_0}(.)$ – нормальний розподіл з відомою дисперсією й невідомим середнім. Тоді статистична модель має вигляд $\mathbb{F} = \left\{F(z,\theta),\ \theta \in \Theta = (-\infty, \infty)\right\}$, де щільність функції розподілу $F(x,\theta)$ становить

$$f(z,\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(z-\theta)^2}{2\sigma^2}}, \quad -\infty < z < \infty.$$

2) Нехай $F_{\xi_0}(.)$ – нормальний розподіл з невідомим середнім і дисперсією. Тоді статистична модель має вигляд

$$\mathbb{F} = \left\{F(z,\theta),\ \theta = (\theta_1, \theta_2) \in \Theta\right\},$$

де $\Theta = \left\{(\theta_1, \theta_2) : -\infty < \theta_1 < \infty,\ 0 < \theta_2 < \infty\right\}$ і $F(z,\theta)$ визначається щільністю

$$f(z,\theta) = \frac{1}{\sqrt{2\pi}\theta_2} e^{-\frac{(z-\theta_1)^2}{2\theta_2^2}}, \quad -\infty < z < \infty.$$

У випадку параметричної моделі розподіл імовірностей на вибірковому просторі X, який відповідає параметру θ, позначається символом P_θ. Аналогічно індекс θ при символах математичного сподівання, дисперсії тощо означає, що відповідні величини підраховуються за розподілом $P_\theta(\cdot)$.

1.3. Емпірична функція розподілу

Нехай $\xi' = (\xi_1, ..., \xi_n)$ – вибірка розміром n із розподілу $F_{\xi_0}(.)$ і $x' = (x_1, ..., x_n)$ – значення ξ, що спостерігалися. Кожній реалізації x вибірки ξ можна поставити у відповідність упорядковану послідовність

$$x_{(1)} \le x_{(2)} \le ... \le x_{(n)}, \tag{1.1}$$

де $x_{(1)} = \min(x_1, ..., x_n)$, $x_{(2)}$ – друге за величиною значення з $x_1, ..., x_n$ і т. д.; $x_{(n)} = \max(x_1, ..., x_n)$. Позначимо через $\xi_{(k)}$ випадкову величину, яка для кожної реалізації x вибірки ξ набуває значення $x_{(k)}$, $k = 1, 2, ..., n$.

За вибіркою ξ визначимо нову послідовність випадкових величин $\xi_{(1)}, ..., \xi_{(n)}$, які називаються ***порядковими статистиками вибірки***. При цьому $\xi_{(k)}$ – k -та порядкова статистика, а $\xi_{(1)}, \xi_{(n)}$ – мінімальне та максимальне значення вибірки, відповідно.

З визначення порядкових статистик випливає, що вони задовольняють нерівності

$$\xi_{(1)} \le \xi_{(2)} \le ... \le \xi_{(n)}. \tag{1.2}$$

Послідовність (1.2) називають варіаційним рядом вибірки. ***Варіаційний ряд вибірки*** – це елементи вибірки, розташовані за зростанням.

Визначимо для кожного дійсного z випадкову величину $\mu_n(z)$, яка дорівнює кількості елементів вибірки $\xi' = (\xi_1, ..., \xi_n)$, значення яких не перевищує z:

$$\mu_n(z) = \left| \{ j : \xi_j \le z \} \right|,$$

де $|\cdot|$ – кількість елементів скінченної множини. Функція, яка задається рівністю $F_n(z) = \dfrac{\mu_n(z)}{n}$, називається ***емпіричною функцією розподілу***. Функція розподілу $F(z)$ випадкової ве-

личини ξ_0, що спостерігається, називається ***теоретичною функцією розподілу***.

За визначенням емпірична функція розподілу – випадковий процес по z: для кожного $z \in (-\infty, \infty)$ значення $F_n(z)$ – випадкова величина, реалізаціями якої є числа $0, \frac{1}{n}, \frac{2}{n}, ..., \frac{n-1}{n}, 1$, при цьому $P\left(F_n(z) = \frac{k}{n}\right) = P(\mu_n(z) = k)$ (рис. 1.1). З визначення $\mu_n(z)$ випливає, що вона розподілена так само, як кількість успіхів у n випробуваннях Бернуллі з імовірністю успіху $p = P(\xi \le z) = F(z)$. Тому

$$P\left(F_n(z) = \frac{k}{n}\right) = C_n^k F^k(z)\left(1 - F(z)\right)^{n-k}, \quad k = 0,1,...,n.$$

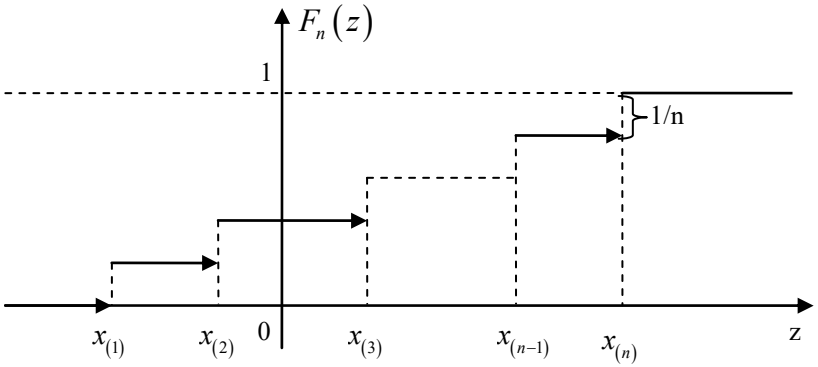

Рис. 1.1

Для кожної реалізації вибірки ξ функція $F_n(z)$ однозначно визначена й має всі властивості функції розподілу: змінюється від 0 до 1, не спадає й неперервна праворуч. При цьому вона стрибкоподібна і зростає тільки в точках послідовності (1.2).

Емпіричну функцію розподілу $F_n(x)$ можна записати у вигляді

$$F_n(z)=\frac{1}{n}\sum_{k=1}^{n} e\left(z-\xi_{(k)}\right),$$

де $e(z)$ – функція одиничного стрибка (Хевісайда):

$$e(z)=\begin{cases}1, & \text{при} \quad z\geq 0,\\ 0, & \text{при} \quad z<0.\end{cases}$$

Найважливіша властивість емпіричної функції розподілу полягає в тому, що при зростанні кількості спостережень над випадковою величиною ξ_0 відбувається зближення цієї функції з теоретичною.

Теорема 1.1. *Нехай $F_n(z)$ – емпірична функція розподілу, яка побудована за вибіркою $\xi'=(\xi_1,...,\xi_n)$, $F(z)$ – відповідна теоретична функція розподілу. Тоді для довільного z $(-\infty<z<\infty)$ і довільного $\varepsilon>0$*

$$\lim_{n\to\infty} P\left(\left|F_n(z)-F(z)\right|<\varepsilon\right)=1.$$

Отже, якщо розмір вибірки великий, то значення емпіричної функції розподілу в кожній точці z може служити оцінкою теоретичної функції розподілу в цій точці. Справедливий також сильніший результат.

Теорема 1.2 (В. І. Глівенко, 1933). *В умовах теореми 1.1*

$$P\left(\lim_{n\to\infty}\sup_{-\infty<z<\infty}\left|F_n(z)-F(z)\right|=0\right)=1.$$

Ця теорема означає, що відхилення $D_n=D_n(\xi)=\sup\limits_{-\infty<z<\infty}\left|F_n(z)-F(z)\right|$ емпіричної функції розподілу від теоретичної на всій осі з імовірністю 1 буде малим за достатньо великого розміру вибірки.

Наведемо ще один результат, який дозволяє для великих n оцінити ймовірність заданих відхилень випадкової величини D_n від 0.

Теорема 1.3 (А. М. Колмогоров, 1933). *Якщо функція* $F(z)$ *неперервна, то при довільному фіксованому* $t > 0$

$$\lim_{n\to\infty} P\left(\sqrt{n}D_n \le t\right) = K(t) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2t^2}.$$

Теорема Колмогорова дозволяє визначити границі, у яких із заданою ймовірністю знаходиться теоретична функція розподілу $F(z)$, якщо вона невідома. Для цього за $\gamma \in (0,1)$ з рівняння $K(t) = \gamma$ знаходимо число t_γ. Тоді за теоремою Колмогорова маємо

$$P\left(\sqrt{n}D_n \le t_\gamma\right) = P\left(F_n(z) - \frac{t_\gamma}{\sqrt{n}} \le F(z) \le F_n(z) + \frac{t_\gamma}{\sqrt{n}}\right) \underset{n\to\infty}{\to}$$

$$\underset{n\to\infty}{\to} K(t_\gamma) = \gamma \text{ для всіх } z.$$

Таким чином, для великих n з імовірністю, близькою до γ, значення функції $F(z)$ для всіх z задовольняє нерівності

$$F_n(z) - \frac{t_\gamma}{\sqrt{n}} \le F(z) \le F_n(z) + \frac{t_\gamma}{\sqrt{n}}.$$

Нехай тепер ξ має абсолютно неперервний розподіл. На практиці вибірки з абсолютно неперервних розподілів часто групують і подають у вигляді ***інтервального статистичного ряду***. При цьому індивідуальні вибіркові значення не наводяться, а вказується лише кількість вибіркових значень, що потрапили в інтервали деякого певного розбиття.

Розглянемо проблему наближення невідомої щільності $p_\xi(z)$ випадкової величини ξ. Розіб'ємо область значень елементів вибірки на інтервали Δ_i довжиною h_i. Нехай n_i^* – кількість елементів вибірки, які потрапили в i-й інтервал; z_i^* – середина i-го інтервалу.

Гістограмою вибірки будемо називати сукупність прямокутників з основами Δ_i і висотами $\frac{n_i^*}{nh_i}$. ***Полігон частот*** – це ла-

мана, яка з'єднує точки $\left(z_i^*, \frac{n_i^*}{nh_i}\right)$, $i = 1, 2, \ldots$. Величини $\frac{n_i^*}{nh_i}$ є оцінками щільності в точках x_i^*.

Теорема 1.4. *Нехай область значень елементів вибірки з генеральної сукупності розбита на інтервали Δ_i однакової довжини h_n і $f_n^*(z) = \frac{n_i^*}{nh_n}$, $z \in \Delta_i$ – гістограма вибірки. Якщо невідома щільність $f(z)$ неперервна, то*

$$f_n^*(z) \underset{n\to\infty}{\overset{P}{\Rightarrow}} f(z),$$

якщо $h_n \to 0$ і $nh_n \to \infty$.

1.4. Вибіркові моменти

Позначимо через $a_k = M\xi_0{}^k = \int_R z^k dF(z)$ теоретичний момент k-го порядку випадкової величини ξ_0, що спостерігається, а через $A_k = A_k(\xi) = \int_R z^k dF_n(z) = \frac{1}{n}\sum_{i=1}^n \xi_i^k$ – емпіричний, або вибірковий, момент k-го порядку. При $k = 1$ величину A_k називають ***вибірковим середнім*** і позначають символом $\overline{\xi}$:

$$\overline{\xi} = A_1 = \frac{1}{n}\sum_{i=1}^n \xi_i .$$

Теорема 1.5. *Якщо існує k-й теоретичний момент, то $A_k \underset{n\to\infty}{\overset{P}{\Rightarrow}} a_k$.*

Доведення цієї теореми спирається на закон великих чисел у формі Хінчина.

Наслідком теореми 1.5 є те, що вибіркові моменти A_k можна розглядати як оцінки відповідного моменту a_k, коли кількість спостережень n достатньо велика.

Позначимо через $b_k = \int\limits_R (z - a_1)^k \, dF(z)$ теоретичний центральний момент k-го порядку, а через $B_k = \int\limits_R (z - \overline{\xi})^k \, dF_n(z)$ – вибірковий центральний момент k-го порядку. При $k = 2$ величину B_k називають ***вибірковою дисперсією*** і позначають символом $S^2 : S^2 = B_2 = \frac{1}{n} \sum\limits_{i=1}^{n} \left(\xi_i - \overline{\xi} \right)^2$.

Щоб для центральних моментів отримати результат, подібний до теореми 1.5, необхідне таке допоміжне твердження.

Лема 1.1. *Нехай випадкові величини* $\eta_1(n), ..., \eta_r(n)$ *збігаються за ймовірністю при* $n \to \infty$ *до деяких сталих* $c_1, ..., c_r$. *Тоді для довільної функції* $\varphi(z_1, ..., z_r)$, *неперервної в точці* $(c_1, ..., c_r)$,

$$\zeta(n) = \phi\left(\eta_1(n), ..., \eta_r(n)\right) \underset{n \to \infty}{\overset{P}{\Rightarrow}} \phi(c_1, ..., c_r).$$

Наслідок 1.1. *Якщо існує* k*-й теоретичний момент, то* $B_k \underset{n \to \infty}{\overset{P}{\Rightarrow}} b_k$.

При $k = 2$ *маємо*:

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\xi_i - \overline{\xi} \right)^2 \underset{n \to \infty}{\overset{P}{\Rightarrow}} D\xi_0{}^2.$$

З центральної граничної теореми для однаково розподілених незалежних доданків випливає такий результат.

Теорема 1.6. *Якщо існує теоретичний момент порядку* $2k$, *то*

$$\frac{A_k - a_k}{\sqrt{\left(a_{2k} - a_k^2\right)/n}} \underset{n \to \infty}{\overset{сл}{\Rightarrow}} \eta,$$

де η *– випадкова величина, що має стандартний нормальний розподіл.*

1.4. Вибіркова медіана та вибіркові квантилі

Розглянемо введений раніше варіаційний ряд $\xi_{(1)} \le \xi_{(2)} \le \ldots \le \xi_{(n)}$, сформований з порядкових статистик.

Нехай $\alpha \in (0,1)$. Для неперервної функції розподілу $F_{\xi_0}(z)$ ***теоретичним*** α ***-квантилем*** z_α називається мінімальний розв'язок рівняння

$$F_{\xi_0}(z) = \alpha. \tag{1.3}$$

Якщо функція $F_{\xi_0}(z)$ строго монотонна, то це рівняння має єдиний корінь.

Вибірковим α ***-квантилем*** $Z_{n,\alpha}$ будемо називати порядкову статистику

$$Z_{n,\alpha} = \xi_{([n\alpha]+1)}.$$

Тут $Z_{n,\alpha}$ – елемент вибірки, лівіше якого знаходиться частка $[n\alpha]/n \le \alpha$ спостережень, при цьому $Z_{n,\alpha}$ – порядкова статистика з максимальним номером, що задовольняє зазначену нерівність. Отже, $Z_{n,\alpha}$ можна розглядати як статистичний аналог z_α.

Вибіркова медіана визначається як

$$Z_{n,1/2} = \xi_{([n/2]+1)}.$$

Вона є оцінкою для теоретичної медіани $z_{1/2}$.

Медіана $z_{1/2}$ є характеристикою, що вказує, де знаходиться "центр" розподілу:

$$P\{\xi_i \le z_{1/2}\} = P\{\xi_i \ge z_{1/2}\} = \frac{1}{2}.$$

Приклад 1.3. За реалізацією вибірки

0, 0, 1, 1, 1, 0, 0, 1, 0, 2, 1, 2, 2, 3, 1, 1, 1, 1, 2, 2

знайти реалізацію емпіричної функції розподілу, вибіркове середнє та вибіркову дисперсію.

Розв'язання. Маємо чотири різні значення, які зустрічаються в реалізації вибірки: 0, 1, 2 та 3. Бачимо, що 0 зустрічається п'ять разів, 1 – дев'ять, 2 – п'ять і 3 – один раз. Реалізація емпіричної функції розподілу $F_n(z)=\frac{1}{n}\sum_{k=1}^{n} e\left(z-x_{(k)}\right)$ матиме стрибки в точках 0, 1, 2 та 3. Величини стрибків становитимуть відповідно 5/20, 9/20, 5/20 та 1/20, а реалізація емпіричної функції розподілу

$$F_{20}(z)=\begin{cases}0, & z<0\\ 1/4, & 0\le z<1\\ 7/10, & 1\le z<2\\ 19/20, & 2\le z<3\\ 1, & z\ge 3\end{cases}.$$

Графік функції матиме такий вигляд (рис. 1.2):

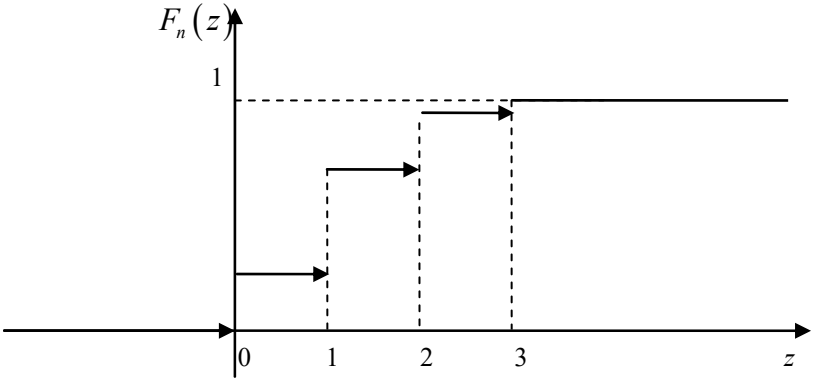

Рис. 1.2

Вибіркове середнє

$$\overline{x}=A_1=\frac{1}{n}\sum_{i=1}^{n}x_i=\frac{0\cdot 5+1\cdot 9+2\cdot 5+3\cdot 1}{20}=\frac{22}{20}=1{,}1.$$

Для вибіркової дисперсії знаходимо

$$S^2 = B_2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 =$$

$$= \frac{\left(0-1.1\right)^2 \cdot 5 + \left(1-1.1\right)^2 \cdot 9 + \left(2-1.1\right)^2 \cdot 5 + \left(3-1.1\right)^2 \cdot 1}{20} = \frac{13.8}{20} = 0.69 .$$

Відповідь: $\overline{x} = 1{,}1$; $S^2 = B_2 = 0.69$.

Приклад 1.4. За реалізацією вибірки побудувати гістограму, розбивши область значень на шість інтервалів.

0,79	0,90	0,99	0,95	0,91	0,80	0,99	0,88	0,86	0,95
0,98	0,90	0,89	0,78	0,88	0,87	1,00	0,84	0,95	0,78
0,94	0,97	0,99	1,00	0,82	1,00	0,87	0,95	0,92	0,92
0,96	0,84	0,96	0,82	0,91	0,76	0,77	0,88	0,83	0,90
0,97	0,99	0,82	0,78	1,00	0,98	0,90	0,87	1,00	0,99
0,86	0,86	0,76	0,95	0,89	0,99	0,79	0,99	1,00	0,77
0,86	0,95	1,00	1,00	0,82	0,99	0,90	0,82	0,98	0,94
0,98	0,92	0,84	1,00	0,88	0,84	1,00	0,86	0,99	0,94
0,94	0,94	0,89	0,98	0,79	0,99	0,94	0,97	0,94	0,94
0,87	0,97	1,00	0,99	0,99	0,93	0,91	0,83	0,79	0,99

Розв'язання. Найменше значення у вибірці становить 0,76, а найбільше – 1. Розіб'ємо відрізок [0,76; 1] на шість підінтервалів і підрахуємо, скільки значень потрапило до кожного з них.

Інтервали	[0,76; 0,80]	(0,80; 0,84]	(0,84; 0,88]	(0,88; 0,92]	(0,92; 0,94]	(0,94; 1]
Кількість елементів(n_i^*)	12	11	13	14	17	33

Гістограму $f_n^*(z)$ побудуємо у вигляді прямокутників з основами Δ_i і висотами $\frac{n_i^*}{0.04n} = \frac{n_i^*}{4}$ (рис. 1.3).

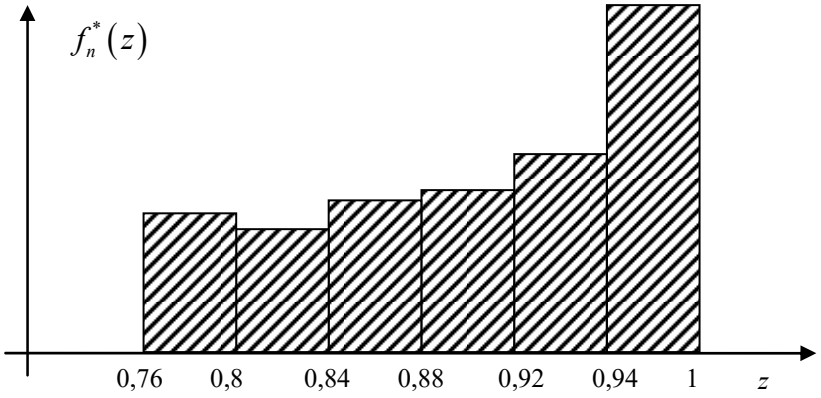

Рис. 1.3

Приклад 1.5. Знайти $M\xi_{(k)}$ та $D\xi_{(k)}$, якщо варіаційний ряд $\xi_{(1)} \le \xi_{(2)} \le ... \le \xi_{(n)}$ отримано для вибірки $\xi_1,...,\xi_n$ з рівномірного на $(0,a)$ розподілу.

Розв'язання. Використовуючи схему незалежних випробувань Бернуллі, де успіхом будемо вважати потрапляння точки ξ_i ліворуч від z, знаходимо функцію розподілу $F_k(z)$ k-ї порядкової статистики

$$F_k(z) = P\{\xi_{(k)} \le z\} = \sum_{i=k}^{n} C_n^i \left(\frac{z}{a}\right)^i \left(1-\frac{z}{a}\right)^{n-i},$$

оскільки ймовірність того, що точно i з n точок опиняться ліворуч (а отже, $n-i$ точок – праворуч) від z, становить

$C_n^i\left(\frac{z}{a}\right)^i\left(1-\frac{z}{a}\right)^{n-i}$, а $P\{\xi_{(k)} \le z\}$ – це ймовірність того, що принаймні k точок опиняться ліворуч від z ($i = k, k+1, ..., n$).

Продиференціювавши $F_k(z)$, отримаємо вираз для щільності:

$$f_{(k)}(z) = \frac{d}{dz}F_k(z) = \sum_{i=k}^{n} C_n^i \left[\frac{i}{a}\left(\frac{z}{a}\right)^{i-1}\left(1-\frac{z}{a}\right)^{n-i} - \frac{n-i}{a}\left(\frac{z}{a}\right)^{i}\left(1-\frac{z}{a}\right)^{n-i-1}\right] =$$

$$= \frac{n}{a}C_{n-1}^{k-1}\left(\frac{z}{a}\right)^{k-1}\left(1-\frac{z}{a}\right)^{n-k}.$$

Тепер неважко підрахувати:

$$\mathrm{M}\xi_{(k)} = \int_0^a z f_k(z)dz = \frac{ka}{n+1}, \mathrm{M}\xi_{(k)}^2 = \int_0^a z^2 f_k(z)dz = \frac{a^2k(k+1)}{(n+2)(n+1)},$$

$$D\xi_{(k)} = \mathrm{M}\xi_{(k)}^2 - \left(\mathrm{M}\xi_{(k)}\right)^2 = \frac{k(n-k+1)a^2}{(n+1)^2(n+2)}.$$

Відповідь: $\mathrm{M}\xi_{(k)} = \frac{ka}{n+1}$; $D\xi_{(k)} = \frac{k(n-k+1)a^2}{(n+1)^2(n+2)}$.

ЗАДАЧІ

1.1. За реалізацією вибірки 3, 7, 9, 6, 7, 8, 4, 8, 2, 4, 2, 5, 7, 3, 4, 2, 0, 9, 4, 7, 3, 5, 1, 8, 3, 6, 0, 7, 5, 7, 4, 3, 2, 1, 4, 3, 8, 6, 9, 9 побудувати реалізацію емпіричної функції розподілу, знайти вибіркове середнє, вибіркову дисперсію, медіану.

1.2. За реалізацією вибірки 3, 7, 5, 6, 4, 1, 3, 4, 8, 6, 5, 8, 3, 4, 9, 0, 7, 6, 5, 7, 3, 5, 3, 1, 8, 5, 9, 3, 2, 1, 6, 8, 6, 5, 4, 3, 6, 5, 6, 7, 8, 9, 3, 4, 5, 6, 3, 2, 7, 6, 8, 4, 5, 1, 3, 4, 5, 6, 3, 1 побудувати реалізацію емпіричної функції розподілу, знайти вибіркове середнє, вибіркову дисперсію, медіану.

1.3. За реалізацією вибірки побудувати гістограму, розбивши область значень на шість інтервалів однакової довжини. Знайти вибіркове середнє, вибіркову дисперсію, медіану.

0,92	0,84	0,91	0,86	0,85	0,86	0,99	0,92	0,84	0,95
0,88	0,84	0,84	0,87	0,88	0,99	0,86	0,85	0,92	0,85
0,92	0,90	0,93	0,89	0,96	0,91	0,95	0,92	0,85	0,93
0,88	0,94	0,93	0,86	0,86	0,90	0,90	1,00	0,89	0,86
0,86	0,85	0,85	0,87	0,91	0,98	0,85	0,86	0,85	0,84
0,90	1,00	0,94	0,91	0,85	0,93	0,84	0,84	0,84	0,92
0,86	0,89	0,90	0,88	0,84	0,89	0,86	0,84	0,86	0,86
0,86	0,94	0,86	0,89	0,86	0,96	0,93	0,96	0,98	0,85
0,88	0,86	0,87	0,98	0,84	0,92	0,96	0,94	0,92	0,99
0,92	0,90	0,91	0,98	0,95	0,86	0,85	0,85	0,87	0,86

1.4. За реалізацією вибірки побудувати гістограму, розбивши область значень на шість інтервалів однакової довжини. Знайти вибіркове середнє, вибіркову дисперсію, медіану.

0,84	1,00	0,94	0,92	0,85	0,93	0,84	0,90	0,84	0,91
0,85	0,84	0,84	0,85	0,88	0,99	0,86	0,88	0,92	0,87
0,92	0,90	0,93	0,93	0,96	0,91	0,95	0,92	0,85	0,89
1,00	0,94	0,93	0,86	0,86	0,90	0,90	0,88	0,89	0,86
0,86	0,85	0,85	0,84	0,91	0,98	0,85	0,86	0,85	0,87
0,92	0,84	0,91	0,95	0,85	0,86	0,99	0,92	0,84	0,86
0,84	0,89	0,90	0,86	0,84	0,89	0,86	0,86	0,86	0,88
0,96	0,94	0,86	0,85	0,86	0,96	0,93	0,86	0,98	0,89
0,94	0,86	0,87	0,99	0,84	0,92	0,96	0,88	0,92	0,98
0,85	0,90	0,91	0,86	0,95	0,86	0,85	0,92	0,87	0,98

1.5. Кожного тижня вівся підрахунок кількості деталей, необхідних для ремонту комп'ютерів фірми. За 30 тижнів були отримані такі значення:

1, 1, 0, 0, 1, 0, 1, 0, 1, 2, 2, 4, 1, 1, 0, 2, 1, 2, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 2.

Побудувати емпіричну функцію розподілу та знайти $\sup_z \left|F(z) - F_n(z)\right|$, вважаючи, що кількість деталей має розпо-

діл Пуассона з параметром 1. Обчислити вибіркове середнє, вибіркову дисперсію й порівняти їх з відповідними теоретичними значеннями.

1.6. Для оцінювання ймовірності настання події було проведено 20 серій послідовних незалежних випробувань до першого успішного випробування. У результаті були отримані такі значення:

7, 7, 3, 10, 10, 0, 0, 6, 3, 2, 3, 2, 0, 2, 2, 7, 2, 6, 2, 2.

Побудувати емпіричну функцію розподілу та знайти $\sup_z \left|F(z) - F_n(z)\right|$ за умови, що $F(z)$ відповідає геометричному розподілу з параметром 0,2. Обчислити вибіркове середнє, вибіркову дисперсію й порівняти їх з відповідними теоретичними значеннями.

1.7. Для знаходження частоти події A здійснюють n незалежних випробувань. Знайти, за якого значення $P(A)$ дисперсія частоти буде найбільшою.

1.8. Виконано n незалежних вимірювань однієї величини. Похибки вимірів мають нормальний розподіл із нульовим математичним сподіванням. За оцінку невідомої дисперсії взяли $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\xi_i - \overline{\xi}\right)^2$, де $\overline{\xi} = n^{-1}\left(\xi_1 + ... + \xi_n\right)$. Знайти дисперсію випадкової величини $\hat{\sigma}^2$.

1.9. Довести, що для вибіркової дисперсії справедлива формула $B_2 = A_2 - A_1^2$, де A_1, A_2 – вибіркові моменти першого і другого порядку, відповідно.

1.10. Довести справедливість таких співвідношень: $B_3 = A_3 - 3A_2A_1 + 2A_1^3$, $B_4 = A_4 - 4A_3A_1 + 6A_2A_1^2 - 2A_1^4$, де A_i – вибірковий момент i-го порядку, B_i – центральний вибірковий момент i-го порядку.

1.11. Довести таку властивість вибіркового середнього:

$\sum_{i=1}^{n}\left(\xi_i - \overline{\xi}\right)^2 < \sum_{i=1}^{n}\left(\xi_i - a\right)^2$ при $a \neq \overline{\xi}$.

1.12. Довести, що для довільного $z \in R$:

а) $MF_n(z) = F(z)$;

б) $DF_n(z)=\dfrac{F(z)\left(1-F(z)\right)}{n}$, де $F_n\left(z\right)$ – емпірична функція розподілу, що побудована за вибіркою з функцією розподілу $F\left(x\right)$.

1.13. Нехай вибірка складається з незалежних у сукупності, однаково розподілених елементів, які мають неперервну функцію розподілу $F(z)$. Довести, що тоді функція розподілу k-ї порядкової статистики $\xi_{(k)}$

$$F_{\xi_{(k)}}\left(z\right)=P\left(\xi_{(k)}\le z\right)=\sum_{i=k}^{n}C_n^iF^i\left(z\right)\left(1-F\left(z\right)\right)^{n-i},$$

а якщо існує щільність розподілу елементів вибірки $f\left(z\right)$, то щільність розподілу k-ї порядкової статистики $\xi_{(k)}$

$$f_{\xi_{(k)}}\left(z\right)=nC_{n-1}^{k-1}F^{k-1}\left(z\right)\left(1-F\left(z\right)\right)^{n-k}f(z).$$

1.14. Нехай вибірка складається з незалежних, однаково розподілених елементів, які мають неперервну функцію розподілу $F(z)$, а величина ξ не залежить від вибірки й має таку саму функцію розподілу. Довести, що тоді:

1) $P\left\{\xi_{(k)}<\xi<\xi_{(k+1)}\right\}=1/(n+1)$, $k=\overline{0,n}$;

2) $P\left\{\xi\le\xi_{(k)}\right\}=k/(n+1)$, $k=\overline{0,n}$.

1.15. Знайти сумісну функцію розподілу порядкових статистик $\xi_{(n)}$, $\xi_{(m)}$ $(n<m)$. У випадку абсолютно неперервного розподілу знайти також сумісну щільність.

1.16. Нехай вибірка складається з незалежних, однаково розподілених елементів, які мають неперервну функцію розподілу $F(z)$. Треба довести, що статистики $\left(F\left(\xi_{(k)}\right)/F\left(\xi_{(k+1)}\right)\right)^k$, $1\le k<n$ незалежні, і знайти їхні розподіли.

1.17. Нехай вибірка складається з n незалежних, однаково розподілених елементів, які мають неперервну функцію розподілу $F(z)$, а $F_n(z)$ – емпірична функція розподілу. Довести, що

розподіл статистики $D_n = \sup\limits_{-\infty<z<\infty} \left|F_n(z) - F(z)\right|$ не залежить від вигляду $F(z)$.

1.18. Нехай $\xi_1, \xi_2, \ldots, \xi_n$, $\eta_1, \eta_2, \ldots, \eta_n$ – взаємно незалежні випадкові величини зі спільною неперервною функцією розподілу. Позначимо через $F_n(z)$ та $G_n(z)$ емпіричні функції розподілу для вибірок $(\xi_1, \xi_2, \ldots, \xi_n)$ та $(\eta_1, \eta_2, \ldots, \eta_n)$, відповідно. Знайти розподіл величини $D_n^* = \sup\limits_{-\infty<z<\infty} \left[F_n(z) - G_n(z)\right]$.

1.19. Нехай $\xi_1, \xi_2, \ldots, \xi_n$, $\eta_1, \eta_2, \ldots, \eta_n$ – взаємно незалежні випадкові величини зі спільною неперервною функцією розподілу. Позначимо через $F_n(z)$ та $G_n(z)$ емпіричні функції розподілу вибірок $(\xi_1, \xi_2, \ldots, \xi_n)$ та $(\eta_1, \eta_2, \ldots, \eta_n)$, відповідно. Знайти $\lim\limits_{n\to\infty} P\left\{\sqrt{\frac{n}{2}} D_n \le t\right\}$, де $D_n = \sup\limits_{-\infty<z<\infty} \left|F_n(z) - F(z)\right|$.

1.20. Нехай $\xi_1, \xi_2, \ldots, \xi_m$, $\eta_1, \eta_2, \ldots, \eta_n$ – взаємно незалежні випадкові величини зі спільною неперервною функцією розподілу. Позначимо через $F_m(z)$ та $G_n(z)$ емпіричні функції розподілу вибірок $(\xi_1, \xi_2, \ldots, \xi_m)$ та $(\eta_1, \eta_2, \ldots, \eta_n)$, відповідно. Знайти ймовірність події $\inf\limits_{0<G_n(z)<1} \left[F_m(z) - G_n(z)\right] > 0$.

1.21. Нехай $\xi_1, \xi_2, \ldots, \xi_n$ – незалежні величини такі, що ξ_i має геометричний розподіл з параметром p_i, тобто $P\{\xi_i = m\} = q_i^{m-1} p_i$, $q_i = 1 - p_i$, $m = 1, 2, \ldots$. Довести, що $\xi_{(1)}$ має геометричний розподіл з параметром $1 - q_1 q_2 \ldots q_n$.

1.22. Довести, що для експоненціального розподілу з параметром $\lambda = 1$ функція розподілу статистики $\xi_{(n)}$ становить $F_{(n)}(z) = \left(1 - e^{-z}\right)^n$. За допомогою цього результату довести, що при $n \to \infty$ $\xi_{(n)} - \ln(n)$ слабко збігається до граничної випадкової величини з функцією розподілу $\exp\left\{-e^{-z}\right\}$ $(-\infty \le z \le \infty)$.

1.23. Довести, що для експоненціального розподілу з параметром $\lambda = 1$ випадкові величини $Y_r = (n - r + 1)\left(\xi_{(r)} - \xi_{(r-1)}\right)$ ($r = 1,\ldots,n$, $\xi_{(0)} = 0$) незалежні й однаково розподілені (теж мають експоненціальний розподіл з параметром $\lambda = 1$).

1.24. Нехай $\xi_1, \xi_2, \ldots, \xi_n$ – вибірка з експоненціального розподілу з параметром $\lambda = 1$.

а) Знайти щільність розподілу статистики $\xi_{(r)}$.

б) Довести, що $\xi_{(r)}$ та $\xi_{(s)} - \xi_{(r)}$ ($s > r$) незалежні.

в) Який розподіл має $\xi_{(r+1)} - \xi_{(r)}$?

1.25. Нехай $\xi_1, \xi_2, \ldots, \xi_n$ – незалежні величини, причому кожна ξ_k має щільність розподілу $p_k(z)$.

а) Знайти щільність розподілу випадкової величини $\xi_{(n)}$.

б) Знайти функцію розподілу, математичне сподівання й дисперсію розмаху вибірки $W = \xi_{(n)} - \xi_{(1)}$.

1.26. Знайти функцію розподілу статистики $M = \frac{1}{2}\left(\xi_{(1)} + \xi_{(n)}\right)$ для вибірки розміром n, елементи якої мають неперервну функцію розподілу $F(z)$.

1.27. Для вибірки розміром $2m+1$, елементи якої мають абсолютно неперервний розподіл зі щільністю $f(z)$ та набувають значень із відрізка $[a,b]$ ($0 \le a \le z \le b$), знайти щільність статистики $\xi_{(2m+1)} / \xi_{(m+1)}$.

1.28. Довести, що для вибірки $(\xi_1,\ldots,\xi_n)$ розміром n $\operatorname{cov}(\overline{\xi}, S^2) = \frac{(n-1)}{n^2} b_3$, де $\overline{\xi} = n^{-1}(\xi_1 + \ldots + \xi_n)$, $S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\xi_i - \overline{\xi}\right)^2$, b_3 – теоретичний момент третього порядку.

1.29. Знайти сумісну щільність усіх порядкових статистик вибірки розміром n з абсолютно неперервного розподілу зі щільністю $f(z)$.

1.30. Довести, що для варіаційного ряду, побудованого за вибіркою з рівномірного на $[a;b]$ розподілу, сумісна щільність екстремальних статистик $\xi_{(1)}$ та $\xi_{(n)}$ має вигляд

$$f(z_1, z_2) = \frac{n(n-1)}{(b-a)^n}\left(z_2 - z_1\right)^{n-2},\ a \le z_1 \le z_2 \le b\ .$$

Для $\xi_{(1)}$ та $\xi_{(n)}$ знайти математичні сподівання, дисперсії та коваріацію.

1.31. Зробіть 50 підкидань грального кубика. За отриманою реалізацією вибірки побудуйте емпіричну функцію розподілу. На цьому ж графіку побудуйте теоретичну функцію розподілу. Визначте максимальне відхилення емпіричної функції розподілу від теоретичної. Якщо зробити ще 50 підкидань кубика і для отриманих 100 чисел знову побудувати емпіричну функцію розподілу, то максимальне відхилення емпіричної функції розподілу від теоретичної: а) збільшиться; б) зменшиться; в) залишиться тим самим; г) інше?

Розділ 2
ТОЧКОВЕ ОЦІНЮВАННЯ НЕВІДОМИХ ПАРАМЕТРІВ

2.1. Статистичні оцінки та загальні вимоги до них. Незсунені оцінки з мінімальною дисперсією

Розглянемо довільну параметричну ймовірнісно-статистичну модель

$$\mathbb{F} = \{F(z;\theta),\ \theta \in \Theta\},$$

яка відповідає схемі повторних незалежних спостережень над деякою випадковою величиною ξ_0. Нехай $\xi' = (\xi_1,...,\xi_n)$ – вибірка з розподілу ξ_0. Апріорна інформація про випадкову величину ξ_0, що спостерігається, полягає в тому, що її функція розподілу $F_{\xi_0}(z,\theta)$ є елементом заданої параметричної сім'ї функцій розподілу $\mathbb{F}$.

У загальному вигляді задача оцінювання параметра θ формулюється так: використовуючи статистичну інформацію, яка міститься у виборці ξ, зробити статистичні висновки про дійсне значення невідомого параметра θ. Отже, задача оцінювання параметра θ полягає в побудові наближених формул

$$\theta \approx T(\xi), \qquad (2.1)$$

де $T(x) = T(x_1,...,x_n)$ – вимірна функція на вибірковому просторі X, яка набуває значення з множини Θ. При цьому функція $T(x)$ має бути тією ж самою для всіх розподілів $F(z,\theta)$ з даної сім'ї $\mathbb{F}$.

Статистикою називається довільна випадкова величина $T = T(\xi)$, яка є функцією від вибірки $\xi' = (\xi_1,...,\xi_n)$. При точко-

вому оцінюванні шукають статистику $\hat{\theta} = T(\xi)$, значення якої при заданій реалізації $x' = (x_1, ..., x_n)$ вибірки ξ приймають за наближене значення параметра θ. У цьому випадку кажуть, що статистика $T(\xi)$ оцінює θ або статистика $T(\xi)$ є оцінкою θ.

Ясно, що для оцінювання θ можна використовувати різні оцінки і, щоб обрати найкращу з них, треба мати критерій якості оцінок. Цей критерій визначається вибором міри близькості оцінки до дійсного значення параметра. Якщо зафіксований клас оцінок і вибрана міра близькості, то оцінка, що мінімізує цю міру близькості, і буде оптимальною.

Довільна оцінка $T = T(\xi)$ є випадковою величиною, тому загальною вимогою до оцінок є вимога концентрації (у тому чи іншому розумінні) розподілу T навколо дійсного значення параметра, що оцінюється. Чим вище степінь цієї концентрації, тим краще відповідна оцінка.

Припустимо, що параметр θ скалярний. Статистика $T(\xi)$ називається ***незсуненою оцінкою*** параметра θ, якщо виконується умова

$$M_\theta T(\xi) = \theta \text{ для будь-якого } \theta \in \Theta. \qquad (2.2)$$

Для оцінок, які не задовольняють умову (2.2), величину

$$b(\theta) = M_\theta T(\xi) - \theta \qquad (2.3)$$

називають ***зсувом*** оцінки $T(\xi)$.

Сама по собі властивість незсуненості не є достатньою для того, щоб оцінка добре наближала невідомий параметр. Наприклад, перший елемент ξ_1 вибірки із закону Бернуллі буде незсуненою оцінкою для θ: $M\xi_1 = 0 \cdot (1-\theta) + 1 \cdot \theta = \theta$. Однак його можливі значення 0 та 1 навіть не належать множині $\Theta = (0,1)$.

Будемо називати середнім квадратом похибки оцінки $T(\xi)$ величину

$$M_\theta \left(T(\xi) - \theta\right)^2 = D_\theta T + b^2(\theta). \qquad (2.4)$$

Для незсунених оцінок середній квадрат похибки збігається з дисперсією $D_\theta T(\xi)$.

Зауваження. При точковому оцінюванні параметрів часто обмежуються класом незсунених оцінок, оскільки вимога незсуненості означає, що принаймні у середньому оцінка, що використовується, обумовлює бажаний результат. Треба зазначити також, що для незсунених оцінок критерієм точності (концентрації) оцінки є її дисперсія. Проте не слід перебільшувати значення незсуненості: у деяких випадках незсунених оцінок не існує; є випадки, коли середній квадрат похибки зсуненої оцінки менше, ніж незсуненої.

Якщо $\alpha_1 = M\xi < \infty$, то $\bar{\xi}$ є незсуненою оцінкою математичного сподівання α_1. Зауважимо, що якщо $D\xi_i = \sigma^2$, то $D\bar{\xi} = \frac{1}{n^2}\sum_{i=1}^{n} D\xi_i = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$. Перевіримо, чи буде незсуненою оцінка S^2 дисперсії:

$$MS^2 = \frac{1}{n}\sum_{i=1}^{n} M(\xi_i - \bar{\xi})^2 = \frac{1}{n}\sum_{i=1}^{n} M(\xi_i - \alpha_1 + \alpha_1 - \bar{\xi})^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n} M(\xi_i - \alpha_1)^2 + \frac{2}{n}\sum_{i=1}^{n} M(\xi_i - \alpha_1)(\alpha_1 - \bar{\xi}) + \frac{1}{n}\sum_{i=1}^{n} M(\alpha_1 - \bar{\xi})^2 =$$

$$= \frac{n\sigma^2}{n} - \frac{2}{n}\cdot nM(\bar{\xi} - \alpha_1)^2 + M(\bar{\xi} - \alpha_1)^2 =$$

$$= \sigma^2 - M(\bar{\xi} - \alpha_1)^2 = \sigma^2 - \frac{\sigma^2}{n} = \frac{\sigma^2(n-1)}{n}.$$

Тут використовувалась рівність

$$\sum_{i=1}^{n}(\xi_i - \alpha_1)(\bar{\xi} - \alpha_1) =$$

$$= (\bar{\xi} - \alpha_1)\sum_{i=1}^{n}(\xi_i - \alpha_1) = (\bar{\xi} - \alpha_1)(n\bar{\xi} - n\alpha_1) = n(\bar{\xi} - \alpha_1)^2.$$

Отже, S^2 – зсунена оцінка дисперсії. Покладемо $\widehat{S^2} = \frac{n}{n-1}S^2$. Тоді $M\widehat{S^2} = \frac{n}{n-1}MS^2 = \frac{n}{n-1}\cdot\sigma^2\cdot\frac{n-1}{n} = \sigma^2$, тобто $\widehat{S^2} = \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - \bar{\xi})^2$ є незсуненою оцінкою дисперсії.

Якщо математичне сподівання α_1 відоме, то незсуненою оцінкою дисперсії є оцінка

$$\overline{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - \alpha_1)^2 .$$

Для розширення застосувань передбачимо можливість оцінювання не тільки параметра θ, а й деякої заданої функції від нього $\tau(\theta)$. У цьому випадку статистика $T = T(\xi)$ є незсуненою оцінкою для $\tau(\theta)$, якщо виконується співвідношення

$$M_\theta T(\xi) = \tau(\theta) \text{ для будь-якого } \theta \in \Theta . \qquad (2.5)$$

Приклад 2.1. Для вибіркового контролю з партії готової продукції відібрано n приладів. Нехай $\xi_1,...,\xi_n$ – час їхньої роботи до відмови, причому ξ_i – незалежні однаково розподілені випадкові величини, які мають показниковий розподіл з невідомим параметром θ: $F_\theta(x) = 1 - e^{-\theta x}$, $x > 0$. Треба оцінити середній час до відмови приладу:

$$\varphi(\theta) = M\xi_1 = \theta\int_0^\infty xe^{-\theta x}dx = \frac{1}{\theta}\int_0^\infty ye^{-y}dy = \frac{1}{\theta} .$$

Використовуючи властивості математичного сподівання, можна показати, що вибіркове середнє $\overline{\xi}$ буде незсуненою оцінкою для функції $\varphi(\theta)$: $M\overline{\xi} = \varphi(\theta)$. Однак якщо спробувати оцінити сам параметр θ за допомогою $\hat{\theta} = 1/\overline{\xi}$, то отримаємо зсунену оцінку.

Дамо тепер формальне визначення поняття ***оптимальної оцінки***. Нехай треба оцінити задану параметричну функцію $\tau = \tau(\theta)$ у моделі $\mathbb{F} = \{F(x;\theta),\ \theta \in \Theta\}$ за статистичною інформацією, яка міститься у вибірці $\xi' = (\xi_1,...,\xi_n)$. Припустимо, що для даної задачі існують незсунені оцінки, тобто статистики $T(\xi)$, що задовольняють (2.5). Позначимо через $\mathfrak{I}_\tau$ клас усіх

незсунених оцінок. Додатково припустимо, що дисперсії всіх оцінок із класу $\mathfrak{I}_\tau$ скінченні,

$$D_\theta T(\xi) = M_\theta\left(T(\xi) - \tau(\theta)\right)^2 < \infty \text{ для } T(\cdot) \in \mathfrak{I}_\tau \text{ та } \theta \in \Theta .$$

Якщо для $T^*(\cdot) \in \mathfrak{I}_\tau$ виконується співвідношення

$$D_\theta T^* \le D_\theta T \text{ для всіх } T(\cdot) \in \mathfrak{I}_\tau \text{ та } \theta \in \Theta , \qquad (2.6)$$

то оцінку $T^*(\cdot)$ називають незсуненою оцінкою з рівномірно мінімальною дисперсією, або оптимальною оцінкою. Щоб підкреслити, що вона належить до функції $\tau(\theta)$, будемо також використовувати позначення τ^*. Таким чином, ***оптимальною*** в класі $\mathfrak{I}_\tau$ буде оцінка $\tau^* \in \mathfrak{I}_\tau$, для якої виконується умова

$$D_\theta \tau^* = \inf_{T \in \mathfrak{I}_\tau} D_\theta T \text{ для всіх } \theta \in \Theta . \qquad (2.7)$$

Зауваження. Вимога рівномірної мінімальності дисперсії досить сильна й не завжди має місце. Може виявитися, що з двох оцінок $T_1, T_2 \in \mathfrak{I}_\tau$ дисперсія $D_\theta T_1$ мінімальна в класі $\mathfrak{I}_\tau$ для одних значень параметра θ, а дисперсія $D_\theta T_2$ – для інших значень θ. У таких випадках за допомогою одного критерію мінімуму дисперсії ці оцінки порівняти неможливо.

Однак умова (2.6) виділяє оптимальну оцінку в класі $\mathfrak{I}_\tau$ однозначно.

Будемо казати, що дві статистики T_1 і T_2 дорівнюють одна одній, $T_1 = T_2$, якщо

$$P_\theta\left(\xi \in \left\{x : T_1(x) \ne T_2(x)\right\}\right) = 0 \text{ для всіх } \theta \in \Theta .$$

Використовуючи це поняття, сформулюємо теорему про єдиність оптимальної оцінки.

Теорема 2.1. *Нехай* $T_i = T_i(\xi), \quad i = 1,2$ *– дві оптимальні оцінки для* $\tau = \tau(\theta)$. *Тоді* $T_1 = T_2$.

Разом із властивостями незсуненості й оптимальності, що були введені при фіксованому обсязі вибірки, розглянемо асимптотичні властивості оцінок.

Оцінку $T(\xi)=T_n(\xi_1,...,\xi_n)=T_n$ можна розглядати як послідовність оцінок при $n\to\infty$. Послідовність оцінок T_n будемо називати ***асимптотично незсуненою*** при оцінюванні параметричної функції $\tau(\theta)$, якщо

$$\lim_{n\to\infty} M_\theta T_n=\tau(\theta) \text{ для всіх } \theta\in\Theta.$$

Послідовність оцінок T_n, $n=1,2,...$, параметричної функції $\tau(\theta)$ називається ***асимптотично нормальною*** з коефіцієнтом $\sigma^2(\theta)$, якщо

$$\sqrt{n}\left(T_n-\tau(\theta)\right)\underset{n\to\infty}{\overset{\text{сл}}{\Rightarrow}}\eta,$$

де випадкова величина η має нормальний розподіл $N\left(0,\sigma^2(\theta)\right)$.

Приклад 2.2. Нехай $(\xi_1,...,\xi_n)$ – вибірка з рівномірного розподілу на відрізку $[0,\theta]$, $\theta>0$. Перевіримо, чи є оцінки $\widehat{\theta_1}=2\cdot\overline{\xi}$ та $\widehat{\theta_2}=\xi_{(n)}=\max\{\xi_1,...,\xi_n\}$ асимптотично нормальними оцінками параметра θ.

Згідно з центральною граничною теоремою

$$\sqrt{n}\left(\widehat{\theta_1}-\theta\right)=\sqrt{n}\left(2\cdot\overline{\xi}-\theta\right)=\sqrt{n}\left(2\cdot\frac{\sum_{i=1}^{n}\xi_i}{n}-\theta\right)=$$

$$=\frac{\sum_{i=1}^{n}2\cdot\xi_i-n\theta}{\sqrt{n}}=\frac{\sum_{i=1}^{n}2\cdot\xi_i-n\cdot M_\theta(2\xi_1)}{\sqrt{n}}\Rightarrow\varsigma,$$

де ς має розподіл $N\left(0;D_\theta(2\xi_1)\right)=N\left(0;4D_\theta\xi_1\right)$, тобто оцінка $\widehat{\theta_1}=2\cdot\overline{\xi}$ є асимптотично нормальною з коефіцієнтом $\sigma^2(\theta)=4D_\theta\xi_1=4\theta^2/12=\theta^2/3$.

Для оцінки $\widehat{\theta_2} = \xi_{(n)}$ маємо:

$\sqrt{n}\left(\widehat{\theta_2} - \theta\right) = \sqrt{n}\left(\xi_{(n)} - \theta\right) < 0$ з імовірністю 1.

За означенням ξ_n слабко збігається до випадкової величини ς, якщо для будь-якої точки x, що є точкою неперервності граничної функції розподілу $F_{\varsigma}(x)$, має місце збіжність $F_{\xi_n}(x) \to F_{\varsigma}(x)$ при $n \to \infty$. Однак $P\left\{\sqrt{n}\left(\xi_{(n)} - \theta\right) < 0\right\} = 1$, а функція нормального розподілу $N\left(0, \sigma^2(\theta)\right)$ усюди неперервна й за нульового значення аргументу дорівнює 0,5. Очевидно, 1 не збігається до 0,5 при $n \to \infty$, тому слабка збіжність $\sqrt{n}\left(\widehat{\theta_2} - \theta\right)$ до нормально $N\left(0, \sigma^2(\theta)\right)$-розподіленої величини не має місця.

Отже, оцінка $\widehat{\theta_2} = \xi_{(n)}$ не є асимптотично нормальною.

Послідовність оцінок T_n параметричної функції $\tau(\theta)$ називається ***конзистентною*** для $\tau(\theta)$, якщо

$$T_n \underset{n\to\infty}{\overset{P}{\Rightarrow}} \tau(\theta) \text{ для всіх } \theta \in \Theta .$$

Якщо збіжність за ймовірністю замінити на збіжність з імовірністю 1, то матимемо сильну конзистентність.

Конзистентність послідовності оцінок означає концентрацію ймовірнісної маси навколо справжнього значення параметра при збільшенні розміру вибірки n.

Як установити, чи буде дана оцінка конзистентною? Зазвичай виявляється корисним один із трьох способів.

1) Конзистентність доводиться на основі безпосереднього обчислення функції розподілу оцінки (приклади 2.2 та 2.3).

2) Перевірка конзистентності спирається на використання закону великих чисел і лему 1.1 (наприклад, оцінка $\hat{\theta} = 1/\overline{\xi}$ із прикладу 2.1 буде конзистентною внаслідок неперервності функції $\varphi(x) = 1/x$ при $x > 0$).

3) При доведенні конзистентності використовується такий допоміжний результат.

Лема 2.1. *Якщо зсув* $b_n(\theta) = M_\theta \widehat{\theta}_n - \theta$ *і дисперсія* $D_\theta \widehat{\theta}_n$ *прямують до нуля при* $n \to \infty$, *то оцінка* $\widehat{\theta}_n$ *конзистентна.*

Приклад 2.3. Для випадкових величин ξ_i, $i = 1,...,n$, рівномірно розподілених на відрізку $[0,\theta]$, доведемо конзистентність оцінки $\widehat{\theta}_1 = \xi_{(n)} = \max\{\xi_1,...,\xi_n\}$ параметра θ:

а) безпосередньо обчислюючи функцію розподілу;

б) застосовуючи лему 2.1.

Доведення. а) Функція розподілу n-ї порядкової статистики становить $F_{\xi_{(n)}}(x) = (x/\theta)^n$ при $0 \le x \le \theta$. Оскільки $P\{\xi_{(n)} \le \theta\} = 1$, то для будь-якого $\varepsilon \in (0,\theta)$ маємо

$$P\left\{\left|\widehat{\theta}_1 - \theta\right| > \varepsilon\right\} = P\{\xi_{(n)} \le \theta - \varepsilon\} = (1 - \varepsilon/\theta)^n \to 0 \text{ при } n \to \infty.$$

б) Використовуючи функцію розподілу n-ї порядкової статистики $\xi_{(n)}$, можна знайти $M\widehat{\theta}_1 = M\xi_{(n)} = \dfrac{n\theta}{n+1} \to \theta$ та $D\widehat{\theta}_1 = \dfrac{n\theta^2}{(n+1)^2(n+2)} \to 0$.

Приклад 2.4. Для статистичної моделі з попереднього прикладу розглянемо оцінку $\widehat{\theta}_2 = (n+1)\xi_{(1)}$, де $\xi_{(1)} = \min\{\xi_1,...,\xi_n\}$. Використовуючи функцію розподілу першої порядкової статистики $\xi_{(1)}$, можемо записати $M\widehat{\theta}_2 = (n+1)M\xi_{(1)} = (n+1)\left(\theta - \dfrac{n\theta}{n+1}\right) = \theta$. Далі з незалежності випадкових величин ξ_i випливає, що при $n \to \infty$

$$P\left\{\widehat{\theta}_2 > \theta + \varepsilon\right\} = \prod_{i=1}^{n} P\left\{\xi_i > \frac{\theta+\varepsilon}{n+1}\right\} = \left(1 - \frac{\theta+\varepsilon}{\theta(n+1)}\right)^n \to e^{-(\theta+\varepsilon)/\theta} > 0.$$

Отже, оцінка $\widehat{\theta_2}$ є незсуненою, але не є конзистентною.

Приклад 2.5. Нехай $\xi_1, \xi_2, ..., \xi_n$ – вибірка з розподілу зі щільністю

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

Розглянемо оцінки $\hat{\theta}_1 = \max\{\xi_1, \xi_2, ..., \xi_n\}$, $\hat{\theta}_2 = \min\{\xi_1, \xi_2, ..., \xi_n\}$, $\hat{\theta}_3 = \dfrac{1}{n}\sum_{i-1}^{n} \xi_i$, $\hat{\theta}_4 = \dfrac{1}{2}(\xi_n + \xi_{n-1})$. Які з оцінок $\hat{\theta}_1, ..., \hat{\theta}_4$ і яких параметрів (можливо, відмінних від *a* та *b*) є незсуненими; конзистентними?

Розв'язання.

1) $\hat{\theta}_1 = \max\{\xi_1, \xi_2, ..., \xi_n\}$. Знайдемо розподіл оцінки $\hat{\theta}_1$.

$$F_{\hat{\theta}_1}(x) = P\left\{\hat{\theta}_1 \le x\right\} = P\left\{\max\{\xi_1, ..., \xi_n\} \le x\right\} =$$

$$= \prod_{i=1}^{n} P\{\xi_i \le x\} = \prod_{i=1}^{n} \int_{-\infty}^{x} f(y)\,dy = \left(\int_{-\infty}^{x} f(y)\,dy\right)^n,$$

звідки випливає, що $\hat{\theta}_1$ – абсолютно неперервна випадкова величина зі щільністю

$$f_{\hat{\theta}_1}(x) = \frac{d}{dx} F_{\hat{\theta}_1} = nf(x)\left(\int_{-\infty}^{x} f(y)\,dy\right)^{n-1} =$$

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

Тоді $M\hat{\theta}_1 = \int_{-\infty}^{\infty} x f_{\hat{\theta}_1}(x)\,dx = \dfrac{a+nb}{n+1}$. Отже, $\hat{\theta}_1$ не є незсуненою оцінкою ні параметра a, ні параметра b, але $M\hat{\theta}_1 \to b$ при

$n \to \infty$. Таким чином, $\hat{\theta}_1$ є асимптотично незсуненою оцінкою параметра b. З'ясуємо, чи є $\hat{\theta}_1$ конзистентною оцінкою параметра b. Для довільного $\varepsilon > 0$

$$P\left\{\left|\hat{\theta}_1 - b\right| > \varepsilon\right\} = \int\limits_{-\infty}^{b-\varepsilon} f_{\hat{\theta}_1}(x)\,dx + \int\limits_{b+\varepsilon}^{\infty} f_{\hat{\theta}_1}(x)\,dx =$$

$$= \int\limits_{a}^{b-\varepsilon} f_{\hat{\theta}_1}(x)\,dx = \left(1 - \frac{\varepsilon}{b-a}\right)^n,$$

що прямує до 0 при $n \to \infty$, тобто $\hat{\theta}_1$ є конзистентною оцінкою параметра b.

2) $\hat{\theta}_2 = \min\{\xi_1, \xi_2, ..., \xi_n\}$. Ця оцінка досліджується аналогічно попередній. Її щільність

$$f_{\hat{\theta}_2}(x) = \begin{cases} \dfrac{n(b-x)^{n-1}}{(b-a)^n}, & \text{якщо } x \in [a,b], \\ 0, & \text{якщо } x \notin [a,b]. \end{cases}$$

Звідси $M\hat{\theta}_2 = \dfrac{an+b}{n+1}$, тобто $\hat{\theta}_2$ – асимптотично незсунена оцінка параметра a.

Аналогічно $\hat{\theta}_1$, $\hat{\theta}_2$ є конзистентною оцінкою параметра a.

3) $\hat{\theta}_3 = \dfrac{1}{n}\sum\limits_{i=1}^{n} \xi_i$. Обчислимо математичне сподівання:

$$M\hat{\theta}_3 = M\frac{1}{n}\sum_{i=1}^{n}\xi_i = \frac{1}{n}\sum_{i=1}^{n} M\xi_i = \frac{a+b}{2},$$

отже $\hat{\theta}_3$ – незсунена оцінка параметра $m = \dfrac{a+b}{2}$. Згідно із законом великих чисел $\hat{\theta}_3$ збігається до m, тобто $\hat{\theta}_3$ є конзистентною оцінкою параметра m.

4) $\hat{\theta}_4 = \frac{1}{2}\left(\xi_n + \xi_{n-1}\right)$. Для цієї оцінки маємо

$$M\hat{\theta}_4 = \frac{1}{2}\left(M\xi_n + M\xi_{n-1}\right) = \frac{a+b}{2},$$

отже, $\hat{\theta}_4$ є незсуненою оцінкою параметра $m = \frac{a+b}{2}$. Щільність $f_{\hat{\theta}_4}(x)$ оцінки $\hat{\theta}_4$ шукаємо як згортку щільностей рівномірних розподілів і одержуємо:

$$f_{\hat{\theta}_4}(x) = \begin{cases} \dfrac{x-a}{(b-a)^2}, & \text{якщо } x \in \left[a, \dfrac{a+b}{2}\right], \\ \dfrac{4(b-x)}{(b-a)^2}, & \text{якщо } x \in \left[\dfrac{a+b}{2}, b\right], \\ 0, & \text{якщо } x \notin [a,b]. \end{cases}$$

Графік щільності $f_{\hat{\theta}_4}(x)$ (рис. 2.1):

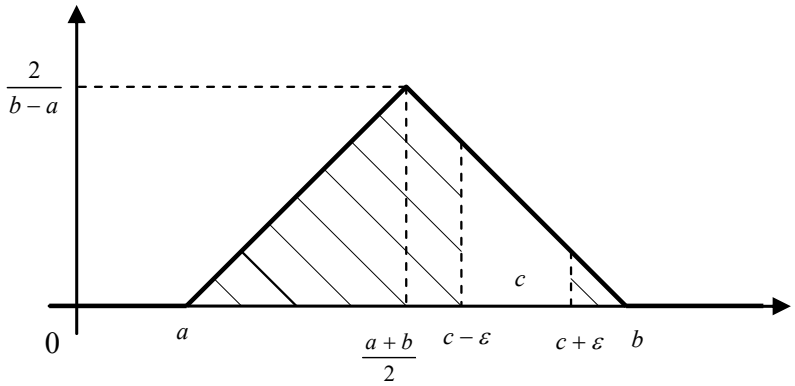

Рис. 2.1

Якщо припустити, що $\hat{\theta}_4$ збігається за ймовірністю до деякої константи c, то для достатньо малого $\varepsilon > 0$

$$P\left\{\left|\hat{\theta}_4 - c\right| > \varepsilon\right\} = \int\limits_{\{x:|x-c|>\varepsilon\}} f_{\hat{\theta}_4}(x)\,dx$$

є константою, що не залежить від n (заштрихована площа на рисунку) і, отже, не збігається до 0 при $n \to \infty$.

2.2. Оптимальна оцінка параметра в схемі Бернуллі

Проаналізуємо з погляду оптимальності оцінки невідомої ймовірності успіху в схемі Бернуллі.

Розглянемо схему незалежних спостережень над випадковою величиною ξ_0, розподіл якої належить множині розподілів $\left\{\theta^x(1-\theta)^{1-x},\ x=0 \text{ або } 1;\quad \theta \in (0,1)\right\}$, а $\xi' = (\xi_1,...,\xi_n)$ – вибірка з розподілу ξ_0. Необхідно оцінити параметр θ. Оскільки $M_\theta \xi_i = \theta$, то вибіркове середнє $\overline{\xi} = \dfrac{\xi_1 + ... + \xi_n}{n}$ є незсуненою оцінкою параметра θ.

Вибіркове середнє $\overline{\xi}$ не є єдиною незсуненою оцінкою. Наприклад, довільна статистика $T(\xi) = \dfrac{1}{n}\sum\limits_{i=1}^{n} b_{in}\xi_i$ при $b_{1n} + b_{2n} + ... + b_{nn} = n$ також є незсуненою оцінкою θ. Якщо $\sup\limits_{n,i}|b_{in}| \le b < \infty$, то ці оцінки є "хорошими" у тому розумінні, що

$$T(\xi) \underset{n\to\infty}{\overset{P}{\Rightarrow}} \theta, \tag{2.8}$$

тобто вони конзистентні. Дійсно,

$$D_\theta T(\xi) = \frac{1}{n^2}\sum_{i=1}^{n} b_i^2 \theta(1-\theta) \le \frac{b^2}{n}\theta(1-\theta) \underset{n\to\infty}{\to} 0$$

і (2.8) справедливо внаслідок нерівності Чебишова.

Таким чином, клас $\Im_\theta$ містить багато оцінок, і виникає задача вибору серед них найкращої.

Покажемо, що оптимальна оцінка T^* існує і $T^* = \overline{\xi}$.

Оскільки $D_\theta \xi_i = \theta(1-\theta)$, то $D_\theta \overline{\xi} = \dfrac{\theta(1-\theta)}{n}$, тобто залишилося показати, що для довільної незсуненої оцінки $T = T(\xi)$ параметра θ

$$D_\theta T \geq \frac{\theta(1-\theta)}{n} \text{ для всіх } \theta \in (0,1). \qquad (2.9)$$

Оскільки розподіл ξ_i задається ймовірностями $f(x;\theta) = \theta^x (1-\theta)^{1-x}$, $x = 0,1$, то розподіл випадкового вектора $\xi' = (\xi_1,...,\xi_n)$ задається ймовірностями

$$L(x;\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{n-\sum_{i=1}^{n} x_i}, \qquad (2.10)$$

де $x' = (x_1,...,x_n)$, $x_i = 0,1$.

Оскільки $1 \equiv \sum_x L(x;\theta)$ і $\theta = M_\theta T(\xi) = \sum_x T(x) L(x;\theta)$, то диференціюванням цих тотожностей за θ отримаємо:

$$0 = \sum_x \frac{\partial L(x;\theta)}{\partial \theta} = \sum_x \frac{\partial \ln L(x;\theta)}{\partial \theta} L(x;\theta) = M_\theta \left(\frac{\partial \ln L(\xi;\theta)}{\partial \theta} \right);$$

$$1 = \sum_x T(x) \frac{\partial \ln L(x;\theta)}{\partial \theta} L(x;\theta) = M_\theta \left(T(\xi) \frac{\partial \ln L(\xi;\theta)}{\partial \theta} \right).$$

Із цих двох рівностей знаходимо

$$M_\theta \left[\left(T(\xi) - \theta \right) \frac{\partial \ln L(\xi;\theta)}{\partial \theta} \right] = 1$$

і, згідно з нерівністю Коші – Буняковського,

$$M_\theta \left(\frac{\partial \ln L(\xi;\theta)}{\partial \theta} \right)^2 M_\theta \left(T(\xi) - \theta \right)^2 \geq 1.$$

Звідси випливає, що

$$D_\theta T(\xi) \geq 1 \Big/ M_\theta \left(\frac{\partial \ln L(\xi;\theta)}{\partial \theta} \right)^2 \quad \text{для всіх} \quad \theta \in (0,1).$$

Неважко підрахувати, що

$$\frac{\partial \ln L(x;\theta)}{\partial \theta} = \frac{1}{\theta} \sum_{i=1}^{n} x_i - \frac{1}{1-\theta} \left(n - \sum_{i=1}^{n} x_i \right) = \frac{1}{\theta(1-\theta)} \sum_{i=1}^{n} (x_i - \theta).$$

Тому

$$M_\theta \left(\frac{\partial \ln L(\xi;\theta)}{\partial \theta} \right)^2 = \frac{1}{\theta^2 (1-\theta)^2} M_\theta \left[\sum_{i=1}^{n} (\xi_i - \theta) \right]^2 =$$

$$= \frac{n}{\theta^2 (1-\theta)^2} D_\theta \xi_1 = \frac{n}{\theta(1-\theta)}.$$

Отже, доведено такий результат.

Теорема 2.2. *Відносна частота події у n незалежних випробуваннях є оптимальною оцінкою для ймовірності цієї події.*

2.3. Нерівність Рао – Крамера й ефективні оцінки

Розглянемо схему повторних незалежних спостережень над випадковою величиною ξ_0. Нехай $f(z,\theta)$, $\theta \in \Theta$, – щільність розподілу ξ_0, $\xi' = (\xi_1,...,\xi_n)$ – вибірка з $\mathbb{F} = \{ f(z,\theta), \theta \in \Theta \}$ і $x' = (x_1,...,x_n)$ – реалізація ξ.

Функція $L(\xi,\theta) = f(\xi_1,\theta) \times ... \times f(\xi_n,\theta)$ називається ***функцією вірогідності***. Реалізація функції вірогідності $L(x,\theta) = f(x_1,\theta) \times ... \times f(x_n,\theta)$ – це щільність розподілу випадкового вектора ξ.

Далі будемо припускати, що $L(x,\theta)>0$ при всіх $x\in X$, $\theta\in\Theta$ і $L(x,\theta)$ диференційована за параметром θ.

Нехай параметр θ – скалярний. Випадкова величина

$$U(\xi;\theta)=\frac{\partial \ln L(\xi;\theta)}{\partial\theta}=\sum_{i=1}^{n}\frac{\partial \ln L(\xi_i;\theta)}{\partial\theta} \qquad (2.11)$$

називається ***внеском*** вибірки ξ, а i-й доданок $\dfrac{\partial \ln L(\xi_i;\theta)}{\partial\theta}$ – внеском i-го спостереження. Будемо вважати, що

$$0<M_\theta U^2(\xi,\theta)<\infty \text{ для всіх } \theta\in\Theta.$$

Далі нам доведеться диференціювати за θ інтеграли від функцій на вибірковому просторі X і міняти місцями порядок інтегрування й диференціювання. Моделі, для яких ця операція коректна, називають коротко ***регулярними***. Точні аналітичні умови, які забезпечують регулярність моделі, відомі з математичного аналізу, а їхній вигляд визначається в кожному конкретному випадку. Обов'язкова умова полягає в тому, що вибірковий простір X не повинен залежати від невідомого параметра θ.

Приклад нерегулярної моделі. Нехай ξ_0 має рівномірний розподіл на $(0,\theta)$. З тотожності $\int_0^\theta \frac{1}{\theta}dz\equiv 1$ не випливає, що $\int_0^\theta \frac{\partial}{\partial\theta}\left(\frac{1}{\theta}\right)dz=0$, оскільки при диференціюванні інтеграла по верхній границі з'явиться ще один доданок. Причина нерегулярності полягає в тому, що вибірковий простір залежить від невідомого параметра θ.

Завжди має місце тотожність

$$\int_X L(x,\theta)dx\equiv 1 \quad (dx=dx_1...dx_n).$$

Якщо модель регулярна, то шляхом диференціювання цієї тотожності за θ маємо

$$0=\int_X \frac{\partial L(x,\theta)}{\partial\theta}dx=\int_X \frac{\partial \ln L(x,\theta)}{\partial\theta}L(x,\theta)dx=M_\theta U(\xi,\theta). \qquad (2.12)$$

Таким чином, для регулярної моделі

$$M_\theta U(\xi,\theta)=0 \text{ для всіх } \theta\in\Theta.$$

Функцію $I_n(\theta)=D_\theta U(\xi;\theta)=M_\theta U^2(\xi;\theta)$ називають ***функцією інформації Фішера*** про параметр θ, яка міститься у вибірці ξ. Величина $I_1(\theta)=I(\theta)=M\left(\frac{\partial \ln f(\xi_1;\theta)}{\partial\theta}\right)^2$ – це кількість фішерівської інформації, яка міститься в одному спостереженні. Для моделі повторних незалежних спостережень

$$I_n(\theta)=nI(\theta).$$

Кількість інформації, яка міститься у виборці, зростає пропорційно розміру вибірки.

Якщо функція $f(x,\theta)$ двічі диференційована за θ, то диференціюванням виразу (2.12) для $n=1$

$$0=\int_X \frac{\partial^2 \ln f(x,\theta)}{\partial\theta^2} f(x,\theta)\,dx+\int_X\left(\frac{\partial \ln f(x,\theta)}{\partial\theta}\right)^2 f(x,\theta)dx,$$

отримаємо таке подання для $I(\theta)$:

$$I(\theta)=-M_\theta\left(\frac{\partial^2 \ln f(\xi_1,\theta)}{\partial\theta^2}\right). \qquad (2.13)$$

Аналогічно $I_n(\theta)=-M_\theta\left(\frac{\partial^2 \ln L(\xi,\theta)}{\partial\theta^2}\right)$

Розглянемо тепер задачу оцінювання заданої параметричної функції $\tau(\theta)$ у моделі $\mathbb{F}=\{F(z,\theta),\theta\in\Theta\}$. Припустимо, що модель $\mathbb{F}$ регулярна, функція $\tau(\theta)$ – диференційована, і нехай $\mathfrak{I}_\tau$ – клас усіх незсунених оцінок $\tau(\theta)$, дисперсія яких скінченна. Тоді має місце таке твердження.

Теорема 2.3 (критерій Рао – Крамера). *Для довільної оцінки* $T = T(\xi) \in \Im_\tau$ *справедлива нерівність*

$$D_\theta T(\xi) \geq \frac{\left[\tau'(\theta)\right]^2}{I_n(\theta)}. \qquad (2.14)$$

Рівність у (2.14) має місце тоді й тільки тоді, коли T *– лінійна функція внеску вибірки:*

$$\frac{I_n(\theta)}{\tau'(\theta)}\left(T(\xi) - \tau(\theta)\right) = \pm U(\xi, \theta). \qquad (2.15)$$

Нерівність (2.14) називається нерівністю Рао – Крамера. Вона визначає нижню границю дисперсій усіх незсунених оцінок заданої параметричної функції $\tau(\theta)$ для регулярної моделі.

Якщо існує оцінка $T^*(\xi) \in \Im_\tau$, для якої нижня границя в нерівності Рао – Крамера досягається, то її називають ***ефективною***. Ефективна оцінка є оптимальною і, згідно з теоремою 2.1, єдиною. Критерієм ефективності оцінки є зображення (2.15). Будемо називати цей критерій оптимальності оцінки ***критерієм Рао – Крамера***.

*Зауваження***.** Внесок вибірки $U(\xi;\theta)$ однозначно визначається моделлю, тому зображення (2.15) (коли воно має місце) єдине й ефективна оцінка може існувати тільки для однієї певної параметричної функції $\tau(\theta)$. Вона не існує для жодної іншої функції параметра θ, яка відрізняється від $a\tau(\theta) + b$, де a та b – константи.

Приклад 2.6. ***Оцінка параметрів нормального розподілу***.

Нехай $\xi' = (\xi_1, \xi_2, ..., \xi_n)$ – вибірка з генеральної сукупності з нормальним розподілом $N(a, \theta)$. Розглянемо два випадки:

1) a – невідомий параметр, а дисперсія θ відома. Тоді

$$L(x, a) = (2\pi\theta)^{-n/2} \exp\left\{\frac{-\sum_{k=1}^{n}(x_k - a)^2}{2\theta}\right\},$$

$$\ln L(x,a) = -\frac{n}{2}\ln(2\pi\theta) - \frac{\sum_{k=1}^{n}(x_k - a)^2}{2\theta},$$

$$\frac{\partial \ln L(x,a)}{\partial a} = \frac{1}{\theta}\sum_{k=1}^{n}(x_k - a) = \frac{n}{\theta}\left(\frac{1}{n}\sum_{k=1}^{n}(x_k - a)\right) = \frac{n}{\theta}(\bar{x} - a),$$

тобто $\frac{\partial}{\partial a}\ln L(\xi,a) = \frac{n}{\theta}(\bar{\xi} - a)$. Середнє $\bar{\xi}$ є незсуненою, сильно конзистентною й асимптотично нормальною оцінкою математичного сподівання.

Підрахуємо $I_n(a)$:

$$I_n(a) = -M\frac{\partial^2}{\partial a^2}\ln L(\xi,a) = \frac{n}{\theta}, \quad D\bar{\xi} = \frac{\theta}{n} \text{ (див. підрозд. 2.1).}$$

Отже, $D\bar{\xi} = I^{-1}(a)$ і $\frac{\partial}{\partial a}\ln L(\xi,a) = I_n(a)(\bar{\xi} - a)$. Таким чином, оцінка $\bar{\xi} = \frac{1}{n}\sum_{k=1}^{n}\xi_k$ параметра a ефективна.

2) Нехай тепер параметр a відомий, будемо оцінювати параметр θ. Аналогічно

$$\ln L(\xi,\theta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\theta) - \frac{\sum_{k=1}^{n}(\xi_k - a)^2}{2\theta},$$

$$\frac{\partial \ln L(\xi,\theta)}{\partial \theta} = -\frac{n}{2\theta} + \frac{1}{2\theta^2}\sum_{k=1}^{n}(\xi_k - a)^2 =$$

$$= \frac{n}{2\theta^2}\left(\frac{1}{n}\sum_{k=1}^{n}(\xi_k - a)^2 - \theta\right) = \frac{n}{2\theta^2}(\bar{S}^2 - \theta).$$

Оцінка $\bar{S}^2$ є незсуненою, сильно конзистентною й асимптотично нормальною оцінкою дисперсії θ.

$$I_n(\theta) = -M\left(\frac{n}{2\theta^2} - \frac{1}{\theta^3}\sum_{k=1}^{n}(\xi_k - a)^2\right) = -\frac{n}{2\theta^2} + \frac{n\theta}{\theta^3} = \frac{n}{2\theta^2}.$$

Отже, $\frac{\partial}{\partial\theta}\ln L(\xi,\theta)=I_n(\theta)\left(\overline{S}^2-\theta\right)$ і $\overline{S}^2$ – ефективна оцінка параметра θ. Її дисперсія $D\overline{S}^2=\frac{2\theta^2}{n}$.

Приклад 2.7. ***Оцінка параметра показникового розподілу.***

Нехай $\xi_1,\xi_2,...,\xi_n$ – незалежні однаково розподілені випадкові величини зі щільністю розподілу $p(x,\theta)=\theta e^{-\theta x}$, $x\ge 0$, $M\xi_k=\frac{1}{\theta}$. Нехай $\hat{\theta}_n^{\,*}=\frac{1}{\overline{\xi}}=\frac{n}{\sum\limits_{k=1}^{n}\xi_k}$. Випадкова величина $\gamma_n=\sum\limits_{k=1}^{n}\xi_k$ має розподіл Ерланга, щільність якого $p_{\gamma_n}(x)=\frac{\theta^n x^{n-1}}{(n-1)!}e^{-\theta x}$, $x\ge 0$. Тоді

$$M\widehat{\theta_n^{\,*}}=\int_0^\infty\frac{n}{x}\frac{\theta^n x^{n-1}}{(n-1)!}e^{-\theta x}dx=\frac{n\theta}{(n-1)!}\int_0^\infty t^{n-2}e^{-t}dt=$$

$$=\frac{n\theta}{(n-1)!}\Gamma(n-1)=\frac{n\theta(n-2)!}{(n-1)!}=\frac{n\theta}{(n-1)}=\theta+\frac{\theta}{n-1},$$

де $\Gamma(\alpha)=\int\limits_0^\infty x^{\alpha-1}e^{-x}dx$ – гамма-функція.

Отже, $\hat{\theta}_n^{\,*}$ – зсунена оцінка параметра θ, $b(\theta)=\frac{\theta}{n-1}$ є зсувом цієї оцінки. Далі розглядатимемо $\hat{\theta}_n=\frac{n-1}{n}\cdot\hat{\theta}_n^{\,*}=\frac{n-1}{\sum\limits_{k=1}^{n}\xi_k}$ – незсунену оцінку параметра θ.

Підрахуємо $I_n(\theta)$:

$$L(\xi,\theta)=\theta^n \exp\left\{-\theta\cdot\sum_{k=1}^{n}\xi_k\right\}$$

$$\ln L(\xi,\theta)=n\cdot\ln\theta-\theta\cdot\sum_{k=1}^{n}\xi_k$$

$$\frac{\partial}{\partial\theta}\ln L(\xi,\theta)=\frac{n}{\theta}-\sum_{k=1}^{n}\xi_k;\quad \frac{\partial^2}{\partial\theta^2}\ln L(\xi,\theta)=-\frac{n}{\theta^2};\quad I_n(\theta)=\frac{n}{\theta^2}.$$

Далі підрахуємо $D\hat{\theta}_n$.

$$M\hat{\theta}_n^2=\int_0^\infty\frac{(n-1)^2}{x^2}\frac{\theta^n x^{n-1}}{(n-1)!}e^{-\theta x}dx=\frac{(n-1)^2\theta^2}{(n-1)!}\int_0^\infty t^{n-3}e^{-t}dt=$$

$$=\frac{(n-1)^2\theta^2}{(n-1)!}\Gamma(n-2)=\frac{(n-1)^2\theta^2}{(n-1)(n-2)}=\frac{(n-1)\theta^2}{n-2},$$

$$D\hat{\theta}_n=\frac{(n-1)\theta^2}{n-2}-\theta^2=\frac{\theta^2}{n-2}.$$

Тепер $\frac{1}{I_n(\theta)}=\frac{\theta^2}{n}$ та $D\hat{\theta}_n=\frac{\theta^2}{n-2}>\frac{\theta^2}{n}$ для $n=3,4,\ldots.$

Отже, оцінка $\hat{\theta}_n=\frac{1}{\bar{\xi}}=\frac{n-1}{\sum_{k=1}^{n}\xi_k}$ не є ефективною оцінкою параметра θ.

2.4. Принцип достатності й оптимальні оцінки

Критерій оптимальності Рао – Крамера має обмежене застосування з двох причин: 1) умова регулярності вихідної моделі; 2) його область застосувань пов'язана з виглядом параметричної функції $\tau(\theta)$.

Достатні статистики, які розглядаються в цьому підрозділі, є ще одним засобом побудови оптимальних оцінок.

Нехай $\mathbb{F}=\{F(x,\theta),\theta\in\Theta\}$ – параметрична модель, яка відповідає схемі повторних незалежних спостережень випадкової величини ξ_0. Статистика $T(\xi)'=(T_1(\xi),...,T_r(\xi))$ називається ***достатньою*** для моделі $\mathbb{F}$ (або достатньою для параметра θ, коли відомо, про яку модель ідеться), якщо умовна щільність (або умовна ймовірність) $L(x|t,\theta)$ випадкового вектора $\xi'=(\xi_1,\xi_2,...,\xi_n)$ за умови $T(\xi)=t$ не залежить від параметра θ.

Зазначена властивість статистики T означає, що ця статистика містить усю інформацію про параметр θ, яка є у вибірці. Тому всі висновки про цей параметр, які можна зробити за спостереженнями x, залежать тільки від $t=T(x)$. Усі статистичні висновки про модель, для якої існує достатня статистика, формулюються в термінах цієї достатньої статистики.

Отже, достатня статистика дає оптимальний спосіб подання статистичних даних, що особливо важливо при обробці великих масивів статистичної інформації. При цьому зазвичай прагнуть знайти достатню статистику мінімальної розмірності, тобто таку, яка подає дані в найбільш стислому вигляді. У цьому розумінні говорять про мінімальну достатню статистику. Очевидно, сама вибірка ξ завжди є достатньою статистикою, але ця статистика називається ***тривіальною***, оскільки не скорочує розмірність необхідних даних.

Наведемо результат, який дозволяє встановити факт існування достатньої статистики, а також знайти її вигляд.

Теорема 2.4 (критерій факторизації). *Для того, щоб статистика $T(\xi)$ була достатньою для θ, необхідно й достатньо, щоб функція вірогідності $L(x,\theta)$ мала вигляд*

$$L(x,\theta)=g(T(x);\theta)h(x), \qquad (2.16)$$

де множник g може залежати від θ, але від x залежить лише через $T(x)$; функції $h(x)$ та $T(x)$ від параметра θ не залежать.

Теорема 2.5 (Рао – Блекуелла – Колмогорова). *Оптимальна оцінка, якщо вона існує, є функцією від достатньої статистики.*

Приклад 2.8. Розглянемо два випадки знаходження достатніх статистик для нерегулярних моделей.

1) Нехай $\mathbb{F}=\{R(0,\theta),\theta\in(0,\infty)\}$, де $R(0,\theta)$ – рівномірний розподіл на відрізку $[0,\theta]$. Тоді реалізація функції вірогідності має вигляд

$$L(x,\theta)=\begin{cases}\dfrac{1}{\theta^n}, & \text{при } x_{(n)}=\max\limits_{1\le i\le n}x_i\le\theta,\\ 0 & -\text{ у протилежному випадку}\end{cases}$$

або $L(x,\theta)=\dfrac{e(\theta-x_{(n)})}{\theta^n}$, де $e(x)$ – функція Хевісайда.

На підставі критерію факторизації звідси маємо, що $T(\xi)=\xi_{(n)}$ – максимальне значення вибірки – є достатньою статистикою для θ.

2) Аналогічно для загальної рівномірної моделі

$$\mathbb{F}=\{R(\theta_1,\theta_2),-\infty<\theta_1<\theta_2<\infty\}$$

функція вірогідності може бути подана у вигляді

$$L(x,\theta)=\frac{e(\theta_2-x_{(n)})e(x_{(1)}-\theta_1)}{(\theta_2-\theta_1)^n}.$$

Таким чином, статистика $T=(T_1,T_2)$, де $T_1=\xi_{(1)}$, $T_2=\xi_{(n)}$, є достатньою для $\theta=(\theta_1,\theta_2)$.

Приклад 2.9. ***Достатня статистика для нормального розподілу*** $N(a,\theta)$.

$$L(x,a,\theta)=(2\pi\theta)^{-n/2}\exp\left\{-\frac{1}{2\theta}\sum_{k=1}^{n}(x_k-a)^2\right\}=$$

$$=(2\pi\theta)^{-n/2}\exp\left\{-\frac{1}{2\theta}\sum_{k=1}^{n}(x_k-\bar{x})^2-\frac{1}{2\theta}\sum_{k=1}^{n}(\bar{x}-a)^2\right\}.$$

Згідно з критерієм факторизації $T(\xi)=\left(\bar{\xi},\sum_{k=1}^{n}\left(\xi_k-\bar{\xi}\right)^2\right)$. Якщо дисперсія θ відома, то достатньою статистикою для a буде $\bar{\xi}$. Якщо ж, навпаки, відоме a, то достатньою статистикою для θ буде $\sum_{k=1}^{n}(\xi_k-a)^2$ або $\left(\sum_{k=1}^{n}\xi_k,\sum_{k=1}^{n}\xi_k{}^2\right)$.

ЗАДАЧІ

2.1. Нехай $\xi'=(\xi_1,...,\xi_n)$ – вибірка з генеральної сукупності з нормальним розподілом $N(a,\sigma^2)$. Довести, що статистика $T(\xi)=\frac{1}{n}\sqrt{\frac{\pi}{2}}\sum_{i=1}^{n}|\xi_i-a|$ є незсуненою оцінкою параметра σ (a – відомий параметр).

2.2. Нехай ξ_1,ξ_2 – два спостереження випадкової величини ξ_0 з нормальним розподілом $N(0,\sigma^2)$. Показати, що статистика $T(\xi)=\frac{\sqrt{\pi}}{2}|\xi_1+\xi_2|$ – незсунена оцінка параметра σ, $\xi'=(\xi_1,\xi_2)$.

2.3. Нехай ξ_1,ξ_2 – два спостереження випадкової величини ξ_0 з нормальним розподілом $N(a,\sigma^2)$. Показати, що статистика $T(\xi)=\frac{\sqrt{\pi}}{2}|\xi_1-\xi_2|$ – незсунена оцінка параметра σ, $\xi'=(\xi_1,\xi_2)$.

2.4. Нехай $\xi_1,...,\xi_n$ – вибірка з рівномірного розподілу на інтервалі $[a,b]$. Які з оцінок $\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n}\xi_i$, $\hat{\theta}_2 = \min\{\xi_i\}$, $\hat{\theta}_3 = \frac{1}{2}(\xi_{n-1} + \xi_n)$, $\hat{\theta}_4 = \max\{\xi_i\}$ є незсуненими оцінками параметрів (можливо, відмінних від a та b)? Яких саме?

2.5. Нехай $\xi_1,...,\xi_n$ – вибірка з розподілу зі щільністю

$$f(x,h,\theta) = \begin{cases} \frac{1}{2h}, \textit{якщо } x \in [\theta - h, \theta + h], \\ 0, \textit{якщо } x \notin [\theta - h, \theta + h]. \end{cases}$$

Які з оцінок $\hat{\theta}_1 = \min\xi_i - \frac{1}{n-1}(\max\xi_i - \min\xi_i)$, $\hat{\theta}_2 = \max\xi_i + \frac{1}{n-1}(\max\xi_i - \min\xi_i)$, $\hat{\theta}_3 = \frac{1}{2}(\max\xi_i + \min\xi_i)$, $\hat{\theta}_4 = \frac{1}{2}(\max\xi_i - \min\xi_i)$ є незсуненими оцінками параметрів θ та h? Можливо, серед них є незсунені оцінки інших параметрів? Яких саме?

2.6. Нехай $\xi_1,...,\xi_n$ – вибірка з розподілу зі щільністю

$$f(x,b,\theta) = \begin{cases} \frac{x}{\theta}\exp\{-x^2/2\theta\}, \textit{якщо } x \geq b, \\ 0, \quad \textit{якщо } x < b. \end{cases}$$

Чи є серед оцінок $\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n}\xi_i$, $\hat{\theta}_2 = \min\xi_i$, $\hat{\theta}_3 = \hat{\theta}_1 - \hat{\theta}_2$ незсунені оцінки параметрів θ та b? Можливо, серед них є незсунені оцінки інших параметрів? Яких саме?

2.7. Нехай $\xi_1,...,\xi_n$ – вибірка з розподілу Парето зі щільністю

$$f(x,\theta)=\begin{cases}\dfrac{\theta\lambda^{\theta}}{x^{\theta+1}}, \textit{ якщо } x\geq\lambda\\ 0, \quad \textit{якщо } x<\lambda,\end{cases}$$

$\lambda>0, \theta>2, \theta$ – відоме. Незсуненими оцінками яких параметрів є оцінки $\hat{\lambda}_1=\dfrac{\theta-1}{\theta n}\sum\limits_{i=1}^{n}\xi_i$, $\hat{\lambda}_2=\dfrac{\theta-2}{\theta n}\sum\limits_{i=1}^{n}\xi_i^{\,2}$?

2.8. Нехай $\xi'=(\xi_1,...,\xi_n)$ – вибірка з генеральної сукупності з рівномірним на відрізку $[a,b]$ розподілом. Побудувати незсунені оцінки параметрів a, b, а також незсунені й конзистентні оцінки параметрів $\dfrac{a+b}{2}$, $b-a$.

2.9. Нехай $\xi'=(\xi_1,...,\xi_n)$ – вибірка з генеральної сукупності з рівномірним на відрізку $[\theta,2\theta]$ розподілом. На який коефіцієнт треба помножити статистику $T(\xi)=\max\limits_{1\leq i\leq n}\xi_i-\min\limits_{1\leq i\leq n}\xi_i=\xi_{(n)}-\xi_{(1)}$, щоб отримати незсунену оцінку параметра θ?

2.10. Нехай $\xi'=(\xi_1,...,\xi_n)$ – вибірка з генеральної сукупності з рівномірним на відрізку $[\theta,\theta+1]$ розподілом. Побудувати незсунену оцінку параметра θ.

2.11. Нехай $\xi'=(\xi_1,...,\xi_n)$ – вибірка з генеральної сукупності зі щільністю

$$p(x,\theta)=\begin{cases}\exp\{-(x-\theta)\}, x\geq\theta\\ 0, \qquad\qquad x<\theta.\end{cases}$$

Показати, що $\xi_{(1)}=\min\limits_{1\leq i\leq n}\xi_i$ – достатня статистика. Довести, що оцінка $\widehat{\theta}_n=\min\limits_{1\leq i\leq n}\xi_i-\dfrac{1}{n}=\xi_{(1)}-\dfrac{1}{n}$ є незсуненою оцінкою параметра θ.

2.12. Нехай $\xi' = (\xi_1, ..., \xi_n)$ – вибірка з генеральної сукупності з розподілом Пуассона, $P\{\xi_0 = k\} = \dfrac{\theta^k}{k!} e^{-\theta}$, $k = 0,1,...$, $\theta > 0$. Показати, що статистика $\widehat{\theta_n} = \overline{\xi}$ є незсуненою ефективною оцінкою параметра θ.

2.13. Нехай $\xi' = (\xi_1, ..., \xi_n)$ – вибірка з генеральної сукупності з розподілом Паскаля, $P\{\xi = k\} = \dfrac{\theta^k}{(1+\theta)^{k+1}}$, $k = 0,1,...$, $\theta > 0$. Показати, що статистика $\widehat{\theta_n} = \overline{\xi}$ є незсуненою ефективною оцінкою параметра θ.

2.14. Нехай $\xi' = (\xi_1, ..., \xi_n)$ – вибірка з генеральної сукупності зі щільністю

$$f(x, \sigma^2) = \begin{cases} \dfrac{1}{\sqrt{2\pi\sigma^2} \cdot x} \exp\left\{-(\ln x - \mu)^2 / 2\sigma^2\right\}, & \textit{якщо } x > 0, \\ 0, & \textit{якщо } x \le 0, \end{cases}$$

μ – відоме, $\sigma^2 \in [a, b]$, $a > 0$ (логарифмічно нормальний розподіл). Чи буде оцінка $\hat{\sigma}^2 = \dfrac{1}{n} \sum\limits_{i=1}^{n} (\ln \xi_i - \mu)^2$ ефективною оцінкою параметра σ^2?

2.15. Нехай ξ_0 – випадкова величина, що має біноміальний розподіл і набуває скінченної кількості значень $0,1,...,n$ з імовірністю $P\{\xi_0 = x\} = C_n^x \theta^x (1-\theta)^{n-x}$, $x = 0,1,...,n$, $\theta \in (0,1)$. Довести, що статистика $\widehat{\theta_n} = \dfrac{\xi_0}{n}$ є незсуненою ефективною оцінкою параметра θ.

2.16. Нехай $\xi' = (\xi_1, ..., \xi_n)$ – вибірка з генеральної сукупності з біноміальним розподілом з параметрами m та p: $P\{\xi_0 = k\} = C_m^k p^k (1-p)^{m-k}$, $k = 0,1,...,m$, m – відоме. Чи є

$\hat{p}_n = \dfrac{1}{mn}\sum_{i=1}^{n}\xi_i$ конзистентною оцінкою параметра p? Чи буде вона ефективною?

2.17. Для показникової вибірки $\xi_1,...,\xi_n$ з функцією розподілу $F_\theta(t) = 1 - \exp\{-\theta t\}$, $t > 0$, покладемо $\eta_n = \xi_1 + ... + \xi_n$. Доведіть незсуненість оцінок:

а) $\hat{\theta} = (n-1)/\eta_n$;

б) $\hat{\varphi} = \left((1-t)/\eta_n\right)^{n-1} I_{\{\eta_n > t\}}$ для функції надійності $\varphi(\theta) = P\{\xi_1 > t\} = e^{-\theta t}$, де $t > 0$.

2.18. Випадкові величини $\xi_1,...,\xi_n$ незалежні та однаково розподілені зі щільністю $p(x,\theta) = \dfrac{1}{\theta}\exp\left\{-\dfrac{x}{\theta}\right\}$, $x \ge 0,\ \theta > 0$. Знайти достатню статистику й записати щільність її розподілу. Чи є оцінка $\hat{\theta}_n = \dfrac{1}{n}\sum_{i=1}^{n}\xi_i$ ефективною оцінкою параметра θ?

2.19. Випадкові величини $\xi_1,...,\xi_n$ незалежні та однаково розподілені зі щільністю $p(x,\theta) = C(\theta)\exp\{-\theta x\}$, $0 \le x \le \theta$. Знайти константу $C(\theta)$. Яке з наведених тверджень вірне:

а) існує лише тривіальна достатня статистика;

б) вектор $\left(\xi_{(n)}, \sum_{i=1}^{n}\xi_i\right)$ – достатня статистика;

в) вектор $\left(\xi_{(1)}, \sum_{i=1}^{n}\xi_i\right)$ – достатня статистика;

г) $\sum_{i=1}^{n}\xi_i$ – достатня статистика?

2.20. Знайти достатню статистику для параметра θ пуассонівського розподілу: $P\{\xi = k\} = \dfrac{\theta^k}{k!}e^{-\theta}$, $k = 0,1,...$, $\theta > 0$. Який розподіл має достатня статистика?

2.21. Випадкова величина ξ має логарифмічно нормальний розподіл з параметрами $\left(a,\sigma^2\right)$, якщо $\eta=\ln\xi$ має нормальний розподіл $N\left(a,\sigma^2\right)$. Знайти щільність розподілу ξ, $M\xi$, $D\xi$. Знайти достатню статистику для векторного параметра $\theta=\left(a,\sigma^2\right)$.

2.22. Знайти достатню статистику для параметра θ розподілу Ерланга:

$$f(x,m,\theta)=\begin{cases}\dfrac{x^{m-1}\cdot\exp\{-x/\theta\}}{\theta^m(m-1)!}, & \textit{якщо } x\geq 0,\\ 0, & \textit{якщо } x<0\end{cases}$$

(m – відоме).

2.5. Методи оцінювання невідомих параметрів

2.5.1. Оцінки максимальної вірогідності

Нехай задана вибірка $\xi'=\left(\xi_1,\xi_2,...,\xi_n\right)$ з розподілу випадкової величини ξ_0, $F_{\xi_0}\left(x\right)\in\mathbb{F}=\left\{F\left(x,\theta\right),\theta\in\Theta\right\}$ і $L\left(\xi,\theta\right)=f\left(\xi_1,\theta\right)\times...\times f\left(\xi_n,\theta\right)$ – функція вірогідності, де $f\left(x,\theta\right)$ – щільність розподілу ξ_0 для абсолютно неперервної моделі F і $f\left(x,\theta\right)=P_\theta\left(\xi_0=x\right)$ для дискретної моделі.

Оцінкою максимальної вірогідності (о.м.в.) $\hat\theta=\hat\theta\left(\xi\right)$ параметра θ називається точка параметричної множини Θ, у якій функція вірогідності досягає максимуму.

Якщо $\hat\theta=\hat\theta\left(x\right)$ – реалізація о.м.в., то

$$L\left(x,\hat\theta\right)\geq L\left(x,\theta\right)\ \text{для всіх}\ \theta\in\Theta\ \text{або}\ L\left(x,\hat\theta\right)=\sup_{\theta\in\Theta}L\left(x,\theta\right).$$

Розглянемо спочатку випадок, коли $\theta \in \Theta \subseteq R^1$ – одновимірний параметр. За умови, що для кожного $x \in X$ максимум $L(x,\theta)$ досягається у внутрішній точці Θ і $L(x,\theta)$ диференційована за θ, о.м.в. $\hat{\theta}$ задовольняє рівняння

$$\frac{\partial L(x,\theta)}{\partial \theta} = 0 \text{ або } \frac{\partial \ln L(x,\theta)}{\partial \theta} = 0, \qquad (2.18)$$

оскільки максимум функції $\ln L(x,\theta)$ досягається в тих самих точках, що й максимум функції $L(x,\theta)$.

Якщо θ – векторний параметр, $\theta' = (\theta_1,...,\theta_r)$, то рівняння (2.18) замінюється на систему рівнянь

$$\frac{\partial \ln L(x,\theta)}{\partial \theta_i} = 0 \quad i = 1,2,...,r. \qquad (2.19)$$

Рівняння (2.18) або (2.19) називаються рівняннями вірогідності.

Властивості оцінок максимальної вірогідності:

1) Якщо існує ефективна оцінка $T(\xi)$ для скалярного параметра θ, то вона збігається з о.м.в.: $\hat{\theta} = T(\xi)$. Це наслідок критерію ефективності Рао – Крамера

$$\frac{\partial \ln L(x,\theta)}{\partial \theta} = \frac{1}{a(\theta)}\left[T(x) - \theta\right].$$

2) Якщо $T = T(\xi)$ – нетривіальна достатня статистика, а оцінка максимальної вірогідності $\hat{\theta}$ існує та єдина, то вона є функцією від $T(\cdot)$. Дійсно, з критерію факторизації випливає, що в даному випадку максимізація $L(x,\theta)$ зводиться до максимізації $g(T(x),\theta)$ за θ. Отже, $\hat{\theta}$ залежить від статистичних даних через $T(x)$.

З 1) випливає, що прикладами оцінок максимальної вірогідності $\hat{\theta}$ для моделей зі скалярними параметрами служать ефективні оцінки.

Приклад 2.11. ***Нормальний розподіл*** $N(a,\theta)$. Знайдемо о.м.в. для параметрів (a,θ) нормального розподілу.

$$L(x,a,\theta)=(2\pi\theta)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\theta}\sum_{i=1}^{n}(x_i-a)^2\right\}.$$

Оскільки на границі Θ виконується $L(x,a,\theta)=0$, то максимум досягається у внутрішній точці. Рівняння вірогідності мають вигляд:

$$\frac{\partial \ln L(x,a,\theta)}{\partial a}=\frac{1}{\theta}\sum_{i=1}^{n}(x_i-a)=0,$$

$$\frac{\partial \ln L(x,a,\theta)}{\partial \theta}=-\frac{n}{2\theta}+\frac{1}{2\theta^2}\sum_{i=1}^{n}(x_i-a)^2=0.$$

Розв'язуючи ці рівняння відносно a та θ, отримаємо

$$\widehat{a_n}=\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\quad \widehat{\theta_n}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2=S^2,$$

тобто точка з координатами $\left(\widehat{a_n},\widehat{\theta_n}\right)$ максимізує значення функції $\ln L(x,a,\theta)$ і для параметрів нормального розподілу (a,θ) оцінкою максимальної вірогідності буде $\left(\overline{\xi},S^2\right)$.

Приклад 2.12. Нехай тепер $\xi'=(\xi_1,\xi_2,...,\xi_n)$ – вибірка з ***рівномірного розподілу*** $R(0,\theta)$. Щільність вектора ξ у цьому випадку

$$L(x,\theta)=\begin{cases}\dfrac{1}{\theta^n} & \text{при } x_{(n)}\le\theta,\\ 0 & \text{при } x_{(n)}>\theta.\end{cases}$$

Використовуючи функцію Хевісайда, матимемо $L(x,\theta)=\theta^{-n}e\left(\theta-x_{(n)}\right)$. З цієї формули видно, що $L(x,\theta)$ моно-

тонно спадає за θ для $\theta \geq x_{(n)}$. При $\theta = x_{(n)}$ $L(x,\theta)$ досягає максимуму. Таким чином, $\hat{\theta} = x_{(n)} = \max\limits_{1\leq i\leq n} x_i$ – реалізація оцінки максимальної вірогідності. У цій точці $L(x,\theta)$ має розрив і похідна не існує. Отже, о.м.в. не є розв'язком рівняння вірогідності, що характерно для випадків, коли вибірковий простір X залежить від невідомого параметра.

Приклад 2.13. ***Показниковий розподіл.*** Розглянемо вибірку $\xi' = (\xi_1, \xi_2, ..., \xi_n)$ з показникового розподілу:

$$p(x,\theta) = \theta e^{-\theta x}, \quad x \geq 0.$$

Функція вірогідності має вигляд $L(x,\theta) = \theta^n \exp\left\{-\theta \sum\limits_{i=1}^{n} x_i\right\}$, а рівняння вірогідності –

$$\frac{\partial \ln L(x,\theta)}{\partial \theta} = \frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0.$$

Отже, $\widehat{\theta_n} = \dfrac{n}{\sum\limits_{i=1}^{n} x_i}$. Оскільки $\dfrac{\partial^2 \ln L(x,\theta)}{\partial \theta^2} = -\dfrac{n}{\theta^2} < 0$, то $\widehat{\theta_n}$ – точка максимуму функції $\ln L(x,\theta)$. Таким чином, оцінкою максимальної вірогідності параметра θ буде $\left(\overline{\xi}\right)^{-1}$. Ця оцінка є зсуненою (див. приклад 2.7). Зауважимо, що коли показниковий розподіл задається у вигляді $p(x,\theta) = \dfrac{1}{\theta} e^{-x/\theta}$, $x \geq 0$, то о.м.в. $\widehat{\theta_n} = \overline{\xi}$ буде незсуненою та ефективною.

2.5.2. Асимптотичні властивості оцінок максимальної вірогідності

Мета цього пункту – показати, що найважливіші властивості оцінок максимальної вірогідності мають асимптотичний характер, тобто справедливі для великих вибірок.

Нехай $\xi' = (\xi_1, \xi_2, ..., \xi_n)$ – вибірка з розподілу випадкової величини ξ_0, $F_{\xi_0}(x) \in \mathbb{F} = \{F(x,\theta), \theta \in \Theta\}$, де множина Θ – невироджений відкритий інтервал; $f(x,\theta)$ – щільність розподілу ξ_0 для абсолютно неперервної моделі $\mathbb{F}$ і $f(x,\theta) = P_\theta(\xi_0 = x)$ для дискретної моделі; $L(x,\theta) = L(x_1,...,x_n;\theta) = f(x_1,\theta) \times ... \times f(x_n,\theta), x \in X$ – реалізація функції вірогідності.

Теорема 2.6. *Припустимо, що модель $\mathbb{F}$ є регулярною, а функція $L(x_1,...,x_n;\theta)$ за всіх $n \geq 1$ та $x \in X$ досягає глобального максимуму по θ усередині Θ. Припустимо також, що $f(x,\theta)$ тричі диференційовна за θ і при цьому існує функція $\mathbf{K}(x)$, яка не залежить від θ і така, що для всіх $\theta \in \Theta$*

$$\left| \frac{\partial^3 \ln f(x,\theta)}{\partial \theta^3} \right| \leq \mathbf{K}(x), \quad M_\theta \mathbf{K}(\xi_0) < \infty .$$

Тоді послідовність оцінок максимальної вірогідності $\hat{\theta}(\xi_1,...,\xi_n) = \hat{\theta}_n$ є конзистентною оцінкою параметра θ і для довільного $\theta \in \Theta$ $\left(\hat{\theta}_n - \theta\right)\sqrt{I_1(\theta)n} \underset{n\to\infty}{\overset{сл}{\Rightarrow}} \eta$, де випадкова величина η має стандартний нормальний розподіл.

Приклад 2.14. Нехай $\xi' = (\xi_1, \xi_2, ..., \xi_n)$ – вибірка з ***рівномірного розподілу*** на відрізку $[0,\theta]$. Як було показано раніше, оцінкою максимальної вірогідності параметра θ є $\hat{\theta}_n = \xi_{(n)} = \max_i \xi_i$. У цьому випадку

$$P\left\{\frac{n}{\theta}\left(\theta-\hat{\theta}_n\right)>x\right\}=P\left\{\xi_{(n)}<\theta\left(1-x/n\right)\right\}=\left[\frac{\theta\left(1-x/n\right)}{\theta}\right]^n\underset{n\to\infty}{\to}e^{-x},$$

тобто $\frac{n}{\theta}\left(\theta-\hat{\theta}_n\right)\underset{n\to\infty}{\overset{сл}{\Rightarrow}}\eta$, де випадкова величина η має показниковий розподіл з параметром $\lambda=1$. Таким чином, твердження теореми про асимптотичну нормальність несправедливе (не виконується умова теореми про регулярність моделі).

2.4.3. Асимптотична ефективність оцінок максимальної вірогідності

Розглянемо випадок незсуненої оцінки параметра $\hat{\theta}=T\left(\xi\right)$ у випадку регулярної моделі.

Згідно з нерівністю Рао – Крамера

$$D_\theta T\left(\xi\right)\geq\frac{1}{nI_1\left(\theta\right)}. \tag{2.23}$$

Права частина (2.23) є нижньою границею для дисперсій усіх незсунених оцінок $T\left(\xi\right)$ параметра θ. ***Ефективністю*** оцінки T $\left(eff\left(T\right)\right)$ будемо називати відношення цього мінімального значення дисперсії до дійсного значення дисперсії T, тобто

$$eff\left(T\right)=\frac{1}{nI_1\left(\theta\right)D_\theta T}. \tag{2.24}$$

Унаслідок нерівності Рао – Крамера $0\leq eff\left(T\right)\leq 1$ і, якщо оцінка T ефективна, то $eff\left(T\right)=1$.

У випадку, коли для послідовності оцінок T_n існує границя $\lim\limits_{n\to\infty}eff\left(T_n\right)=eff_0$, вона називається ***граничною ефективністю*** послідовності оцінок T_n. Якщо $eff_0=1$, то послідовність оцінок T_n називається ***асимптотично ефективною***.

Розглянемо тепер оцінки максимальної вірогідності $\hat{\theta}_n$ у ситуації, коли виконується попередня теорема. Тоді

$$\eta_n = \sqrt{I_1(\theta)n}\left(\hat{\theta}_n - \theta\right) \underset{n\to\infty}{\overset{сл}{\Rightarrow}} \eta, \tag{2.25}$$

де η – випадкова величина, що має нормальний розподіл з параметрами 0, 1.

Нехай окрім збіжності (2.25) має місце збіжність

$$D_\theta \eta_n \underset{n\to\infty}{\to} D\eta = 1. \tag{2.26}$$

Тоді з (2.26) маємо $D_\theta \hat{\theta}_n = \dfrac{1}{nI_1(\theta)} + o\left(\dfrac{1}{n}\right)$, тобто

$$\lim_{n\to\infty} eff\left(\hat{\theta}_n\right) = \lim_{n\to\infty} \frac{1}{nI_1(\theta) D_\theta \hat{\theta}_n} = 1.$$

Отже, оцінки максимальної вірогідності в регулярних моделях при виконанні умов теореми 2.6 і умови (2.26) асимптотично ефективні.

2.5.4. Метод моментів

Нехай $\xi' = (\xi_1, \xi_2, ..., \xi_n)$ – вибірка з розподілу випадкової величини ξ_0 і $F_{\xi_0}(x) \in \mathbb{F} = \{F(x,\theta), \theta \in \Theta\}$, де $\theta' = (\theta_1, \theta_2, ..., \theta_r)$ і $\Theta \subseteq R^r$. Припустимо, що для випадкової величини ξ_0, що спостерігається, існують перші r моментів

$$a_k = M_\theta \xi_0^k, \quad k = 1, 2, ..., r.$$

Ці моменти є функціями від невідомих параметрів: $a_k = a_k(\theta) = a_k(\theta_1, \theta_2, ..., \theta_r)$.

Нехай $x' = (x_1, x_2, ..., x_n)$ – реалізація вибірки ξ. Метод моментів полягає у прирівнюванні теоретичних і вибіркових моментів (точніше реалізацій вибіркових моментів), а реалізації оцінок

параметрів $\theta_1, \theta_2, ..., \theta_r$ за методом моментів знаходять як розв'язок системи рівнянь

$$a_k(\theta) = \frac{1}{n}\sum_{i=1}^{n} x_i^k, \quad k = 1, 2, ..., r. \tag{2.28}$$

Оцінки параметрів $\theta_1, \theta_2, ..., \theta_r$, знайдені за методом моментів, зазвичай є конзистентними, але часто неефективними. Тому їх можна використовувати як перше наближення, а далі шукати оцінки з меншою дисперсією. Зазначимо також, що метод моментів не можна застосовувати, коли теоретичні моменти необхідного порядку не існують.

Приклад 2.15. У випадку ***показникової моделі*** $M_\theta \xi_0 = a_1(\theta) = \frac{1}{\theta}$. Тому рівняння (2.28) має вигляд $\theta^{-1} = \overline{x}$. Звідси $\theta = \overline{x}^{-1}$ – реалізація оцінки методу моментів.

Приклад 2.16. Нехай $\xi_1, \xi_2, ..., \xi_n$ – незалежні однаково розподілені випадкові величини зі щільністю $p(x, \theta) = \frac{2x}{\theta^2} e^{-x^2/\theta^2}$, $x \geq 0$, $\theta > 0$. Знайдемо методом моментів оцінку параметра θ. Спочатку підрахуємо

$$M\xi = \int_0^\infty \frac{2x}{\theta^2} e^{-x^2/\theta^2} dx = 2\int_0^\infty te^{-t} \frac{\theta}{2\sqrt{t}} dt = \theta \int_0^\infty \sqrt{t} e^{-t} dt = \theta \cdot \Gamma\left(\frac{3}{2}\right) = \frac{\theta}{2}\sqrt{\pi}.$$

Прирівнюємо цей вираз до вибіркового середнього

$$\frac{\theta}{2}\sqrt{\pi} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Звідси знаходимо оцінку параметра θ: $\hat{\theta}_n = \frac{2\overline{\xi}}{\sqrt{\pi}} = \frac{2}{n\sqrt{\pi}}\sum_{i=1}^{n} \xi_i$.

Оцінка $\hat{\theta}_n$ є незсуненою та конзистентною.

Приклад 2.17. ***Рівномірний розподіл.*** Нехай тепер

$$p(x,\theta)=\begin{cases}\dfrac{1}{2\theta} & \text{при } x\in[-\theta,\theta],\\ 0 & \text{при } x\notin[-\theta,\theta],\end{cases}\qquad \theta>0.$$

Тоді $M\xi_k=\int\limits_{-\theta}^{\theta}\dfrac{x}{2\theta}dx=0$ і не залежить від θ.

Підрахуємо $M\xi_k^2=\int\limits_{-\theta}^{\theta}\dfrac{x^2}{2\theta}dx=\dfrac{\theta^2}{3}$.

Отже, рівняння для знаходження оцінки має вигляд $\dfrac{\theta^2}{3}=\dfrac{1}{n}\sum\limits_{i=1}^{n}x_i{}^2$ і оцінка параметра θ буде $\hat\theta_n=\sqrt{\dfrac{3}{n}\sum\limits_{i=1}^{n}\xi_i{}^2}$. При цьому оцінка $\hat\theta_n^2=\dfrac{3}{n}\sum\limits_{i=1}^{n}\xi_i^2$ є незсуненою та конзистентною оцінкою θ^2.

Приклад 2.18. Розглянемо модель зсуненого показникового закону зі щільністю $p(x,\theta)=e^{-(x-\theta)}$, $x\ge\theta$, $\theta>0$. Тоді

$$M\xi_k=\int\limits_{\theta}^{\infty}xe^{-(x-\theta)}dx=\int\limits_{\theta}^{\infty}(y+\theta)e^{-y}dy=1+\theta.$$

З рівняння $1+\theta=\overline{\xi}$ знаходимо оцінку методу моментів $\hat\theta_n=\overline{\xi}-1$.

Знайдемо тепер оцінку максимальної вірогідності параметра θ. У цьому випадку функція вірогідності

$$L(x,\theta)=\prod_{i=1}^{n}p(x_i,\theta)=e^{-(x_1+\ldots+x_n)}e^{n\theta}I\left\{x_{(1)}\ge\theta\right\}.$$

Звідси отримуємо як о.м.в. $\tilde\theta_n=\xi_{(1)}$, яка відрізняється від оцінки методу моментів. Зауважимо, що тут $L(x,\theta)$ не є гладкою функцією, а отже, о.м.в. не можна шукати, прирівнюючи до нуля похідну функції вірогідності.

ЗАДАЧІ

2.23. Використовуючи метод моментів, знайти за вибіркою $\xi_1,...,\xi_n$, де $P\{\xi_k = m\} = e^{-\theta}\dfrac{\theta^m}{m!}$, $m = 0,1,...$, оцінку $\widehat{\theta_n}$ параметра θ. Чи буде ця оцінка незсуненою, конзистентною? Знайти також оцінку максимальної вірогідності параметра θ.

2.24. Нехай $\xi_1,...,\xi_n$ – вибірка з геометричного розподілу з параметром p. Оцінити параметр p методом моментів.

2.25. Нехай $\xi_1,...,\xi_n$ – вибірка з рівномірного на інтервалі $[a,b]$ розподілу:

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

Знайти оцінки параметрів a та b методом максимальної вірогідності. Чи будуть вони незсуненими? Конзистентними?

2.26. Побудувати за допомогою методу моментів конзистентні оцінки параметрів a та b за результатами n незалежних спостережень $\xi_1,...,\xi_n$, кожне з яких має нормальний розподіл $N(0,1)$ або $N(a,1)$ з імовірністю b та $1-b$, відповідно.

2.27. Нехай $\xi_1,...,\xi_n$ – вибірка з генеральної сукупності зі щільністю $p(x,\theta) = k(\theta)x^2e^{-x^3/\theta^3}$, $x \geq 0$, $\theta > 0$. Знайти функцію $k(\theta)$ та оцінку параметра θ за допомогою методу моментів. Чи буде оцінка незсуненою й конзистентною? Чи збігається вона з оцінкою максимальної вірогідності?

2.28. Методом максимальної вірогідності знайти за вибіркою $\xi_1,...,\xi_n$, де $P\{\xi_k = m\} = \dfrac{\theta^m}{(1+\theta)^{m+1}}$, $m = 0,1,2,...,\theta > 0$, оцінку параметра θ. Які властивості задовольняє ця оцінка?

2.29. Методом максимальної вірогідності знайти за вибіркою $\xi_1,...,\xi_n$, де

$$P\{\xi_k = m\} = \frac{(\theta-1)^m}{\theta^{m+1}},\ m = 0,1,2,...,\theta > 1,$$

оцінку параметра θ. Чи буде ця оцінка незсуненою, ефективною? Знайти достатню статистику для параметра θ.

2.30. Нехай $\xi_1,...,\xi_n$ – вибірка з генеральної сукупності зі щільністю $p(x,\beta,m) = \frac{\beta^m}{\Gamma(m)} x^{m-1} e^{-\beta x},\ x > 0,\ \beta > 0,\ m > 0$.

Знайти оцінки невідомих параметрів β та m за допомогою методу моментів. Нехай при $n = 10$ $\xi_1,...,\xi_n$ набули таких значень: 0,1; 0,4; 0,5; 0,7; 0,6; 0,1; 0,05; 0,8; 0,15; 0,1. Обчислити реалізації оцінок.

2.31. Методом максимальної вірогідності за вибіркою $\xi' = (\xi_1,...,\xi_n)$ з генеральної сукупності з розподілом $N(\theta, 2\theta)$ знайти оцінку параметра θ.

2.32. Методом максимальної вірогідності знайти оцінку параметра θ біноміального розподілу. Які властивості задовольняє ця оцінка? Чи буде вона збігатися з оцінкою методу моментів?

2.33. Методом максимальної вірогідності знайти за вибіркою $\xi_1,...,\xi_n$, що має розподіл Ерланга

$$f(x,m,\theta) = \begin{cases} \dfrac{x^{m-1} \cdot \exp\{-x/\theta\}}{\theta^m (m-1)!}, & \textit{якщо } x \geq 0, \\ 0, & \textit{якщо } x < 0, \end{cases}$$

оцінку параметра θ (m – відоме). Чи буде ця оцінка незсуненою, ефективною?

2.34. Нехай $\xi_1,...,\xi_n$ – вибірка з розподілу зі щільністю

$$f(x,b,a) = \begin{cases} \dfrac{1}{a} \exp\left\{-\dfrac{x-b}{a}\right\}, & \textit{якщо } x \geq b, \\ 0, & \textit{якщо } x < b. \end{cases}$$

Знайти оцінки параметрів a та b методом моментів.

2.35. Методом максимальної вірогідності знайти за вибіркою $\xi_1,...,\xi_n$, що має розподіл Релея

$$f(x,\theta)=\begin{cases}\dfrac{x}{\theta}\exp\left\{-x^2/2\theta\right\}, & \textit{якщо } x>0,\\ 0, & \textit{якщо } x\le 0,\end{cases}$$

оцінку параметра θ. Чи буде вона незсуненою, ефективною? Знайти достатню статистику для параметра θ.

2.36. Методом максимальної вірогідності та методом моментів знайти за вибіркою $\xi_1,...,\xi_n$, що має логарифмічно нормальний розподіл зі щільністю

$$f(x,a,\sigma^2)=\begin{cases}\dfrac{1}{\sqrt{2\pi\sigma^2}\cdot x}\exp\left\{-(\ln x-a)^2/2\sigma^2\right\}, & \textit{якщо } x>0,\\ 0, & \textit{якщо } x\le 0,\end{cases}$$

оцінку параметрів $\left(a,\sigma^2\right)$.

2.37. Нехай $\xi_1,...,\xi_n$ – вибірка з розподілу зі щільністю

$$f(x,\theta)=\frac{1}{2\theta}\exp\left\{-\frac{|x|}{\theta}\right\},\quad \theta>0.$$

Знайти оцінку дисперсії методом моментів.

2.48. Нехай випадкова величина ξ_0 – кількість невдач до появи r-го успіху в необмеженій послідовності незалежних випробувань Бернуллі з імовірністю успіху p в одному випробуванні. Випадкова величина ξ_0 має від'ємний біноміальний розподіл (розподіл Паскаля) з параметрами r, p: $P\{\xi=k\}=C_{r-1+k}^{r-1}p^r(1-p)^k$, $k=0,1,2,...$ Нехай $\xi_1,...,\xi_n$ – вибірка з такого розподілу, r – відоме. Оцінити параметр p методом моментів.

Вказівка. Загальну кількість невдач до r-го успіху можна подати у вигляді суми $\xi = \eta_1 + ... + \eta_r$, де $\eta_1, ..., \eta_r$ – незалежні випадкові величини, що мають геометричний розподіл з параметром p.

2.41. Методом максимальної вірогідності знайти за вибіркою $\xi_1, ..., \xi_n$, де

$$P\{\xi_k = m\} = C_{r-1+m}^{r-1} \left(\frac{1}{1+\theta}\right)^r \left(\frac{\theta}{1+\theta}\right)^m,$$

$m = 0, 1, 2, ..., \theta > 0$, r – відоме, оцінку параметра θ. Чи буде ця оцінка незсуненою, ефективною? Знайти достатню статистику для параметра θ.

Розділ 3
ІНТЕРВАЛЬНЕ ОЦІНЮВАННЯ

3.1. Розподіли математичної статистики, пов'язані з нормальним розподілом

Для зручності подальшого викладення матеріалу наведемо деякі розподіли, що використовуються у статистиці.

1) ***Нормальний розподіл*** з параметрами $\mu \in (-\infty, +\infty)$, $\sigma^2 > 0$ має щільність

$$f_\xi(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad -\infty < x < +\infty;$$

при цьому $M\xi = \mu$, $D\xi = \sigma^2$.

Розподіл $N(0,1)$ називають стандартним нормальним, його функцію розподілу позначають $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} \exp\left\{-\frac{t^2}{2}\right\} dt$. Її значення можна знайти в табл. 1.

Рівняння $\Phi(x_\alpha) = \alpha$, $\alpha \in (0,1)$, однозначно визначає квантиль x_α, при цьому $x_{1-\alpha} = -x_\alpha$. Квантилі для деяких значень α наведено в табл. 2 додатка.

2) ***Розподіл суми квадратів*** n незалежних випадкових величин $\xi_1, \xi_2, ..., \xi_n$, кожна з яких розподілена за стандартним нормальним законом розподілу, тобто розподіл випадкової величини $\chi^2(n) = \xi_1^2 + \xi_2^2 + ... + \xi_n^2$, називається ***розподілом Пірсона***, або χ^2 (***хі-квадрат***)***-розподілом*** із n ступенями сво-

боди. Якщо через $k_n(x)$ позначити його щільність, то можна підрахувати, що

$$k_n(x)=\frac{x^{\frac{n}{2}-1}}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}e^{-\frac{x}{2}},\quad x>0,$$

де $\Gamma(\lambda)=\int\limits_0^\infty x^{\lambda-1}e^{-x}dx\quad(\lambda>0)$ – гамма-функція.

Наведемо деякі властивості χ^2-розподілу.

1. *Стійкість за операцією додавання.* Випадкова величина $\chi^2(n)+\chi^2(m)$ має розподіл χ^2 із $n+m$ ступенями свободи.

2. *Моменти χ^2-розподілу.* Математичне сподівання та дисперсія становлять відповідно $M\chi^2(n)=n$ та $D\chi^2(n)=2n$.

3. *Асимптотична нормальність.* Випадкова величина при $n\to\infty$ асимптотично нормальна з $M\chi^2(n)=n$, $D\chi^2(n)=2n$:

$$\lim_{n\to\infty}P\left\{\frac{\chi^2(n)-n}{\sqrt{2n}}\le x\right\}=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{x}e^{-t^2/2}dt$$

для будь-якого $x\in\mathbb{R}^1$.

4. Якщо $\xi_1,\xi_2,...,\xi_n$ – незалежні, однаково розподілені випадкові величини з нормальним $N(a,\sigma^2)$-розподілом, то випадкова величина

$$\chi^2(n)=\sum_{i=1}^{n}\left(\frac{\xi_i-a}{\sigma}\right)^2$$

має χ^2-розподіл із n ступенями свободи.

Вправа. Показати, що при $n\ge 2$ максимум щільності χ^2-розподілу з n ступенями свободи досягається в точці $n-2$.

3) Нехай ξ та $\chi^2(n)$ – незалежні випадкові величини, причому ξ має розподіл $N(0,1)$, $\chi^2(n)$ – χ^2-розподіл із n ступенями свободи.

Розподілом Стьюдента, або t ***-розподілом*** із n ступенями свободи називається розподіл випадкової величини

$$t(n)=\frac{\xi}{\sqrt{\chi^2(n)/n}}.$$

Випадкова величина $t(n)$ має щільність

$$s_n(x)=\frac{1}{\sqrt{\pi n}}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\frac{1}{\left(1+\frac{x^2}{n}\right)^{\frac{n+1}{2}}},\quad -\infty<x<\infty.$$

Наведемо деякі властивості t-розподілу.

1. *Симетричність.* Якщо випадкова величина $t(n)$ має розподіл Стьюдента з n ступенями свободи, то і випадкова величина $-t(n)$ має такий самий розподіл.

2. *Асимптотична нормальність.* Розподіл Стьюдента з n ступенями свободи при $n\to\infty$ наближається до нормального розподілу з параметрами (0,1):

$$\lim_{n\to\infty}P\{t(n)\le x\}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-t^2/2}dt$$

для будь-якого $x\in\mathbb{R}^1$.

Зауважимо, що для невеликих значень n розподіл Стьюдента помітно відрізняється від нормального розподілу. Імовірності великих відхилень від середнього значення більші для розподілу Стьюдента, ніж для нормального розподілу.

3. У розподілі Стьюдента існують лише моменти порядку $m<n$, причому всі існуючі моменти непарного порядку дорівнюють нулю.

4) ***Розподілом Фішера (Фішера – Снедекора),*** або ***F-розподілом*** із (n,m) ступенями свободи називається розподіл випадкової величини

$$F(n,m)=\frac{\frac{1}{n}\chi^2(n)}{\frac{1}{m}\chi^2(m)},$$

де $\chi^2(n)$ та $\chi^2(m)$ – незалежні випадкові величини, що мають χ^2-розподіл відповідно з n та m ступенями свободи. Цей розподіл називають ще ***розподілом дисперсійного відношення***.

У табл. 6 додатка наведені величини α-квантилів розподілу Фішера для деяких значень α, n та m.

3.2. Визначення надійного інтервалу

Раніше ми розглядали точкові оцінки параметра θ у моделі

$$\mathbb{F}=\{F(z,\theta),\theta\in\Theta\}.$$

Довільна точкова оцінка – це функція $\hat{\theta}=T(\xi)$ вибірки $\xi'=(\xi_1,\xi_2,...,\xi_n)$, яка при кожній реалізації x вибірки ξ визначає одне значення параметра θ.

Можна задачу оцінювання поставити інакше: необхідно вказати такий інтервал, усередині якого з високою ймовірністю γ міститься справжнє значення параметра, що оцінюється. Імовірність γ відображує ступінь готовності миритися з можливістю похибки. При заданому γ довжина надійного інтервалу характеризує точність локалізації параметра, тому бажано обирати найменший інтервал.

При інтервальному оцінюванні шукають такі дві статистики

$$T_1=T_1(\xi)\ \text{і}\ T_2=T_2(\xi),\quad T_1<T_2,$$

для яких при заданому $\gamma \in (0,1)$ виконується умова

$$P_\theta\left(T_1(\xi) < \theta < T_2(\xi)\right) \geq \gamma \text{ для всіх } \theta \in \Theta. \qquad (3.1)$$

$\left(T_1(\xi), T_2(\xi)\right)$ називають γ-надійним інтервалом для θ, γ – рівень надійності, $T_1(\xi), T_2(\xi)$ – нижня й верхня границі інтервалу.

Отже, γ-надійний інтервал – це випадковий інтервал у параметричній множині $\Theta\left((T_1, T_2) \subset \Theta\right)$, який залежить тільки від вибірки ξ (не від θ) і охоплює значення невідомого параметра θ з імовірністю, яка не менша за γ.

Якщо ймовірність у лівій частині нерівності (3.1) прямує до γ при $n \to \infty$, то інтервал називається асимптотичним. Зазвичай довжина надійного інтервалу зростає при збільшенні коефіцієнта надійності γ та прямує до нуля зі збільшенням розміру вибірки n.

3.3. Побудова надійного інтервалу за допомогою центральної статистики

Нехай $\mathbb{F} = \{F(z,\theta), \theta \in \Theta\}$ – абсолютно неперервна модель та існує випадкова величина $G(\xi,\theta)$, яка залежить від θ і така, що:

1) розподіл $G(\xi,\theta)$ не залежить від θ;

2) при кожному $x \in X$ функція $G(x,\theta)$ неперервна і строго монотонна за θ.

Таку випадкову величину $G(\xi,\theta)$ називають ***центральною статистикою*** для θ.

Нехай для моделі $\mathbb{F}$ побудована центральна статистика $G(\xi,\theta)$ і $f_G(g)$ – її щільність розподілу. Функція $f_G(g)$ від параметра θ не залежить (умова 1)), тому для довільного $\gamma \in (0,1)$ можна обрати $g_1 < g_2$ (багатьма способами) так, що

$$P_\theta\{g_1 < G(\xi,\theta) < g_2\} = \int_{g_1}^{g_2} f_G(g)\,dg = \gamma \text{ для всіх } \theta \in \Theta. \qquad (3.2)$$

Визначимо при кожному $x \in X$ числа $T_1(x), T_2(x)$ $\left(T_1(x) < T_2(x)\right)$ як розв'язки за θ рівнянь

$$G(x,\theta) = g_i, \quad i = 1,2. \tag{3.3}$$

Однозначність визначення цих чисел забезпечує умова 2), яка накладається на функцію $G(x,\theta)$. Тоді нерівності

$$g_1 < G(x,\theta) < g_2$$

еквівалентні нерівностям

$$T_1(x) < \theta < T_2(x)$$

і (2.32) можна переписати у вигляді

$$P_\theta\left\{T_1(\xi) < \theta < T_2(\xi)\right\} = \gamma \text{ для всіх } \theta \in \Theta.$$

Отже, інтервал $\left(T_1(\xi), T_2(\xi)\right)$ є γ-надійним інтервалом для θ (рис. 3.1).

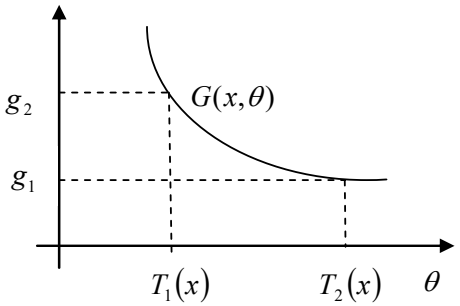

Рис. 3.1

3.4. Інтервальне оцінювання в нормальній моделі

3.4.1. Надійний інтервал для середнього, коли відома дисперсія

Нехай $\xi' = (\xi_1, \xi_2, ..., \xi_n)$ – вибірка з нормального розподілу $N(\theta, \sigma^2)$. Параметр θ невідомий, а σ^2 – відоме. Цю модель часто застосовують до даних, отриманих при незалежних вимірюваннях деякої величини θ за допомогою приладу (або методу), який має відому середню похибку (стандартну) σ. Для цієї моделі випадкова величина

$$G(\xi, \theta) = \sqrt{n}\,\frac{\overline{\xi} - \theta}{\sigma}, \quad \overline{\xi} = \frac{\xi_1 + \xi_2 + ... + \xi_n}{n},$$

буде центральною статистикою.

Дійсно, $G(\xi, \theta) = \dfrac{\sum_{i=1}^{n} (\xi_i - \theta)}{\sqrt{n}\,\sigma}$ як сума незалежних нормально розподілених випадкових величин буде розподілена нормально. Оскільки

$$M_\theta G(\xi, \theta) = 0, \quad D_\theta G(\xi, \theta) = 1,$$

то $G(\xi, \theta)$ має нормальний розподіл з параметрами (0,1), який, очевидно, не залежить від параметра θ.

При фіксованому x функція $G(x, \theta)$ неперервна й монотонно спадна за θ. Таким чином, умови 1), 2) виконуються.

Розв'язком рівнянь (2.33) будуть функції

$$T_1(x) = \overline{x} - \frac{\sigma}{\sqrt{n}} g_2, \quad T_2(x) = \overline{x} - \frac{\sigma}{\sqrt{n}} g_1,$$

а γ-надійний інтервал для θ можна записати формулою

$$\Delta_\gamma(\xi) = \left(\overline{\xi} - \frac{\sigma}{\sqrt{n}} g_2, \overline{\xi} - \frac{\sigma}{\sqrt{n}} g_1 \right),$$

де $g_1 < g_2$ – довільні числа, що задовольняють умову

$$\Phi(g_2) - \Phi(g_1) = \gamma,$$

$\Phi(\cdot)$ – функція розподілу стандартного нормального закону.

Хоч інтервал $\Delta_\gamma(\xi)$ випадковий, його довжина стала і становить

$$l_\gamma(g_1, g_2) = \frac{\sigma(g_2 - g_1)}{\sqrt{n}}.$$

Тому, щоб побудувати інтервал $\Delta_\gamma^*(\xi)$ мінімальної довжини, треба розв'язати задачу на умовний екстремум

$$\begin{cases} \dfrac{\sigma(g_2 - g_1)}{\sqrt{n}} \to \min, \\ \Phi(g_2) - \Phi(g_1) = \gamma. \end{cases}$$

Використовуючи метод невизначених множників Лагранжа, можна довести, що мінімум досягається при $g_1 = -g_2$.

Тепер g_2 визначається однозначно:

$$\Phi(g_2) - \Phi(-g_2) = \Phi(g_2) - (1 - \Phi(g_2)) = \gamma,$$

$$\Phi(g_2) = \frac{1+\gamma}{2}, \quad g_2 = c_{\frac{1+\gamma}{2}} = \Phi^{-1}\left(\frac{1+\gamma}{2}\right),$$

де $c_{\frac{1+\gamma}{2}} = \Phi^{-1}\left(\dfrac{1+\gamma}{2}\right)$ – квантиль порядку $\dfrac{1+\gamma}{2}$ для стандартного нормального розподілу (див. табл. 2 додатка).

У результаті отримуємо оптимальний γ -надійний інтервал

$$\Delta_\gamma^*(\xi) = \left(\overline{\xi} - \frac{\sigma}{\sqrt{n}} c_{\frac{1+\gamma}{2}}, \overline{\xi} + \frac{\sigma}{\sqrt{n}} c_{\frac{1+\gamma}{2}}\right).$$

3.4.2. Надійний інтервал для дисперсії, коли відоме середнє

Нехай $\xi' = (\xi_1, \xi_2, ..., \xi_n)$ – вибірка з нормального розподілу $N(\mu, \theta^2)$. Параметр μ тепер вважатимемо відомим, а θ^2 – невідомим. Таку модель можна використовувати для визначення середньої точності приладу (або методу) шляхом багатократних вимірювань еталона.

За центральну статистику візьмемо $G(\xi, \theta) = \frac{1}{\theta^2} \sum_{i=1}^{n} (\xi_i - \mu)^2$.

Розподіл випадкової величини $G(\xi, \theta)$ збігається, очевидно, з розподілом суми квадратів n незалежних випадкових величин, кожна з яких розподілена згідно зі стандартним нормальним законом розподілу.

За параметром θ $G(x, \theta)$ неперервна й монотонно спадна.

Застосовуючи метод центральної статистики, знаходимо нижню й верхню границі надійного інтервалу у вигляді

$$T_1(\xi) = \frac{1}{g_2} \sum_{i=1}^{n} (\xi_i - \mu)^2, \quad T_2(\xi) = \frac{1}{g_1} \sum_{i=1}^{n} (\xi_i - \mu)^2,$$

де числа $0 < g_1 < g_2 < \infty$ задовольняють умову

$$\gamma = \int_{g_1}^{g_2} k_n(x)dx = P_\theta(g_1 < G(\xi, \theta) < g_2).$$

Зазвичай g_1 та g_2 обирають так, щоб

$$\int_0^{g_1} k_n(x)dx = \frac{1-\gamma}{2}; \quad \int_{g_2}^{\infty} k_n(x)dx = \frac{1-\gamma}{2}. \tag{3.4}$$

У цьому випадку площа, що залишилася за межами криволінійної трапеції з основою $[g_1, g_2]$, ділиться навпіл. Тепер g_1 та g_2 визначаються однозначно через γ:

$$g_1 = \chi^2_{\frac{1-\gamma}{2}}(n), \qquad g_2 = \chi^2_{\frac{1+\gamma}{2}}(n),$$

де $\chi^2_p(n)$ – p-квантиль розподілу $\chi^2(n)$.

Надійний інтервал, побудований за умови (2.34), називають ***центральним***. Для $N\left(\mu,\theta^2\right)$-моделі центральний надійний інтервал має вигляд

$$\Delta_\gamma(\xi)=\left[\frac{1}{\chi^2_{\frac{1+\gamma}{2},n}}\sum_{i=1}^{n}(\xi_i-\mu)^2,\ \frac{1}{\chi^2_{\frac{1-\gamma}{2},n}}\sum_{i=1}^{n}(\xi_i-\mu)^2\right].$$

3.4.3. Загальна нормальна модель. Надійний інтервал для дисперсії

Нехай $\xi'=\left(\xi_1,\xi_2,...,\xi_n\right)$ – вибірка з нормального розподілу $N\left(\theta_1,\theta_2^2\right)$. Щоб побудувати центральну статистику для дисперсії, потрібен такий результат.

Теорема 3.1. *Нехай* $\xi'=\left(\xi_1,\xi_2,...,\xi_n\right)$ *– вибірка з нормального розподілу* $\mathbf{N}\left(\theta_1,\theta_2^2\right)$. *Тоді* $\dfrac{nS^2}{\theta_2^2}$, *де* $S^2\left(\xi\right)=\dfrac{1}{n}\sum_{i=1}^{n}\left(\xi_i-\overline{\xi}\right)^2$, *має* χ^2*-розподіл із* $\left(n-1\right)$ *ступенями свободи.*

Звідси випливає, що розподіл випадкової величини

$$G\left(\xi,\theta_2^2\right)=\frac{1}{\theta_2^2}\sum_{i=1}^{n}\left(\xi_i-\overline{\xi}\right)^2$$

збігається з розподілом $\chi^2\left(n-1\right)$ і не залежить від параметра $\theta=\left(\theta_1,\theta_2^2\right)$. Неперервність і монотонність за θ_2 очевидна. Отже, $G\left(\xi,\theta_2^2\right)$ – центральна статистика для θ_2^2. Ураховуючи попередній аналіз для $N\left(\mu,\theta^2\right)$-моделі, робимо висновок, що центральним γ-надійним інтервалом для θ_2^2 є інтервал

$$\Delta_\gamma(\xi)=\left(\frac{nS^2(\xi)}{\chi^2_{\frac{1+\gamma}{2},n-1}},\frac{nS^2(\xi)}{\chi^2_{\frac{1-\gamma}{2},n-1}}\right),$$

де $\chi^2_{\frac{1\pm\gamma}{2},n-1}-\dfrac{1\pm\gamma}{2}$ – квантиль розподілу χ^2 із $(n-1)$ ступенями свободи.

3.4.4. Загальна нормальна модель. Надійний інтервал для середнього

Побудуємо центральну статистику для середнього θ_1 у загальній моделі $N(\theta_1,\theta_2^2)$.

Розглянемо

$$G(\xi,\theta_1)=\sqrt{n-1}\frac{\overline{\xi}-\theta_1}{S(\xi)}=\frac{\sqrt{n}(\overline{\xi}-\theta_1)/\theta_2}{\sqrt{\frac{1}{\theta_2^2}\sum_{i=1}^{n}(\xi_i-\overline{\xi})^2\Big/(n-1)}}.$$

Оскільки вибіркове середнє $\overline{\xi}=\dfrac{1}{n}\sum_{i=1}^{n}\xi_i$ і дисперсія $S^2(\xi)=\dfrac{1}{n}\sum_{i=1}^{n}(\xi_i-\overline{\xi})^2$ незалежні, то $G(\xi,\theta_1)$ має розподіл Стьюдента з $(n-1)$ ступенями свободи. Таким чином, $G(\xi,\theta_1)$ – центральна статистика для θ_1. Ураховуючи схожість розподілу Стьюдента і стандартного нормального розподілу, маємо для θ_1 надійний інтервал мінімальної довжини

$$\Delta_\gamma^*(\xi)=\left(\overline{\xi}-\frac{S(\xi)}{\sqrt{n-1}}t_{\frac{1+\gamma}{2},n-1},\overline{\xi}+\frac{S(\xi)}{\sqrt{n-1}}t_{\frac{1+\gamma}{2},n-1}\right),$$

де $t_{\frac{1+\gamma}{2},n-1}-\dfrac{1+\gamma}{2}$ – квантиль розподілу Стьюдента з $(n-1)$ ступенями свободи.

3.5. Побудова надійних інтервалів на основі точкових оцінок

Якщо є деяка точкова оцінка параметра θ $T(\xi)$ і відома її функція розподілу $F_T(t,\theta)$, то надійний інтервал можна побудувати за умови, що $F_T(t,\theta)$ неперервна й монотонна за θ.

Обираючи різні оцінки $T(\xi)$, отримаємо різні надійні інтервали. Кінцева мета – при фіксованому рівні надійності γ отримати якомога коротший інтервал. Припустимо, що використовуються незсунені та приблизно нормальні оцінки. Тоді інтервали тим коротші, чим менша дисперсія оцінки. Таким чином, ефективні й асимптотично ефективні оцінки зумовлюють найменші або асимптотично найменші надійні інтервали. Ці вимоги задовольняють оцінки максимальної вірогідності $\hat{\theta}_n$ у моделях, для яких виконуються відповідні умови регулярності.

Зафіксуємо $0<\gamma<1$ і $\hat{c}_\gamma$ визначимо з рівняння $2\Phi(\hat{c}_\gamma)-1=\gamma$.

Отже, $\hat{c}_\gamma = c_{\frac{1+\gamma}{2}} = \Phi^{-1}\left(\frac{1+\gamma}{2}\right)$ – квантиль порядку $\frac{1+\gamma}{2}$ для стандартного нормального розподілу (табл. 2 додатка). При виконанні умов теореми про асимптотичну нормальність оцінок максимальної вірогідності

$$P_\theta\left\{\left|\hat{\theta}_n-\theta\right|\sqrt{nI_1\left(\hat{\theta}_n\right)}\le\hat{c}_\gamma\right\}\underset{n\to\infty}{\to}\Phi(\hat{c}_\gamma)-\Phi(-\hat{c}_\gamma)=2\Phi(\hat{c}_\gamma)-1=\gamma.$$

Отже, $\left(\hat{\theta}_n-\frac{c_{\frac{1+\gamma}{2}}}{\sqrt{nI_1\left(\hat{\theta}_n\right)}},\hat{\theta}_n+\frac{c_{\frac{1+\gamma}{2}}}{\sqrt{nI_1\left(\hat{\theta}_n\right)}}\right)$ – асимптотично найменший γ-надійний інтервал для параметра θ.

Приклад 3.1. Нехай $\xi' = (\xi_1, ..., \xi_n)$ – вибірка з генеральної сукупності з розподілом Пуассона з параметром θ. Відповідно $P\{\xi_0 = x\} = \frac{\theta^x}{x!} e^{-\theta}$, $x = 0,1,...$. Користуючись наведеною вище методикою, побудуємо асимптотичний γ-надійний інтервал для невідомого параметра θ. Якщо $x' = (x_1, ..., x_n)$ – реалізація вектора ξ, то функція вірогідності

$$L(x,\theta) = e^{-n\theta} \frac{\theta^{\sum_{k=1}^{n} x_k}}{\prod_{k=1}^{n} x_k!}.$$

Далі, $\frac{\partial \ln L(x,\theta)}{\partial \theta} = \frac{n}{\theta}(\overline{x} - \theta)$, $\frac{\partial^2 \ln L(x,\theta)}{\partial \theta^2} = -\frac{n}{\theta^2}\overline{x}$.

З рівняння $\frac{n}{\theta}(\overline{\xi} - \theta) = 0$ знаходимо оцінку максимальної вірогідності $\hat{\theta}_n = \overline{\xi} = \frac{1}{n}\sum_{i=1}^{n} \xi_i$. $I(\theta) = -M \frac{\partial^2 \ln L(\xi,\theta)}{\partial \theta^2} = \frac{n\theta}{\theta^2} = \frac{n}{\theta}$.

Тоді асимптотично найкоротший γ-надійний інтервал для параметра θ буде таким:

$$\left(\overline{\xi} - c_{\frac{1+\gamma}{2}} \sqrt{\frac{\overline{\xi}}{n}},\ \overline{\xi} + c_{\frac{1+\gamma}{2}} \sqrt{\frac{\overline{\xi}}{n}} \right).$$

ЗАДАЧІ

3.1. Надійний інтервал для біноміального розподілу. Нехай $P\{\xi_0 = x\} = C_n^x \theta^x (1-\theta)^{n-x}$, $x = 0,1,...,n$, $\theta \in (0;1)$. Побудувати надійний інтервал для параметра θ.

3.2. Виконано 100 незалежних випробувань, у результаті яких подія A спостерігалась 40 разів. Визначити надійний інтервал для ймовірності події A за рівнів надійності 0,95 та 0,99, якщо кількість появ події A має біноміальний розподіл.

3.3. На телефонній станції проводились спостереження за кількістю невірних з'єднань за хвилину. Спостереження протягом години дали такі результати:

x_i	0	1	2	3	4	5	7
n_i^*	8	17	16	10	6	2	1

Припускаючи, що кількість невірних з'єднань за хвилину має пуассонівський розподіл, знайти надійний інтервал для невідомого параметра з надійністю 0,99.

3.4. Побудувати надійний інтервал для параметра θ нормального розподілу $N\left(\theta, 4\theta^2\right)$.

3.5. Побудувати надійний інтервал для параметра θ за вибіркою $\xi_1, \xi_2, \ldots, \xi_n$ із генеральної сукупності з розподілом:

а) $P_\theta\{\xi_0 = x\} = \dfrac{\theta^x}{(1+\theta)^{x+1}}$; $x = 0,1,\ldots$; $\theta > 0$;

б) $P_\theta\{\xi_0 = x\} = \dfrac{(\theta-1)^x}{\theta^{x+1}}$; $x = 0,1,\ldots$; $\theta > 1$.

3.6. Знайти надійний інтервал для параметрів a та σ^2 нормального розподіл у за вибіркою:

а) 0,6; 2,4; 2,1; 1,4; 1,2; 4,8; 0,9; 1,1; 3,5; 3,0;

б) 0,2; 0,5; 1,0; 1,5; 0,8; 1,0; 2,0.

Для параметра a покласти $\gamma = 0.95$, для σ^2 – $\gamma = 0.9$.

3.7. Знайти надійний інтервал для математичного сподівання нормального розподілу $N\left(a, \sigma^2\right)$, якщо $n = 25$, $\overline{x} = 16.8$, $\sigma^2 = 25$, $\gamma = 0.99$.

3.8. Знайти надійний інтервал для дисперсії нормального розподілу $N\left(a, \sigma^2\right)$, якщо $n = 20$, $\hat{S}^2 = 10$, $\gamma = 0.9$.

3.9. За вибіркою з нормального розподілу

x_i	[-2;0)	[0;1)	[1;2)	[2;3)	[3;4)	[4;5)	[5;6]
n_i^*	3	3	4	6	2	1	1

з надійністю $\gamma = 0.95$ знайти інтервальні оцінки для параметра a, коли:

а) $\sigma^2 = 4$;

б) σ^2 – невідоме.

Знайти також надійний інтервал для σ^2, коли:

в) $a = 2$;

г) a – невідоме. Надійність γ покласти 0,9.

Розділ 4
ПЕРЕВІРКА СТАТИСТИЧНИХ ГІПОТЕЗ

4.1. Поняття статистичної гіпотези та статистичного критерію

Статистична гіпотеза – це довільне твердження про тип або властивості розподілів випадкових величин, що спостерігаються в експерименті.

Приклад 4.1. Нехай експеримент полягає в багаторазовому вимірюванні деякої фізичної величини, точне значення якої a невідоме й у процесі вимірювань не змінюється. На результати вимірювань впливають багато факторів: точність налагодження приладу, похибка заокруглення тощо, тому результат i-го вимірювання ξ_i можна записати у вигляді

$$\xi_i = a + \varepsilon_i,$$

де ε_i – випадкова похибка вимірювання.

Будемо вважати, що загальна похибка ε_i складається з великої кількості похибок, кожна з яких невелика. На основі центральної граничної теореми припускатимемо, що випадкові величини ξ_i мають нормальний розподіл. Таке припущення є статистичною гіпотезою про тип розподілу випадкових величин, що спостерігаються.

Наведемо кілька типів статистичних гіпотез.

1. Гіпотеза про тип розподілу. Нехай виконано n незалежних спостережень над деякою випадковою величиною ξ_0 з

невідомою функцією розподілу $F_{\xi_0}(z)$. Гіпотеза, яка підлягає перевірці:

$$H_0 : F_{\xi_0}(z) = F(z),$$

де функція $F(z)$ повністю задана, або H_0: $F_{\xi_0}(z) \in \mathbb{F} = \{F(z,\theta)\theta \in \Theta\}$ – задана сім'я функцій розподілу.

2. Гіпотеза однорідності. Нехай виконано k серій незалежних спостережень $(\xi_{i1}, \xi_{i2}, \ldots, \xi_{in_i})$, $i = 1,2,\ldots,k$ з генеральних сукупностей з функціями розподілу $F_i(z)$, $i = 1,2,\ldots,k$ (узагалі кажучи, невідомими). Чи є підстава розглядати ці дані як результати спостережень над тією самою випадковою величиною? Якщо це так, то кажуть, що статистичні дані однорідні. Відповідно перевіряється гіпотеза однорідності

$$H_0 : F_1(z) \equiv \ldots \equiv F_k(z).$$

3. Гіпотеза незалежності. В експерименті спостерігається двовимірна випадкова величина (ξ, η) з невідомою сумісною функцією розподілу $F_{\xi,\eta}(z_1, z_2)$ і є підстава вважати, що компоненти ξ та η незалежні. У цьому випадку треба перевірити гіпотезу незалежності, тобто

$$H_0 : F_{\xi,\eta}(z_1, z_2) = F_\xi(z_1) F_\eta(z_2).$$

Якщо гіпотеза H_0 однозначно фіксує розподіл спостережень, то її називають ***простою***, у протилежному випадку – ***складною***. У наведених вище прикладах лише гіпотеза про тип розподілу $H_0 : F_{\xi_0}(z) = F(z)$ є простою.

Статистичний критерій – це правило, згідно з яким гіпотеза, що перевіряється, приймається або відкидається.

Розглянемо методи перевірки гіпотез описаних вище типів. Нехай про розподіл вибірки $\xi' = (\xi_1, \xi_2, \ldots, \xi_n)$, що описує результати експерименту, сформульована гіпотеза H_0. Необхідно перевірити, узгоджуються чи ні статистичні дані з цією гіпоте-

зою. Відповідні критерії називаються ***критеріюми згоди***. Наведемо методику побудови критеріїв згоди.

Обирається статистика $T = T(\xi)$, яка характеризує відхилення емпіричних даних від гіпотетичних значень, що відповідають гіпотезі H_0. Вона є мірою розбіжності статистичного та гіпотетичного законів розподілу і називається ***статистикою критерію***. Розподіл статистики $T = T(\xi)$ треба знати точно або наближено в припущенні, що розподіл спостережень збігається з гіпотетичним.

Нехай таку статистику знайдено. Визначимо для фіксованого достатньо малого числа $\alpha > 0$ число t_α так, щоб у випадку справедливості гіпотези H_0 імовірність настання події $P\{T(\xi) \geq t_\alpha / H_0\} = \alpha$. Число α називається ***рівнем значущості критерію***.

Нехай $x' = (x_1, ..., x_n)$ – реалізація вибірки, а $\hat{T} = T(x)$ – відповідне значення статистики T, обчислене за статистичними даними. Якщо $\hat{T} \geq t_\alpha$, то відхилення від гіпотетичного закону розподілу вважається значущим і гіпотеза відхиляється. У протилежному випадку немає підстав відмовлятися від висунутої гіпотези і слід вважати, що спостереження не суперечать гіпотезі (на рівні α). Область $\{t : t \geq t_\alpha\}$ називається критичною областю для гіпотези H_0.

4.2. Гіпотези про тип розподілу

4.2.1. Критерій згоди Колмогорова

Нехай $\xi' = (\xi_1, \xi_2, \ldots, \xi_n)$ – вибірка з генеральної сукупності з невідомою функцією розподілу $F_{\xi_0}(z)$, про яку висунута проста гіпотеза H_0: $F_{\xi_0}(z) = F(z)$, $F(z)$ – неперервна функція.

Статистика критерію Колмогорова – це величина

$$D_n = D_n(\xi) = \sup_{-\infty<z<\infty} \left|F_n(z) - F(z)\right|,$$

яка є максимальним відхиленням емпіричної функції розподілу $F_n(z)$ від гіпотетичної $F(z)$.

У тих випадках, коли гіпотеза H_0 справедлива, зі збільшенням розміру вибірки n відбувається зближення $F_n(z)$ із $F(z)$. Тому принаймні для великих n значення D_n не повинно істотно відрізнятися від 0.

Особливості статистики D_n:

1) Її розподіл за справедливості гіпотези H_0 не залежить від вигляду функції $F(z)$:

$$D_n = \sup_{-\infty<z<\infty} \left|F_n(z) - F(z)\right| \stackrel{d}{=} \sup_{0\le u\le 1} \left|\Phi_n(u) - u\right|, \qquad (4.1)$$

де $\Phi_n(u)$ – емпірична функція розподілу для вибірки з рівномірного на інтервалі $[0,1]$ розподілу, $\stackrel{d}{=}$ – рівність за розподілом.

Дійсно, поклавши у (4.1) $z = F^{-1}(u)$, $0 \le u \le 1$, де $F^{-1}(u)$ – функція, обернена до $F(z)$, отримаємо

$$D_n = \sup_{0\le u\le 1} \left|F_n(F^{-1}(u)) - u\right|.$$

Перейдемо до нових випадкових величин, використовуючи формулу $U_i = F(\xi_i)$ $i = 1,2,\ldots,n$.

Нехай $U_{(1)} \le U_{(2)} \le \ldots \le U_{(n)}$ – їх варіаційний ряд. Функція $F(z)$ – монотонна, тому $U_{(k)} = F(\xi_{(k)})$ $k = 1,2,\ldots,n$ і нерівності $F^{-1}(u) \ge \xi_{(k)}$ еквівалентні нерівностям $u \ge U_{(k)}$. Оскільки $F_n(x) = \dfrac{1}{n}\sum_{k=1}^{n} e(x - \xi_{(k)})$, то маємо

$$F_n(F^{-1}(u)) = \frac{1}{n}\sum_{k=1}^{n} e(F^{-1}(u) - \xi_{(k)}) = \frac{1}{n}\sum_{k=1}^{n} e(u - U_{(k)}) = \Phi_n(u).$$

Розподіл U_i збігається з розподілом $R(0,1)$ і $\Phi_n(u)$ – емпірична функція розподілу для вибірки з рівномірного розподілу $R(0,1)$.

Цей факт дуже корисний, оскільки достатньо обчислити розподіл D_n тільки один раз, а саме для вибірки з рівномірного $R(0,1)$ розподілу, і використовувати його для перевірки гіпотези щодо довільної неперервної функції розподілу $F(z)$.

2) Друга особливість полягає в тому, що

$$\sqrt{n}D_n \underset{n\to\infty}{\overset{сл}{\Rightarrow}} \eta,$$

де випадкова величина η має функцію розподілу Колмогорова $K(t)=\sum_{j=-\infty}^{\infty}(-1)^j e^{-2j^2t^2}$. Цей граничний розподіл і використовується як розподіл D_n уже при $n \geq 20$.

Використаємо останній факт. За заданим рівнем значущості α підбираємо число λ_α так, що

$$P\left\{\sqrt{n}D_n \geq \lambda_\alpha / H_0\right\} \approx 1-K(\lambda_\alpha)=\alpha.$$

У табл. 5 додатка наведені значення λ_α для різних α.

Базуючись на цьому, будуємо ***правило перевірки гіпотези*** H_0. Нехай $\lambda_n=\sqrt{n}D_n\left(x\right)$ – значення статистики критерію, обчислене за реалізацією вибірки $x'=\left(x_1,...,x_n\right)$. Якщо $\lambda_n \geq \lambda_\alpha$, то гіпотеза відхиляється, а при $\lambda_n < \lambda_\alpha$ – приймається, тобто робиться висновок, що статистичні дані не суперечать гіпотезі.

Приклад 4.2. Нехай маємо реалізацію вибірки x: 0,7; 2,3; 4,8; 9,7; 5,3; 6,8; 5,9; 8,7; 1,4; 3,2. Треба перевірити гіпотезу про те, що величина, яка спостерігається, має рівномірний розподіл на [0;10) з рівнем значущості $\alpha=0,05$:

$$H_0: F_{\xi_0}\left(z\right)=\begin{cases}0, & z<0,\\ \dfrac{z}{10}, & z\in[0,10],\\ 1, & z>10.\end{cases}$$

Значення емпіричної та гіпотетичної функцій розподілу наведемо у таблиці ($n=10$):

Значення	0,7	1,4	2,3	3,2	4,8	5,3	5,9	6,8	8,7	9,7
$F_n(z_i)$	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
$F_{\xi_0}(z_i)$	0,07	0,14	0,23	0,32	0,48	0,53	0,59	0,68	0,87	0,97
$\lvert F_n(z_i)-F_{\xi_0}(z_i)\rvert$	0,03	0,06	0,07	0,08	0,02	0,07	0,11	0,12	0,03	0,03

Значення статистики критерію

$$\lambda_n=\sqrt{n}\sup_z\left|F_n(z)-F_{\xi_0}(z)\right|=\sqrt{10}\cdot 0{,}12\approx 0{,}38,$$

критична область визначається значенням $\lambda_\alpha=1{,}358$. Оскільки $\lambda_n<\lambda_\alpha$, то гіпотеза H_0 приймається.

4.2.2. Критерій χ^2 К. Пірсона

Нехай, як і раніше, $\xi'=(\xi_1,\xi_2,\ldots,\xi_n)$ – вибірка з невідомою функцією розподілу $F_{\xi_0}(z)$, про яку висунута проста гіпотеза

$$H_0:\ F_{\xi_0}(z)=F(z).$$

Про властивості гіпотетичної $F(z)$ у даному випадку нічого не відомо, тобто цей критерій можна використовувати як для неперервних, так і для дискретних розподілів.

Задамо $\mathrm{E}_1,\mathrm{E}_2,\ldots,\mathrm{E}_N$ – інтервали групування даних, що не перетинаються. Якщо спостерігається дискретна випадкова величина, то $\mathrm{E}_1,\mathrm{E}_2,\ldots,\mathrm{E}_N$ – її різні значення. Нехай $\nu'=(\nu_1,\nu_2,\ldots,\nu_N)$ – вектор частот потрапляння елементів вибірки до відповідних інтервалів групування. Позначимо $p_i=P\{\xi_0\in \mathrm{E}_i/H_0\}$, $i=1,2,\ldots,N$. Очевидно, що $M(\nu_i/H_0)=np_i$.

За міру відхилення емпіричних даних від їх гіпотетичних значень візьмемо статистику

$$\hat{\chi}^2_n=\sum_{i=1}^{N}\frac{(\nu_i-np_i)^2}{np_i}. \qquad (4.2)$$

Теорема 4.1. *Якщо* $0 < p_i < 1$, $i = 1,2,\ldots,N$, *то при* $n \to \infty$ *розподіл величини* $\hat{\chi}_n^2$ *слабко збігається до* χ^2*-розподілу з* $(N-1)$ *ступенями свободи.*

Використовуючи апроксимацію розподілу статистики $\hat{\chi}_n^2$ розподілом хі-квадрат, маємо

$$\alpha = \int\limits_{\chi^2_{1-\alpha,N-1}}^{\infty} k_{N-1}(x)dx \approx P\left\{\hat{\chi}_n^2 \geq \chi^2_{\alpha,N-1} / H_0\right\}.$$

Таким чином, критична область – це множина $\left\{t : t \geq \chi^2_{1-\alpha,N-1}\right\}$.

На практиці граничний розподіл $\chi^2(N-1)$ можна використовувати з непоганим наближенням уже при $n \geq 50$ і $\nu_i \geq 5$.

Критерій перевірки гіпотези H_0 будується таким чином. Обчисливши значення статистики критерію $\hat{\chi}_n^2 = \sum\limits_{i=1}^{N} \dfrac{(\nu_i - np_i)^2}{np_i}$ і вибравши рівень значущості α, за таблицею значень квантилів χ^2-розподілу (табл. 3 додатка) визначимо величину $\chi^2_{1-\alpha,N-1}$ таку, що $P\left\{\chi^2(N-1) \geq \chi^2_{1-\alpha,N-1}\right\} = \alpha$ (або $P\left\{\chi^2(N-1) < \chi^2_{1-\alpha,N-1}\right\} = 1-\alpha$). Якщо $\hat{\chi}_n^{\,2} \geq \chi^2_{1-\alpha,N-1}$, то гіпотеза H_0 відхиляється, якщо ж $\hat{\chi}_n^{\,2} < \chi^2_{1-\alpha,N-1}$, – то приймається.

Зауваження 1. Гіпотетичний розподіл, що не залежить від параметрів, на практиці зустрічається рідко. Зазвичай гіпотетичний розподіл залежить від невідомих параметрів, стосовно значень яких є лише та інформація, що міститься у виборці. У такій ситуації маємо задачу перевірки складних гіпотез, для яких також можна використовувати критерій χ^2. Нехай за вибіркою $\xi' = (\xi_1,\ldots,\xi_n)$ треба перевірити гіпотезу H_0: $F_{\xi_0}(z) \in \mathbb{F}$, де $\mathbb{F} = \{F(z,\theta), \theta \in \Theta\}$ – задана сім'я функцій розподілу. Значення

параметрів, а отже, і ймовірностей $p_i(\theta)$, невідомі. Тому природно оцінити невідомий параметр θ за вибіркою й у статистику (4.2) підставити ймовірності p_i, підраховані через $F(z,\hat{\theta}_n)$, де $\widehat{\theta_n}$ – оцінка θ. У даному випадку величини $p_i(\widehat{\theta_n})$ уже не сталі, вони є функціями від вибірки, а отже, випадковими величинами. Тому теорема 4.1 не може бути застосована. Окрім того, розподіл цієї статистики залежить від методу побудови оцінки $\widehat{\theta_n}$. Р. Фішер показав, що існують методи оцінювання параметра θ, за яких граничним розподілом для статистики критерію буде χ^2-розподіл з $(N-l-1)$ ступенями свободи, де l – розмірність параметра θ. Одним з таких методів є метод мінімуму χ^2 ([10, с. 79 – 80]; [14, с. 460 – 470]).

Далі критерій перевірки гіпотези будуємо аналогічно наведеному вище. Якщо $\hat{\chi}_n^{\,2} \geq \chi^2_{1-\alpha,N-l-1}$, то гіпотеза H_0 відхиляється, якщо ж $\hat{\chi}_n^{\,2} < \chi^2_{1-\alpha,N-l-1}$, – то приймається.

Зауваження 2. У [14, с. 470 – 474] за допомогою методу мінімуму χ^2 знайдено оцінки для параметра λ розподілу Пуассона й параметрів a та σ^2 нормального розподілу. Доведено, що оцінки параметрів λ та a близькі до вибіркового середнього, а оцінкою σ^2 є вибіркова дисперсія $\overline{S}^2 = \dfrac{1}{n}\sum_i \nu_i \left(\xi_i - \overline{\xi}\right)^2$

Приклад 4.3. При 50 підкиданнях монети герб з'явився 20 разів. Чи можна вважати, що монета симетрична? Прийняти $\alpha = 0{,}1$.

Експеримент із підкиданням монети можна описати в термінах незалежних спостережень випадкової величини ξ_0, яка набуває двох значень: $x=1$, якщо випав герб, та 0, якщо випала решка. Гіпотеза про симетричність монети в термінах розподілу ξ_0 формулюється так: розподілом ξ_0 є

$$P\{\xi = x\} = \frac{1}{2^x}\cdot\frac{1}{2^{1-x}} = \frac{1}{2}, \quad x = 0,1.$$

Імовірності потрапляння вибіркових значень у підмножини вибіркових значень, якими є дві підмножини $X_0 = \{0\}$ та $X_1 = \{1\}$, підраховані за гіпотетичним розподілом, становлять $p_0 = p_1 = 1/2$.

Підрахуємо значення статистики критерію χ^2:

$$\hat{\chi}^2_{50} = \frac{\left(30 - \frac{1}{2}50\right)^2}{\frac{1}{2}50} + \frac{\left(20 - \frac{1}{2}50\right)^2}{\frac{1}{2}50} = 2 .$$

У таблиці значень χ^2-розподілу (див. табл. 3 додатка) шукаємо значення $\chi^2_{1-\alpha,N-1} = \chi^2_{0.9,1} = 2,71$. Оскільки значення статистики критерію не перевищує табличне, то гіпотеза приймається. Таким чином, гіпотеза узгоджується з даними експерименту, або, іншими словами, гіпотеза про симетричність монети не суперечить експериментальним даним.

Приклад 4.4. За спостереженнями, наведеними у таблиці, за допомогою критерію χ^2 з рівнем значущості $\alpha = 0,05$ перевірити гіпотезу, що випадкова величина має нормальний розподіл.

Інтервал	$[-4;0)$	$[0;2)$	$[2;4)$	$[4;6)$
ν_i	20	40	30	10

Обчислимо оцінки параметрів нормального розподілу за вибіркою. Згідно із зауваженням 2

$$\overline{x} = \frac{1}{100}\left(-40 + 40 + 90 + 50\right) = 1,4 ;$$

$$\overline{S}^2 = \frac{1}{n}\sum_i \nu_i \left(x_i - \overline{x}\right)^2 =$$

$$= \frac{1}{100}\left(\left(-2 - 1,4\right)^2 20 + \left(1 - 1,4\right)^2 40 + \left(3 - 1,4\right)^2 30 + \left(5 - 1,4\right)^2 10\right) =$$

$$= 4.48 ; \ \overline{S} = 2,1 .$$

Тепер підрахуємо ймовірності p_i, $i=\overline{1,4}$, використовуючи функцію розподілу випадкової величини, що має стандартний нормальний розподіл $\Phi(z)=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{z} e^{-t^2/2}dt$ (див. табл. 1 додатка).

$$p_1 = P\{-4 \le \xi < 0\} = P\left\{\frac{-4-1,4}{2,1} \le \frac{\xi - a}{\sigma} < \frac{0-1,4}{2,1}\right\} =$$

$$= \Phi(-0,66) - \Phi(-2,57) = 0,2546 - 0,0051 = 0,2495;$$

$$p_2 = P\{0 \le \xi < 2\} = P\left\{\frac{0-1,4}{2,1} \le \frac{\xi - a}{\sigma} < \frac{2-1,4}{2,1}\right\} =$$

$$= \Phi(0,29) - \Phi(-0,66) = 0,6141 - 0,2546 = 0,3595;$$

$$p_3 = P\{2 \le \xi < 4\} = P\left\{\frac{2-1,4}{2,1} \le \frac{\xi - a}{\sigma} < \frac{4-1,4}{2,1}\right\} =$$

$$= \Phi(1,24) - \Phi(0,29) = 0,8925 - 0,6141 = 0,2784;$$

$$p_4 = P\{4 \le \xi < 6\} = P\left\{\frac{4-1,4}{2,1} \le \frac{\xi - a}{\sigma} < \frac{6-1,4}{2,1}\right\} =$$

$$= \Phi(2,19) - \Phi(1,24) = 0,9857 - 0,8925 = 0,0932.$$

Обчислені результати заносимо в таблицю:

Інтервал	$[-4;0)$	$[0;2)$	$[2;4)$	$[4;6)$
ν_i	20	40	30	10
p_i	0,2495	0,3595	0,2784	0,0932
np_i	24,95	35,95	27,84	9,32
$(\nu_i - np_i)^2$	24,50	16,40	4,67	0,64
$\frac{(\nu_i - np_i)^2}{np_i}$	0,98	0,46	0,17	0,07

Значення статистики критерію $\hat{\chi}_n^{\,2} = 1{,}68$. Кількість інтервалів $N = 4$, кількість невідомих параметрів $l = 2$. Кількість ступенів свободи $N - l - 1 = 1$. З таблиці беремо критичне значення $\chi^2_{0.95;1} = 3.841$. Оскільки 1,68<3,841, то гіпотеза приймається.

Приклад 4.5. За час Другої світової війни на Лондон впало 535 літаків-снарядів. Уся територія Лондона була розділена на 576 ділянок площею по 0,25 км². Нижче наведені кількості ділянок n_k, на які впало k снарядів:

k	0	1	2	3	4	5
n_k	229	211	93	35	7	1

Чи узгоджуються ці дані з гіпотезою про те, що кількість снарядів, які впали на ділянку, мають розподіл Пуассона? Прийняти $\alpha = 0{,}05$.

Маємо $n = \sum_{k=0}^{5} n_k = 576$ незалежних спостережень випадкової величини ξ_0 – кількості літаків-снарядів, що впали на ділянку.

Відносно невідомого розподілу випадкової величини ξ_0 висувається гіпотеза

$$H_0 : P\{\xi_0 = k\} = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0,1,\ldots,$$

яку треба перевірити.

Параметр λ гіпотетичного розподілу невідомий. Згідно із зауваженням 2 за його оцінку ми можемо взяти вибіркове середнє:

$$\hat{\lambda} = \frac{1}{576}\left(0 \cdot 229 + 1 \cdot 211 + 2 \cdot 93 + 3 \cdot 35 + 4 \cdot 7 + 5 \cdot 1\right) \approx 0{,}9,$$

отже, гіпотетичний розподіл має вигляд

$$P\{\xi_0 = k\} = \frac{0{,}9^k}{k!} e^{-0{,}9}, \quad k = 0,1,\ldots.$$

Користуючись наведеною вище методикою, шукаємо значення статистики критерію $\hat{\chi}_n^{\,2}$:

$$\hat{\chi}_n^{\,2} = \frac{(229-576e^{-0,9})^2}{576e^{-0,9}} + \frac{(211-576\cdot 0,9e^{-0,9})^2}{576\cdot 0,9e^{-0,9}} +$$

$$+\frac{(93-576\frac{(0,9)^2}{2!}e^{-0,9})^2}{576\frac{(0,9)^2}{2!}e^{-0,9}} + \frac{(35-576\frac{(0,9)^3}{3!}e^{-0,9})^2}{576\frac{(0,9)^3}{3!}e^{-0,9}} +$$

$$+\frac{\left(7-576\frac{(0,9)^4}{4!}e^{-0,9}\right)^2}{576\frac{(0,9)^4}{4!}e^{-0,9}} + \frac{\left(1-576\left(1-\sum_{i=0}^{4}\frac{(0,9)^i}{i!}e^{-0,9}\right)\right)^2}{576\left(1-\sum_{i=0}^{4}\frac{(0,9)^i}{i!}e^{-0,9}\right)} \approx 1,171.$$

Отримане значення порівнюємо з табличним значенням $\chi^2_{1-\alpha;N-1-l}$, де l – кількість параметрів, оцінених за вибіркою, N – кількість підмножин, на які розбито вибірковий простір. Маємо $\chi^2_{1-\alpha;N-2} = \chi^2_{0,95;4} = 9,49$. Таким чином, експериментальні дані узгоджуються з гіпотезою.

4.3. Гіпотези однорідності

Нехай є дві незалежні вибірки $\xi' = (\xi_1, \xi_2, \ldots, \xi_n)$ і $\eta' = (\eta_1, \ldots, \eta_m)$, які описують те саме явище, однак отримані вони в різний час і в різних умовах. Треба перевірити, чи будуть ці вибірки з одного розподілу, чи закон розподілу від вибірки до вибірки змінювався.

Подібна задача може виникнути при контролі якості деякої продукції. У загальному вигляді її можна сформулювати так.

Нехай $\xi' = (\xi_1, \xi_2, \ldots, \xi_n)$ – вибірка з деякою невідомою функцією розподілу $F_1(z)$, а $\eta' = (\eta_1, \ldots, \eta_m)$ – вибірка з невідомою функцією розподілу $F_2(z)$. Треба перевірити гіпотезу однорідності

$$H_0 : F_1(z) \equiv F_2(z).$$

4.3.1. Критерій Смірнова – Колмогорова

Одним із критеріїв перевірки гіпотези однорідності є критерій Смірнова – Колмогорова, який застосовують лише у випадку неперервних розподілів. Цей критерій основується на статистиці $D_{n,m} = \sup\limits_{-\infty<z<\infty} \left|F_{1n}(z) - F_{2m}(z)\right|$, де $F_{1n}(z)$ і $F_{2n}(z)$ – емпіричні функції розподілу, побудовані за першою та другою вибірками.

Критичну границю знаходять на основі відомого за гіпотези H_0 граничного розподілу статистики $D_{n,m}$.

Теорема 4.2 (Н. В. Смірнов, 1944). *Нехай $F_{1n}(z)$ і $F_{2n}(z)$ – дві емпіричні функції розподілу, що побудовані на основі двох незалежних вибірок розмірами n та m того самого розподілу, який має неперервну функцію розподілу $F(z)$. Тоді для довільного фіксованого $t > 0$*

$$\lim_{n,m\to\infty} P\left(\sqrt{nm/(n+m)}D_{nm} \le t\right) = K(t) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2t^2} .$$

Відповідно за міру розбіжності приймається величина $\lambda_{n,m} = \sqrt{\dfrac{nm}{n+m}}D_{n,m} = \sqrt{\dfrac{nm}{n+m}} \sup\limits_{-\infty<z<\infty} \left|F_{1n}(z) - F_{2m}(z)\right|$, розподіл якої збігається, згідно з теоремою 4.2, до розподілу Колмогорова. Далі критерій будується аналогічно критерію Колмогорова: за заданим рівнем значущості α знайдемо за таблицею критичне значення λ_α (табл. 5 додатка). Якщо $\lambda_{n,m} \ge \lambda_\alpha$, то гіпотеза H_0 відхиляється, якщо ж $\lambda_{n,m} < \lambda_\alpha$, – то приймається.

4.3.2. Критерій однорідності χ^2

Цей критерій можна використовувати для перевірки дискретних даних. Окрім того, за його допомогою можна перевіряти однорідність будь-якої скінченної кількості вибірок (критерій Смірнова – Колмогорова може аналізувати лише дві вибірки).

Припустимо, що виконано k послідовних серій незалежних спостережень, які включають $n_1,\ldots,n_k$ спостережень, відповідно. При цьому в кожному експерименті може з'являтися один із s результатів.

Нехай ν_{ij} – кількість реалізацій i-го результату в j-й серії, $\sum_{i=1}^{s}\nu_{ij}=n_j \quad j=1,2,\ldots,k$, $n_1+\ldots+n_k=n$ – загальний обсяг спостережень. Треба перевірити гіпотезу H_0 про те, що всі спостереження були над однією випадковою величиною.

Оскільки $M(\nu_{ij}/H_0)=n_j p_j$, то, спираючись на принцип χ^2, за міру відхилення дослідних даних від їх гіпотетичних значень у даному випадку слід брати статистику

$$\hat{\chi}^2_n=\sum_{i=1}^{s}\sum_{j=1}^{k}\frac{\left(\nu_{ij}-\dfrac{n_j\nu_i}{n}\right)^2}{\dfrac{n_j\nu_i}{n}},\quad \nu_i=\sum_{j=1}^{k}\nu_{ij}.$$

Для знаходження критичної границі застосовують таку граничну теорему, аналогічну теоремі 4.1:

$$\chi^2_n \underset{n\to\infty}{\overset{сл}{\Rightarrow}} \chi^2((s-1)(k-1)).$$

У таблиці χ^2-розподілу (табл. 3 додатка) за заданим рівнем значущості α та кількістю ступенів свободи $(s-1)(k-1)$ знаходимо число $\chi^2_{1-\alpha,(s-1)(k-1)}$ таке, що

$P\left\{\chi^2(s-1)(k-1)\ge\chi^2_{1-\alpha,(s-1)(k-1)}\right\}=\alpha$ (або $P\left\{\chi^2(s-1)(k-1)<\chi^2_{1-\alpha,(s-1)(k-1)}\right\}=1-\alpha$). Якщо значення статистики критерію $\hat{\chi}^2_n\ge\chi^2_{1-\alpha,(s-1)(k-1)}$, то гіпотеза H_0 відхиляється. У протилежному випадку – приймається.

Приклад 4.6. За допомогою критерію χ^2 перевірити гіпотезу однорідності двох вибірок при $\alpha = 0{,}05$.

x_i	1	2	3	4
ν_{i1}	40	26	24	10
ν_{i2}	30	20	30	20
ν_i	70	46	54	30

$$n_1 = 100,\ n_2 = 100,\ n = 200.$$

$$\chi_n^2 = \frac{(40-35)^2}{35} + \frac{(26-23)^2}{23} + \frac{(24-27)^2}{27} + \frac{(10-15)^2}{15} + \frac{(30-35)^2}{35} +$$

$$+ \frac{(20-23)^2}{23} + \frac{(30-27)^2}{27} + \frac{(20-15)^2}{15} = 6{,}2;\ \chi^2_{0{,}95;3} = 7{,}815.$$

Оскільки 6,2<7,815, то гіпотеза приймається.

4.4. Гіпотези незалежності. Критерій незалежності χ²

Є n пар незалежних спостережень (ξ_1, η_1) ,…, (ξ_n, η_n) випадкових величин $(\xi_0; \eta_0)$ з невідомою функцією розподілу $F_{\xi_0,\eta_0}(z_1, z_2)$, для якої треба перевірити гіпотезу

$$H_0:\ F_{\xi_0,\eta_0}(z_1, z_2) = F_{\xi_0}(z_1) F_{\eta_0}(z_2),$$

де $F_{\xi_0}(\cdot)$, $F_{\eta_0}(\cdot)$ – одновимірні функції розподілу.

Припускатимемо, що випадкова величина ξ_0 набуває скінченної кількості значень s, які будемо позначати літерами $a_1,\ldots,a_s$, а друга компонента η_0 – k значень $b_1,\ldots,b_k$.

Якщо модель має іншу структуру, то групують усі можливі значення випадкових величин окремо за першою та другою компонентами. У цьому випадку множина значень ξ_0 розбивається на s інтервалів $\mathrm{E}_1^{(1)},\ldots,\mathrm{E}_s^{(1)}$, множина значень η_0 – на k інтервалів $\mathrm{E}_1^{(2)},\ldots,\mathrm{E}_k^{(2)}$, а множина значень вектора $(\xi_0; \eta_0)$ – на $N = s \cdot k$ прямокутників $\mathrm{E}_i^{(1)} \times \mathrm{E}_j^{(2)}$.

Позначимо через ν_{ij} кількість спостережень пари (a_i, b_j) (або кількість елементів вибірки, які належать прямокутнику $\mathrm{E}_i^{(1)} \times \mathrm{E}_j^{(2)}$, якщо дані групуються), $\sum_{i=1}^{s}\sum_{j=1}^{k}\nu_{ij} = n$.

Результати спостережень зручно подати у вигляді таблиці спряженості двох ознак:

ξ_0	η_0				Сума
	b_1	b_2	…	b_k	
a_1	ν_{11}	ν_{12}	…	ν_{1k}	$\nu_{1\cdot}$
a_{21}	ν_{21}	ν_{22}	…	ν_{2k}	$\nu_{2\cdot}$
⋮	⋮	⋮	⋮	⋮	⋮
a_s	ν_{s1}	ν_{s2}	…	ν_{sk}	$\nu_{s\cdot}$
Сума	$\nu_{\cdot 1}$	$\nu_{\cdot 2}$	…	$\nu_{\cdot k}$	n

Відстань між емпіричними даними та їх гіпотетичними значеннями має вигляд

$$\hat{\chi}_n^2 = \sum_{i=1}^{s}\sum_{j=1}^{k}\frac{\left(\nu_{ij} - \dfrac{\nu_{i\cdot}\nu_{\cdot j}}{n}\right)^2}{\dfrac{\nu_{i\cdot}\nu_{\cdot j}}{n}},\ \nu_{\cdot j} = \sum_{i=1}^{s}\nu_{ij},\ \nu_{i\cdot} = \sum_{j=1}^{k}\nu_{ij}.$$

При $n \to \infty$ розподіл відхилення $\hat{\chi}_n^2$ збігається до χ^2-розподілу із $(s-1)(k-1)$ ступенями свободи.

Вибір табличного значення $\chi^2_{1-\alpha;(s-1)(k-1)}$ і прийняття рішення про допустимість гіпотези робиться аналогічно описаній вище процедурі для критерію однорідності χ^2. При цьому з імовірністю α гіпотеза H_0 буде відхилятися, коли вона вірна.

Приклад 4.7. Нижче наведені результати опитування 100 студентів перших трьох курсів, яким ставилося одне запитання: "Чи вважаєте ви, що куріння заважає навчанню?"

З'ясувати, чи підтверджують ці дані припущення про те, що ставлення до куріння студентів на різних курсах різне? Прийняти $\alpha = 0,05$.

Відповідь	Курс			Разом	
	Перший	Другий	Третій		
Так	-	30	25	55	$\nu_{1\cdot}$
Не знаю	8	5	7	20	$\nu_{2\cdot}$
Ні	15	10	-	25	$\nu_{3\cdot}$
Разом	23	45	32	100	
	$\nu_{\cdot 1}$	$\nu_{\cdot 2}$	$\nu_{\cdot 3}$		

$$\chi_n^2 = \frac{\left(0 - 100\frac{23}{100}\frac{55}{100}\right)^2}{100\frac{23}{100}\frac{55}{100}} + \frac{\left(30 - 100\frac{55}{100}\frac{45}{100}\right)^2}{100\frac{55}{100}\frac{45}{100}} + \frac{\left(25 - 100\frac{55}{100}\frac{32}{100}\right)^2}{100\frac{55}{100}\frac{32}{100}} + \cdots = 44,2.$$

$$\chi^2_{1-\alpha;(s-1)(k-1)} = \chi^2_{0,95;4} = 9,488.$$

Гіпотеза про незалежність відкидається і сформульована теза приймається.

Приклад 4.8. Проведено 300 спостережень одночасно над випадковими величинами ξ_0 та η_0, які набувають значень 1, 2 та 1, 2, 3, відповідно. Кількості спостережень пар наведені в таблиці:

ξ_0	η_0			$\nu_{i\cdot}$	
	1	2	3		
1	32	68	50	150	$\nu_{1\cdot}$
2	40	70	40	150	$\nu_{2\cdot}$
$\nu_{\cdot j}$	72	138	90	300	
	$\nu_{\cdot 1}$	$\nu_{\cdot 2}$	$\nu_{\cdot 3}$		

Перевірити за допомогою критерію χ^2, чи є незалежними випадкові величини ξ_0 та η_0 при рівні значущості 0,01.

Знайдемо величини $m_{ij} = \dfrac{\nu_{i\cdot}\nu_{\cdot j}}{n}$: $\left\| m_{ij} \right\| = \begin{pmatrix} 36 & 69 & 45 \\ 36 & 69 & 45 \end{pmatrix}$. Тепер шукаємо матрицю з елементами $\left(\nu_{ij} - m_{ij}\right)^2$:

$$\left\| \left(\nu_{ij} - m_{ij}\right)^2 \right\| = \begin{pmatrix} 16 & 1 & 25 \\ 16 & 1 & 25 \end{pmatrix},$$

і матрицю $\left\| \dfrac{\left(\nu_{ij} - m_{ij}\right)^2}{m_{ij}} \right\| = \begin{pmatrix} 0,44 & 0,014 & 0,55 \\ 0,44 & 0,014 & 0,55 \end{pmatrix}$.

Підсумувавши елементи останньої матриці, отримаємо значення статистики критерію $\chi_n^{\,2} = 2,008$. Кількість ступенів свободи $(s-1)\cdot(k-1) = 2$, у таблиці знаходимо $\chi^2_{0,99;2} = 9,21$. Оскільки $\chi^2_n < \chi^2_{0,99;2}$, то гіпотеза про незалежність приймається.

ЗАДАЧІ

4.1. За спостереженнями, наведеними в таблиці, за допомогою критерію χ^2 перевірити гіпотезу, що випадкова величина має пуассонівський розподіл.

а) $\alpha = 0,05$,

x_i	0	1	2	3	4
m_i	109	65	22	3	1

б) $\alpha = 0,05$,

x_i	0	1	2	3	4	5	6	7
m_i	112	168	130	68	32	5	1	1

в) $\alpha = 0{,}05$,

x_i	0	1	2	3	4	5
m_i	229	211	93	35	7	1

г) $\alpha = 0{,}01$,

x_i	0	1	2	3	4	5	6	7
m_i	8	17	16	10	6	2	0	1

д) $\alpha = 0{,}1$,

x_i	0	1	2	3	4	5
m_i	376	100	81	35	7	1

4.2. За спостереженнями, наведеними в таблиці, за допомогою критерію χ^2 перевірити узгодженість із рівномірним розподілом. У першому рядку таблиці вказана ліва границя інтервалу (i – номер інтервалу $[i;i+1)$).

а) $\alpha = 0.05$,

x_i	0	1	2	3	4	5	6	7	8	9	10	11
m_i	48	42	36	54	39	43	41	33	37	41	47	39

б) $\alpha = 0.1$,

x_i	0	1	2	3	4	5	6	7	8	9
m_i	69	89	83	79	80	73	77	75	76	91

в) $\alpha = 0.01$,

x_i	0	1	2	3	4	5	6	7	8	9
m_i	16	15	19	13	14	19	14	12	17	13

4.3. За спостереженнями, наведеними в таблиці, за допомогою критерію χ^2 перевірити узгодженість із нормальним розподілом.

а) $\alpha = 0.05$,

Інтервал	[0;5)	[5;10)	[10;15)	[15;20)	[20;25)
m_i	15	75	100	50	10

б) $\alpha = 0.01$,

Інтервал	[3,0;3,6)	[3,6;4,2)	[4,2;4,8)	[4,8;5,4)	[5,4;6,0)	[6,0;6,6)
m_i	2	8	35	43	22	10

в) $\alpha = 0.05$,

Інтервал	[-3;-1)	[-1;0)	[0;1)	[1;2)	[2;3)	[3;5)
m_i	13	15	24	25	13	10

г) $\alpha = 0.1$,

Інтервал	[-8;-2)	[-2;4)	[4;10)	[10;16)
m_i	10	50	30	10

д) $\alpha = 0.05$,

Інтервал	[-4;0)	[0;2)	[2;4)	[4; 6)
m_i	25	40	25	10

4.4. Для таких 50 реалізацій випадкової величини:

2,03	2,25	2,94	2,30	1,00	2,18	1,93	1,60	1,52	2,42
2,32	1,43	1,79	2,07	1,89	1,49	1,31	2,58	2,17	1,53
2,55	2,46	2,65	1,68	1,81	1,21	2,34	2,00	1,35	2,53
2,49	1,30	2,79	2,76	2,60	1,25	1,71	2,57	1,70	1,65
1,58	1,93	2,84	1,03	2,85	2,25	2,85	2,45	1,37	1,90

за критерієм Колмогорова – Смірнова або χ^2 перевірити гіпотезу про рівномірний розподіл на рівні значущості 0,1.

4.5. Для таких 50 реалізацій випадкової величини:

9	8	8	6	10	6	4	7	6	10
6	5	7	6	6	4	12	2	12	4
5	6	10	7	8	3	9	6	4	8
7	4	8	4	6	2	10	5	6	3
9	8	5	4	7	8	7	4	3	7

за критерієм χ^2 перевірити: а) гіпотезу про пуассонівський розподіл; б) гіпотезу про пуассонівський розподіл з параметром $\lambda = 7$ із надійністю 0,9.

4.6. Для таких 50 реалізацій випадкової величини:

2,30	2,04	3,62	2,21	1,82	2,82	2,77	1,44	1,72	2,69
1,72	4,08	2,11	0,91	2,82	1,82	2,91	0,14	0,76	1,45
1,59	1,80	2,33	2,25	1,76	3,31	1,92	3,32	0,60	1,81
2,96	1,33	3,01	-1,23	0,76	2,19	2,82	2,95	3,22	2,48
1,56	5,06	0,61	1,83	2,04	1,57	1,49	1,07	3,06	1,77

за критерієм χ^2 перевірити: а) гіпотезу про нормальний розподіл; б) гіпотезу про нормальний розподіл $N(2;1)$ з надійністю 0,95.

4.7. Для таких 50 реалізацій випадкової величини:

9,68	4,31	27,89	5,67	12,62	0,68	16,79	7,05	9,10	6,59
10,87	2,07	4,84	6,65	3,42	25,28	3,40	61,42	3,51	5,53
26,68	9,85	13,96	1,34	5,72	3,12	17,00	17,03	2,79	8,36
33,51	7,21	129,97	47,03	3,22	15,63	7,78	22,26	5,26	8,26
1,53	33,77	0,87	2,91	19,83	36,54	4,93	5,64	3,44	2,86

за критерієм χ^2 перевірити: а) гіпотезу про логнормальний розподіл; б) гіпотезу про логнормальний розподіл $\log N(2; 1)$ з надійністю 0,9.

4.8. Для таких 50 реалізацій випадкової величини:

0	7	0	8	0	4	0	0	0	2
1	0	0	2	1	0	2	2	0	0
1	0	3	10	1	0	4	3	1	0
1	3	0	1	2	0	0	0	0	3
0	1	6	0	4	6	4	3	2	0

за критерієм χ^2 перевірити: а) гіпотезу про геометричний розподіл; б) гіпотезу про геометричний розподіл з параметром $p = 1/3$ з надійністю 0,95.

4.9. Для таких 50 реалізацій випадкової величини:

0,57	0,95	0,15	0,40	0,56	0,29	1,64	0,18	0,62	0,00
0,92	0,43	0,16	1,22	0,56	0,21	0,02	0,22	0,15	0,49
0,27	0,38	0,05	0,31	0,20	0,09	0,28	0,05	0,30	0,32
0,41	0,49	0,81	0,33	0,71	1,59	0,58	0,59	0,18	0,14
0,04	0,07	0,03	0,45	0,16	0,78	0,25	0,08	0,02	0,31

за критерієм χ^2 перевірити: а) гіпотезу про експоненціальний розподіл; б) гіпотезу про експоненціальний розподіл з параметром $\lambda = 2$ з надійністю 0,95.

4.10. Рентгенівське випромінювання викликає в органічних клітинах певну перебудову хромосом. У таблиці наведено результати експерименту (підраховувалась кількість перебудов хромосом під впливом рентгенівських променів).

i	0	1	2	3	4 і більше	Усього
n_i	434	195	44	9	0	682

Тут i – кількість змін у клітині, n_i – кількість клітин з i змінами.

Чи узгоджується з наведеними даними гіпотеза про пуассонівський розподіл кількості перебудов у клітині?

4.11. Із продукції двох верстатів зробили дві вибірки по 38 виробів:

Розмір деталі	24	26	28	30	32	34	36	38	40	42	44	46	48	50	52	54	56
$m_i^{(1)}$	2	1	2	2	1	1	4	1	1	0	5	0	0	6	4	3	5
$m_i^{(2)}$	2	0	0	2	3	0	1	6	0	5	3	1	5	3	3	2	4

Перевірити, використовуючи критерій Смірнова – Колмогорова, гіпотезу про те, що ці вибірки належать одній генеральній сукупності при рівні значущості $\alpha = 0{,}1$.

4.12. У першому потоці з 300 абітурієнтів оцінку "2" отримало 33 особи, "3" – 43 особи, "4" – 80 осіб, "5" – 144. У другому потоці інші 300 абітурієнтів мали такий результат: "2" – 39 осіб, "3" – 35, "4" – 72, "5" – 154. Чи можна вважати обидва потоки однорідними при рівні значущості 0,05?

4.13. У таблиці наведено результати обстеження 697 школярів. Хлопці були впорядковані за рівнем IQ та відповідно до умов їх проживання вдома. При цьому використано позначення: A – дуже здібний, B – досить здібний, C – має середні здібності, D – недостатньо розвинутий, E – розумово відсталий. Чи можна вважати, що умови життя (забезпеченість) дітей впливають на їхні здібності?

Забезпеченість	Здібність хлопців					Усього
	A	B	C	D	E	
Хороша	33	137	125	47	8	350
Погана	21	127	129	61	9	347
Усього	54	264	254	108	17	697

4.14. За допомогою критерію χ^2 для $\alpha = 0{,}05$ перевірити гіпотезу однорідності двох вибірок, наведених у таблиці:

x_i	1	2	3	4	5	6	7	8
$m_i^{(1)}$	4	4	15	51	22	3	1	0
$m_i^{(2)}$	1	1	8	43	34	7	3	3

4.15. Двовимірна випадкова величина $(\xi_0;\eta_0)$ може набувати чотирьох значень: $(0;0)$, $(0;1)$, $(1;0)$, $(1;1)$. 180 незалежних спостережень дали такі результати: значення $(0;0)$ з'явилося 39 разів, $(0;1)$ – 50, $(1;0)$ – 53, $(1;1)$ – 38 разів. Чи можна вважати, що ξ_0 та η_0 – незалежні? Рівень значущості прийняти $\alpha = 0{,}1$.

4.16. Проведено 200 спостережень над випадковими величинами ξ_0 та η_0, які набувають значень 1, 2 та 1, 2, 3, відповідно. Результати спостережень наведені у таблиці:

ξ_0 \ η_0	1	2	3	$\nu_{i.}$
1	25	50	25	100
2	52	41	7	100
$\nu_{.j}$	77	91	32	200

Перевірити за допомогою критерію χ^2, чи будуть незалежними випадкові величини ξ_0 та η_0 при $\alpha = 0.05$.

4.17. Серед 300 осіб, які вступали до університету, 97 мали оцінку "5" у школі, 48 отримали "5" на вступних іспитах з тієї самої дисципліни, причому лише 18 осіб мали "5" і в школі, і на вступних іспитах. На рівні значущості 0,1 перевірити гіпотезу незалежності оцінок "5" у школі й на вступних іспитах.

4.18. У таблиці (Грінвуд, Юл, 1915) наведено дані про 818 випадків, класифікованих за двома ознаками: наявність щеплення проти холери та відсутність захворювання.

Чи можна на основі цих даних дійти висновку про залежність між відсутністю захворювання та наявністю щеплення?

Наявність щеплення	Наявність захворювання		Усього
	Не захворіли	Захворіли	
Щеплені	276	3	279
Нещеплені	473	66	539
Усього	749	69	818

Розділ 5
ПАРАМЕТРИЧНІ ГІПОТЕЗИ

5.1. Поняття параметричної гіпотези

Нехай $\xi' = (\xi_1, \xi_2, \ldots, \xi_n)$ – незалежні спостереження випадкової величини ξ_0. Параметричні гіпотези – це гіпотези про справжнє значення невідомого параметра, який визначає сім'ю розподілів $F_{\xi_0}(x) \in \mathbb{F} = \{F(x,\theta), \theta \in \Theta\}$, $\theta' = (\theta_1, \theta_2, \ldots, \theta_r) \in \Theta \subseteq R^r$. У загальному випадку параметричну гіпотезу можна подати таким чином: $H_0 : \theta \in \Theta_0$. Альтернативна гіпотеза має вигляд $H_1 : \theta \in \Theta_1 = \Theta \setminus \Theta_0$. Точки $\theta \in \Theta_1$ називають альтернативами. Якщо множина Θ_0 складається з однієї точки, то гіпотеза H_0 називається простою, у протилежному випадку – складною. За вибіркою $\xi' = (\xi_1, \xi_2, \ldots, \xi_n)$ треба перевірити, чи вірна гіпотеза H_0 відносно альтернативи H_1.

Приклади параметричних гіпотез:

1) $H_0 : \theta = \theta_0$, де $\theta_0 \in \Theta$ – деяке фіксоване значення параметра;

2) $H_0 : \theta_1 = \theta_2 = \ldots = \theta_r$;

3) $H_0 : g(\theta) = g_0$, де g_0 – фіксоване значення, а $g(\theta)$ – функція параметра θ.

Тут 1) – проста гіпотеза; 2) – складна; 3) – може бути і простою, і складною.

Приклад 5.1. Нехай $\mathbb{F} = N(\theta_1, \theta_2^2)$. Тоді

$H_0 : \theta_1 = \theta_{10}$, $\theta_2 = \theta_{20}$ – проста гіпотеза;

$H_0 : \theta_1 = \theta_{10}$ – складна гіпотеза.

5.2. Критерії перевірки гіпотези

Для перевірки сформульованої гіпотези $H_0 : \theta \in \Theta_0$ потрібен критерій (правило), який давав би можливість для кожної реалізації x вибірки ξ обрати одне з двох рішень: прийняти гіпотезу H_0 (відхилити H_1) або відхилити її (прийняти H_1). Тому вибірковий простір $\mathbf{X}$ розбиваємо на дві підмножини: $\mathbf{X}_0$ та $\mathbf{X}_1$:

$$\mathbf{X}_0 \cup \mathbf{X}_1 = \mathbf{X} \ (\mathbf{X}_0 \cap \mathbf{X}_1 = \varnothing).$$

При $x \in \mathbf{X}_0$ приймається H_0, при $x \in \mathbf{X}_1$ приймається H_1 (тобто H_0 відхиляється). $\mathbf{X}_0$ – область прийняття (ухвалення) гіпотези, $\mathbf{X}_1$ – область її відхилення (критична область).

Отже, критерій повністю визначається заданням критичної області $\mathbf{X}_1$. Його часто називають $\mathbf{X}_1$ ***-критерієм.***

Загальний принцип вибору критичної області критерію. При виборі критичної області треба мати на увазі, що, приймаючи або відхиляючи гіпотезу, можна допустити похибки двох видів.

Похибка першого роду*:* відхилення H_0, коли вона справедлива. Будемо позначати ймовірність цієї похибки через $\alpha = P(H_1 / H_0)$ і називати ***рівнем значущості критерію***.

Похибка другого роду*:* прийняти H_0, коли справедлива гіпотеза H_1. $\beta = P(H_0 / H_1)$ – імовірність похибки другого роду. Імовірність $1-\beta$ називається ***потужністю критерію***.

Бажано було б побудувати такий критерій перевірки гіпотези, щоб імовірності α та β були мінімальними. Очевидно, що за заданої кількості випробувань (спостережень) n одночасно зменшити похибки першого та другого роду неможливо.

Уведемо $W(\theta) = W(\mathbf{X}_1;\theta) = P_\theta(\xi \in \mathbf{X}_1)$, $\theta \in \Theta$ – ***функцію потужності критерію*** $\mathbf{X}_1$. $P(H_1 / H_0) = W(\theta)$, $\theta \in \Theta_0$ – імовір-

ність похибки першого роду; $P(H_0 / H_1) = 1 - W(\theta),\ \theta \in \Theta_1$ – імовірність похибки другого роду.

Раціональний принцип: за заданої кількості випробувань n фіксується ймовірність похибки першого роду. При цьому обирається та критична область $\mathbf{X}_1$, для якої ймовірність похибки другого роду мінімальна. Таким чином, за фіксованого α обирається така критична область $\mathbf{X}_1$, що

$$\begin{cases} W(\theta) \le \alpha & \text{для всіх } \theta \in \Theta_0, \\ 1 - W(\theta) \to \min & \text{для всіх } \theta \in \Theta_1, \end{cases}$$

або

$$\begin{cases} W(\theta) \le \alpha & \text{для всіх } \theta \in \Theta_0, \\ W(\theta) \to \max & \text{для всіх } \theta \in \Theta_1. \end{cases}$$

Зазвичай $\alpha = 0{,}005;\ 0{,}01;\ 0{,}05$.

Похибки другого роду на практиці призводять до більших втрат, ніж похибки першого роду. Це слід ураховувати, обираючи гіпотезу H_0 або H_1.

Нехай $\mathbf{X}_{1\alpha}$ і $\mathbf{X}^*_{1\alpha}$ – два критерії однакового рівня значущості α для гіпотези H_0. Якщо

$$W(\mathbf{X}^*_{1\alpha}; \theta) \le W(\mathbf{X}_{1\alpha}; \theta) \text{ для всіх } \theta \in \Theta_0 \qquad (5.1)$$

і

$$W(\mathbf{X}^*_{1\alpha}; \theta) \ge W(\mathbf{X}_{1\alpha}; \theta) \text{ для всіх } \theta \in \Theta_1, \qquad (5.2)$$

причому строга нерівність у (5.2) має місце хоча б за одного значення θ, то кажуть, що критерій $\mathbf{X}^*_{1\alpha}$ рівномірно потужніший, ніж критерій $\mathbf{X}_{1\alpha}$.

Якщо співвідношення (5.1) – (5.2) виконуються для довільного критерію $\mathbf{X}_{1\alpha}$, то $\mathbf{X}^*_{1\alpha}$ називають рівномірно найпотужнішим критерієм для перевірки гіпотези H_0.

Уведемо клас незсунених критеріїв:

$$W(\theta) \le \alpha \text{ для всіх } \theta \in \Theta_0;\ W(\theta) \ge \alpha \text{ для всіх } \theta \in \Theta_1.$$

У деяких задачах, для яких рівномірно найпотужніші критерії не існують, можуть існувати рівномірно найпотужніші незсунені критерії.

Часто критична область $\mathbf{X}_1$ задається у вигляді

$$\mathbf{X}_1 = \{x : T(x) \geq c\}.$$

У цьому випадку $T(\xi)$ – ***статистика критерію***.

5.3. Вибір із двох простих гіпотез. Критерій Неймана – Пірсона

Розглянемо випадок $\Theta = \{\theta_0, \theta_1\}$,

$$H_0 : \theta = \theta_0, \; H_1 : \theta = \theta_1,$$

$\mathbb{F} = \{F(x,\theta_0), F(x,\theta_1)\}$, α – задане,

$$\begin{cases} W(\mathbf{X}_{1\alpha};\theta_0) = \alpha, \\ W(\mathbf{X}_{1\alpha};\theta_1) \to \max. \end{cases}$$

Із цих співвідношень шукаємо $\mathbf{X}^*_{1\alpha}$.

Припустимо, що функції розподілу $F(x,\theta_0)$ і $F(x,\theta_1)$ абсолютно неперервні, а щільності $f_0(x)$ та $f_1(x)$ задовольняють умову $f_j(x) > 0, \; j = 0,1$.

Розглянемо статистику ***відношення вірогідності***

$$l(\xi) = \frac{L(\xi,\theta_1)}{L(\xi,\theta_0)} = \frac{\prod\limits_{i=1}^{n} f_1(\xi_i)}{\prod\limits_{i=1}^{n} f_0(\xi_i)}$$

і визначимо функцію $\psi(c) = P_{\theta_0}(l(\xi) \geq c)$:

1) $\psi(0) = 1$;

2) $\psi(c) \to 0$ при $c \to \infty$.

Дійсно,

$$P_{\theta_1}(l(\xi) \geq c) = \int\limits_{x:l(x)\geq c} L(x;\theta_1)dx \geq$$

$$\geq c \int\limits_{x:l(x)\geq c} L(x;\theta_0)dx = cP_{\theta_0}(l(\xi) \geq c) = c\psi(c),$$

тому $\psi(c) \leq 1/c,$ звідки й випливає 2).

Теорема 5.1 (Неймана – Пірсона). *За зроблених припущень існує найпотужніший критерій перевірки гіпотези* H_0. *Цей критерій задається критичною областю* $X^*_{1\alpha} = \{x : l(x) \geq c\}$, *де критична границя* c *визначається з умови* $\psi(c) = \alpha$.

Побудований критерій перевірки гіпотези H_0 називають ***критерієм Неймана – Пірсона***.

5.4. Перевірка гіпотези про математичне сподівання в нормальній моделі

Розглянемо ситуацію, коли щодо нормально розподіленої випадкової величини ξ_0 з невідомим середнім θ і відомої дисперсії σ^2 існують дві гіпотези:

$$H_0 : \theta = \theta_0; \quad H_1 : \theta = \theta_1.$$

Для визначеності вважатимемо, що $\theta_1 > \theta_0$.

Нехай $\xi' = (\xi_1, \xi_2, \ldots, \xi_n)$ – вибірка з розподілу випадкової величини ξ_0 та x – реалізація ξ, що спостерігалася. Тоді

$$l(x) = \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[(x_i - \theta_1)^2 - (x_i - \theta_0)^2]\right\} =$$

$$= \exp\left\{\frac{n}{\sigma^2}(\theta_1 - \theta_0)\bar{x} - \frac{n}{2\sigma^2}(\theta_1^2 - \theta_0^2)\right\}.$$

Нерівність $l(x) \geq c$ еквівалентна нерівності

$$\overline{x} \geq \frac{\sigma^2 \ln c}{n(\theta_1 - \theta_0)} + \frac{\theta_1 + \theta_0}{2},$$

яку можна подати у вигляді

$$\frac{\sqrt{n}}{\sigma}(\overline{x} - \theta_0) \geq \frac{\sigma}{\sqrt{n}(\theta_1 - \theta_0)} \ln c + \frac{\sigma}{2\sqrt{n}}(\theta_1 - \theta_0) = t(c).$$

При $\theta = \theta_0$ випадкова величина $\frac{\sqrt{n}}{\sigma}(\overline{\xi} - \theta_0)$ має стандартний нормальний розподіл, тому

$$\psi(c) = P_{\theta_0}(l(\xi) \geq c) = P_{\theta_0}\left(\frac{\sqrt{n}}{\sigma}(\overline{\xi} - \theta_0) \geq t(c)\right) =$$

$$= 1 - \Phi(t(c)) = \Phi(-t(c)).$$

Для довільного $\alpha \in (0,1)$ послідовно визначаємо величини t_α і c_α такі, що $\Phi(-t_\alpha) = \alpha$ та $t(c_\alpha) = t_\alpha$.

Умови теореми Неймана – Пірсона виконуються, отже, найпотужніший критерій для перевірки гіпотези H_0 проти альтернативи H_1 задаємо критичною областю

$$\mathsf{X}^*_{1\alpha} = \{x : \frac{\sqrt{n}}{\sigma}(\overline{x} - \theta_0) \geq t_\alpha\}, \ \Phi(-t_\alpha) = \alpha.$$

Обчислимо потужність критерію:

$$W(\mathsf{X}^*_{1\alpha}; \theta_1) = P_{\theta_1}\left(\frac{\sqrt{n}}{\sigma}(\overline{\xi} - \theta_0) \geq t_\alpha\right) = P_{\theta_1}\left(\overline{\xi} \geq \theta_0 + \frac{\sigma}{\sqrt{n}} t_\alpha\right) =$$

$$= P_{\theta_1}\left(\frac{\sqrt{n}}{\sigma}(\overline{\xi} - \theta_1) \geq -\frac{\sqrt{n}}{\sigma}(\theta_1 - \theta_0) + t_\alpha\right) =$$

$$= 1 - \Phi\left(-\frac{\sqrt{n}}{\sigma}(\theta_1 - \theta_0) + t_\alpha\right) = \Phi\left(\frac{\sqrt{n}}{\sigma}(\theta_1 - \theta_0) - t_\alpha\right).$$

Звідси випливає, що ймовірність похибки другого роду становить

$$\beta = \beta(\alpha, n) = \Phi(t_\alpha - \sqrt{n}(\theta_1 - \theta_0)/\sigma).$$

Розглянемо **задачу.** Нехай заздалегідь задані ймовірності похибок першого та другого роду. Визначити, якою має бути мінімальна кількість $n^* = n^*(\alpha, \beta)$ випробувань, щоб хибні висновки могли бути зроблені з імовірностями, які не перевищують α та β.

Для визначення n маємо два рівняння:

$$\Phi(-t_\alpha) = \alpha,\ \Phi(t_\alpha - \sqrt{n}(\theta_1 - \theta_0)/\sigma) = \beta.$$

Нехай k_p – це квантиль p-го порядку, або розв'язок рівняння $\Phi(k_p) = p$. Тоді

$$-t_\alpha = k_\alpha;\ t_\alpha - \sqrt{n}(\theta_1 - \theta_0)/\sigma = k_\beta \text{ і}$$

$$n = \frac{\sigma^2(k_\alpha + k_\beta)^2}{(\theta_1 - \theta_0)^2}, \text{ або } n^* = \left[\frac{\sigma^2(k_\alpha + k_\beta)^2}{(\theta_1 - \theta_0)^2}\right] + 1.$$

Звідси випливає, що за фіксованих похибок кількість спостережень пропорційна дисперсії й обернено пропорційна квадрату різниці між середніми значеннями.

Підсумовуючи вищевказане, наведемо деякі алгоритми перевірки гіпотез про параметри нормальної моделі.

Гіпотеза $H_0: m = m_0$ ($H_0: m = m_0$ невідома).

Зазначимо, що така гіпотеза може мати різні альтернативи ($m \neq m_0$, $m > m_0$, $m < m_0$).

Оскільки незалежно від того, справедлива гіпотеза H_0: $m = m_0$ чи ні, оцінка $\bar{\xi} = \frac{1}{n}\sum_{k=1}^{n} \xi_k$ є конзистентною та незсуненою для m, то логічно прийняти гіпотезу $H_0: m = m_0$ за невеликих значень $\bar{\xi} - m_0$ і відхилити її за великих значень $\bar{\xi} - m_0$. Межі, що відділяють великі значення $\bar{\xi} - m_0$ від невеликих, будуються

на основі того факту, що для вибірки $\xi' = (\xi_1, \xi_2, \ldots, \xi_n)$ величина $\dfrac{\overline{\xi} - m_0}{\hat{S}/\sqrt{n}}$ має розподіл Стьюдента із $n-1$ ступенем свободи, де $\hat{S}^2 = \dfrac{1}{n-1}\sum_{k=1}^{n}\left(\xi_k - \overline{\xi}\right)^2$.

Критерій Стьюдента перевірки гіпотези $H_0: m = m_0$.

а) За альтернативи $H_1: m \neq m_0$ гіпотеза $H_0: m = m_0$ відхиляється при

$$\frac{\left|\overline{\xi} - m_0\right|}{\hat{S}/\sqrt{n}} > t_{1-\frac{\alpha}{2},n-1},$$

де $t_{1-\frac{\alpha}{2},n-1}$ – квантиль рівня $1-\dfrac{\alpha}{2}$ розподілу Стьюдента з $n-1$ ступенями свободи. У протилежному випадку гіпотезу $H_0: m = m_0$ приймаємо. При цьому з імовірністю α (рівень значущості) гіпотеза H_0 буде відхилятися, коли вона справедлива.

б) За альтернативи $H_1: m > m_0$ гіпотеза $H_0: m = m_0$ відхиляється при

$$\frac{\overline{\xi} - m_0}{\hat{S}/\sqrt{n}} > t_{1-\alpha,n-1}.$$

У протилежному випадку гіпотезу $H_0: m = m_0$ приймаємо (рівень значущості α).

в) За альтернативи $H_1: m < m_0$ гіпотеза $H_0: m = m_0$ відхиляється при

$$\frac{\overline{\xi} - m_0}{\hat{S}/\sqrt{n}} < t_{\alpha,n-1}.$$

У протилежному випадку гіпотезу $H_0: m = m_0$ приймаємо (рівень значущості α).

Приклад 5.2. Для перевірки твердження виробника про те, що генератор за зміну споживає в середньому 20 л пального, здійснили 10 випробувань. За 10 змін споживання генератора встановили: 19,1; 18,6; 19,1; 18,1; 16,6; 20,1; 19,8; 21,1; 24,4; 21,6. Перевірити твердження виробника при рівні значущості 0.05.

Розв'язання. $H_0: m = m_0 = 20$, $H_1: m \neq m_0$,

$$\bar{\xi} = \frac{1}{10}\big(19{,}1 + 18{,}6 + 19{,}1 + 18{,}1 + 16{,}6 +$$

$$+20{,}1 + 19{,}8 + 21{,}1 + 24{,}4 + 21{,}6\big) = 19{,}85.$$

$$\hat{S}^2 = \frac{1}{9}\Big((19{,}1-19{,}85)^2 + (18{,}6-19{,}85)^2 + \ldots + (21{,}6-19{,}85)^2\Big) =$$

$$= \frac{41.705}{9} = \frac{8341}{1800}.$$

$$\frac{\left|\bar{\xi} - m_0\right|}{\hat{S}/\sqrt{n}} \approx 0{,}22 \quad < \quad t_{1-\frac{\alpha}{2},n-1} = t_{1-\frac{0.05}{2},10-1} = t_{0.975,9} = 2{,}262.$$

Це означає, що дані не суперечать твердженню виробника про те, що генератор за зміну споживає в середньому 20 л пального, і гіпотеза $H_0: m = m_0 = 20$ приймається.

Гіпотеза $H_0: \sigma^2 = \sigma_0^2$ (m невідоме).

Альтернативна гіпотеза при цьому може бути як однобічною ($H_1: \sigma^2 > \sigma_0^2$ чи $H_1: \sigma^2 < \sigma_0^2$), так і двобічною ($H_1: \sigma^2 \neq \sigma_0^2$).

Незалежно від того, справедлива чи ні гіпотеза $H_0: \sigma^2 = \sigma_0^2$, оцінка $\hat{S}^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(\xi_k - \bar{\xi}\right)^2$ є конзистентною та незсуненою оцінкою параметра σ^2, тому гіпотезу $H_0: \sigma^2 = \sigma_0^2$ слід приймати, якщо $\frac{\hat{S}^2}{\sigma_0^2}$ не дуже відхиляться від 1. Межі, що відділяють значення $\frac{\hat{S}^2}{\sigma_0^2}$, які дуже відхиляються від 1, від значень $\frac{\hat{S}^2}{\sigma_0^2}$, які

мало відхиляються від 1, будують на базі того, що $\frac{(n-1)\hat{S}^2}{\sigma_0^2}$ має χ^2-розподіл із $(n-1)$ ступенями свободи.

Критерій перевірки гіпотези $H_0: \sigma^2 = \sigma_0^2$**.**

а) За альтернативи $H_1: \sigma^2 \neq \sigma_0^2$ гіпотеза $H_0: \sigma^2 = \sigma_0^2$ приймається при

$$\hat{S}^2 \in \left(\frac{\sigma_0^2}{n-1}\chi^2_{\frac{\alpha}{2},n-1}; \frac{\sigma_0^2}{n-1}\chi^2_{1-\frac{\alpha}{2},n-1} \right).$$

У протилежному випадку гіпотеза $H_0: \sigma^2 = \sigma_0^2$ відхиляється (рівень значущості α).

б) За альтернативи $H_1: \sigma^2 > \sigma_0^2$ гіпотеза $H_0: \sigma^2 = \sigma_0^2$ приймається при

$$\hat{S}^2 < \frac{\sigma_0^2}{n-1}\chi^2_{1-\alpha,n-1}.$$

У протилежному випадку гіпотеза $H_0: \sigma^2 = \sigma_0^2$ відхиляється (рівень значущості α).

в) За альтернативи $H_1: \sigma^2 < \sigma_0^2$ гіпотеза $H_0: \sigma^2 = \sigma_0^2$ приймається при

$$\hat{S}^2 > \frac{\sigma_0^2}{n-1}\chi^2_{\alpha,n-1}.$$

У протилежному випадку гіпотеза $H_0: \sigma^2 = \sigma_0^2$ відхиляється (рівень значущості α).

Нехай $\xi^{(1)\prime} = \left(\xi_1^{(1)}, \xi_2^{(1)}, \ldots, \xi_{n_1}^{(1)}\right)$ – вибірка з нормального розподілу $N\left(m_1, \sigma_1^2\right)$, а $\xi^{(2)\prime} = \left(\xi_1^{(2)}, \xi_2^{(2)}, \ldots, \xi_{n_2}^{(2)}\right)$ – вибірка з нормального розподілу $N\left(m_2, \sigma_2^2\right)$, причому параметри $m_1, m_2, \sigma_1^2, \sigma_2^2$ невідомі.

Гіпотеза $H_0: m_1 - m_2 = c$ (коли $\sigma_1^2 = \sigma_2^2 = \sigma^2$).

Позначимо $m = m_1 - m_2$ та $\Delta = \overline{\xi}^{(1)} - \overline{\xi}^{(2)}$. Незалежно від того, справедлива чи ні гіпотеза $H_0: m_1 - m_2 = c$, Δ є конзистентною та незсуненою оцінкою для c, тому гіпотезу H_0 слід приймати за незначних відхилень $\Delta - c$. Межі, які відділяють великі відхилення від малих, будуються на основі того, що $\dfrac{\Delta - c}{\hat{S}\sqrt{\dfrac{n_1 + n_2}{n_1 n_2}}}$ має розподіл Стьюдента з $n_1 + n_2 - 2$ ступенями свободи, де $\hat{S}^2 = \dfrac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{n_1 + n_2 - 2}$, $\hat{S}_i^2 = \dfrac{1}{n_i - 1}\sum_{k=1}^{n_i}\left(\xi_k^{(i)} - \overline{\xi^{(i)}}\right)^2$, $(i = 1, 2)$.

Критерій Стьюдента перевірки гіпотези $H_0: m_1 - m_2 = c$.

а) За альтернативи $H_1: m_1 - m_2 \neq c$ гіпотеза $H_0: m_1 - m_2 = c$ відхиляється при

$$\frac{|\Delta - c|}{\hat{S}\sqrt{\dfrac{n_1 + n_2}{n_1 n_2}}} > t_{1-\frac{\alpha}{2}, n_1 + n_2 - 2}.$$

У протилежному випадку гіпотезу $H_0: m_1 - m_2 = c$ приймаємо (рівень значущості α).

б) За альтернативи $H_1: m_1 - m_2 > c$ гіпотеза $H_0: m_1 - m_2 = c$ відхиляється при

$$\frac{\Delta - c}{\hat{S}\sqrt{\dfrac{n_1 + n_2}{n_1 n_2}}} > t_{1-\alpha, n_1 + n_2 - 2}.$$

У протилежному випадку гіпотезу $H_0: m_1 - m_2 = c$ приймаємо (рівень значущості α).

в) За альтернативи $H_1:\ m_1 - m_2 < c$ гіпотеза $H_0:\ m_1 - m_2 = c$ відхиляється при

$$\frac{\Delta - c}{\hat{S}\sqrt{\dfrac{n_1 + n_2}{n_1 n_2}}} < t_{\alpha, n_1+n_2-2}\,.$$

У протилежному випадку гіпотезу $H_0:\ m_1 - m_2 = c$ приймаємо (рівень значущості α).

Приклад 5.3. Нижче наведено дані вимірювання рівня чутливості однорідного фотоматеріалу за допомогою двох фотосенсометрів. Чи можна вважати, що між показниками приладів немає систематичного розходження (рівень значущості 0,05)?

Перший прилад: 5,0; 5,1; 5,5; 6,0; 5,5; 3,6; 1,8; 7,8; 6,9; 2,7;

Другий прилад: 6,1; 5,4; 5,8; 2,7; 2,8; 2,8; 4,2; 2,4.

Розв'язання. $H_0:\ m_1 - m_2 = 0$, $H_1: m_1 - m_2 \neq 0$,

$$\overline{\xi}^{(1)} = \frac{1}{10}\left(5{,}0 + 5{,}1 + 5{,}5 + 6{,}0 + 5{,}5 + 3{,}6 + 1{,}8 + 7{,}8 + 6{,}9 + 2{,}7\right) = 4{,}99,$$

$$\overline{\xi}^{(2)} = \frac{1}{8}\left(6{,}1 + 5{,}4 + 5{,}8 + 2{,}7 + 2{,}8 + 2{,}8 + 4{,}2 + 2{,}4\right) = 4{,}025\,,$$

$$\Delta = \overline{\xi}^{(1)} - \overline{\xi}^{(2)} = 4{,}99 - 4{,}025 = 0{,}965,$$

$$\hat{S}_1^2 = \frac{1}{9}\left(\left(5{,}0 - 4{,}99\right)^2 + \left(5{,}1 - 4{,}99\right)^2 + \ldots + \left(2{,}7 - 4{,}99\right)^2\right) \approx 3{,}38,$$

$$\hat{S}_2^2 = \frac{1}{7}\left(\left(6{,}1 - 4{,}025\right)^2 + \left(5{,}4 - 4{,}025\right)^2 + \ldots + \left(2{,}4 - 4{,}025\right)^2\right) \approx 2{,}4,$$

$$\hat{S}^2 = \frac{\left(n_1 - 1\right)\hat{S}_1^2 + \left(n_2 - 1\right)\hat{S}_2^2}{n_1 + n_2 - 2} \approx \frac{9 \cdot 3{,}38 + 7 \cdot 2{,}4}{10 + 8 - 2} = \frac{47{,}22}{16} \approx 2{,}95,$$

$$\frac{\left|\Delta - c\right|}{\hat{S}\sqrt{\dfrac{n_1 + n_2}{n_1 n_2}}} \approx \frac{\left|0{,}965 - 0\right|}{\sqrt{2{,}95 \cdot \dfrac{10 + 8}{10 \cdot 8}}} \approx 1{,}1844 \quad < \quad t_{1 - \frac{0{,}05}{2},\, 10+8-2} = t_{0.975,16} = 2{,}12,$$

а це означає, що дані не суперечать гіпотезі про те, що між показниками приладів немає систематичного розходження, і ми приймаємо цю гіпотезу.

Гіпотеза $H_0: \sigma_1^2 = \sigma_2^2$.

Незалежно від того, справедлива чи ні гіпотеза $H_0: \sigma_1^2 = \sigma_2^2$, $\hat{S}_1^2$ та $\hat{S}_2^2$ є конзистентними та незсуненими оцінками для σ_1^2 та σ_2^2, відповідно, тому, як відомо, $\frac{\hat{S}_1^2}{\hat{S}_2^2}$ збігається за ймовірністю до $\frac{\sigma_1^2}{\sigma_2^2}$. Отже, за справедливості гіпотези $H_0: \sigma_1^2 = \sigma_2^2$ $\frac{\hat{S}_1^2}{\hat{S}_2^2}$ має не дуже відхилятися від 1. Межі, що відділяють значні відхилення від незначних, отримують, виходячи з того, що $\frac{\hat{S}_1^2}{\hat{S}_2^2}$ має F-розподіл (Фішера) із $n_1 - 1$ та $n_2 - 1$ ступенями свободи.

Критерій для перевірки гіпотези $H_0: \sigma_1^2 = \sigma_2^2$.

а) За альтернативи $H_1: \sigma_1^2 \neq \sigma_2^2$ гіпотеза $H_0: \sigma_1^2 = \sigma_2^2$ приймається при

$$\frac{\hat{S}_1^2}{\hat{S}_2^2} \in \left[\frac{1}{F_{1-\frac{\alpha}{2}, n_2-1, n_1-1}};\ F_{1-\frac{\alpha}{2}, n_1-1, n_2-1} \right],$$

де $F_{\alpha,m,n}$ – квантиль розподілу Фішера рівня α зі ступенями свободи m, n.

У протилежному випадку гіпотезу $H_0: \sigma_1^2 = \sigma_2^2$ відхиляємо (рівень значущості α).

б) За альтернативи $H_1: \sigma_1^2 > \sigma_2^2$ гіпотеза $H_0: \sigma_1^2 = \sigma_2^2$ приймається при

$$\frac{\hat{S}_1^2}{\hat{S}_2^2} \leq F_{1-\alpha, n_1-1, n_2-1}.$$

У протилежному випадку гіпотезу $H_0: \sigma_1^2 = \sigma_2^2$ відхиляємо (рівень значущості α).

в) За альтернативи $H_1\colon\ \sigma_1^2 < \sigma_2^2$ гіпотеза $H_0\colon\ \sigma_1^2 = \sigma_2^2$ приймається при

$$\frac{\hat{S}_1^2}{\hat{S}_2^2} \geq \frac{1}{F_{1-\alpha,n_2-1,n_1-1}}.$$

У протилежному випадку гіпотезу $H_0\colon\ \sigma_1^2 = \sigma_2^2$ відхиляємо (рівень значущості α).

Приклад 5.4. Чи можна в умовах попереднього прикладу вважати, що точність вимірювань на обох фотосенсометрах однакова (рівень значущості 0,05)?

Розв'язання. $H_0\colon\ \sigma_1^2 = \sigma_2^2$, $H_1\colon\ \sigma_1^2 \neq \sigma_2^2$, $\dfrac{\hat{S}_1^2}{\hat{S}_2^2} \approx \dfrac{3{,}38}{2{,}4} \approx 1.4$,

$$F_{1-\frac{\alpha}{2};n_2-1;n_1-1} = F_{1-\frac{0{,}05}{2};8-1;10-1} = F_{0{,}975;7;9} = 4{,}197,$$

$$F_{1-\frac{\alpha}{2};n_1-1;n_2-1} = F_{0{,}975;9;7} = 4{,}82,$$

$$\frac{\hat{S}_1^2}{\hat{S}_2^2} \approx 1{,}4 \in \left[\frac{1}{F_{1-\frac{\alpha}{2},n_2-1,n_1-1}};\ F_{1-\frac{\alpha}{2},n_1-1,n_2-1}\right] = \left[\frac{1}{4{,}197};4{,}82\right] = \left[0{,}238;4{,}82\right],$$

а це означає, що дані не суперечать гіпотезі про те, що точність вимірювань на обох фотосенсометрах однакова, і ми приймаємо цю гіпотезу.

5.5. Перевірка гіпотез про рівність математичних сподівань і дисперсій двох нормальних вибірок

Нехай ξ_0 та η_0 – дві незалежні випадкові величини, кожна з яких має нормальний розподіл $N\left(a_1,\sigma_1^2\right)$ і $N\left(a_2,\sigma_2^2\right)$, відповід-

но. У результаті спостережень цих випадкових величин отримано дві вибірки: $\xi' = \left(\xi_1, ..., \xi_{n_1}\right)$ та $\eta' = \left(\eta_1, ..., \eta_{n_2}\right)$.

1. Гіпотеза про рівність математичних сподівань за відомих дисперсій.

Необхідно перевірити гіпотезу $H_0 : a_1 = a_2$ проти альтернативної гіпотези $H_1 : |a_1 - a_2| > 0$; σ_1^2 і σ_2^2 – відомі.

Випадкові величини $\bar{\xi} = \frac{1}{n_1}\sum_{i=1}^{n_1}\xi_i$, $\bar{\eta} = \frac{1}{n_2}\sum_{i=1}^{n_2}\eta_i$ мають нормальний розподіл $N\left(a_1, \frac{\sigma_1^2}{n_1}\right)$ і $N\left(a_2, \frac{\sigma_2^2}{n_2}\right)$, відповідно. Тоді $\bar{\xi} - \bar{\eta}$ має нормальний розподіл $N\left(a_1 - a_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$. Якщо справедлива гіпотеза H_0, то $\frac{\bar{\xi} - \bar{\eta}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ має стандартний нормальний розподіл $N(0,1)$. Критична область задається нерівністю

$$R_{n_1} = \left\{ (x, y) : \frac{\left|\bar{x} - \bar{y}\right|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq c_{1-\frac{\alpha}{2}} \right\},$$

$x' = \left(x_1, ..., x_{n_1}\right)$, $y' = \left(y_1, ..., y_{n_2}\right)$ – реалізації вибірок ξ' та η', $\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i$, $\bar{y} = \frac{1}{n_2}\sum_{i=1}^{n_2} y_i$ і $\Phi\left(c_{1-\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}$ (див. табл. 2 додатка).

2. Гіпотеза про рівність математичних сподівань за невідомих дисперсій.

Нехай треба перевірити ту саму гіпотезу, що в попередньому випадку, але $\sigma_1^2 = \sigma_2^2 = \sigma^2$ і величина σ^2 не відома. Аналогічно за справедливості гіпотези H_0 величина $\dfrac{\overline{\xi} - \overline{\eta}}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ має нормальний розподіл $N(0,1)$, а величини $\dfrac{(n_1 - 1)\hat{S}_1^2}{\sigma^2} = \sum\limits_{i=1}^{n_1} \dfrac{\left(\xi_i - \overline{\xi}\right)^2}{\sigma^2}$, $\dfrac{(n_2 - 1)\hat{S}_2^2}{\sigma^2} = \sum\limits_{i=1}^{n_2} \dfrac{\left(\eta_i - \overline{\eta}\right)^2}{\sigma^2}$ мають χ^2-розподіл із $n_1 + n_2 - 2$ ступенями свободи.

Оскільки

$$M\left[\frac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{n_1 + n_2 - 2}\right] = \frac{(n_1 - 1)\sigma^2 + (n_2 - 1)\sigma^2}{n_1 + n_2 - 2} = \sigma^2,$$

то величина $\hat{\sigma}^2 = \dfrac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{n_1 + n_2 - 2}$ є незсуненою оцінкою σ^2. Таким чином, випадкова величина $\dfrac{\overline{\xi} - \overline{\eta}}{\hat{\sigma}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ за справедливості гіпотези H_0 має розподіл Стьюдента з $n_1 + n_2 - 2$ ступенями свободи. Тоді критична область задається нерівністю

$$R_{n_1} = \left\{ (x, y) : \frac{\left|\overline{x} - \overline{y}\right|}{\sqrt{\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)\dfrac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{n_1 + n_2 - 2}}} \geq t_{n_1 + n_2 - 2;1 - \frac{\alpha}{2}} \right\},$$

де $\hat{S}_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \overline{x})^2$, $\hat{S}_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_i - \overline{y})^2$. Кількість $t_{n_1+n_2-2;1-\frac{\alpha}{2}}$ знаходимо за таблицею розподілу Стьюдента (табл. 4 додатка) при кількості ступенів свободи $n_1 + n_2 - 2$ на основі умови $P\left\{t_{(n_1+n_2-2)} < t_{n_1+n_2-2;1-\frac{\alpha}{2}}\right\} = 1 - \frac{\alpha}{2}$.

3. Гіпотеза про рівність дисперсій за невідомих математичних сподівань.

Нехай тепер необхідно перевірити гіпотезу $H_0 : \sigma_1^2 = \sigma_2^2$ проти альтернативної гіпотези $H_1 : \sigma_1^2 > \sigma_2^2$.

За справедливості гіпотези H_0 випадкова величина $F = \frac{\hat{S}_1^2}{\hat{S}_2^2}$ має розподіл Снедекора – Фішера з $(n_1 - 1, n_2 - 1)$ ступенями свободи.

Критична область задається нерівністю

$$R_{n_1} = \left\{(x, y) : \frac{\hat{S}_1^2}{\hat{S}_2^2} \geq F_{(1-\alpha, n_1-1, n_2-1)}\right\},$$

де $\hat{S}_1^2 \geq \hat{S}_2^2$, що завжди можна зробити, помінявши індекси. Величину $F_{(1-\alpha, n_1-1, n_2-1)}$ знаходимо за таблицею розподілу Снедекора – Фішера (табл. 6 додатка) на основі умови $P\left\{F_{n_1-1, n_2-1} < F_{(1-\alpha, n_1-1, n_2-1)}\right\} = 1 - \alpha$.

4. Гіпотеза про рівність дисперсій за відомих математичних сподівань.

Ця гіпотеза перевіряється аналогічно попередній, але в даному випадку $F = \frac{\overline{S}_1^2(\xi)}{\overline{S}_2^2(\eta)}$, де $\overline{S}_1^2(\xi) = \frac{1}{n_1}\sum_{i=1}^{n_1}(\xi_i - a_1)^2$,

$\overline{S}_2^2(\eta)=\frac{1}{n_2}\sum_{i=1}^{n_2}(\eta_i-a_2)^2$. Якщо справедлива гіпотеза $H_0:\sigma_1^2=\sigma_2^2$, то випадкова величина F має розподіл Снедекора – Фішера з (n_1,n_2) ступенями свободи. Критична область задається нерівністю

$$R_{n_1}=\left\{(x,y):\frac{\overline{S}_1^2(x)}{\overline{S}_2^2(y)}\geq F_{(1-\alpha,n_1,n_2)}\right\},$$

$$\overline{S}_1^2(x)=\frac{1}{n_1}\sum_{i=1}^{n_1}(x_i-a_1)^2,\ \ \overline{S}_2^2(y)=\frac{1}{n_2}\sum_{i=1}^{n_2}(y_i-a_2)^2.$$

Приклад 5.5. Нижче наведено дані про вимірювання нерівностей поверхні однакової чистоти обробки за допомогою двох подвійних мікроскопів.

Мікроскоп 1: 1,3; 1,9; 3,0; 3,5; 3,7; 2,5; 1,7; 0,9; 1,0; 2,3; 3,3.

Мікроскоп 2: 1,4; 2,1; 3,1; 3,3; 2,7; 1,7; 1,1; 0,2; 1,6; 2,8; 3,4.

Чи можна вважати, що між показниками приладів немає систематичної розбіжності?

У термінах математичної статистики цю задачу можна переформулювати таким чином. Маємо реалізації двох незалежних вибірок $\xi'=(\xi_1,...,\xi_{n_1})$ – дані вимірювань, отримані за допомогою першого мікроскопа, та $\eta'=(\eta_1,...,\eta_{n_2})$ – другого мікроскопа з розподілів $N(a_1,\sigma_1^2)$ і $N(a_2,\sigma_2^2)$, відповідно (припущення про нормальний розподіл результатів спостережень у більшості випадків справджується). Відносно параметрів a_1 та a_2 висувається гіпотеза $H_0:a_1=a_2$. Це гіпотеза про відсутність систематичних розбіжностей між показаннями приладів. Відхилення гіпотези H_0 інтерпретується як наявність таких розбіжностей.

Згідно із критерієм Стьюдента для перевірки гіпотези $H_0:a_1=a_2$ необхідно порівняти значення $|t|=\frac{|\overline{x}-\overline{y}|}{\hat{\sigma}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ із

$t_{n_1+n_2-2;1-\frac{\alpha}{2}}$ – $(1-\frac{\alpha}{2})$ – квантилем розподілу Стьюдента з n_1+n_2-2 ступенями свободи.

Отже, у даному прикладі $n_1=12$, $n_2=12$,

$$\overline{x}=\frac{1}{12}\left(1,3+1,9+3,0+...+2,3+3,3\right)=2,1;$$

$$\overline{y}=1,4+2,1+3,1+...+2,8+3,4=2,0;$$

$$\hat{S}_1^2=\frac{1}{n_1-1}\sum_{i=1}^{n_1}\left(x_i-\overline{x}\right)^2=\frac{1}{11}\left(\left(1,3-2,1\right)^2+...+\left(3,3-2,1\right)^2\right)=$$

$$=\frac{1}{11}\cdot 10,46=0,951;$$

$$\hat{S}_2^2=\frac{1}{n_2-1}\sum_{i=1}^{n_2}\left(y_i-\overline{y}\right)^2=\frac{1}{11}\left(\left(1,4-2,0\right)^2+...+\left(3,4-2,0\right)^2\right)=$$

$$=\frac{1}{11}\cdot 10,75=0,977;$$

$$\hat{\sigma}^2=\frac{\left(n_1-1\right)\hat{S}_1^2+\left(n_2-1\right)\hat{S}_2^2}{n_1+n_2-2}=\frac{11\cdot 0,951+11\cdot 0,977}{22}=0,964,$$

$$\left|t\right|=\frac{\left|\overline{x}-\overline{y}\right|}{\hat{\sigma}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}=\frac{\left|2,1-2,0\right|}{\sqrt{0,964}\sqrt{\frac{1+1}{12}}}=0,25.$$

Маємо $\left|t\right|=0,25<2,074=t_{22;0,975}$. Згідно з критерієм Стьюдента гіпотеза про рівність математичних сподівань на рівні $\alpha=0,05$ приймається, тобто припущення про відсутність систематичних розбіжностей між показаннями мікроскопів не суперечать експериментальним даним (експеримент не дає підстав говорити про наявність систематичних розбіжностей між показаннями мікроскопів).

Приклад 5.6. За двома вибірками розміром $n_1=25$ та $n_2=50$ з генеральних сукупностей випадкових величин ξ та η, що ма-

ють нормальні розподіли $N\left(a_1,\sigma_1^2\right)$ і $N\left(a_2,\sigma_2^2\right)$, підраховано $\overline{x}=9,79,\quad \overline{y}=9,60$. Перевірити гіпотезу $H_0: a_1=a_2$ при $\sigma_1^2=\sigma_2^2=0,09$ та $\alpha=0,01$.

Маємо $c=\dfrac{\left|\overline{x}-\overline{y}\right|}{\sigma\sqrt{1/n_1+1/n_2}}=\dfrac{9,79-9,60}{0,3\sqrt{1/25+1/50}}=2,59$.

Порівняємо значення c із табличним значенням $c_{1-\frac{\alpha}{2}}$ (табл. 2 додатка). $2,59>2,57,$ отже, гіпотезу про рівність математичних сподівань відхиляємо.

Приклад 5.7 (точність вимірювань). Визначається границя міцності на розрив матеріалу на двох різних стендах: А та В. Отримано такі вибірки значень границі міцності на розрив:

Стенд А: 1,32; 1,35; 1,32; 1,35; 1,30; 1,30; 1,37; 1,31; 1,39; 1,39.

Стенд В: 1,35; 1,31; 1,31; 1,41; 1,39; 1,37; 1,32; 1,34.

Чи можна вважати, що точність вимірювань границі міцності на розрив на стендах А та В однакова?

У термінах математичної статистики цю задачу можна переформулювати таким чином. Маємо реалізації двох незалежних вибірок: $\xi'=\left(\xi_1,...,\xi_{n_1}\right)$ – вибірка, отримана на стенді А, та $\eta'=\left(\eta_1,...,\eta_{n_2}\right)$ – вибірка, отримана на стенді В, з нормальних розподілів $N\left(a_1,\sigma_1^2\right)$ і $N\left(a_2,\sigma_2^2\right)$. Відносно невідомих параметрів σ_1^2 і σ_2^2 необхідно перевірити гіпотезу $H_0:\sigma_1^2=\sigma_2^2$ (гіпотеза про однакову точність вимірювань на стендах А та В) проти альтернативної гіпотези $H_1:\sigma_1^2>\sigma_2^2$ або $\sigma_2^2>\sigma_1^2$.

Згідно із критерієм для перевірки такої гіпотези відхиляємо гіпотезу H_0, якщо $\dfrac{\hat{S}_1^2}{\hat{S}_2^2}\geq F_{(1-\alpha,n_1-1,n_2-1)}$, де вибираємо $\hat{S}_1^2\geq\hat{S}_2^2$, а величину $F_{(1-\alpha,n_1-1,n_2-1)}$ знаходимо за таблицею розподілу Снедекора – Фішера на основі умови $P\left\{F_{n_1-1,n_2-1}\leq F_{(1-\alpha,n_1-1,n_2-1)}\right\}=1-\alpha$.

У даному випадку $n_1 = 10$, $n_2 = 8$, $\overline{x} = 1{,}34$, $\overline{y} = 1{,}35$.

$$\hat{S}_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(x_i - \overline{x}\right)^2 = \frac{1}{9}\left(\left(1{,}32 - 1{,}34\right)^2 + \ldots + \left(1{,}39 - 1{,}34\right)^2\right) =$$

$$= 0{,}0012;$$

$$\hat{S}_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}\left(y_i - \overline{y}\right)^2 = \frac{1}{7}\left(\left(1{,}35 - 1{,}35\right)^2 + \ldots + \left(1{,}34 - 1{,}35\right)^2\right) =$$

$$= 0{,}0014.$$

Отримуємо $\dfrac{\hat{S}_1^2}{\hat{S}_2^2} = \dfrac{0{,}0014}{0{,}0012} = 1{,}167$. Порівнюємо це значення з табличним: $\dfrac{\hat{S}_1^2}{\hat{S}_2^2} = 1{,}167 < 3{,}293 = F_{(0{,}95;7;9)}$. Отже, гіпотеза H_0 на рівні значущості $\alpha = 0{,}05$ приймається. Це означає: припущення, що стенди А та В мають однакову точність вимірювань границі міцності на розрив, не суперечить експериментальним даним. Іншими словами, експеримент не дає підстав стверджувати, що точність вимірювань границі міцності на розрив на стендах А та В різна.

Приклад 5.8. За двома вибірками розміром $n_1 = 10$, $n_2 = 15$ із генеральних сукупностей випадкових величин ξ та η, що мають нормальний розподіл, підраховані вибіркові дисперсії $\hat{S}_1^2 = 9{,}6$ і $\hat{S}_2^2 = 5{,}7$. Перевірити гіпотезу про рівність дисперсій випадкових величин ξ та η, $\alpha = 0{,}05$.

Обчислимо $F = \dfrac{\hat{S}_1^2}{\hat{S}_2^2} = \dfrac{9{,}6}{5{,}7} = 1{,}68$, $k_1 = 9$, $k_2 = 14$. Порівняємо значення $F = 1{,}68$ із табличним значенням $F_{(0.95;9;14)} = 2{,}65$. Маємо $1{,}68 < 2{,}65$. Отже, гіпотеза про рівність дисперсій ξ та η приймається.

ЗАДАЧІ

5.1. В експериментальній групі людей, що випробовували новий препарат для схуднення, проводили зважування до та після курсу лікування. За результатами зважувань необхідно визначити, чи істотно змінювалась вага під впливом нових ліків. Рівень значущості обрати 0,05.

До лікування (кг): 120, 112, 116, 113, 119, 130, 122, 125, 127, 120.

Після лікування (кг): 131, 129, 91, 98, 122, 114, 108, 121, 128, 123.

5.2. За документацією дозатор наливає в кожну ємність у середньому 10 л розчину. Для перевірки роботи дозатора зробили тестові розливання. Перевірити з рівнем значущості 0,05 відповідність роботи дозатора до документації.

Значення (л): 10,1; 10; 10,2; 9,9; 9,8; 9,9; 10; 10,1; 10,1; 10; 9,9; 10.

5.3. За значеннями генератора випадкових гауссівських величин перевірити з рівнем значущості 0,05 гіпотезу про одиничну дисперсію.

Значення: 4,4; 4,7; 5,5; 5,2; 5,4; 3,8; 3,9; 3,9; 4,6; 3,7.

5.4. Мікроскопічні нерівності однорідної поверхні спочатку вимірювали старим приладом, а потім – новим. Чи можна стверджувати, що новий прилад має вищу точність вимірювання (при рівні значущості 0,1), якщо виміри були такими:

старий прилад (мкм): 4,56; 4,32; 4,53; 4,05; 3,31; 5,04; 4,88; 5,56; 7,19; 5,81;

новий прилад (мкм): 4,42; 4,59; 5,45; 4,80; 5,07; 6,01; 4,44; 5; 5,89; 5,71.

5.5. Для перевірки гіпотези про те, що людина припиняє рости у віці 19 років, вимірювали ріст у 100 осіб у віці 19 років та у 100 осіб – у віці 20 років. За даними вимірювань були розраховані такі характеристики: середні значення росту в першій та другій групах виявилися $\overline{\xi}^{(1)} = 1,71$ та $\overline{\xi}^{(2)} = 1,72$ см, а вибіркові дисперсії встановили $S_1^2 = 16$ та $S_2^2 = 26$ (см^2), відповідно. Перевірити висунуту гіпотезу при рівні значущості 0.1. Примітка: $t_{0.1,198} = -1.286$.

5.6. За нормативами дисперсія розсіювання ваги пакунка сухої шпаклівки становить 0,1 кг^2. Із часом у фасувального авто-

мата дисперсія розсіювання ваги зростає, що вимагає його ремонту. За даними контрольного зважування визначити, чи потрібно ремонтувати автомат (рівень значущості 0,1).

Результати зважувань (кг): 25,1; 25,09; 25,09; 25,07; 24,9; 25,01; 24,92; 25,07; 24,98; 24,94.

5.7. Середня вага дорослої людини становить 85 кг. Для того, щоб довести, що вживання гамбургерів і картоплі фрі не сприяє збільшенню ваги, контрольну групу зі 100 дорослих людей годували гамбургерами та картоплею фрі тривалий час, а потім вимірювали їхню вагу. За результатами зважувань зробили такі розрахунки: середня вага $\bar{\xi} = 91$ кг, вибіркова дисперсія $S^2 = 12$ кг^2. Перевірити висунуту гіпотезу при рівні значущості 0.1, урахувавши, що $t_{0.9,99} = 1.29$.

5.8. Виміри росту хлопців двох класів наведено нижче. Висунути й перевірити (при рівні значущості 0,1) гіпотезу про те, що середній ріст хлопців у обох класах відрізняється несуттєво.

Клас А (см): 163,9; 163,4; 163,6; 162,6; 155,8; 160,2; 156,9; 162,7; 159,2; 157,4.

Клас Б (см): 157,1; 157,9; 162,2; 159; 160,3; 165; 157,1; 160; 164,4; 163,5.

5.9. Нижче наведено дані вимірів діаметрів стандартних втулок двома різними штангенциркулями. Чи можна стверджувати (при рівні значущості 0,05), що обидва прилади мають однакову точність вимірів?

Перший штангенциркуль (мм): 9,1; 9,8; 12,2; 11,4; 10,8; 12,1; 12,8; 12,3; 12; 11,2.

Другий штангенциркуль (мм): 11,6; 11,3; 11,9; 13; 12,2; 12,5; 12,7; 12; 12,3; 12,2.

5.10. За документацією середній час спрацьовування запалу гранати Ф-6 становить 8 с. Зменшення цього часу викликає загрозу для життя. За наведеними нижче даними контрольних вимірів часу спрацьовування запалу висунути та перевірити відповідну гіпотезу з рівнем значущості 0,1.

Час спрацьовування: 7,28; 7,48; 8,56; 7,76; 8,09; 9,26; 7,29; 8; 9,11; 8,89; 5,1.

5.11. Вважається, що жінки живуть у середньому на вісім років довше за чоловіків. Для перевірки цього припущення були

проаналізовані дати життя та смерті 50 чоловіків та 60 жінок. За результатами обробки статистичних даних було підраховано, що середня тривалість життя в чоловіків становить $\overline{\xi}^{(1)} = 56$ років, а в жінок – $\overline{\xi}^{(2)} = 65$ років. При цьому вибіркова дисперсія в першій вибірці встановила $S_1^2 = 3$ (роки^2), а за другою $-S_2^2 = 2.5$ (роки^2). Перевірити висунуту гіпотезу при рівні значущості 0,1. Примітка: $t_{0.95,108} \approx 1.66$.

5.12. У результаті спостережень над випадковими величинами ξ та η отримали такі вибірки:

$$\xi: 45, 48, 53, 44, 59, 60, 41, 43, 57;$$
$$\eta: 51, 50, 42, 44, 39, 40, 48, 38, 59, 55, 51.$$

Чи можна вважати, що випадкові величини ξ та η мають однакові математичні сподівання? Похибка першого роду дорівнює $0,05$. Припускається, що випадкові величини ξ та η мають нормальний розподіл з рівними дисперсіями.

5.13. Нехай

$$\xi: 2,50; 2,50; 2,60; 2,75; 2,80; 2,80; 2,95;$$
$$\eta: 2,50; 2,80; 2,85; 2,90; 2,90; 2,95; 3,40.$$

Чи можна вважати, що ξ та η мають однакові математичні сподівання? $\alpha = 0,05$.

5.14. В одному класі з 20 дітей навмання відібрали 10, яким щодня почали видавати апельсиновий сік. Інші 10 учнів щодня отримували молоко. Через деякий час зафіксували збільшення маси дітей у фунтах:

сік: 4,0; 2,5; 3,5; 4.0; 1,5; 1,0; 3,5; 3,0; 2,5; 3,5;
молоко: 1,5; 3,5; 2,5; 3,0; 2,5; 2,0; 2,0; 2,5; 1,5; 3,0.

Чи суттєво відрізняється середнє збільшення ваги дітей у групах?

5.15. З нормальної генеральної сукупності із $\sigma^2 = 25$ отримано дві вибірки розміром $n_1 = n_2 = 9$. Середнє першої вибірки $\overline{x} = 2$, другої – $\overline{y} = 3$. Чи можна пояснити цю розбіжність випадковими причинами при похибці першого роду $\alpha = 0,05$?

5.16. Одним приладом було зроблено дві серії вимірів:

1) 2,5; 3,2; 3,5; 3,8; 3,5;

2) 2,0; 2,7; 2,5; 2,9; 2,3; 2,6.

а) Припускаючи, що виміри мають нормальний розподіл з однаковими дисперсіями, перевірити гіпотезу про рівність математичних сподівань при $\alpha = 0,05$.

б) Перевірити гіпотезу про те, що дисперсії однакові для цих вимірів, $\alpha = 0,05$.

в) Перевірити гіпотезу про те, що дисперсії однакові для цих вимірів, якщо $\alpha_1 = 3,\ \alpha_2 = 2,5,\ \alpha = 0,05$.

5.17. На станку-автоматі виготовляється один вид продукції. Критичним розміром виробів є зовнішній діаметр. Після налагодження станка відібрали 20 виробів. При цьому виявилось, що вибіркова дисперсія розміру зовнішнього діаметра становить 0,84 мм^2. Через деякий проміжок часу з метою контролю точності роботи станка (а за необхідності – його налагодження) відібрали 15 виробів. Вибіркова дисперсія, обчислена за ними, дорівнює 1,07 мм^2.

Чи свідчать наведені дані про зміну точності роботи станка? Узяти $\alpha = 0,05$.

5.18. Для зменшення дисперсії відбиваючої властивості фарби внесено зміни в технологію її виробництва. Щоб переконатись, що такі зміни насправді дають ефект, виготовили 10 пробних зразків і визначили відбиваючу властивість фарби, що виготовлена з використанням звичайної (А) та нової (В) технологій (в умовних одиницях). Отримано такі дані:

технологія А: 40, 45, 195, 65, 145;

технологія В: 110, 55, 120, 50, 80.

Чи свідчать ці дані про зміну дисперсії відбиваючої властивості фарби? Узяти $\alpha = 0,05$.

5.19. За двома вибірками з генеральних сукупностей розміром $n_1 = 11,\ n_2 = 15$ підраховано $\overline{S}_1^2(x) = \frac{1}{n_1}\sum_{i=1}^{n_1}\left(x_i - \overline{x}\right)^2 = 0,76$ і $\overline{S}_2^2(y) = \frac{1}{n_2}\sum_{i=1}^{n_2}\left(y_i - \overline{y}\right)^2 = 0,38$. При $\alpha = 0,05$ перевірити гіпотезу про рівність дисперсій двох нормальних сукупностей.

Розділ 6
ЕЛЕМЕНТИ РЕГРЕСІЙНОГО АНАЛІЗУ

Нехай у прямокутній декартовій системі координат маємо n точок $(x_1, y_1), \ldots, (x_n, y_n)$ і бажаємо підібрати функцію, яка відома з точністю до параметрів $b_1, \ldots, b_m$,

$$y = y(x, b_1, \ldots, b_m),$$

так, щоб її графік проходив близько до всіх указаних точок.

Метод найменших квадратів (МНК) дозволяє підібрати таку функцію, для якої сума квадратів відхилень заданих точок площини від точок графіка функції з тими самими абсцисами є найменшою з усіх можливих.

МНК як обчислювальна процедура був запропонований Лагранжем у 1806 р., а першим, хто пов'язав МНК із теорією ймовірностей, був Гаусс (1809). Термін ***регресія*** ввів Френсіс Гальтон у 1886 р., досліджуючи зв'язки між ростом батьків і дітей (він виявив, що в середньому діти високих батьків нижчі за них, а низьких – вищі за батьків. Це Гальтон інтерпретував як ***регресію до посередності***).

Підбір параметрів $b_1, \ldots, b_m$ здійснюється з метою мінімізації виразу

$$Q = Q(b_1, \ldots, b_m) = \sum_{i=1}^{n} \left(y_i - y(x_i, b_1, \ldots, b_m) \right)^2,$$

для чого параметри $b_1, \ldots, b_m$ шукають як розв'язки системи

$$\frac{\partial Q}{\partial b_j} = 0, \quad j = \overline{1,m}.$$

Підбір прямої за МНК. Нехай для точок $(x_1, y_1), \ldots, (x_n, y_n)$ треба підібрати лінійну функцію вигляду

$$y(x, b_0, b_1) = b_0 + b_1 x$$

за МНК. Для цього потрібно мінімізувати вираз

$$Q = Q(b_0, \mathrm{b}_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 ,$$

для чого розглянемо систему

$$\begin{cases} \dfrac{\partial Q}{\partial b_0} = 0 \\ \dfrac{\partial Q}{\partial b_1} = 0 \end{cases} \Leftrightarrow \begin{cases} \displaystyle\sum_{i=1}^{n} 2(y_i - b_0 - b_1 x_i)(-1) = 0 \\ \displaystyle\sum_{i=1}^{n} 2(y_i - b_0 - b_1 x_i)(-x_i) = 0 \end{cases} \Leftrightarrow$$

$$\Leftrightarrow \begin{cases} \displaystyle\sum_{i=1}^{n} y_i - b_0 n - b_1 \sum_{i=1}^{n} x_i = 0 \\ \displaystyle\sum_{i=1}^{n} x_i y_i - b_0 \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x_i^2 = 0. \end{cases}$$

Позначимо

$$S_x = \sum_{i=1}^{n} x_i, \quad S_y = \sum_{i=1}^{n} y_i, \quad S_{xy} = \sum_{i=1}^{n} x_i y_i, \quad S_{x^2} = \sum_{i=1}^{n} x_i^2.$$

Тоді

$$\begin{cases} S_y - n b_0 - S_x b_1 = 0 \\ S_{xy} - S_x b_0 - S_{x^2} b_1 = 0 \end{cases} \Rightarrow$$

$$\Rightarrow \begin{cases} b_1 = \dfrac{n S_{xy} - S_x S_y}{n S_{x^2} - S_x^2} \equiv \dfrac{S_{xy} - n \cdot \overline{x} \cdot \overline{y}}{S_{x^2} - n \cdot \overline{x}^2} \equiv \dfrac{\displaystyle\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\displaystyle\sum_{i=1}^{n} (x_i - \overline{x})^2} \\ b_0 = \dfrac{S_y - S_x b_1}{n} = \overline{y} - b_1 \overline{x} \end{cases} \qquad (6.1)$$

Зауваження 1. Пряма МНК проходить через точку $(\overline{x},\overline{y})$ і часто її рівняння пишуть у вигляді

$$y=\overline{y}+r\frac{\sigma_y}{\sigma_x}(x-\overline{x}), \tag{6.2}$$

де $\sigma_x^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$, $\sigma_y^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2$ – вибіркові дисперсії, а

$r=\dfrac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sigma_x\sigma_y}$ – вибірковий коефіцієнт кореляції.

Приклад 6.1. Побудуємо пряму МНК для точок (1;10), (3;20), (4;18), (5;20).

Оскільки $S_{xy}=\sum_{i=1}^{4}x_iy_i=242$, $\overline{x}=\frac{1}{4}\sum_{i=1}^{4}x_i=\frac{13}{4}$, $\overline{y}=\frac{1}{4}\sum_{i=1}^{4}y_i=17$,

$S_{x^2}=\sum_{i=1}^{4}x_i^2=51$, $\overline{x}^2=\left(\frac{13}{4}\right)^2=10\frac{9}{16}$, то із (6.1)

$$b_1=\frac{242-4\cdot\frac{13}{4}\cdot17}{51-4\cdot10\frac{9}{16}}=2{,}4 \text{ та } b_0=17-2{,}4\cdot\frac{13}{4}=9{,}2.$$

Отже,

$$y=b_0+b_1x=9.2+2.4x.$$

Дане рівняння можна також знайти, використовуючи (6.2):

$$\sigma_x^2=\frac{1}{n}S_{x_2}-(\overline{x})^2=\frac{1}{4}\cdot51-\left(\frac{13}{4}\right)^2=2\frac{3}{16},$$

$$\sigma_y^2=\frac{1}{n}S_{y_2}-(\overline{y})^2=\frac{1}{4}\cdot1224-17^2=17,$$

$$r=\frac{S_{xy}-n\bullet\overline{x}\bullet\overline{y}}{n\bullet\sigma_x\sigma_y}=\frac{242-4\cdot\frac{13}{4}\cdot17}{4\cdot\sqrt{2\frac{3}{16}\cdot17}}=\frac{21}{\sqrt{35\cdot17}}.$$

Отримаємо

$$y = 17 + \frac{21}{\sqrt{35 \cdot 17}} \cdot \frac{\sqrt{17}}{\sqrt{\frac{35}{16}}} \left(x - \frac{13}{4} \right) = 2.4x + 9.2 .$$

Вправа 1. Рівняння (6.2) називається вибірковим рівнянням прямої лінії регресії y на x. Довести, що якщо аналогічно розглянути $x = x(y)$, то вибіркове рівняння прямої лінії регресії x на y набуде вигляду

$$x = \overline{x} + r \frac{\sigma_x}{\sigma_y} (y - \overline{y}).$$

6.1. Статистична модель парної лінійної регресії (ПЛР)

У статистиці МНК застосовують до статистичних даних, які зазнають впливу похибок вимірювань. Припустимо, що наші спостереження (x_i, y_i) мають вигляд

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \qquad (6.4)$$

де β_0 та β_1 – невідомі сталі, а ε_j – незалежні випадкові величини з гауссівським розподілом $N(0, \sigma^2)$ (дисперсія σ^2 невідома). Змінна x розглядається як незалежна змінна, яка є невипадковою, а y – залежна змінна. Значення $\hat{\beta}_0$ та $\hat{\beta}_1$, які ми підрахуємо, є оцінками невідомих параметрів β_0 та β_1. Якщо $\beta_1 = 0$, то лінійного зв'язку між x та y не існує.

Незсуненою та ефективною оцінкою параметра σ^2 є [8]

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n-2},$$

де $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_2 x_i, i = \overline{1,n}.$

6.2. Критерій значущості лінії регресії

Навіть у випадку, коли *x* та *y* незалежні, спостережені значення можна нанести на площину у вигляді точок і підібрати до них пряму за МНК. Така пряма зазвичай матиме нульовий коефіцієнт нахилу. Ненульовий коефіцієнт нахилу тоді буде результатом випадковості й не відображатиме лінійну залежність між *x* та *y*. Як перевірити, чи є знайдена регресія значущою?

Розглянемо суму квадратів відхилень значень y_i від $\overline{y}$ (англ. total sum of squares – TSS):

$$TSS = \sum_{i=1}^{n} (y_i - \overline{y})^2. \tag{6.5}$$

Неважко показати, що для МНК має місце такий розклад:

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{6.6}$$

Отже, суму (6.5) розбито на дві складові:

а) сума квадратів відхилень значень регресії відносно $\overline{y}$ (***регресійна сума квадратів,*** англ. Estimated sum of squares– ESS)

$$ESS = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2;$$

б) сума квадратів відхилень спостережень відносно лінії регресії (***залишкова сума квадратів,*** англ. Residual Sum of Squares)

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

Скорочено можна записати:

$$TSS = ESS + RSS.$$

Якщо підібрана функція проходить через усі точки (ідеальний випадок), то $RSS = 0$ і $TSS = ESS$. Якщо ж вихідні дані не містять знайденої залежності, то ESS буде малою, а $TSS = RSS$.

Формально треба перевірити гіпотезу $H_0 : \beta_1 = 0$ за альтернативи $H_0 : \beta_1 \neq 0$. Якщо з'ясуємо, що $\beta_1 \neq 0$, то регресію ви-

знаємо ***значущою***. Усі підрахунки заносимо в ***таблицю дисперсійного аналізу:***

Джерело варіації	Сума квадратів	Ступені свободи	Середній квадрат
(1)	(2)	(3)	(4) = (2) / (3)
Регресія	$ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	1	ESS
Залишок	$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$n-2$	$\frac{RSS}{n-2}$
Загальна варіація	$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$n-1$	$\frac{TSS}{n-2}$

Статистикою критерію є

$$\frac{ESS}{RSS/(\mathrm{n}-2)},$$

яка за справедливості H_0 має F-розподіл (Фішера) зі ступенями свободи 1 та $n-2$ [8]. Критична область, за якої відхиляється основна гіпотеза H_0, задається нерівністю

$$R = \left\{ \frac{ESS}{RSS/(\mathrm{n}-2)} \geq F_{1-\alpha}(1, \mathrm{n}-2) \right\},$$

де $F_{1-\alpha}(1, \mathrm{n}-2)$ – квантиль F-розподілу порядку $1-\alpha$ зі ступенями свободи 1 та $n-2$.

Отже, регресійна модель вважається значущою з рівнем надійності $1-\alpha$, якщо гіпотеза H_0 відхиляється.

Запропонований критерій перевірки іноді називають F-критерієм.

Приклад 2. Перевіримо знайдену в прикладі 1 регресію на значущість (при $\alpha = 0.05$).

$$ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{4} (2.4x_i + 9.2 - 17)^2 = 50.4,$$

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 17.6.$$

Джерело варіації	Сума квадратів	Ступені свободи	Середній квадрат
(1)	(2)	(3)	(4)
Регресія	*ESS* = 50,4	1	*ESS* = 50,4
Залишок	*RSS*=7,6	2	*RSS/2* =8,8
Загальна варіація	*TSS*=68	3	—

$$50.4 / 8.8 \approx 5.7 < F_{0,95}(1,2) = 18.51,$$

тому немає підстав відхиляти основну гіпотезу H_0, а отже, регресія є незначущою.

Зауваження 2. Величину $R^2 = \dfrac{ESS}{TSS}$ називають ***коефіцієнтом детермінації*** (частка суми квадратів, що пояснюється регресією). Це значення є квадратом коефіцієнта кореляції між спостереженнями y_i та розрахованими $\hat{y}_i$. Звідси $R^2 = 50,4 / 68 = 0,7412.$

Зауваження 3. Після побудови лінії регресії може виникнути запитання: а чи правильно підібрана модель? Таке запитання може виникати як для парної, так і для криволінійної та множинної регресій. Коли ми складали модель (6.4), то вважали, що ε_i мають стандартний нормальний розподіл $N(0,1)$, тому можна очікувати, що відхилення $e_i = y_i - \hat{y}_i$ теж матимуть розподіл $N(0,1)$. Отже, слід перевірити гіпотезу про те, що Δ_i мають розподіл $N(0,1)$ (як перевіряти такі гіпотези, ми вже знаємо). Якщо перевірка гіпотези дає негативний результат, то можна спробувати підібрати іншу модель.

Вправа 4. Довести, що:

1) регресійна пряма проходить через "середню точку" $(\overline{x}, \overline{y})$;

2) залишки $e_i = y_i - \hat{y}_i$ мають нульову коваріацію зі спостереженнями x_i та оціненими значеннями $\hat{y}_i$, тобто $\sum_{i=1}^{n} e_i x_i = 0$ та $\sum_{i=1}^{n} e_i \hat{y}_i = 0$;

3) сума квадратів залишків є функцією від кута нахилу.

6.3. Множинна лінійна регресія

Під лінійною (множинною) регресійною моделлю розуміють таку ситуацію, коли спостережувані випадкові величини $Y_1,...,Y_n$ у середньому лінійно залежать від деяких невипадкових факторів $x_1,...,x_k$ ($k < n$), значення яких може змінюватись від досліду до досліду. У цьому випадку початкові статистичні дані складаються із множини значень $Y_1,...,Y_n$, що спостерігались, і відповідних значень факторів, тобто мають вигляд $(y_i, x_{i1}, x_{i2},..., x_{ik})$, $i = 1,...n$. При цьому вважають, що

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i, \quad i = \overline{1,n}, \tag{6.7}$$

де $\beta_1, \beta_2,..., \beta_k$ – невідомі параметри, які називаються ***коефіцієнтами регресії***; ε_i – незалежні випадкові величини (похибки вимірювань) $i = \overline{1,n}$.

Зауваження 5. Доволі часто розглядають випадок залежності з вільним членом, тобто коли $x_{i1} = 1$.

Позначимо $Y = (Y_1,..., Y_n)^T$ – вектор-стовпчик залежних змінних; $\beta = (\beta_1,..., \beta_k)^T$ – вектор-стовпчик невідомих коефіцієнтів; $\varepsilon = (\varepsilon_1,..., \varepsilon_n)^T$ – вектор-стовпчик похибок; $X = \begin{pmatrix} x_{11} & x_{12} & ... & x_{1k} \\ x_{21} & x_{22} & ... & x_{2k} \\ ... & ... & ... & .... \\ x_{n1} & x_{n2} & ... & x_{nk} \end{pmatrix}$ – матриця невипадкових факторів (матриця плану).

Зауваження 6. У випадку залежності з вільним членом матриця X має вигляд

$$X = \begin{pmatrix} 1 & x_{12} & \dots & x_{1k} \\ 1 & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots. \\ 1 & x_{n2} & \dots & x_{nk} \end{pmatrix}.$$

Тоді можна записати лінійну регресію в матричному вигляді:

$$Y = X \cdot \beta + \varepsilon. \tag{6.8}$$

Розглянемо такі умови:

1. $rank(X) = k$.

2.1. $M\varepsilon_i = 0$ для всіх $i = \overline{1,n}$ та $M\varepsilon_i^2 = D\varepsilon_i = \sigma^2$ не залежить від i.

2.2. $M\varepsilon_i\varepsilon_j = 0$ при $i \neq j$, отже, похибки некорельовані для різних спостережень.

2.3. Випадкові величини $\varepsilon_i, i = \overline{1,n}$, є незалежними з розподілом $N(0,\sigma^2)$. Якщо ця умова виконується, то модель називають нормальною лінійною регресією.

Умови 2.1., 2.2. еквівалентні тому, що коваріаційна матриця вектора ε має вигляд

$$\mathrm{D}(\varepsilon) = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \sigma^2 \end{pmatrix}$$

Оцінку вектора невідомих параметрів β шукають методом МНК, тобто мінімізують суму квадратів залишків регресії, як у випадку лінійної парної регресії:

$$Q(\beta) = (y - X\beta)^T (y - X\beta) = \sum_{i=1}^{n} (y_i - \beta_1 x_{i1} - \dots - \beta_k x_{ik})^2 \to \min.$$

Необхідною умовою того, щоб $Q(\beta)$ мала мінімум у точці $\hat{\beta}$, є рівність нулю частинних похідних за $\beta_1,...,\beta_k : \left.\frac{\partial Q}{\partial \beta_i}\right|_{\beta_i=\widehat{\beta_i}} = 0,\ i=\overline{1,\mathrm{k}}.$

Розв'язавши дану систему рівнянь, отримаємо, що при виконанні умови 1 оцінка параметра β методом найменших квадратів становитиме

$$\hat{\beta} = (X^T X)^{-1} X^T y. \tag{6.9}$$

Незсуненою та ефективною оцінкою параметра σ^2 (похибкою вимірювань) буде [8]

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{(\mathrm{y}_i - \hat{\mathrm{y}}_i)^2}{n-k},$$

де $\hat{y}_i = \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_k x_{ik}, i=\overline{1,n}.$

Теорема [8]. *Нехай для лінійної множинної регресії вигляду (6.7) виконуються умови 1, 2.1, 2.2. Тоді оцінка (6.9) є ефективною (у сенсі мінімальної дисперсії) серед усіх лінійних незсунених оцінок параметра* β.

Якщо ще виконується умова 2.3, то $\hat{\beta} \sim \mathrm{N}(\beta, \sigma^2(\mathrm{X}^T\mathrm{X})^{-1})$. Для коефіцієнтів β_j можна побудувати надійні інтервали з рівнем надійності $1-\alpha$:

$$\hat{\beta}_j - \hat{\sigma}_j \cdot t_{1-\alpha/2}(n-k) < \beta_j < \hat{\beta}_j + \hat{\sigma}_j \cdot t_{1-\alpha/2}(n-k),$$

де $t_p(n-k)$ – квантиль порядку p розподілу Стьюдента з $n-k$ ступенями свободи, $\hat{\sigma}_j^2$ – j-й діагональний елемент матриці $\hat{\sigma}^2(\mathrm{X}^T\mathrm{X})^{-1}$.

Для перевірки значущості лінії регресії у випадку нормальної множинної регресії з вільним членом використовують F-критерій, який подібний тому, що був запропонований для парної лінійної регресії.

Перевіряємо гіпотезу $H_0: \beta_2 = \beta_3 = ... = \beta_k = 0$ за альтернативи $H_1:$ існує j таке, що $\beta_j \neq 0$.

Статистикою критерію в цьому випадку буде

$$\frac{ESS/(k-1)}{RSS/(n-k)}.$$

Критична область, за якої відхиляється основна гіпотеза H_0, задається нерівністю

$$R = \left\{ \frac{ESS/(k-1)}{RSS/(n-k)} \geq F_{1-\alpha}(k-1, n-k) \right\}.$$

Регресійну модель вважають значущою з рівнем надійності $1-\alpha$, якщо гіпотеза H_0 відхиляється.

У деяких випадках нелінійну залежність можна звести до лінійної, зробивши певні заміни змінних. Приклади такої заміни наведено в таблиці:

№	Початкова функція	До якого вигляду зводиться	Заміна змінної
1	$y = Ae^{kx}$	$z = \alpha_0 + \alpha_1 x$	$z = \ln y, \alpha_0 = \ln A,$ $\alpha_1 = k$
2	$y = Bx^{\beta}$	$z = \alpha_0 + \alpha_1 u$	$z = \ln y, u = \ln x,$ $\alpha_0 = \ln B,\ \alpha_1 = \beta$
3	$y = \alpha_0 + \frac{\alpha_1}{x}$	$y = \alpha_0 + \alpha_1 u$	$u = \frac{1}{x}$
4	$y = \alpha_0 + \frac{\alpha_1}{x^{\beta}}$	$y = \alpha_0 + \alpha_1 u$	$u = \frac{1}{x^{\beta}}$
5	$y = Ae^{-(x-a)^2/(2\sigma^{2x})}$	$z = \alpha_0 + \alpha_1 x + \alpha_2 x^2$	$z = \ln y,$ $\alpha_0 = \ln A - \frac{a^2}{2\sigma^2},$ $\alpha_1 = \frac{a}{\sigma^2}, \alpha_2 = -\frac{1}{2\sigma^2}$

Продовження

№	Початкова функція	До якого вигляду зводиться	Заміна змінної
6	$y=\alpha_0+\frac{\alpha_1}{x}+\frac{\alpha_2}{x^2}+...$	$y=\alpha_0+\alpha_1 u+$ $+\alpha_2 u^2+...$	$u=\frac{1}{x}$
7	$y=\alpha_0+\alpha_1 x^{\beta}$ $+\alpha_2 x^{2\beta}+...$	$y=\alpha_0+\alpha_1 u$ $+\alpha_2 u^2+...$	$u=x^{\beta}$
8	$y=\alpha_0 x^{-m}+\alpha_1 x^k$	$z=\alpha_0+\alpha_1 u$	$z=yx^m, u=x^{m+k}$

ЗАДАЧІ

6.1. Припускаючи лінійну залежність величин x та y (тобто розглядаємо регресію вигляду $y=\beta_0+\beta_1 x+\varepsilon$), за результатами спостережень, наведених у таблиці, знайти коефіцієнти $\hat{\beta}_0, \hat{\beta}_1$, похибку вимірювань, надійні інтервали для $\hat{\beta}_0, \hat{\beta}_1$ з рівнем надійності $1-\alpha=0,95$ та вибірковий коефіцієнт кореляції. Перевірити модель на значущість:

а)

x_i	1	2	3	4	5
y_i	4,5	7	8	7,5	9

б)

x_i	2	4	6	8	9
y_i	10	8	7	5	2

в)

x_i	-1	0	1	2	3	4
y_i	2	3	4	6	5	7

6.2. Припускаючи залежність у вигляді $y=\beta_0+\beta_1 x+\beta_2 x^2+\varepsilon$, за результатами спостережень, наведених у таблиці, знайти коефіцієнти $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2$, похибку вимірювань, надійні інтервали для них з рівнем надійності $1-\alpha=0,99$ та вибірковий коефіцієнт кореляції:

а)

x_i	-3	-2	-1	0	1	2	3
y_i	-0,7	0	0,5	0,7	0,9	0,8	0,5

б)

x_i	0	2	4	6	8	10
y_i	5	-1	0,5	1,5	4,5	8,5

6.3. Припускаючи залежність у вигляді $y=\beta_0+\frac{\beta_1}{x}+\varepsilon$, за результатами спостережень, наведених у таблиці, знайти коефіцієнти $\hat{\beta}_0,\hat{\beta}_1$, похибку вимірювань, надійні інтервали для них з рівнем надійності $1-\alpha=0,99$ та вибірковий коефіцієнт кореляції:

а)

x_i	0,25	0,5	1	2
y_i	6	4	3	2,5

б)

x_i	1	2	5	10
y_i	10	5	2	1

6.4. У таблиці представлено 50 пар спостережень із досліду докторів Л. Матера та М. Уілсона. Розглядались змінні: x – довжина "лінії життя" на лівій руці в сантиметрах (з точністю до найближчих 0,15 см), y – тривалість життя людини (округлення

до найближчого цілого року). Чи вірно, що y та x пов'язані лінійною регресійною залежністю?

x	y	x	y	x	y	x	y
9,75	19	7,20	61	9,00	68	10,20	75
9,00	40	7,95	62	7,80	69	6,00	76
9,60	42	8,85	62	10,05	69	8,85	77
9,75	42	8,25	65	10,50	70	9,00	80
11,25	47	8,85	65	9,15	71	9,75	82
9,45	49	9,75	65	9,45	71	10,65	82
11,25	50	8,85	66	9,45	71	13,20	82
9,00	54	9,15	66	9,45	72	7,95	83
7,95	56	10,20	66	8,10	73	7,95	86
12,00	56	9,15	67	8,85	74	9,15	88
8,10	57	7,95	68	9,60	74	9,75	88
10,20	57	8,85	68	6,45	75	9,00	94
8,55	58			9,75	75		

6.5. У таблиці наведено дані щодо щорічного світового видобутку нафти з 1880 по 1984 роки. Збільшення видобутку добре описується моделлю з експоненційним зростанням:

$$Mbbl = a \cdot e^{b\cdot(\mathrm{Year}-1880)},$$

де *Year* – рік видобутку, *Mbbl* –кількість видобутої нафти (у млн барелів). Знайти оцінки параметрів $\hat{a},\hat{b}$ та за даною моделлю зробити прогноз на 1986 та 1990 роки.

Year	*Mbbl*	*Year*	*Mbbl*	*Year*	*Mbbl*
1880	30	1935	1655	1968	14104
1890	77	1940	2150	1970	16690
1900	149	1945	2595	1972	18584
1905	215	1950	3803	1974	20389
1910	328	1955	5626	1976	20188
1915	432	1960	7674	1978	21922
1920	689	1962	8882	1980	21732
1925	1069	1964	10310	1982	19403
1930	1412	1966	12016	1984	19608

ДОДАТКИ

Таблиця 1
Функція розподілу Ф(x) стандартної нормальної випадкової величини N(0,1)

x	0	1	2	3	4	5	6	7	8	9
-0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
-0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
-0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
-0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
-0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
-0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
-0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
-0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
-0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
-1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
-1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
-1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
-1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
-1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
-1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
-1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
-1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
-1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
-1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
-2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
-2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
-2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
-2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
-2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
-2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
-2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
-2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026

Функція розподілу Φ(x) стандартної нормальної випадкової величини N(0,1) (продовження)

x	0	1	2	3	4	5	6	7	8	9
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
x	3,0	3,1	3,2	3,3	3,4	3,5	3,6	3,7	3,8	3,9
$\Phi(x$	0,9987	0,9990	0,9993	0,9995	0,9997	0,9998	0,9998	0,9999	0,9999	1,0000

Таблиця 2
Квантилі стандартного гауссівського розподілу

α	0,010	0,025	0,050	0,100	0,900	0,950	0,975	0,990
c_α	-2,3263	-1,96	-1,6449	-1,2816	1,2816	1,6449	1,96	2,3263

Таблиця 3
Квантилі розподілу Пірсона Хі-квадрат ($\chi^2_{\alpha;k}$)
залежно від імовірності $P=\left\{\chi^2(k)\le\chi^2_{\alpha;k}\right\}=1-\alpha$
і кількості ступенів свободи k

k \ $1-\alpha$	**0,010**	**0,025**	**0,050**	**0,100**	**0,900**	**0,950**	**0,975**	**0,990**
1	0,000	0,001	0,004	0,016	2,706	3,841	5,024	6,635
2	0,020	0,051	0,103	0,211	4,605	5,991	7,378	9,210
3	0,115	0,216	0,352	0,584	6,251	7,815	9,348	11,345
4	0,297	0,484	0,711	1,064	7,779	9,488	11,143	13,277
5	0,554	0,831	1,145	1,610	9,236	11,070	12,832	15,086
6	0,872	1,237	1,635	2,204	10,645	12,592	14,449	16,812
7	1,239	1,690	2,167	2,833	12,017	14,067	16,013	18,475
8	1,647	2,180	2,733	3,490	13,362	15,507	17,535	20,090
9	2,088	2,700	3,325	4,168	14,684	16,919	19,023	21,666
10	2,558	3,247	3,940	4,865	15,987	18,307	20,483	23,209
11	3,053	3,816	4,575	5,578	17,275	19,675	21,920	24,725
12	3,571	4,404	5,226	6,304	18,549	21,026	23,337	26,217
13	4,107	5,009	5,892	7,041	19,812	22,362	24,736	27,688
14	4,660	5,629	6,571	7,790	21,064	23,685	26,119	29,141
15	5,229	6,262	7,261	8,547	22,307	24,996	27,488	30,578
16	5,812	6,908	7,962	9,312	23,542	26,296	28,845	32,000
17	6,408	7,564	8,672	10,085	24,769	27,587	30,191	33,409
18	7,015	8,231	9,390	10,865	25,989	28,869	31,526	34,805
19	7,633	8,907	10,117	11,651	27,204	30,144	32,852	36,191
20	8,260	9,591	10,851	12,443	28,412	31,410	34,170	37,566
21	8,897	10,283	11,591	13,240	29,615	32,671	35,479	38,932
22	9,542	10,982	12,338	14,041	30,813	33,924	36,781	40,289
23	10,196	11,689	13,091	14,848	32,007	35,172	38,076	41,638
24	10,856	12,401	13,848	15,659	33,196	36,415	39,364	42,980
25	11,524	13,120	14,611	16,473	34,382	37,652	40,646	44,314
26	12,198	13,844	15,379	17,292	35,563	38,885	41,923	45,642
27	12,878	14,573	16,151	18,114	36,741	40,113	43,195	46,963
28	13,565	15,308	16,928	18,939	37,916	41,337	44,461	48,278
29	14,256	16,047	17,708	19,768	39,087	42,557	45,722	49,588
30	14,953	16,791	18,493	20,599	40,256	43,773	46,979	50,892

Таблиця 4

Квантилі *t*-розподілу Стьюдента t_α залежно від імовірності $P\{t_k \le t_\alpha\} = \alpha$ і кількості ступенів свободи k

α

n \ *α*	0,900	0,950	0,975	0,990	0,995
1	3,078	6,314	12,706	31,821	63,656
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
25	1,316	1,708	2,060	2,485	2,787
30	1,310	1,697	2,042	2,457	2,750
40	1,303	1,684	2,021	2,423	2,704
60	1,296	1,671	2,000	2,390	2,660
120	1,289	1,658	1,980	2,358	2,617
∞	1,282	1,645	1,960	2,326	2,576

Таблиця 5

Критичні значення λ_α для розподілу Колмогорова

$$P\{\lambda_n > \lambda_\alpha\} = \alpha$$

α	0,2	0,1	0,05	0,02	0,01	0,001
λ_α	1,073	1,224	1,358	1,520	1,627	1,950

Таблиця 6

Квантилі F-розподілу Фішера – Снедекора $F_{0,95}(n_1, n_2)$

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10
1	161,446	199,499	215,707	224,583	230,160	233,988	236,767	238,884	240,543	241,882
2	18,513	19,000	19,164	19,247	19,296	19,329	19,353	19,371	19,385	19,396
3	10,128	9,552	9,277	9,117	9,013	8,941	8,887	8,845	8,812	8,785
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041	5,999	5,964
5	6,608	5,786	5,409	5,192	5,050	4,950	4,876	4,818	4,772	4,735
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147	4,099	4,060
7	5,591	4,737	4,347	4,120	3,972	3,866	3,787	3,726	3,677	3,637
8	5,318	4,459	4,066	3,838	3,688	3,581	3,500	3,438	3,388	3,347
9	5,117	4,256	3,863	3,633	3,482	3,374	3,293	3,230	3,179	3,137
10	4,965	4,103	3,708	3,478	3,326	3,217	3,135	3,072	3,020	2,978
11	4,844	3,982	3,587	3,357	3,204	3,095	3,012	2,948	2,896	2,854
12	4,747	3,885	3,490	3,259	3,106	2,996	2,913	2,849	2,796	2,753
15	4,543	3,682	3,287	3,056	2,901	2,790	2,707	2,641	2,588	2,544
20	4,351	3,493	3,098	2,866	2,711	2,599	2,514	2,447	2,393	2,348
24	4,260	3,403	3,009	2,776	2,621	2,508	2,423	2,355	2,300	2,255
30	4,171	3,316	2,922	2,690	2,534	2,421	2,334	2,266	2,211	2,165
40	4,085	3,232	2,839	2,606	2,449	2,336	2,249	2,180	2,124	2,077
60	4,001	3,150	2,758	2,525	2,368	2,254	2,167	2,097	2,040	1,993
100	3,936	3,087	2,696	2,463	2,305	2,191	2,103	2,032	1,975	1,927
120	3,920	3,072	2,680	2,447	2,290	2,175	2,087	2,016	1,959	1,910

151

Продовження таб. 6

n_2 \ n_1	12	14	16	18	20	30	40	50	60	100
1	243,905	245,363	246,466	247,324	248,016	250,096	251,144	251,774	252,196	253,043
2	19,412	19,424	19,433	19,440	19,446	19,463	19,471	19,476	19,479	19,486
3	8,745	8,715	8,692	8,675	8,660	8,617	8,594	8,581	8,572	8,554
4	5,912	5,873	5,844	5,821	5,803	5,746	5,717	5,699	5,688	5,664
5	4,678	4,636	4,604	4,579	4,558	4,496	4,464	4,444	4,431	4,405
6	4,000	3,956	3,922	3,896	3,874	3,808	3,774	3,754	3,740	3,712
7	3,575	3,529	3,494	3,467	3,445	3,376	3,340	3,319	3,304	3,275
8	3,284	3,237	3,202	3,173	3,150	3,079	3,043	3,020	3,005	2,975
9	3,073	3,025	2,989	2,960	2,936	2,864	2,826	2,803	2,787	2,756
10	2,913	2,865	2,828	2,798	2,774	2,700	2,661	2,637	2,621	2,588
11	2,788	2,739	2,701	2,671	2,646	2,570	2,531	2,507	2,490	2,457
12	2,687	2,637	2,599	2,568	2,544	2,466	2,426	2,401	2,384	2,350
15	2,475	2,424	2,385	2,353	2,328	2,247	2,204	2,178	2,160	2,123
20	2,278	2,225	2,184	2,151	2,124	2,039	1,994	1,966	1,946	1,907
24	2,183	2,130	2,088	2,054	2,027	1,939	1,892	1,863	1,842	1,800
30	2,092	2,037	1,995	1,960	1,932	1,841	1,792	1,761	1,740	1,695
40	2,003	1,948	1,904	1,868	1,839	1,744	1,693	1,660	1,637	1,589
60	1,917	1,860	1,815	1,778	1,748	1,649	1,594	1,559	1,534	1,481
100	1,850	1,792	1,746	1,708	1,676	1,573	1,515	1,477	1,450	1,392
120	1,834	1,775	1,728	1,690	1,659	1,554	1,495	1,457	1,429	1,369

Продовження табл. 6

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10
1	4052,185	4999,340	5403,534	5624,257	5763,955	5858,950	5928,334	5980,954	6022,397	6055,925
2	98,502	99,000	99,164	99,251	99,302	99,331	99,357	99,375	99,390	99,397
3	34,116	30,816	29,457	28,710	28,237	27,911	27,671	27,489	27,345	27,228
4	21,198	18,000	16,694	15,977	15,522	15,207	14,976	14,799	14,659	14,546
5	16,258	13,274	12,060	11,392	10,967	10,672	10,456	10,289	10,158	10,051
6	13,745	10,925	9,780	9,148	8,746	8,466	8,260	8,102	7,976	7,874
7	12,246	9,547	8,451	7,847	7,460	7,191	6,993	6,840	6,719	6,620
8	11,259	8,649	7,591	7,006	6,632	6,371	6,178	6,029	5,911	5,814
9	10,562	8,022	6,992	6,422	6,057	5,802	5,613	5,467	5,351	5,257
10	10,044	7,559	6,552	5,994	5,636	5,386	5,200	5,057	4,942	4,849
11	9,646	7,206	6,217	5,668	5,316	5,069	4,886	4,744	4,632	4,539
12	9,330	6,927	5,953	5,412	5,064	4,821	4,640	4,499	4,388	4,296
15	8,683	6,359	5,417	4,893	4,556	4,318	4,142	4,004	3,895	3,805
20	8,096	5,849	4,938	4,431	4,103	3,871	3,699	3,564	3,457	3,368
24	7,823	5,614	4,718	4,218	3,895	3,667	3,496	3,363	3,256	3,168
30	7,562	5,390	4,510	4,018	3,699	3,473	3,305	3,173	3,067	2,979
40	7,314	5,178	4,313	3,828	3,514	3,291	3,124	2,993	2,888	2,801
60	7,077	4,977	4,126	3,649	3,339	3,119	2,953	2,823	2,718	2,632
100	6,895	4,824	3,984	3,513	3,206	2,988	2,823	2,694	2,590	2,503
120	6,851	4,787	3,949	3,480	3,174	2,956	2,792	2,663	2,559	2,472

Закінчення табл. 6

n_2 \ n_1	12	14	16	18	20	30	40	50	60	100
1	6106,682	6143,004	6170,012	6191,432	6208,662	6260,350	6286,427	6302,260	6312,970	6333,925
2	99,419	99,426	99,437	99,444	99,448	99,466	99,477	99,477	99,484	99,491
3	27,052	26,924	26,826	26,751	26,690	26,504	26,411	26,354	26,316	26,241
4	14,374	14,249	14,154	14,079	14,019	13,838	13,745	13,690	13,652	13,577
5	9,888	9,770	9,680	9,609	9,553	9,379	9,291	9,238	9,202	9,130
6	7,718	7,605	7,519	7,451	7,396	7,229	7,143	7,091	7,057	6,987
7	6,469	6,359	6,275	6,209	6,155	5,992	5,908	5,858	5,824	5,755
8	5,667	5,559	5,477	5,412	5,359	5,198	5,116	5,065	5,032	4,963
9	5,111	5,005	4,924	4,860	4,808	4,649	4,567	4,517	4,483	4,415
10	4,706	4,601	4,520	4,457	4,405	4,247	4,165	4,115	4,082	4,014
11	4,397	4,293	4,213	4,150	4,099	3,941	3,860	3,810	3,776	3,708
12	4,155	4,052	3,972	3,910	3,858	3,701	3,619	3,569	3,535	3,467
15	3,666	3,564	3,485	3,423	3,372	3,214	3,132	3,081	3,047	2,977
20	3,231	3,130	3,051	2,989	2,938	2,778	2,695	2,643	2,608	2,535
24	3,032	2,930	2,852	2,789	2,738	2,577	2,492	2,440	2,403	2,329
30	2,843	2,742	2,663	2,600	2,549	2,386	2,299	2,245	2,208	2,131
40	2,665	2,563	2,484	2,421	2,369	2,203	2,114	2,058	2,019	1,938
60	2,496	2,394	2,315	2,251	2,198	2,028	1,936	1,877	1,836	1,749
100	2,368	2,265	2,185	2,120	2,067	1,893	1,797	1,735	1,692	1,598
120	2,336	2,234	2,154	2,089	2,035	1,860	1,763	1,700	1,656	1,559

Таблиця 7

Розподіл Пуассона. Значення функції $p_k(\lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$

k\λ	0,1	0,2	0,3	0,4	0,5	0,6	0,8	1	1,5	2
0	0,904837	0,818731	0,740818	0,670320	0,606531	0,548812	0,449329	0,367879	0,22313	0,135335
1	0,090484	0,163746	0,222245	0,268128	0,303265	0,329287	0,359463	0,367879	0,334695	0,270671
2	0,004524	0,016375	0,033337	0,053626	0,075816	0,098786	0,143785	0,18394	0,251021	0,270671
3	0,000151	0,001092	0,003334	0,007150	0,012636	0,019757	0,038343	0,061313	0,125511	0,180447
4	0,000004	0,000055	0,000250	0,000715	0,001580	0,002964	0,007669	0,015328	0,047067	0,090224
5	0	0,000002	0,000015	0,000057	0,000158	0,000356	0,001227	0,003066	0,014120	0,036089
6	0	0	0,000001	0,000004	0,000013	0,000036	0,000164	0,000511	0,003530	0,012030
7	0	0	0	0	0,000001	0,000003	0,000019	0,000073	0,000756	0,003437
8	0	0	0	0	0	0	0,000002	0,000009	0,000142	0,000859
9	0	0	0	0	0	0	0	0,000001	0,000024	0,000191
10	0	0	0	0	0	0	0	0	0,000004	0,000038
11	0	0	0	0	0	0	0	0	0	0,000007
12	0	0	0	0	0	0	0	0	0	0,000001
13	0	0	0	0	0	0	0	0	0	0

Закінчення табл. 7

$k\backslash\lambda$	2,5	3	3,5	4	4,5	5	5,5	6	6,5	7
0	0,082085	0,049787	0,030197	0,018316	0,011109	0,006738	0,004087	0,002479	0,001503	0,000912
1	0,205212	0,149361	0,105691	0,073263	0,04999	0,033690	0,022477	0,014873	0,009772	0,006383
2	0,256516	0,224042	0,184959	0,146525	0,112479	0,084224	0,061812	0,044618	0,03176	0,022341
3	0,213763	0,224042	0,215785	0,195367	0,168718	0,140374	0,113323	0,089235	0,068814	0,052129
4	0,133602	0,168031	0,188812	0,195367	0,189808	0,175467	0,155819	0,133853	0,111822	0,091226
5	0,066801	0,100819	0,132169	0,156293	0,170827	0,175467	0,171401	0,160623	0,145369	0,127717
6	0,027834	0,050409	0,077098	0,104196	0,12812	0,146223	0,157117	0,160623	0,157483	0,149003
7	0,009941	0,021604	0,038549	0,059540	0,082363	0,104445	0,123449	0,137677	0,146234	0,149003
8	0,003106	0,008102	0,016865	0,029770	0,046329	0,065278	0,084871	0,103258	0,118815	0,130377
9	0,000863	0,002701	0,006559	0,013231	0,023165	0,036266	0,051866	0,068838	0,085811	0,101405
10	0,000216	0,00081	0,002296	0,005292	0,010424	0,018133	0,028526	0,041303	0,055777	0,070983
11	0,000049	0,000221	0,000730	0,001925	0,004264	0,008242	0,014263	0,022529	0,032959	0,045171
12	0,000010	0,000055	0,000213	0,000642	0,001599	0,003434	0,006537	0,011264	0,017853	0,026350
13	0,000002	0,000013	0,000057	0,000197	0,000554	0,001321	0,002766	0,005199	0,008926	0,014188
14	0	0,000003	0,000014	0,000056	0,000178	0,000472	0,001087	0,002228	0,004144	0,007094
15	0	0,000001	0,000003	0,000015	0,000053	0,000157	0,000398	0,000891	0,001796	0,003311
16	0	0	0,000001	0,000004	0,000015	0,000049	0,000137	0,000334	0,00073	0,001448
17	0	0	0	0,000001	0,000004	0,000014	0,000044	0,000118	0,000279	0,000596
18	0	0	0	0	0,000001	0,000004	0,000014	0,000039	0,000101	0,000232

ЛІТЕРАТУРА

1. Анісімов В. В. Математична статистика / В. В. Анісімов, О. І. Черняк. – К. : МП "Леся", 1995.

2. Афифи А. Статистический анализ. Подход с использованием ЭВМ / А. Афифи, С. Эйзен. – М. : Мир, 1982.

3. Беляев Ю. К. Основы математической статистики. В 3 ч. / Ю. К. Беляев, Е. В. Чепурин. – М. : Изд-во МГУ, 1982 – 1983.

4. Большев Л. Н. Таблицы математической статистики / Л. Н. Большев, Н. В. Смирнов. – М. : Наука, 1983.

5. Боровков А. А. Математическая статистика. Оценка параметров. Проверка гипотез / А. А. Боровков. – М. : Наука, 1984.

6. Боровков А. А. Математическая статистика. Дополнительные главы / А. А. Боровков. – М. : Наука, 1984.

7. Гулкс С. Математическая статистика / С. Гулкс. – М. : Наука, 1967.

8. Ивченко Г. И. Математическая статистика / Г. И. Ивченко, Ю. И. Медведев. – М. : Высшая школа, 1984.

9. Карташов М. В. Імовірність, процеси, статистика / М. В. Карташов. – К. : ВПЦ "Київ. ун-т", 2007.

10. Козлов М. В. Введение в математическую статистику / М. В. Козлов, А. В. Прохоров. – М. : Изд. МГУ, 1987.

11. Крамер Г. Математические методы статистики / Г. Крамер. – М. : Мир, 1975.

12. Лагутин М. Б. Наглядная математическая статистика / М. Б. Лагутин. – М. : БИНОМ, 2007.

13. Лаккин Г. Ф. Биометрия / Г. Ф. Лаккин. – М. : Высшая школа, 1980.

14. Носов В. Н. Компьютерная биометрика / В. Н. Носов. – М. : Изд-во МГУ, 1990.

15. Себер Дж. Линейный регрессионный анализ / Дж. Себер. – М. : Мир, 1982.

16. Турчин В. М. Теорія ймовірностей і математична статистика / В. М. Турчин. – Дніпропетровськ : Вид-во ДНУ, 2006.

17. Харин Ю. С. Практикум на ЭВМ по математической статистике / Ю. С. Харин, М. Д. Степанова. – Минск : Изд-во "Университетское", 1987.

18. Энслейн К. Статистические методы для ЭВМ / К. Энслейн, Э. Рэлстон, Г. С. Уилф. – М. : Наука, 1986.

ЗМІСТ

Навчальне видання

ЛЕБЄДЄВ Є. О.
ЛІВІНСЬКА Г. В.
РОЗОРА І. В.
ШАРАПОВ М. М.

МАТЕМАТИЧНА СТАТИСТИКА

Начальний посібник

Оригінал-макет виготовлено Видавничо-поліграфічним центром "Київський університет"

Формат 60x84$^{1/16}$. Ум. друк. арк. 9,3. Наклад 100. Зам. № 216-7709.
Гарнітура Times New Roman. Папір офсетний. Друк офсетний. Вид. № Гр-3.
Підписано до друку 28.03.16

Видавець і виготовлювач
ВПЦ "Київський університет"
б-р Т. Шевченка, 14, 01601, м. Київ
☎ (044) 239 32 22; (044) 239 31 72; тел./факс (044) 239 31 28
e-mail: vpc_div.chief@univ.kiev.ua
http: vpc.univ.kiev.ua
Свідоцтво суб'єкта видавничої справи ДК № 1103 від 31.10.02