Міністерство освіти і науки України
Тернопільський національний економічний університет

Методичні вказівки до вивчення розділу
«Математична статистика» дисципліни ТІМС
для студентів всіх спеціальностей

Тернопіль – 2017

УДК 519.2

Рецензенти:

В.З. Чорний – к. ф.-м. н., доцент кафедри математики та методики її викладання Тернопільського національного педагогічного університету імені Володимира Гнатюка

Л.М. Буяк – к. е. н., доцент кафедри економічної кібернетики та інформатики Тернопільського національного економічного університету;

атверджено на засіданні кафедри економіко-математичних методів, протокол № 1 від 28.08.2017 р.

В пропонованих методичних вказівках висвітлюється матеріал з розділу «Математична статистика» дисципліни ТІМС. Вони містять теоретичні питання та розв'язування типових задач. Відбір теоретичного матеріалу здійснювався з метою «автономності» вказівок (відсутності потреби, як правило, користуватися додатковими джерелами). Символи ○ та ● означають відповідно початок і завершення розв'язування задачі. В додатках наведені необхідні для розв'язування задач таблиці, пов'язані із розподілами: нормальним, Стюдента, хі-квадрат, Фішера-Снедокора.

Єрьоменко В. О., Шинкарик М.І., Мартинюк О. М., Березька К.М., Пласконь С.А., Сенів Г.В., Дзюбановська Н.В. Методичні вказівки до вивчення розділу «Математична статистика» дисципліни ТІМС для студентів всіх спеціальностей, 2017. – 116 с.

УДК 5192

Відповідальний за випуск: О.М. Мартинюк, кандидат фіз..-мат. наук, доцент кафедри ЕММ ТНЕУ

МАТЕМАТИЧНА СТАТИСТИКА

§ 1.ВСТУП В МАТЕМАТИЧНУ СТАТИСТИКУ. ВИБІРКОВИЙ МЕТОД

Задачі математичної статистики. Генеральна та вибіркова сукупності. Способи утворення вибіркової сукупності. Статистичний розподіл вибірки. Емпірична функція розподілу та її властивості. Графічне зображення статистичних розподілів (полігон та гістограма). Числові характеристики вибірки. Метод добутків обчислення зведених характеристик вибірки. Числові характеристики сукупностей, що складаються із груп.

КОРОТКІ ТЕОРЕТИЧНІ ВІДОМОСТІ

Перша задача математичної статистики — вказати метод відбору і групування статистичних даних, а також знаходження числа необхідних випробувань (статистичних даних).

Друга задача математичної статистики полягає в розробці методів аналізу статистичних даних в залежності від цілей дослідження. Одна з них — оцінка невідомих:

— імовірності випадкової події;

— функції розподілу імовірностей (густини розподілу);

— параметрів розподілу, вид якого відомий;

— залежності випадкової величини від однієї або кількох випадкових величин.

Друга мета — це перевірка статистичних гіпотез про вид невідомого розподілу або про величину параметрів розподілу, вид якого відомий.

Генеральною називається вся сукупність однотипних об'єктів, яка підлягає вивченню. Множину цих об'єктів позначатимемо через Ω.

Об'єкти множини Ω можуть характеризуватися однією або кількома ознаками. Ці ознаки можуть бути кількісними або якісними.

Іноді проводять **суцільне** обстеження, тобто обстежують **кожний** елемент множини Ω відносно ознаки, яка досліджується. Проте на практиці суцільне обстеження застосовують порівняно рідко. Це зумовлено тим, що при перевірці об'єкт частково або повністю знищується, або дослідження об'єктів вимагає великих матеріальних витрат. Деколи провести суцільне обстеження фізично неможливо. В таких випадках природно відібрати із генеральної сукупності обмежене число об'єктів, дослідити їх на кількісну чи якісну ознаку, а потім робити висновки про всю генеральну сукупність.

Вибірковою сукупністю або просто **вибіркою** називається сукупність **випадково** відібраних об'єктів із генеральної сукупності. Вибірка утворює підмножину V множини Ω ($V \subset \Omega$).

Обсягом сукупності (генеральної або вибіркової) називається число об'єктів цієї сукупності. Надалі обсяг генеральної сукупності позначатимемо літерою N, а вибіркової — n.

Для правомірності висновків про досліджувану ознаку об'єктів генеральної сукупності на підставі опрацювання вибірки необхідно, щоб об'єкти вибірки правильно представляли генеральну сукупність, тобто вибірка повинна володіти властивістю репрезентативності (представницькості). Випадковість відбору об'єктів у вибіркову сукупність і використання закону великих чисел дозволяють вирішити питання про репрезентативність вибірки.

Точність результатів вибіркового спостереження, в кінцевому підсумку, буде залежати від способу відбору об'єктів, ступеня коливання досліджуваної ознаки в генеральній сукупності та від обсягу вибірки.

На практиці використовуються різні способи утворення вибірки, які принципово розподіляються на два види:

1) відбір, що не вимагає розчленування генеральної сукупності на частини (**простий** (**власне випадковий**) **відбір**);

2) відбір, при якому генеральна сукупність розбивається на частини (**типовий** відбір, **механічний** відбір, **серійний** відбір, **комбінований** відбір).

Простим випадковим (**власне випадковим**) називається такий відбір, при якому об'єкти відбираються по одному випадковим чином із усієї генеральної сукупності. Проста випадкова вибірка може бути **повторною** або **безповторною. Повторною** називається вибірка, при утворенні якої відібраний об'єкт (перед відбором наступного) повертається в генеральну сукупність. **Безповторною** називається вибірка, в процесі утворення якої відібраний об'єкт в генеральну сукупність не повертається.

Нехай досліджується кількісна ознака X об'єктів генеральної сукупності Ω. Для скорочення припустимо, що вона є одновимірною випадковою величиною. Після опрацювання n об'єктів вибіркової сукупності отримуються n чисел $x_1, x_2, \ldots, x_n$, які називаються **варіантами** і утворюють **ряд варіант** або **простий статистичний ряд**.

Первинна обробка ряду варіант полягає у групуванні рівних варіант цього ряду. Ряд варіант розташуємо в порядку зростання і у відповідності з цим перенумеруємо їх. В результаті одержиться послідовність чисел $x_1 \le x_2 \le x_3 \le \ldots \le x_n$, яка називається варіаційним рядом. Якщо серед цієї послідовності є однакові варіанти, то ми їх знову перенумеруємо, залишаючи один і той самий номер однаковим варіантам. Нехай у варіаційному ряді варіанта x_1 повторюється n_1 разів, x_2 — n_2 разів, ..., x_k — n_k разів. Числа n_i називаються **частотами** (**абсолютними частотами**), а їх відношення до обсягу вибірки $n_i / n = w_i$ — **відносними частотами**. Із цих означень випливають рівності:

$$\sum_{i=1}^{k} n_i = n, \quad \sum_{i=1}^{k} w_i = 1. \tag{1.1}$$

Статистичним розподілом вибірки називається відповідність між варіантами та частотами або відносними частотами:

$$\begin{array}{c|cccc} x_i & x_1 & x_2 & \dots & x_k \\ \hline n_i & n_1 & n_2 & \dots & n_k \end{array}, \tag{1.2}$$

$$\begin{array}{c|cccc} x_i & x_1 & x_2 & \dots & x_k \\ \hline w_i = n_i/n & w_1 & w_2 & \dots & w_k \end{array}. \tag{1.3}$$

В більшості випадків статистичний розподіл вибірки у вигляді (1.2) або (1.3) використовується тоді, коли ряд варіант є реалізацією **дискретної** випадкової величини *X*. Якщо ж *X* є неперервною випадковою величиною, тоді статистичний розподіл вибірки задається у вигляді відповідності між інтервалами і частотами або відносними частотами тих варіант, які потрапляють у ці інтервали, тобто у вигляді таблиць:

$$\begin{array}{c|cccc} [x_i, x_{i+1}) & [x_1, x_2) & [x_2, x_3) & \dots & [x_k, x_{k+1}] \\ \hline n_i & n_1 & n_2 & \dots & n_k \end{array}, \tag{1.4}$$

$$\begin{array}{c|cccc} [x_i, x_{i+1}) & [x_1, x_2) & [x_2, x_3) & \dots & [x_k, x_{k+1}] \\ \hline w_i = n_i/n & w_1 & w_2 & \dots & w_k \end{array}. \tag{1.5}$$

Ці таблиці називаються **інтервальним статистичним розподілом вибірки**. При побудові інтервального статистичного розподілу на основі ряду варіант розглядається *k* інтервалів однакової довжини. Кількість інтервалів можна визначати наближено за формулою Стерджеса: $k = 1+3{,}322 \lg n$. При цьому $k \in [5;6]$ при $20 \le n \le 30$; $k \in [7;8]$ при $60 \le n \le 70$; $k \in [8;9]$ при $70 \le n \le 200$; $k \in [9;15]$ при $n > 200$.

Статистичні розподіли вибірки (1.3) та (1.5) називаються **емпіричними (дослідними) розподілами випадкової величини *X*** (кількісної ознаки об'єктів генеральної сукупності).

Одна із задач математичної статистики — оцінка (наближене знаходження) невідомої функції розподілу $F(x)$ імовірностей кількісної ознаки X об'єктів генеральної сукупності. За означенням $F(x) = P(X < x)$. В розпорядженні дослідника є статистичні дані, згруповані в статистичному розподілі частот або відносно частот. Тому, врахувавши властивість стійкості відносної частоти, доцільно імовірність події $(X < x)$ наближати відносною частотою цієї ж події.

Емпіричною функцією розподілу (функцією розподілу вибірки) називається функція $F^*(x)$ детермінованого аргумента *x*, яка дорівнює відносній частоті появи події $(X < x)$ для даної вибірки значень випадкової величини *X*, тобто

$$F^*(x) = W(X < x) = n_x/n, \tag{1.6}$$

де n_x — сума частот тих варіант, які менші від x,

n — обсяг вибірки.

На противагу $F^*(x)$ функцію $F(x)$ називають теоретичною функцією розподілу.

Із означення емпіричної функції випливають такі її властивості, аналогічні властивостям теоретичної функції розподілу:

1) $D(F^*) = R$, $E(F^*) = [0, 1]$;

2) $F^*(x)$ — неспадна функція;

3) якщо x_1 — найменша варіанта, тоді $F^*(x) = 0$ для $x \le x_1$; якщо x_k — найбільша варіанта, тоді $F^*(x) = 1$ для $x > x_k$.

В процесі аналізу статистичних даних суттєву роль відіграє геометрична ілюстрація цих даних. Для наочності будують різні графіки статистичних розподілів, зокрема полігон і гістограму.

Нехай статистичний розподіл вибірки визначається таблицями (1.2) або (1.3).

Полігоном частот (**частотним многокутником**) називається ламана, прямолінійні відрізки якої з'єднують сусідні точки (x_1, n_1), (x_2, n_2), …, (x_k, n_k). Для побудови полігона на осі абсцис відкладають варіанти, а на осі ординат — відповідні їм частоти.

Полігоном відносних частот називається ламана, прямолінійні відрізки якої з'єднують сусідні точки (x_1, w_1), (x_2, w_2), …, (x_k, w_k). При побудові полігона відносних частот на осі абсцис відкладають варіанти x_i, а на осі ординат — відповідні їм відносні частоти w_i.

У випадку неперервної ознаки доцільно будувати гістограму. Нехай розподіли (1.4) та (1.5) такі, що довжина кожного із частинних інтервалів дорівнює одному і тому ж числу h.

Гістограмою частот називається сходинкова фігура, що складається із прямокутників, основами яких є частинні інтервали довжиною h, а висоти дорівнюють відношенню n_i/h (**густина частоти**). Для побудови гістограми частот на осі абсцис відкладаються частинні інтервали, а над ними проводяться прямолінійні відрізки, паралельні осі абсцис на віддалі n_i/h. Площа i-го частинного прямокутника дорівнює $hn_i/h = n_i$, тобто сумі частот тих варіант, що потрапляють в i-ий інтервал. Тому площа гістограми частот дорівнює сумі всіх частот вибірки n.

Порівняння двох гістограм дозволяє зробити висновок про те, що виразність гістограми суттєво залежить від обрання довжини h частинних інтервалів.

Гістограмою відносних частот називається сходинкова фігура, що складається із прямокутників, основами яких є частинні інтервали довжиною h, а висоти дорівнюють відношенню w_i/h (**густини відносної частоти**). Для побудови гістограми відносних частот на осі абсцис відкладаються частинні інтервали, а над ними проводяться відрізки, паралельні осі абсцис на віддалі w_i/h. Площа i-ого частинного прямокутника дорівнює $hw_i/h = w_i$, тобто відносній

частоті тих варіант, що потрапили в *i*-ий інтервал. Отже, площа гістограми відносних частот дорівнює одиниці.

Побудова статистичних розподілів вибірки (1.2) або (1.4) та їх графічне зображення — це тільки перший крок на шляху розв'язування задач математичної статистики. Наступний крок передбачає знаходження числових характеристик, які у компактній формі виражають найбільш суттєві особливості статистичного розподілу вибірки і слугують оцінками (наближеними значеннями) невідомих параметрів розподілу кількісної ознаки генеральної сукупності.

Середня вибіркова (**середня арифметична варіант**) статистичного розподілу (1.2) визначається формулою

$$\bar{x}_в = \left(\sum_{i=1}^{k} x_i n_i \right) \Big/ n. \qquad (1.7)$$

Якщо всі *n* варіант різні, тоді (1.7) набирає такого виду:

$$\bar{x}_в = \left(\sum_{i=1}^{n} x_i \right) \Big/ n. \qquad (1.7^*)$$

Якщо вибірка задається інтервальним статистичним розподілом частот (1.4), тоді при знаходженні $\bar{x}_в$ потрібно перейти до дискретного розподілу (1.2), "нові" варіанти якого є серединами інтервалів, а потім використати формулу (1.7).

Крім вказаної середньої, у статистиці застосовують ще й **структурні середні**, які не залежать від значень варіант, що розташовані на краях розподілу, а пов'язані із рядом частот. До структурних середніх належать медіана та мода.

Медіаною Me^* дискретного статистичного розподілу вибірки (1.2) називається таке число, яке ділить варіаційний ряд, що "породжує" цей розподіл, на дві рівні за кількістю варіант частини. Якщо число варіант непарне, тобто $n = 2m + 1$, тоді $Me^* = x_{m+1}$. Якщо ж обсяг вибірки є парним числом, тобто $n = 2m$, тоді медіана дорівнює середньому арифметичному "середньої" (медіанної) пари варіант:

$$Me^* = (x_m + x_{m+1})/2.$$

Медіаною для інтервального статистичного розподілу називається таке число Me^*, для якого виконується рівність:

$$F^*(Me^*) = 0,5, \qquad (1.8)$$

де $F^*(x)$ — емпірична функція цього розподілу.

Формула для обчислення медіани має такий вид:

$$Me^* = x_m + \frac{0,5 - F^*(x_m)}{F^*(x_{m+1}) - F^*(x_m)} (x_{m+1} - x_m), \qquad (1.9)$$

де $[x_m, x_{m+1})$ – так званий медіанний частинний інтервал ($1 \le m \le k$) для якого виконуються нерівності $F^*(x_m) < 0,5, \;\; F^*(x_{m+1}) > 0,5$.

Модою Mo^* дискретного статистичного розподілу (1.2) називається варіанта, якій відповідає найбільша частота.

Мода для інтервального статистичного розподілу обчислюється таким чином. Спочатку визначається модальний інтервал $[x_m, x_{m+1})$, тобто такий інтервал, для якого $n_m/h_m = \max\limits_{1\le i\le k}\{n_i/h_i\}$, де h_i — довжина частинного інтервалу $[x_m, x_{m+1})$, n_i — число варіант з цього інтервалу. Значення Mo^* міститься всередині модального інтервалу і обчислюється за інтерполяційною формулою

$$Mo^* = x_m + \frac{n_m - n_{m-1}}{2n_m - n_{m-1} - n_{m+1}} h_m. \tag{1.10}$$

Розглянемо деякі числові характеристики розсіювання варіант навколо середньої вибіркової.

Дисперсією вибірковою статистичного розподілу (1.2) називається середнє арифметичне квадратів відхилень варіант від середньої вибіркової:

$$D_в = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x}_в)^2 n_i. \tag{1.11}$$

$D_в$ характеризує середню величину розкиду варіант навколо $\bar{x}_в$ в квадратних одиницях.

На практиці зручніше користуватися так званою **розрахунковою формулою для обчислення дисперсії**:

$$D_в = \overline{x^2} - (\bar{x}_в)^2 = \frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i - (\bar{x}_в)^2. \tag{1.12}$$

Недоліком $D_в$ є її розмірність. Для виправлення цього недоліку використовується інша числова характеристика:

середнє квадратичне відхилення вибіркове

$$\sigma_в = \sqrt{D_в}. \tag{1.13}$$

Коливність окремих значень варіант характеризують показники варіації. Найпростішим із них є показник **розмаху варіації** R, який дорівнює різниці між найбільшою та найменшою варіантами розподілу: $R = x_{\max} - x_{\min}$. Розмах варіації використовується при статистичному вивченні якості продукції.

Якщо $\bar{x}_в$ відмінна від нуля, тоді для порівняння двох статистичних розподілів з точки зору їх розмірності відносно середньої вибіркової вводиться показник **коефіцієнт варіації**, який дорівнює відношенню середнього квадратичного відхилення до середньої вибіркової і виражений у відсотках:

$$V = \frac{\sigma_в}{\bar{x}_в} \cdot 100\%. \tag{1.14}$$

Узагальнюючими характеристиками статистичних розподілів є статистичні моменти розподілу.

Початковим емпіричним моментом *m*-го порядку ν_m^* розподілу (1.2) називається середнє значення варіант у степені *m*:

$$\nu_m^* = \frac{1}{n}\sum_{i=1}^{k}(x_i)^m n_i. \qquad (1.15)$$

Тоді:

при $m = 0$ $\quad \nu_0^* = \frac{1}{n}\sum_{i=1}^{k}(x_i)^0 n_i = 1;$

при $m = 1$ $\quad \nu_1^* = \frac{1}{n}\sum_{i=1}^{k} x_i n_i = \bar{x}$ — середнє вибіркове;

при $m = 2$ $\quad \nu_2^* = \frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i = \overline{x^2}$ — середнє квадратичне;

при $m = 3$ $\quad \nu_3^* = \frac{1}{n}\sum_{i=1}^{k} x_i^3 n_i = \overline{x^3};$

при $m = 4$ $\quad \nu_4^* = \frac{1}{n}\sum_{i=1}^{k} x_i^4 n_i = \overline{x^4}.$

На практиці використовуються моменти перших чотирьох порядків.

З урахуванням рівності (1.12) отримується такий вираз дисперсії вибіркової через початкові емпіричні моменти першого та другого порядків: $D_в = \nu_2^* - (\nu_1^*)^2.$

Центральним емпіричним моментом *m*-го порядку розподілу (1.2) називається середня величина відхилення варіант (від середньої вибіркової) у степені *m*:

$$\mu_m^* = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x}_в)^m n_i. \qquad (1.16)$$

За означенням $\mu_0^* = 1,\ \mu_1^* = 0,$ за означенням $D_в$ (1.11) $\mu_2^* = D_в.$

На практиці використовуються центральні емпіричні моменти третього та четвертого порядку, які з допомогою бінома Ньютона можна виразити через відповідні початкові моменти:

$$\mu_3^* = \nu_3^* - 3\nu_2^*\nu_1^* + 2(\nu_1^*)^3; \qquad (1.17)$$

$$\mu_4^* = \nu_4^* - 4\nu_3^*\nu_1^* + 6\nu_2^*(\nu_1^*)^2 - 3(\nu_1^*)^4. \qquad (1.18)$$

Центральний емпіричний момент третього порядку використовується для обчислення **коефіцієнта асиметрії**

$$A_s^* = \mu_3^* / \sigma_в^3. \qquad (1.19)$$

Для строго симетричного розподілу варіант статистичного розподілу $A_s^* = 0.$ Вважається, якщо $|A_s^*| < 0{,}25,$ то асиметрія низька, якщо $|A_s^*| \le 0{,}5$ — середня, якщо $|A_s^*| > 0{,}5$ — висока.

Якщо $A_s^* < 0,$ то у статистичному розподілі вибірки переважають варіанти, значення яких менші від $\bar{x}_в$. Така **асиметрія** називається **від'ємною** або **лівосторонньою**. Коли $A_s^* > 0,$ тоді у варіаційному ряді переважають варіанти, значення яких більші від $\bar{x}_в$ (**додатна** або **правостороння асиметрія**).

Центральний емпіричний момент четвертого порядку μ_4^* використовується при обчисленні **ексцесу**:

$$E_s^* = \frac{\mu_4^*}{\sigma_в^4} - 3, \tag{1.20}$$

який дає оцінку крутості досліджуваного статистичного розподілу в порівнянні із нормальним.

Для нормально розподіленої ознаки $\mu_4^* / \sigma_в^4 = 3,$ а тому $E_s^* = 0.$ Якщо $E_s^* > 0,$ тоді розподіл вважається гостровершинним, якщо ж $E_s^* < 0,$ тоді — плосковершинним.

Досі розглядалися числові характеристики кількісної ознаки об'єктів. Наведемо означення числової характеристики, пов'язаної із якісною ознакою об'єктів вибірки.

Вибірковою часткою називається відношення числа m об'єктів вибірки з ознакою α до обсягу вибірки: $w = m/n$. Ознакою α може бути стандартність виробу, сортність продукції, стать людини тощо. За змістом w є відносною частотою випадкової події, яка полягає в тому, що навмання відібраний об'єкт із генеральної сукупності має ознаку α.

На практиці часто досліджують вибірки значних обсягів, при цьому варіанти можуть мати велике число знаків або в цілій частині, або в дробовій. В таких випадках знаходження числових характеристик вибірки проводиться з використанням комп'ютерної техніки. Разом з тим, в більшості випадків обсяг обчислень можна суттєво зменшити, використавши спеціальні методи. Одним із них є **метод добутків**. Цей метод дає можливість спрощено обчислювати емпіричні моменти різних порядків для статистичного розподілу (1.2) із **рівновіддаленими варіантами**, тобто варіантами, що утворюють арифметичну прогресію.

Під **зведеними характеристиками** вибірки будемо розуміти середню, дисперсію та початкові емпіричні моменти.

Отже, нехай статистичний розподіл

$$\begin{array}{c|cccc} x_i & x_1 & x_2 & \dots & x_k \\ \hline n_i & n_1 & n_2 & \dots & n_k \end{array} \tag{1.21}$$

має рівновіддалені варіанти, тобто для будь-якого $i = \overline{1, k-1}$ виконується рівність

$$x_{i+1} - x_i = h \neq 0.$$

Метод добутків передбачає здійснення переходу від початкових варіант розподілу (1.21) до **умовних** варіант, які визначаються рівностями

$$u_i = (x_i - C)/h, \quad i = \overline{1,k}, \tag{1.22}$$

де C — хибний нуль (новий початок відліку).

В якості C береться одна із варіант розподілу (1.21), яка є модою або медіаною.

Однією із переваг умовних варіант є їх цілочисельність.

Для умовних варіант розглянемо емпіричні моменти. **Умовним емпіричним моментом порядку *m*** називається початковий момент порядку m, обчислений для умовних варіант:

$$M_m^* = \frac{1}{n}\sum_{i=1}^{k} u_i^m n_i. \tag{1.23}$$

З урахуванням рівності (1.22) для $m_1 = 1$ отримаємо:

$$M_1^* = \frac{1}{n}\sum_{i=1}^{k}\frac{x_i - C}{h} n_i = \frac{1}{h}\left[\frac{1}{n}\sum_{i=1}^{k} x_i n_i - C\frac{1}{n}\sum_{i=1}^{k} n_i\right] = (\bar{x}_в - C)/h, \tag{1.24}$$

звідки

$$\bar{x}_в = M_1^* h + C. \tag{1.25}$$

Аналогічно отримуються центральні емпіричні моменти μ_m^* через умовні емпіричні моменти:

$$D_в = \left[\, M_2^* - (M_1^*)^2 \right] h^2, \tag{1.26}$$

$$\begin{aligned} \mu_3^* &= \left[\, M_3^* - 3M_2^* M_1^* + 2(M_1^*)^3 \right] h^3, \\ \mu_4^* &= \left[\, M_4^* - 4M_3^* M_1^* + 6M_2^*(M_1^*)^2 - 3(M_1^*)^4 \right] h^4. \end{aligned} \tag{1.27}$$

Таким чином, **знаходження $\bar{x}_в$, $D_в$, μ_3^* та μ_4^* зводиться до обчислення умовних емпіричних моментів першого-четвертого порядків**. Останні можна знайти за допомогою методу добутків.

Припустимо, що деяка сукупність розбита на групи, що **не перетинаються**, тобто кожний об'єкт сукупності **належить тільки одній групі**. Для кожної із груп можна знайти середню (арифметичну), яка називається **груповою середньою**. Виникає питання: яким чином пов'язані середня всієї сукупності (загальна середня) із середніми груповими. Відповідь на це питання дає наступне твердження.

Теорема 1.1. Загальна середня дорівнює середній арифметичній групових середніх, зваженій за обсягами цих груп:

$$\bar{x} = \left(\overline{x^{(1)}} N_1 + \overline{x^{(2)}} N_2 + \ldots + \overline{x^{(k)}} N_k \right) \Big/ n = \sum_{i=1}^{k} \overline{x^{(i)}} N_i \Big/ n, \tag{1.28}$$

де $\bar{x}$ — загальна середня; $\overline{x^{(i)}}$ — середня групова i-ої групи $(i=\overline{1,k})$; N_i — обсяг i-ої групи; $n=\sum_{i=1}^{k} N_i$.

Аналогічне питання вивчимо і стосовно дисперсії всієї сукупності S, попередньо розглянувши додаткові означення.

Груповою дисперсією j-ої групи S_j називається дисперсія D_j розподілу значень ознаки, що складає цю сукупність, відносно групової середньої $\overline{x^{(j)}}$:

$$D_j=\left(\sum_{i=1}^{m}(x_i-\overline{x^{(j)}})^2 n_i\right)\Big/ N_j,$$

де n_i — частоти варіант $x_i \in S_j$, $N_j=\sum_{i=1}^{m} n_i$ — обсяг групи S_j.

Внутрігруповою дисперсією називається середнє арифметичне всіх групових дисперсій, зважених по обсягах груп:

$$D_{внгр}=\left(\sum_{j=1}^{k} N_j D_j\right)\Big/ n, \qquad (1.29)$$

де N_j — обсяг j-ої групи S_j; $n=\sum_{j=1}^{k} N_j$ — обсяг всієї сукупності S.

Міжгруповою дисперсією називається дисперсія групових середніх відносно загальної середньої:

$$D_{міжгр}=\left(\sum_{j=1}^{k} N_j(\overline{x^{(j)}}-\bar{x})^2\right)\Big/ n, \qquad (1.30)$$

де $\overline{x^{(j)}}$ — групова середня групи S_j; N_j — обсяг групи S_j; $\bar{x}$ — загальна середня (сукупності S), $n=\sum_{j=1}^{k} S_j$ — обсяг всієї сукупності S.

Загальною дисперсією називається дисперсія значень ознаки всієї сукупності S відносно загальної середньої:

$$D_{заг}=\left(\sum_{i=1}^{m} n_i(x_i-\bar{x})^2\right)\Big/ n, \qquad (1.31)$$

де n_i — частота значення x_i, $\bar{x}$ — загальна середня сукупності S, n — обсяг сукупності S.

Зв'язок між $D_{заг}$, $D_{внгр}$ та $D_{міжгр}$ встановлює:

Теорема 1.2. Якщо сукупність складається із кількох груп, що не перетинаються, тоді загальна дисперсія дорівнює сумі внутрігрупової і міжгрупової дисперсій:

$$D_{заг} = D_{внгр} + D_{міжгр}.$$

РОЗВ'ЯЗУВАННЯ ТИПОВИХ ЗАДАЧ

Задача 1.1. При вивченні питання про норми виробітку тканьспостерігалася така частота обривів пряжі на однотипних ткацьких станках на різних проміжках часу однакової довжини t:

7, 1, 2, 1, 2, 5, 4, 4, 3, 2, 2, 6, 0, 1, 6,
5, 3, 2, 0, 1, 4, 3, 2, 1, 5, 3, 0, 4, 2, 3.

1) Скласти статистичний розподіл частот та відносних частот числа обривів пряжі на станках;
2) побудувати полігон частот та відносних частот;
3) знайти емпіричну функцію розподілу та побудувати її графік;
4) з'ясувати питання, чи можна використати метод добутків для знаходження зведених числових характеристик; у випадку позитивної відповіді виконати розрахунки зведених характеристик в 5) цим методом;
5) обчислити вибіркові: середню, дисперсію, середнє квадратичне відхилення, моду, медіану, розмах варіацій, коефіцієнт варіацій, коефіцієнт асиметрії та ексцес;
6) здійснити розбиття вибіркової сукупності на групи, що не перетинаються, з тим, щоб використати теореми 1.1 та 1.2 для обчислення $\bar{x}_в$ та $D_в$, співставивши отримані відповіді із результатами з 5).

○ 1) Обсяг вибірки $n = 30$. Даний ряд варіант запишемо у вигляді варіаційного ряду:

0, 0, 0, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 6, 6, 7.

Варіанти: $x_1 = 0$, $x_2 = 1$, $x_3 = 2$, $x_4 = 3$, $x_5 = 4$, $x_6 = 5$, $x_7 = 6$, $x_8 = 7$.

Частоти: $n_1 = 3$, $n_2 = 5$, $n_3 = 7$, $n_4 = 5$, $n_5 = 4$, $n_6 = 3$, $n_7 = 2$, $n_8 = 1$.

В підсумку отримується статистичний розподіл частот:

x_i	0	1	2	3	4	5	6	7
n_i	3	5	7	5	4	3	2	1

.

Контроль: $\sum n_i = 3 + 5 + 7 + 5 + 4 + 3 + 2 + 1 = 30 = n$.

За формулою $w_i = n_i / n$ послідовно обчислюємо відносні частоти: $w_1 = 3/30$, $w_2 = 5/30$, $w_3 = 7/30$, $w_4 = 5/30$, $w_5 = 4/30$, $w_6 = 3/30$, $w_7 = 2/30$, $w_8 = 1/30$. Контроль: $\sum w_i = 3/30 + 5/30 + 7/30 + 5/30 + 4/30 + 3/30 + 2/30 + 1/30 = 1$. Отже, статистичний розподіл відносних частот числа обривів пряжі на станках має такий вид:

x_i	0	1	2	3	4	5	6	7
w_i	3/30	5/30	7/30	5/30	4/30	3/30	2/30	1/30

.

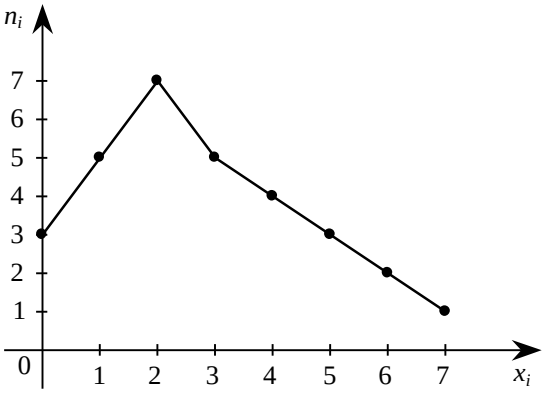
2) Полігон частот зобразимо на мал.1.1, а полігон відносних частот – на мал.1.2.

3) Обсяг вибірки $n = 30$. Якщо $x \le 0$, тоді немає жодної варіанти, меншої від x, тобто $n_x = 0$, а тому і $F^*(x) = n_x/30 = 0$.

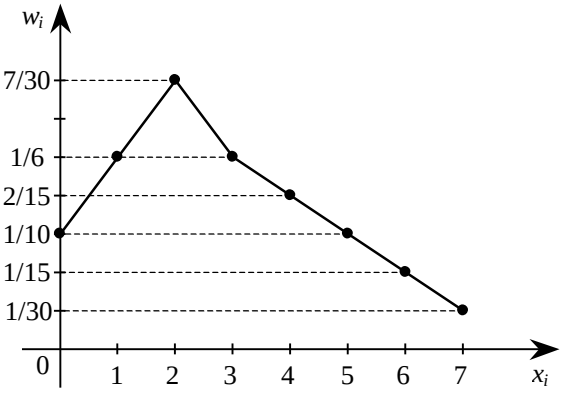
Нехай $x \in (0; 1]$. Тоді варіанта $x = 0$ є меншою від x, тому $n_x = 3$ і $F^*(x) = 3/30 = 0{,}1$.

Якщо x задовільняє подвійній нерівності $1 < x \le 2$, тоді меншими від x є варіанти 0 та 1, сума частот яких $n_x = 3 + 5 = 8$. Тому $F^*(x) = 8/30 = 4/15$ для $x \in (1; 2]$.

Якщо x таке, що виконується подвійна нерівність $2 < x \le 3$, тоді меншими від x є варіанти 0, 1, 2, суми частот яких $n_x = 3 + 5 + 7 = 15$. Тому для $x \in (2; 3]$ $F^*(x) = 15/30 = 0{,}5$.



Мал.1.1.



Мал.1.2.

Аналогічно знаходимо значення $F^*(x)$ для інтервалів (3; 4], (4; 5], (5; 6], (6; 7], (7; ∞]. В підсумку отримаємо шукану емпіричну функцію розподілу:

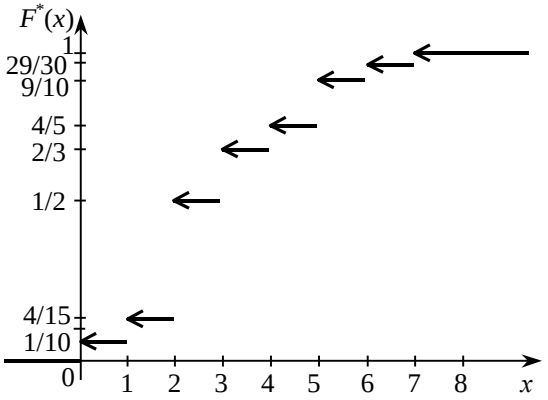
$$F^*(x)=\begin{cases} 0, & \text{якщо } x \le 0, \\ 1/10, & \text{якщо } 0 < x \le 1, \\ 4/15, & \text{якщо } 1 < x \le 2, \\ 1/2, & \text{якщо } 2 < x \le 3, \\ 2/3, & \text{якщо } 3 < x \le 4, \\ 4/5, & \text{якщо } 4 < x \le 5, \\ 9/10, & \text{якщо } 5 < x \le 6, \\ 29/30, & \text{якщо } 6 < x \le 7, \\ 1, & \text{якщо } x > 7. \end{cases}$$

Графік цієї функції наведено на мал. 1.3.

4) Метод добутків для знаходження зведених числових характеристик можна використати для статистичного розподілу із рівновіддаленими варіантами. В нашому випадку отриманий статистичний розподіл володіє цією властивістю.

5) Для знаходження зведених характеристик (середньої, дисперсії та початкових емпіричних моментів), використаємо розрахункову таблицю методу добутків, заповнюючи її в такому порядку:

1) Варіанти даного розподілу запишемо в перший стовпець.



Мал. 1.3.

2) Відповідні частоти поміщаємо у другий стовпець; суму частот (30) запишемо в нижню клітинку стовпця.

3) В якості хибного нуля C виберемо варіанту 3, яка знаходиться приблизно в середині варіаційного ряду. В третьому стовпці в

рядку, що містить цю варіанту, записуємо 0; над ним записуємо послідовно –1, –2, –3, а під ним — числа 1, 2, 3, 4.

4) Добутки частот на умовні варіанти записуємо в четвертий стовпець; суму цих чисел $\left(\sum n_i u_i\right)$ поміщаємо в нижню клітинку стовпця.

5) В п'ятий стовпець запишемо добутки вмістимих клітинок третього та четвертого стовпців; їх суму $\left(\sum n_i u_i^2\right)$ поміщаємо в нижню клітинку стовпця.

6) В шостий контрольний стовпець записуємо добуток частот на квадрат умовних варіант, збільшений на одиницю. Сума їх $\left(\sum n_i (u_i+1)^2\right)$ записується в нижню клітинку. Суми другого, четвертого, п'ятого та шостого стовпців при правильних розрахунках повинні задовольняти рівність

$$\sum n_i (u_i+1)^2 = \sum n_i u_i^2 + 2\sum n_i u_i + n.$$

7) В сьомий стовпець записуються добутки вмістимих третього та п'ятого стовпців; їх сума $\left(\sum n_i u_i^3\right)$ поміщається в нижній клітинці стовпця.

8) У восьмий стовпець поміщаються добутки чисел сьомого та третього стовпців. Сума $\left(\sum n_i u_i^4\right)$ записується в нижній клітинці стовпця.

У підсумку отримаємо розрахункову табл. 1.1.

Таблиця 1.1

1	2	3	4	5	6	7	8
x_i	n_i	u_i	$n_i u_i$	$n_i u_i^2$	$n_i (u_i+1)^2$	$n_i u_i^3$	$n_i u_i^4$
0	3	–3	–9	27	12	–81	243
1	5	–2	–10	20	5	–40	80
2	7	–1	–7	7	0	–7	7
3	5	0	0	0	5	0	0
4	4	1	4	4	16	4	4
5	3	2	6	12	27	24	48
6	2	3	6	18	32	54	162
7	1	4	4	16	25	64	256
	n = = 30		$\sum n_i u_i$ = = –6	$\sum n_i u_i^2$ = = 104	$\sum n_i (u_i+1)^2$ = = 122	$\sum n_i u_i^3$ = = 18	$\sum n_i u_i^4$ = = 80

Контроль: $\sum n_i u_i^2 + 2\sum n_i u_i + n = 104 + 2\cdot(-6) + 30 = 122,$

$\sum n_i (u_i+1)^2 = 122.$

Обчислимо умовні моменти першого-четвертого порядків:

$$M_1^* = \left(\sum n_i u_i\right)/n = -6/30 = -0{,}2;$$

$$M_2^* = \left(\sum n_i u_i^2\right)/n = 104/30 \approx 3{,}4667;$$

$$M_3^* = \left(\sum n_i u_i^3\right)/n = 18/30 = 0{,}6;$$

$$M_4^* = \left(\sum n_i u_i^4\right)/n = 800/30 = 26{,}6667.$$

За умовою $h = 1$, за вибором хибного нуля $C = 3$.

Обчислимо вибіркові середню, дисперсію і середнє квадратичне відхилення, використавши рівності (1.25) та (1.26):

$$\bar{x}_в = M_1^* h + C = -0{,}2 \cdot 1 + 3 = 2{,}8;$$

$$D_в = [M_2^* - (M_1^*)^2]h^2 = [3{,}4667 - (0{,}2)^2] \cdot 1^2 = 3{,}4267;$$

$$\sigma_в = \sqrt{D_в} = \sqrt{3{,}4267} = 1{,}8511.$$

Знайдемо центральні емпіричні моменти третього та четвертого порядків за формулами (1.27):

$$\mu_3^* = \left[\, M_3^* - 3M_2^*M_1^* + 2(M_1^*)^3 \right] h^3 =$$

$$= [0{,}6 - 3 \cdot 3{,}4667(-0{,}2) + 2(-0{,}2)^3] \cdot 1^3 = 2{,}66402;$$

$$\mu_4^* = \left[\, M_4^* - 4M_3^*M_1^* + 6M_2^*(M_1^*)^2 - 3(M_1^*)^4 \right] h^4 =$$

$$= [26{,}6667 - 4 \cdot 0{,}6 \cdot (-0{,}2) + 6 \cdot 3{,}4667 \cdot (-0{,}2)^2 - 3(-0{,}2)^4] \cdot 1^4 = 27{,}973908.$$

Мода $Mo^* = 2$, оскільки варіанті 2 відповідає найбільша частота 7.

Медіану Me^* можна знайти або за варіаційним рядом, отриманим в 1), або безпосередньо із статистичного розподілу. Сума частот перших трьох варіант цього розподілу дорівнює 15 (половині обсягу вибірки), а наступних п'яти — також 15. Тому медіана знаходиться між варіантами 2 та 3: $Me^* = (2+3)/2 = 2{,}5$. Неспівпадання $\bar{x}_в$, Mo^* та Me^* свідчить про відсутність строгої симетричності розподілу.

Коефіцієнт варіації (згідно із формулою (1.14))

$$V = \frac{\sigma_в}{\bar{x}_в} \cdot 100\% = \frac{1{,}8511}{2{,}8} \cdot 100\% = 66{,}11\%.$$

Коефіцієнт асиметрії та ексцес обчислимо за формулами (1.19) та (1.20) відповідно:

$$A_s^* = \mu_3^* / \sigma_в^3 = 2{,}66402/(1{,}8511)^3 = 0{,}420;$$

$$E_s^* = \frac{\mu_4^*}{\sigma_в^4} - 3 = 27{,}973908/(1{,}8511)^4 - 3 = -0{,}6175.$$

Отже, даний статистичний розподіл числа розривів пряжі на станках характеризується середньою правосторонньою асиметрією і є плосковершинним.

Відмітимо, що для обчислення $\bar{x}_в$ та $D_в$ безпосередньо можна було використати формули (1.7) і (1.12):

$$\bar{x}_в = (\sum_{i=1}^{k} x_i n_i)/n = (0\cdot 3+1\cdot 5+2\cdot 7+3\cdot 5+4\cdot 4+5\cdot 3+6\cdot 2+7\cdot 1)/30 =$$

$$= 84/30 = 2{,}8;$$

$$D_в = \overline{x^2} - (\bar{x}_в)^2 = \frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i - (\bar{x}_в)^2 =$$

$$= (0^2\cdot 3+1^2\cdot 5+2^2\cdot 7+3^2\cdot 5+4^2\cdot 4+5^2\cdot 3+6^2\cdot 2+7^2\cdot 1)/30 -$$

$$-(2{,}8)^2 = 338/30 - 7{,}84 = 3{,}4267.$$

Висновок: середнє число обривів пряжі на станках за проміжок часу складає 2,8, а середня величина розкиду чисел розривів пряжі навколо середньої 2,8 дорівнює 1,8511.

6) Розіб'ємо статистичний розподіл на дві групи, що не перетинаються :

перша група

x_i	0	1	2	3
n_i	3	5	7	5

;

друга група

x_i	4	5	6	7
n_i	4	3	2	1

.

Тоді обсяги груп: $N_1 = 3+5+7+5=20$, $N_2 = 4+3+2+1=10$ і $n = N_1 + N_2 = 30$.

Обчислимо групові середні та загальну середню (в даному випадку середню вибіркову):

$$\overline{x^{(1)}} = \left(\sum x_i n_i\right)/N_1 = (0\cdot 3+1\cdot 5+2\cdot 7+3\cdot 5)/20 = 1{,}7;$$

$$\overline{x^{(2)}} = \left(\sum x_i n_i\right)/N_2 = (4\cdot 4+5\cdot 3+6\cdot 2+7\cdot 1)/10 = 5,$$

$$\bar{x} = \left(\sum N_i \overline{x^{(i)}}\right)/n = (20\cdot 1{,}7+10\cdot 5)/30 = 2{,}8.$$

Знайдемо групові дисперсії:

$$D_1 = \overline{x^2} - \left(\overline{x^{(1)}}\right)^2 = (0^2\cdot 3+1^2\cdot 5+2^2\cdot 7+3^2\cdot 5)/20 - (1{,}7)^2 = 1{,}01;$$

$$D_2 = \overline{x^2} - \left(\overline{x^{(2)}}\right)^2 = (4^2\cdot 4+5^2\cdot 3+6^2\cdot 2+7^2\cdot 1)/10 - 5^2 = 1.$$

Внутрігрупова дисперсія:

$$D_{внгр} = \frac{N_1 D_1 + N_2 D_2}{N_1 + N_2} = \frac{20\cdot 1{,}01+10\cdot 1}{30} = 1{,}0067;$$

міжгрупова дисперсія:

$$D_{міжгр} = \frac{N_1\left(\overline{x^{(1)}} - \bar{x}\right)^2 + N_2\left(\overline{x^{(2)}} - \bar{x}\right)^2}{n} = \frac{20(1{,}7-2{,}8)^2 + 10(5-2{,}8)^2}{30} = 2{,}42.$$

Згідно із теоремою 1.2 загальна дисперсія:

$$D_{заг} = D_{внгр} + D_{міжгр} = 1{,}0067 + 2{,}42 = 3{,}4267. \bullet$$

Розв'язування наступної задачі ілюструє:

1) перехід від дискретного розподілу до інтервального;

2) ефективність використання методу добутків при обчисленні зведених характеристик вибірки;

3) мізерність похибок при обчисленні числових характеристик вибірки, які виникають внаслідок переходу до інтервального розподілу.

Задача 1.2. За період між черговими переналадками обладнання здійснено контрольні виміри товщини (в міліметрах) 200 вкладишів шатунних підшипників. Отримані дані наведені в табл. 1.2.

Таблиця 1.2

1,754	1,739	1,743	1,764	1,733	1,736	1,743	1,742
1,728	1,732	1,731	1,752	1,737	1,747	1,758	1,737
1,724	1,737	1,733	1,713	1,740	1,740	1,729	1,740
1,751	1,739	1,747	1,748	1,730	1,750	1,740	1,732
1,740	1,730	1,748	1,757	1,741	1,733	1,743	1,745
1,748	1,723	1,737	1,748	1,741	1,751	1,714	1,750
1,744	1,748	1,758	1,756	1,727	1,731	1,738	1,753
1,735	1,738	1,743	1,729	1,743	1,737	1,731	1,734
1,741	1,742	1,744	1,756	1,744	1,752	1,739	1,740
1,729	1,745	1,742	1,753	1,743	1,734	1,731	1,734
1,732	1,732	1,746	1,748	1,755	1,738	1,742	1,729
1,731	1,725	1,729	1,745	1,739	1,754	1,752	1,720
1,750	1,734	1,749	1,738	1,747	1,757	1,751	1,746
1,723	1,736	1,746	1,744	1,759	1,728	1,751	1,750
1,746	1,759	1,748	1,740	1,735	1,745	1,740	1,746
1,737	1,726	1,743	1,755	1,740	1,726	1,745	1,744
1,735	1,746	1,739	1,732	1,758	1,744	1,754	1,724
1,742	1,750	1,761	1,758	1,753	1,757	1,720	1,733
1,738	1,728	1,758	1,732	1,763	1,733	1,745	1,766
1,745	1,743	1,734	1,733	1,755	1,756	1,769	1,750
1,740	1,762	1,738	1,742	1,740	1,740	1,760	1,752
1,746	1,728	1,743	1,718	1,738	1,762	1,728	1,734
1,753	1,751	1,748	1,735	1,739	1,729	1,754	1,736
1,762	1,748	1,738	1,726	1,757	1,738	1,726	1,720
1,751	1,734	1,724	1,741	1,752	1,732	1,738	1,739

1) Скласти інтервальний статистичний розподіл частот та відносних частот вибірки.

На основі отриманого інтервального статистичного розподілу:

2) побудувати гістограми частот та відносних частот; 3) знайти емпіричну функцію розподілу та побудувати її графік; 4) методом добутків обчислити зведені характеристики вибірки, моду, медіану, а також зробити висновки про симетричність чи асиметричність вибіркового розподілу та його гостровершинність чи плосковершинність на підставі коефіцієнта асиметрії та ексцесу; 5)

порівняти числові характеристики $\bar{x}_в$, $D_в$ і $\sigma_в$, знайдені за статистичними даними із табл.1.2 та обчисленими на підставі відповідного інтервального розподілу.

○ 1) Всі варіанти знаходяться у проміжку [1,713; 1,769] довжиною 0,056 мм. Розіб'ємо його на 8 рівних за довжиною інтервалів:
[1,713; 1,720), [1,720; 1,727), [1,727; 1,734), [1,734; 1,741),
[1,741; 1,748), [1,748; 1,755), [1,755; 1,762), [1,762; 1,769].
Кожну із варіант табл. 1.2 віднесемо до одного із цих частинних інтервалів і просумуємо число варіант, що потрапляють в перший, другий, …, восьмий інтервали. В результаті отримаємо:

$n_1 = 3$, $n_2 = 13$, $n_3 = 32$, $n_4 = 49$, $n_5 = 42$, $n_6 = 35$, $n_7 = 19$, $n_8 = 7$, $n = \sum_{i=1}^{8} n_i = 200$.

Підсумком є інтервальний статистичний розподіл частот, наведений в табл. 1.3.

Таблиця 1.3

$[x_i; x_{i+1})$	n_i	$[x_i; x_{i+1})$	n_i
[1,713; 1,720)	3	[1,741; 1,748)	42
[1,720; 1,727)	13	[1,748; 1,755)	35
[1,727; 1,734)	32	[1,755; 1,762)	19
[1,734; 1,741)	49	[1,762; 1,769]	7

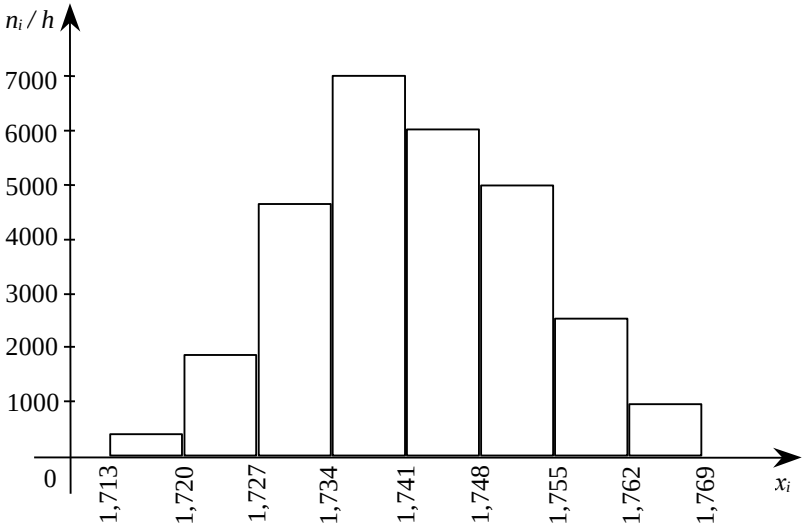
З урахуванням того, що обсяг вибірки $n = 200$, запишемо інтервальний статистичний розподіл відносних частот вибірки (табл. 1.4):

Таблиця 1.4

$[x_i; x_{i+1})$	w_i	$[x_i; x_{i+1})$	w_i
[1,713; 1,720)	3/200	[1,741; 1,748)	42/200
[1,720; 1,727)	13/200	[1,748; 1,755)	35/200
[1,727; 1,734)	32/200	[1,755; 1,762)	19/200
[1,734; 1,741)	49/200	[1,762; 1,769]	7/200

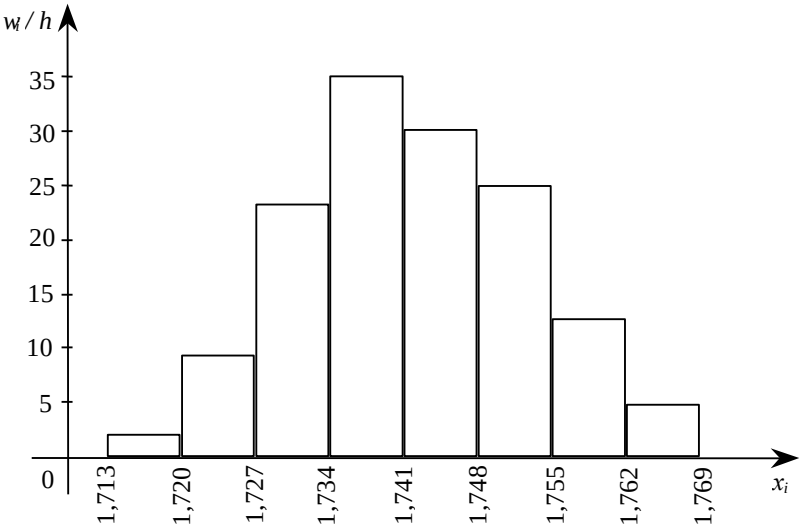
2) Гістограма частот розподілу із h=0,007 зображена на мал.1.4, а гістограма відносних частот на мал.1.5.

3) Досліджувана кількісна ознака X є непервною випадковою величиною. Тому і функція розподілу імовірностей $F(x)$, і емпірична функція $F^*(x)$ є неперервними функціями детермінованого аргумента x.



Мал.1.4.

Нехай $x \le 1{,}713$. Тоді $n_x = 0$, оскільки спостережених значень кількісної ознаки, менших від x, немає. Отже, і $F^*(x) = 0$ для всіх $x \le 1{,}713$.



Мал.1.5.

Знайдемо значення емпіричної функції розподілу для $x = 1{,}720$ — лівого кінця другого інтервалу. При цьому вважатимемо, що ми не можемо зробити

цього для кожної внутрішньої точки першого інтервалу. При $x = 1{,}720$ $n_x = 3$ і $F^*(1{,}720) = 3/200 = 0{,}015$.

Для $x = 1{,}727$ $n_x = 3 + 13 = 16$, $F^*(1{,}727) = 16/200 = 0{,}08$.

Для $x = 1{,}734$ $n_x = 16 + 32 = 48$, $F^*(1{,}734) = 48/200 = 0{,}24$.

Аналогічно знаходимо:

$F^*(1{,}741) = (48 + 49)/200 = 97/200 = 0{,}485$;

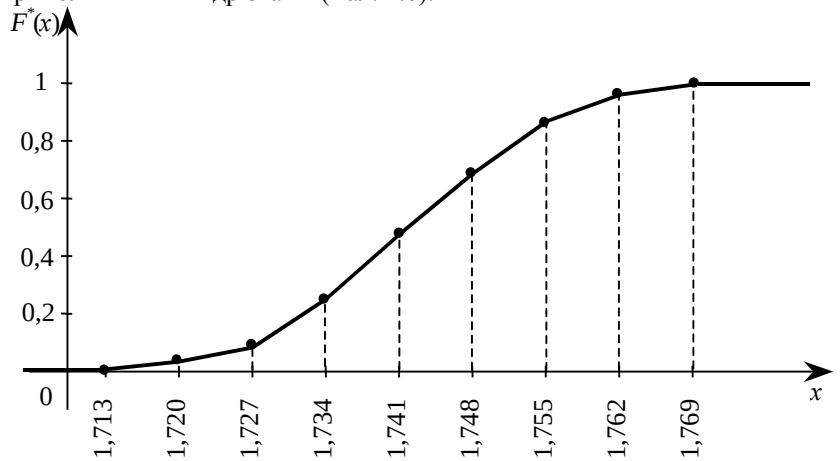
$F^*(1{,}748) = (97 + 42)/200 = 139/200 = 0{,}695$;

$F^*(1{,}755) = (139 + 35)/200 = 174/200 = 0{,}87$;

$F^*(1{,}762) = (174 + 19)/200 = 193/200 = 0{,}965$;

Нарешті, для $x > 1{,}769$ всі 200 спостережених значень кількісної ознаки менші від x, тобто $n_x = 200$. Тому $F^*(x) = 1$ для $x > 1{,}769$.

Побудуємо графік отриманої емпіричної функції розподілу: спочатку на інтервалах $(-\infty; 1{,}713]$ і $(1{,}769; \infty)$, а потім у вказаних точках. Для того, щоб показати неперервність зміни $F^*(x)$, отримані сусідні точки з'єднаємо прямолінійними відрізками (мал. 1.6).



Мал 1.6.

4) Для знаходження числових характеристик вибірки (в тому числі зведених), заданої інтервальним статистичним розподілом із п.1, перейдемо до дискретного розподілу :

x_i	1,7165	1,7235	1,7305	1,7375	1,7445	1,7515	1,7585	1,7655
n_i	3	13	32	49	42	35	19	7

При цьому "нові" варіанти є серединами частинних інтервалів.

Складемо розрахункову таблицю методу добутків (табл.1.5)

Таблиця 1.5

1	2	3	4	5	6	7	8
x_i	n_i	u_i	n_iu_i	$n_iu_i^2$	$n_i(u_i+1)^2$	$n_iu_i^3$	$n_iu_i^4$
1,7165	3	–3	–9	27	12	–81	243
1,7235	13	–2	–36	72	13	–144	288
1,7305	32	–1	–32	32	0	–32	32
1,7375	49	0	0	0	49	0	0
1,7445	42	1	42	42	168	42	42
1,7515	35	2	70	140	315	280	560
1,7585	19	3	57	171	304	513	1539
1,7655	7	4	28	112	175	448	1792
	n = = 200		$\sum n_iu_i$ = = 120	$\sum n_iu_i^2$ = = 596	$\sum n_i(u_i+1)^2$ = = 1036	$\sum n_iu_i^3$ = = 1026	$\sum n_iu_i^4$ = = 4496

Контроль:

$$\sum n_iu_i^2 + 2\sum n_iu_i + n = 596 + 2\cdot 120 + 200 = 1036.$$

$$\sum n_i(u_i+1)^2 = 1036.$$

Обчислимо умовні моменти першого-четвертого порядків:

$$M_1^* = \left(\sum n_iu_i\right)/n = 120/200 = 0{,}6;$$

$$M_2^* = \left(\sum n_iu_i^2\right)/n = 596/200 = 2{,}98;$$

$$M_3^* = \left(\sum n_iu_i^3\right)/n = 16/200 = 5{,}13;$$

$$M_4^* = \left(\sum n_iu_i^4\right)/n = 4496/200 = 22{,}48.$$

За умовою $h = 0{,}007$, за вибором хибного нуля $C = 1{,}7375$.

Обчислимо вибіркові середню, дисперсію і середнє квадратичне відхилення, використавши рівності (1.30) та (1.31):

$$\bar{x}_в = M_1^*h + C = 0{,}6\cdot 0{,}007 + 1{,}7375 = 1{,}7417;$$

$$D_в = [M_2^* - (M_1^*)^2]h^2 = [2{,}98 - (0{,}6)^2](0{,}007)^2 = 0{,}0001283;$$

$$\sigma_в = \sqrt{D_в} = 0{,}0113269.$$

Знайдемо центральні емпіричні моменти третього та четвертого порядків за формулами (1.27):

$$\mu_3^* = \left[\, M_3^* - 3M_2^*M_1^* + 2(M_1^*)^3 \right] h^3 =$$

$$= [5{,}13 - 3\cdot 2{,}98\cdot 0{,}6 + 2(0{,}6)^3](0{,}007)^3 =$$

$$= (5{,}13 - 5{,}364 + 0{,}216)(0{,}007)^3 = -0{,}018\cdot(0{,}007)^3.$$

$$\mu_4^* = \left[M_4^* - 4M_3^* M_1^* + 6M_2^*(M_1^*)^2 - 3(M_1^*)^4 \right] h^4 =$$

$$= [22{,}48 - 4 \cdot 5{,}13 \cdot 0{,}6 + 62{,}98 \cdot (0{,}6)^2 - 3(0{,}6)^4](0{,}007)^4 = = 16{,}216 \cdot (0{,}007)^4.$$

$$= (22{,}48 - 12{,}312 + 6{,}4368 - 0{,}3888)(0{,}007)^4 =$$

Зауваження. Форма відповіді для μ_3^* та μ_4^* зумовлена, з одного боку, тим, що вони є "комп'ютерними нулями". По-друге, при обчисленні коефіцієнта асиметрії та ексцесу вони діляться на дуже малі числа: $\sigma_в^3$ та $\sigma_в^4$ ($\sigma_в \approx 0{,}011$).

Оскільки всі частинні інтервали даного розподілу мають однакову довжину, то модальним є інтервал [1,734; 1,741), якому відповідає найбільша частота 49.

Значення Mo^* міститься всередині цього інтервалу і обчислюється за формулою (1.10):

$$Mo^* = x_m + \frac{n_m - n_{m-1}}{2n_m - n_{m-1} - n_{m+1}} h = 1{,}734 + \frac{49 - 32}{2 \cdot 49 - 32 - 42} \cdot 0{,}007 =$$

$$= 1{,}734 + \frac{17}{24} \cdot 0{,}007 \approx 1{,}73896.$$

Для обчислення медіани знайдемо спочатку медіанний частинний інтервал $[x_m; x_{m+1})$, для якого виконуються нерівності:

$$F^*(x_m) < 0{,}5, \qquad F^*(x_{m+1}) > 0{,}5.$$

Згідно із 3) $F^*(1{,}741)=0{,}485<0{,}5$, $F^*(1{,}748)=0{,}695>0{,}5$.

Отже, $x_m = 1{,}741$, $x_{m+1} = 1{,}748$. За формулою (1.9)

$$Me^* = x_m + \frac{0{,}5 - F^*(x_m)}{F^*(x_{m+1}) - F^*(x_m)}(x_{m+1} - x_m) =$$

$$= 1{,}741 + \frac{0{,}5 - 0{,}485}{0{,}695 - 0{,}485}(1{,}748 - 1{,}741) =$$

$$= 1{,}741 + \frac{0{,}015}{0{,}21} \cdot 0{,}007 = 1{,}741 + 0{,}0005 = 1{,}7415.$$

Коефіцієнт варіації обчислюємо за формулою (1.14):

$$V = \frac{\sigma_в}{\overline{x}_в} \cdot 100\% = \frac{0{,}0113269}{1{,}7417} \cdot 100\% = 0{,}65\%.$$

Коефіцієнт асиметрії та ексцес знайдемо за формулами (1.19) та (1.20) відповідно:

$$A_s^* = \mu_3^* / \sigma_в^3 = -0{,}018 \cdot (0{,}007)^3 / (0{,}0113269)^3 =$$

$$= -0{,}018(0{,}007/0{,}0113269)^3 \approx -0{,}0043.$$

$E_s^* = \mu_4^* / \sigma_в^4 - 3 = 16{,}216 \cdot (0{,}007)^4 / (0{,}0113269)^4 - 3 \approx$

$\approx 2{,}365 - 3 = -0{,}6347.$

На підставі отриманих даних можна зробити висновок, що даний інтервальний розподіл є практично симетричним із мізерною лівосторонньою асиметрією і плосковершинним.

5) Використання формул (1.7), (1.12) та (1.13) для початкового ряду варіант з табл. 1.2 дає такі результати:

$$\bar{x}_в = 1{,}7417, \qquad \sigma_в = 0{,}0113269.$$

Порівняння цих результатів із числовими характеристиками з 4) дозволяє зробити висновок про достатній рівень точності результатів, отриманих внаслідок переходу до інтервального розподілу.

§ 2. СТАТИСТИЧНЕ ОЦІНЮВАННЯ

Точкові статистичні оцінки параметрів розподілу та їх властивості. Оцінка середньої генеральної для простої вибірки (повторної та безповторної). Оцінка генеральної частки для простої вибірки (повторної та безповторної). Середні квадратичні помилки (СКП) простої вибірки. Виправлена дисперсія вибіркова. Інтервальні статистичні оцінки (Довірчі інтервали для оцінок $\bar{x}_г$ та p для немалих вибірок. Знаходження мінімального обсягу вибірки. Довірчі інтервали для оцінки $\bar{x}_г = a$ для малої вибірки. Довірчі інтервали для $D_г$ та $\sigma_г$ у випадку малої вибірки).

КОРОТКІ ТЕОРЕТИЧНІ ВІДОМОСТІ

Точкові статистичні оцінки параметрів розподілу та їх властивості

Нехай досліджується неперервна кількісна ознака X об'єктів генеральної сукупності з метою знаходження невідомого закону розподілу. В розпорядженні дослідника є статистичні дані вибірки. В задачі 1.2 для статистичних даних знайдено коефіцієнт асиметрії $A_S^* = -0{,}0043,$ значення якого дозволяє висунути гіпотезу про нормальний закон розподілу ознаки X (дослідження питання про правильність цієї гіпотези або хибність буде проведене в наступному параграфі). Оскільки нормальний розподіл повністю визначається двома параметрами a та σ, то виникає необхідність оцінити їх, тобто знайти наближені значення, використовуючи **тільки** спостережені варіанти $x_1, x_2, ..., x_n$ вибірки обсягом n. Параметр $a = M(X)$, математичне сподівання характеризує середнє арифметичне спостережених можливих значень випадкової величини X. З другого боку, $\bar{x}_в$ — це середнє арифметичне варіант. Тому (поки що інтуїтивно) доцільно наближати a середнім вибірковим:

$$a \approx \bar{x}_{\text{в}} = \left(\sum_{i=1}^{k} x_i n_i \right) \Big/ n. \qquad (2.1)$$

Аналогічно

$$\sigma \approx \sigma_{\text{в}} = \sqrt{\sum_{i=1}^{k} (x_i - \bar{x}_{\text{в}})^2 n_i \Big/ n}. \qquad (2.2)$$

Ще раз відмітимо, що праві частини рівностей (2.1), (2.2) є **випадковими величинами**, оскільки об'єкти у вибірку потрапляють **випадковим чином**.

Нехай тепер кількісна ознака X об'єктів генеральної сукупності є дискретною випадковою величиною. Припустимо, що статистичні розподіли генеральної та вибіркової сукупності описуються таблицями:

$$\begin{array}{c|cccc} x_i & x_1 & x_2 & \dots & x_m \\ \hline N_i & N_1 & N_2 & \dots & N_m \end{array}, \qquad N = \sum_{i=1}^{m} N_i, \qquad (2.3)$$

$$\begin{array}{c|cccc} x_i & x_1 & x_2 & \dots & x_m \\ \hline n_i & n_1 & n_2 & \dots & n_m \end{array}, \qquad n = \sum_{i=1}^{m} n_i, \qquad (2.4)$$

де N та n — обсяги генеральної та вибіркової сукупностей відповідно. Розподіл (2.3) є гіпотетичним — він завжди буде для нас невідомим, бо в протилежному випадку відпала б необхідність в дослідженні вибіркової сукупності. Нарешті, зауважимо, що деякі із частот розподілу (2.4) можуть дорівнювати нулю, що відповідає ситуації, коли значення кількісної ознаки об'єктів генеральної сукупності не зустрілися серед варіант вибірки (порівняйте розподіл (2.4) з (1.2)).

За аналогією із числовими характеристиками вибірки наступні формули визначають **числові характеристики генеральної сукупності**:

$$\bar{x}_{\text{г}} = \left(\sum_{i=1}^{m} x_i N_i \right) \Big/ N, \qquad (2.5)$$

$$D_{\text{г}} = \left(\sum_{i=1}^{m} (x_i - \bar{x}_{\text{г}})^2 N_i \right) \Big/ N, \qquad (2.6)$$

$$\sigma_{\text{г}} = \sqrt{D_{\text{г}}}. \qquad (2.7)$$

Ці **числа** невідомі, і оцінки (наближення) їх дають такі рівності:

$$\bar{x}_{\text{г}} \approx \bar{x}_{\text{в}}, \qquad D_{\text{г}} \approx D_{\text{в}}, \qquad \sigma_{\text{г}} \approx \sigma_{\text{в}}. \qquad (2.8)$$

Наведені приклади дозволяють зробити деякі висновки. Ліві частини **наближених** рівностей (2.1), (2.2) та (2.8) є невідомими параметрами "відомого" закону розподілу або числовими характеристиками генеральної сукупності; вони є невідомими числами для конкретної генеральної сукупності. Праві частини цих рівностей є функціями випадкових величин, які для **фіксованої** вибірки статистичних даних набирають числові значення, що можна зобразити точками. Це дозволяє назвати їх **точковими статистичними оцінками** відповідних параметрів або числових характеристик генеральної сукупності.

Позначимо узагальнено символом Θ ліві частини наближених рівностей (2.1), (2.2), (2.8) (а також багатьох інших, що можна отримати для інших законів розподілу), а символом Θ^* — праві частини цих рівностей.

Точковою статистичною оцінкою, вибірковою функцією або **статистикою** числового параметра Θ називається функція вибіркових значень (варіант) $\Theta^* = \Theta^*(x_1, x_2, ..., x_n)$, яка в певному статистичному сенсі є близькою до справжнього значення цього параметра.

Незміщеною називається точкова статистична оцінка Θ^*, математичне сподівання якої дорівнює оцінюваному параметру Θ при довільному обсязі вибірки, тобто

$$M(\Theta^*) = \Theta. \qquad (2.9)$$

Зміщеною називається оцінка, для якої не виконується рівність (2.9).

Нехай для оцінювання параметра Θ можуть бути використані незміщені точкові оцінки $\Theta_1^*, \Theta_2^*, ..., \Theta_k^*$. Оцінка Θ_m^*, $1 \le m \le k$, називається **ефективною**, якщо при заданому обсязі n вибірки для неї виконується рівність

$$D(\Theta_m^*) = \min_{1 \le i \le k} D(\Theta_i^*).$$

Оцінка Θ^* називається **спроможною** оцінкою параметра Θ, якщо при $n \to \infty$ вона збігається по імовірності до Θ, тобто для як завгодно малого $\varepsilon > 0$ має місце граничний перехід

$$\lim_{n \to \infty} P\left(\left| \Theta - \Theta^* \right| < \varepsilon \right) = 1.$$

Оцінки середньої генеральної та генеральної частки для простої вибірки

Теорема 2.1. **Для повторної вибірки обсягом n середня вибіркова $\bar{x}_в$ є незміщеною і спроможною оцінкою невідомої середньої генеральної $\bar{x}_г$. Якщо n досить велике, тоді $\bar{x}_в$ з достатнім ступенем точності розподілена за нормальним законом з параметрами:**

$$a = \bar{x}_г, \qquad \sigma = \sqrt{D_г / n}. \qquad (2.10)$$

Теорема 2.2. **Для безповторної вибірки обсягом n середня вибіркова $\bar{x}_в$ є незміщеною оцінкою невідомої середньої генеральної $\bar{x}_г$. Для досить великих n $\bar{x}_в$ з достатнім ступенем точності розподілена за нормальним законом з параметрами:**

$$a = \bar{x}_г, \qquad \sigma = \sqrt{\frac{D_г}{n} \cdot \frac{N-n}{N-1}}, \qquad (2.11)$$

де N — обсяг генеральної сукупності.

Зауваження. Обсяг генеральної сукупності N, як правило, дуже великий. Тому заміна в знаменнику другої рівності (2.11) $N-1$ на N невідчутна для σ. В зв'язку із цим надалі будемо користуватися рівністю

$$\sigma = \sqrt{\frac{D_{\text{г}}}{n} \cdot \left(1 - \frac{n}{N}\right)}. \qquad (2.12)$$

Нехай досліджується якісна ознака об'єктів генеральної сукупності, число яких є скінченним. **Генеральною часткою** будемо називати відношення числа M об'єктів генеральної сукупності, що володіють ознакою α, до обсягу N генеральної сукупності:

$$p = M/N.$$

При такому дослідженні знову будемо виходити із положення про неможливість суцільної перевірки всієї генеральної сукупності. Тоді подія A – навмання відібраний об'єкт із генеральної сукупності має ознаку α є випадковою і $P(A)=p$.

Точковою статистичною оцінкою невідомого числа p є вибіркова частка w, отримана внаслідок дослідження об'єктів вибірки. Очевидно, що за означенням w є відносною частотою випадкової події A, тобто випадковою величиною для нефіксованої вибірки.

Наступні твердження з'ясовують властивості відносної частоти w як оцінки генеральної частки p, а також закон розподілу, якому вона підпорядковується.

Теорема 2.3. **Для повторної вибірки обсягом n середня вибіркова частка w є незміщеною і спроможною точковою оцінкою невідомої генеральної частки p. Якщо n є досить великим, тоді w з достатнім ступенем точності розподілена за нормальним законом з параметрами:**

$$a = p, \quad \sigma = \sqrt{pq/n}, \qquad (2.13)$$

де $q = 1 - p$.

Теорема 2.4. **Для безповторної вибірки обсягом n вибіркова частка w є незміщеною точковою оцінкою невідомої генеральної частки p. Якщо n є досить великим, тоді w з достатнім ступенем точності розподілена за нормальним законом з параметрами:**

$$a = p, \quad \sigma = \sqrt{\frac{pq}{n} \cdot \frac{N-n}{N-1}}, \qquad (2.14)$$

Зауваження. Оскільки обсяг генеральної сукупності N в більшості випадків дуже великий, то заміна в знаменнику другої рівності (2.14) $N-1$ на N невідчутна для σ. З огляду на це надалі будемо користуватися рівністю

$$\sigma = \sqrt{\frac{pq}{n}\left(1 - \frac{n}{N}\right)}. \qquad (2.15)$$

Випадкові величини $\bar{x}_в$ та w є незміщеними точковими статистичними оцінками невідомих $\bar{x}_г$ та p відповідно. Їх реалізація або можливі значення, знайдені на основі даних простої вибірки (повторної або безповторної), не співпадають із оцінюваними параметрами. І кожне таке неспівпадання або відхилення природно називати **помилкою репрезентативності оцінки**, зумовленою тим, що досліджується не вся генеральна сукупність, а лише її частина (вибіркова сукупність). Для статистики дуже важливою є інформація про середню величину таких помилок.

Середньою квадратичною помилкою (**СКП**) при оцінюванні невідомих середньої генеральної $\bar{x}_г$ та генеральної частки p називається середнє квадратичне відхилення середньої вибіркової $\bar{x}_в$ та вибіркової частки w відповідно.

З урахуванням результатів теорем 2.1-2.4 можна вказати наступні формули для визначення середніх квадратичних помилок:

СКП середньої вибіркової повторної вибірки (див. (2.10))

$$\bar{\sigma}_{\bar{x}} = \sqrt{D_г / n}; \qquad (2.16)$$

СКП середньої вибіркової безповторної вибірки (див. (2.12))

$$\bar{\sigma}'_{\bar{x}} = \sqrt{\frac{D_г}{n}\left(1 - \frac{n}{N}\right)}; \qquad (2.17)$$

СКП вибіркової частки повторної вибірки (див. (2.13))

$$\bar{\sigma}_w = \sqrt{pq / n}; \qquad (2.18)$$

СКП вибіркової частки безповторної вибірки (див. (2.15))

$$\bar{\sigma}'_w = \sqrt{\frac{pq}{n}\left(1 - \frac{n}{N}\right)}. \qquad (2.19)$$

На практиці користуватися формулами (2.16)-(2.19) неможливо, оскільки для цього необхідно знати або дисперсію генеральну, або генеральну частку (проаналізуйте, наскільки обтяжливою є ця умова, згадавши початкову умову задачі оцінювання). Цю принципову трудність можна усунути за рахунок заміни у формулах СКП дисперсії генеральної $D_г$ на дисперсію вибіркову $D_в$, а генеральної частки p — на вибіркову частку w. Проте при такій заміні потрібно попередньо впевнитися, чи не з'явиться ще додаткова систематична помилка за рахунок зміщеності оцінок $D_в$ в обох задачах оцінювання. Якби вони виявилися зміщеними, то при заміні слід було б ввести "виправлені" оцінки невідомих дисперсій генеральних. На шляху реалізації вказаного вище підходу корисними є такі твердження.

Теорема 2.5. Математичне сподівання дисперсії вибіркової в задачі про оцінювання невідомої середньої генеральної для повторної вибірки визначається рівністю

$$M(D_{\text{в}}) = \frac{n-1}{n} D_{\text{г}}, \tag{2.20}$$

а для безповторної —

$$M(D_{\text{в}}) = \frac{n-1}{n} \cdot \frac{N}{N-1} D_{\text{г}}, \tag{2.21}$$

де n та N відповідно обсяги вибіркової та генеральної сукупностей.

Теорема 2.6. Математичне сподівання дисперсії вибіркової в задачі про оцінювання невідомої генеральної частки для повторної вибірки визначається рівністю

$$M(D_{\text{в}}) = \frac{n-1}{n} pq, \tag{2.22}$$

а для безповторної —

$$M(D_{\text{в}}) = \frac{n-1}{n} \cdot \frac{N}{N-1} pq. \tag{2.23}$$

Зміст теорем 2.5 та 2.6 полягає в тому, що дисперсія вибіркова є зміщеною оцінкою дисперсії генеральної в задачах оцінювання невідомих $\bar{x}_{\text{г}}$ та p у випадку простої вибірки (повторної та безповторної). При цьому формули (2.20)-(2.23) вказують на заниження значень дисперсії генеральної за рахунок наявності в кожній із них множника $(n-1)/n$.

Проте цю зміщеність легко "виправити": достатньо дисперсію вибіркову помножити на дріб $n/(n-1)$.

Виправленою дисперсією називається числова характеристика S^2, яка визначається рівністю

$$S^2 = \frac{n}{n-1} D_{\text{в}}.$$

Для **немалих** вибірок (обсягом $n \geq 30$), замінивши у формулах (2.16)-(2.19) дисперсію генеральну $D_{\text{г}}$ вибірковою $D_{\text{в}}$, а генеральну частку p — вибірковою w, отримаємо використовувані на практиці формули для знаходження СКП:

СКП середньої вибіркової повторної вибірки

$$\bar{\sigma}_{\bar{x}} = \sqrt{D_{\text{в}}/n} = \sigma_{\text{в}}/\sqrt{n}; \tag{2.24}$$

СКП середньої вибіркової безповторної вибірки

$$\bar{\sigma}'_{\bar{x}} = \sqrt{\frac{D_{\text{в}}}{n}\left(1 - \frac{n}{N}\right)}; \tag{2.25}$$

СКП вибіркової частки повторної вибірки

$$\overline{\sigma}_w = \sqrt{\frac{w(1-w)}{n}}; \qquad (2.26)$$

СКП вибіркової частки безповторної вибірки

$$\overline{\sigma}'_w = \sqrt{\frac{w(1-w)}{n}\left(1-\frac{n}{N}\right)}. \qquad (2.27)$$

Інтервальні статистичні оцінки.

Точкова статистична оцінка Θ^* не співпадає (за виключанням рідкісних випадків) із справжнім значенням невідомого параметра Θ. Тому завжди виникає похибка при заміні невідомого параметра його оцінкою, тобто $\left|\Theta-\Theta^*\right|>0$. Величина похибки при цьому невідома, хоча потрібно знати, до яких помилок може призвести вказана вище заміна.

Інтервальною називається статистична оцінка, яка визначається **двома числами** — кінцями інтервалу. Перевагою інтервальних оцінок є те, що вони дозволяють встановити точність і надійність оцінок.

Число $\delta>0$, яке фігурує у нерівності

$$\left|\Theta-\Theta^*\right|<\delta, \qquad (2.28)$$

природно назвати **точністю** оцінки Θ^*. Проте говорити про виконання нерівності (2.28) можна тільки в імовірносному сенсі, бо Θ^* — випадкова величина. Тобто, для достатньо малого δ ця нерівність є випадковою подією.

Надійністю (довірчою імовірністю) оцінки Θ по Θ^* називається імовірність γ виконання нерівності (2.28):

$$P\left(\left|\Theta-\Theta^*\right|<\delta\right)=\gamma. \qquad (2.29)$$

Довірчим називається інтервал $(\Theta^*-\delta;\ \Theta^*+\delta)$, який із заданою надійністю γ покриває невідомий параметр Θ.

Довірчі інтервали для оцінок $\overline{x}_г$ та p для немалих вибірок

Найбільше відхилення середньої вибіркової (або вибіркової частки) від середньої генеральної (або генеральної частки), яке можливе для заданої довірчої імовірності γ, називається **граничною помилкою** Δ.

Гранична помилка знаходиться за формулою

$$\Delta = t\overline{\sigma}, \qquad (2.30)$$

де t — корінь рівняння $2\Phi(t)=\gamma$.

Підставивши в рівність (2.30) вирази СКП (2.24)-(2.27), отримаємо **придатні для практики формули граничної помилки**:

середньої вибіркової повторної вибірки

$$\Delta = t\sqrt{D_в/n}; \qquad (2.31)$$

середньої вибіркової безповторної вибірки

$$\Delta = t\sqrt{\frac{D_{\text{в}}}{n}\left(1-\frac{n}{N}\right)}; \tag{2.32}$$

частки вибіркової повторної вибірки

$$\Delta = t\sqrt{w(1-w)/n}; \tag{2.33}$$

частки вибіркової безповторної вибірки

$$\Delta = t\sqrt{\frac{w(1-w)}{n}\left(1-\frac{n}{N}\right)}. \tag{2.34}$$

В результаті $(\bar{x}_{\text{в}} - \Delta;\ \bar{x}_{\text{в}} + \Delta)$ є довірчим інтервалом, який з надійністю γ покриває невідому середню генеральну $\bar{x}_{\text{г}}$. Аналогічно $(w-\Delta;\ w+\Delta)$ — довірчий інтервал, який з тією ж надійністю покриває невідому генеральну частку p.

Зауваження. В деяких випадках може бути відомою числова характеристика $\sigma_{\text{г}}$ (наприклад, як інформація технологічного процесу). Тоді $D_{\text{в}}$ у формулах (2.32), (2.33) слід замінити на $\sigma_{\text{г}}^2$.

Знаходження мінімального обсягу вибірки

Перед утворенням вибіркової сукупності необхідно з'ясувати, яким повинен бути її обсяг.

Наступні формули визначають мінімальні обсяги вибірки при оцінюванні невідомих:

а) середньої генеральної

$$n = \frac{t^2 D_{\text{г}}}{\Delta^2} \tag{2.35}$$

для повторної вибірки,

$$n' = \frac{nN}{n+N} \tag{2.36}$$

для безповторної вибірки (n визначається (2.35));

б) генеральної частки

$$n = \frac{t^2 pq}{\Delta^2} \tag{2.37}$$

для повторної вибірки,

$$n' = \frac{nN}{n+N} \tag{2.38}$$

для безповторної вибірки (n визначається (2.37));

де N – обсяг генеральної сукупності.

Довірчі інтервали для оцінки $\bar{x}_г = a$ для малої вибірки

Нехай про досліджувану кількісну ознаку X відомо тільки те, що вона розподілена за нормальним законом. Ставиться задача: побудувати довірчий інтервал для оцінки невідомого параметра $a = M(X)$ (або $\bar{x}_г$ у випадку скінченності обсягу генеральної сукупності) за даними повторної вибірки малого обсягу n, заданою довірчою імовірністю γ і якщо невідомий параметр $\sigma = \sigma_г$.

У [8] доведено, що довірчий інтервал для невідомого параметра a ($\bar{x}_г$) при невідомому $\sigma(\sigma_г)$ з надійністю γ має такий вид:

$$x_в - t(\gamma, n-1)\frac{S}{\sqrt{n}} < a < x_в + t(\gamma, n-1)\frac{S}{\sqrt{n}}, \tag{2.39}$$

де

$$S^2 = \sum_{i=1}^{n}(x_i - x_в)^2 \Big/ (n-1), \tag{2.40}$$

параметр $t=t(\gamma, n\text{-}1)$ – корінь рівняння

$$P\left(\left|\frac{\bar{x}_в - a}{S/\sqrt{n}}\right| < t\right) = P\left(|T| < t\right) = 2\int_0^t g_k(t)dt = \gamma, \tag{2.41}$$

який можна знайти за табл.4 додатків в залежності від заданої довірчої імовірності γ і числа ступенів вільності $k = n - 1$; $g_k(t)$ – густина розподілу Ст'юдента.

Довірчі інтервали для $D_г$ та $\sigma_г$ у випадку малої вибірки

При знаходженні мінімального обсягу вибірки необхідною є інформація про дисперсію генеральну (середнє квадратичне відхилення генеральне). Часто в розпорядженні дослідника є тільки вибірка малого обсягу.

Можна довести (див. [8]), що довірчий інтервал для оцінки невідомої дисперсії генеральної $D_г=\sigma^2$ з надійністю γ має такий вид:

$$\frac{(n-1)S^2}{\chi_2^2} < \sigma^2 < \frac{(n-1)S^2}{\chi_1^2}, \tag{2.42}$$

де S визначається рівністю (2.40), значення χ_1^2 і χ_2^2 знаходяться за табл.6 додатків з використанням рівнянь

$$P\left(\chi^2(k) > \chi_1^2(p;k)\right) = p, \quad p = \frac{1+\gamma}{2}; \tag{2.43}$$

$$P\left(\chi^2(k) > \chi_2^2(p;k)\right) = p, \quad p = \frac{1-\gamma}{2}; \tag{2.44}$$

$k = n - 1$ – число ступенів вільності закону розподілу χ^2.

Із подвійної нерівності (2.42) отримаємо рівносильну подвійну нерівність

$$\frac{S\sqrt{n-1}}{\chi_2} < \sigma < \frac{S\sqrt{n-1}}{\chi_1}, \tag{2.45}$$

яка визначає довірчий інтервал для оцінки невідомої $\sigma(\sigma_г)$.

Зауваження. В табл. 6 додатків наведені значення $\chi^2(p;k)$, що задовольняють рівняння $P\left(\chi^2(k) > \chi^2(p;k)\right) = p$ тільки для $k = n - 1 \le 30$, а також для $k = 40, 50, 100$. Це зумовлено тим, що при зростанні k закон розподілу χ^2 наближається до нормального.

РОЗВ'ЯЗУВАННЯ ТИПОВИХ ЗАДАЧ

Задача 2.1. Результати випробувань на міцність стальних дротів однакової довжини наведені в табл. 2.1 інтервальним розподілом.

Таблиця 2.1

Розривне зусилля, (кг/мм^2)	[30; 32)	[32; 34)	[34; 36)	[36; 38)	[38; 40)	[40; 42)	[42; 44]
Кількість дротів	6	12	17	29	19	13	4

1) Знайти довірчий інтервал, який з надійністю $\gamma = 0{,}95$ покриває середнє розривне всієї партії 4000 дротів.
2) Знайти довірчу імовірність того, що середня вибіркова відхилиться від середньої генеральної за абсолютною величиною не більше, ніж на 0,5 кг/мм^2.

○ 1) Утворена (немала, бо $n = 100 > 30$) вибірка є безповторною, оскільки після перевірки об'єкт (дріт) не може бути повернутий в генеральну сукупність (дріт розривається внаслідок перевірки його міцності). Для знаходження меж довірчого інтервалу граничну помилку Δ знайдемо за формулою (2.32), попередньо знайшовши числові характеристики вибірки $\bar{x}_в$ та $D_в$ за допомогою методу добутків, використаного до дискретного розподілу, варіанти x_i якого є серединами частинних інтервалів.

Результати допоміжних обчислень помістимо в табл. 2.2, де $h = 2$, $C = 37$, $u_i = (x_0 - C)/h$.

Таблиця 2.2

x_i	n_i	u_i	$n_i u_i$	$n_i u_i^2$	$n_i (u_i+1)^2$
31	6	–3	–18	54	24
33	12	–2	–24	48	12
35	17	–1	–17	17	0
37	29	0	0	0	29
39	19	1	19	19	76
41	13	2	26	52	117
43	4	3	12	36	64
$\sum$	100		–2	226	322

Контроль:

$$n + 2\sum n_i u_i + \sum n_i u_i^2 = 100 - 4 + 226 = 322 = \sum n_i (u_i + 1)^2.$$

Тоді $M_1^* = \sum n_i u_i / n = -0{,}02$, $M_2^* = \sum n_i u_i^2 / n = 2{,}26$ і за формулами (1.25), (1.26)

$$\bar{x}_в = M_1^* h + C = -0{,}02 \cdot 2 + 37 = 36{,}96,$$

$$D_в = \left[M_2^* - (M_1^*)^2 \right] h^2 = \left[2{,}26 - (-0{,}02)^2 \right] 2^2 = 9{,}0384.$$

За таблицями значень функції Лапласа (табл. 3 додатків) знайдемо значення t з рівняння $\Phi(t) = \gamma/2 = 0{,}95/2 = 0{,}475$, $t = 1{,}96$. За формулою (2.32), де $N = 4000$,

$$\Delta = 1{,}96 \sqrt{\frac{9{,}0384}{100}\left(1 - \frac{100}{4000}\right)} \approx 0{,}5818,$$

тоді ліва межа довірчого інтервалу

$$\bar{x}_в - \Delta = 36{,}96 - 0{,}5818 = 36{,}3782,$$

права –

$$\bar{x}_в + \Delta = 36{,}96 + 0{,}5818 = 37{,}5418.$$

Остаточно шуканий довірчий інтервал має такий вид (36,3782; 37,5418).

2) Шуканою є імовірність $P\left(|\bar{x}_в - \bar{x}_г| \le 0{,}5\right)$, яку можна знайти за формулою

$$P\left(|\bar{x}_в - \bar{x}_г| < \varepsilon \right) = 2\Phi\left(\frac{\varepsilon}{\bar{\sigma}}\right),$$

оскільки $\bar{x}_в$ є нормально розподіленою випадковою величиною (згідно з теоремою 2.2), $M(\bar{x}_в) = \bar{x}_г$. За формулою (2.25)

$$\bar{\sigma} = \bar{\sigma}'_{\bar{x}} = \sqrt{\frac{D_в}{n}\left(1 - \frac{n}{N}\right)} = \sqrt{\frac{9{,}0384}{100}\left(1 - \frac{100}{4000}\right)} \approx 0{,}2968.$$

Тоді $\varepsilon/\overline{\sigma} = 0{,}5/0{,}2968 = 1{,}69$ і за табл. 3 додатків

$$P\left(\left|\overline{x}_{\text{в}} - \overline{x}_{\text{г}}\right| \le 0{,}5\right) = 2\Phi\left(0{,}5/0{,}2968\right) =$$
$$= 2\Phi\left(1{,}69\right) = 2 \cdot 0{,}45449 = 0{,}90898. \quad \bullet$$

Задача 2.2. Із партії 9000 однотипних деталей перевірено 400 деталей. Серед них виявилося 360 першосортних деталей. Знайти межі, в яких з імовірністю 0,9542 міститься частка деталей першого сорту всієї партії, якщо вибірка: а) повторна; б) безповторна.

○ За табл. 3 додатків знаходимо, що коренем рівняння $2\Phi(t) = 0{,}9542$ є $t = 2$. Частка вибіркова $w = 360/400 = 0{,}9$. Граничну помилку повторної вибірки знайдемо за формулою (2.33) при $t = 2$, $w = 0{,}9$ і $n = 400$:

$$\Delta = 2 \cdot \sqrt{\frac{0{,}9(1-0{,}9)}{400}} = 0{,}03.$$

Тоді для повторної вибірки довірчим є інтервал (0,9 – 0,03; 0,9 + 0,03) = (0,87; 0,93).

Для безповторної вибірки граничну помилку знайдемо за формулою (2.34) (при тих самих значеннях *t*, *w* та *n*):

$$\Delta = 2\sqrt{\frac{0{,}9 \cdot 0{,}1}{400}\left(1 - \frac{400}{9000}\right)} \approx 0{,}029.$$

В результаті отримаємо для безповторної вибірки такий довірчий інтервал (0,871; 0,929). ●

Задача 2.3. Партія виробів в кількості 10000 шт. перевірена на відповідність стандарту. Для цього відібрано 10% виробів, серед яких виявилось 94% стандартних. Знайти довірчу імовірність того, що відсоток таких виробів у всій партії відрізняється від відсотку їх у вибірці не більше, ніж на одиницю за абсолютною величиною, якщо вибірка: а) повторна; б) безповторна.

○ За умовою $w = 0{,}94$ (або 94%), $n = 1000$ (10% від 10000), $N = 10000$. СКП вибіркової частки у випадку повторної вибірки обчислимо за формулою (2.26)

$$\overline{\sigma}_w = \sqrt{\frac{w(1-w)}{n}} = \sqrt{\frac{0{,}94 \cdot 0{,}06}{1000}} \approx 0{,}0075,$$

а у випадку безповторної — за формулою (2.27)

$$\overline{\sigma}'_w = \sqrt{\frac{w(1-w)}{n}\left(1 - \frac{n}{N}\right)} = \sqrt{\frac{0{,}94 \cdot 0{,}06}{1000}\left(1 - \frac{1000}{10000}\right)} \approx 0{,}0071.$$

Довірчу імовірність знайдемо за формулою $P\left(\left|\Theta^* - \Theta\right| < \Delta\right) = 2\Phi\left(\frac{\Delta}{\overline{\sigma}}\right)$,

де $\Theta^* = w$, $\Theta = p$. Для повторної вибірки при $\Delta = 0{,}01$ (або 1%) і $\overline{\sigma}_w = 0{,}0075$ одержимо

$$P\left(\left|0{,}94 - p\right| \le 0{,}01\right) = 2\Phi\left(\frac{0{,}01}{0{,}0075}\right) = 2\Phi(1{,}333) \approx 0{,}8164,$$

а для безповторної вибірки за тією ж формулою при $\overline{\sigma}'_w = 0{,}0071$ знайдемо:

$$P\left(\left|0{,}94 - p\right| \le 0{,}01\right) = 2\Phi\left(\frac{0{,}01}{0{,}0071}\right) = 2\Phi(1{,}408) \approx 0{,}8414. \quad \bullet$$

Задача 2.4. Із партії однотипних деталей здійснено просту вибірку обсягом 100, при цьому з'ясовано, що 90 деталей виявилися першосортними. Знайти довірчий інтервал, в яких з імовірністю 0,9542 міститься частка деталей першого сорту всієї партії, користуючись "практичною" формулою граничної помилки (2.33).

○ За умовою задачі обсяг N генеральної сукупності невідомий, тому слід вважати, що вибірка є повторною. За умовою $n = 100$, $w = 90/100 = 0{,}9$, $\gamma = 0{,}9542$.

Знайдемо t із рівняння $\Phi(t) = \gamma/2 = 0{,}9542/2 = 0{,}4771$; за таблицями значень функції Лапласа $t = 2$.

Використання формули (2.33) дає:

$$\Delta = t\sqrt{\frac{w(1-w)}{n}} = 2\sqrt{\frac{0{,}9 \cdot 0{,}1}{100}} = 0{,}06,$$

звідки довірчий інтервал для цього випадку має такий вид: (0,9 – 0,06; 0,9 + 0,06) або остаточно (0,84; 0,96). ●

Задача 2.5. Визначити, якими повинні бути обсяги повторної і безповторної вибірок, щоб з імовірністю 0,95 частка деталей другого сорту в партії із 10000 деталей відрізнялась від частки у вибірці не більше, ніж на 0,03 за абсолютною величиною, якщо: а) про частку деталей другого сорту в усій партії деталей немає даних; б) деталей другого сорту не більше 5%.

○ Довірча імовірність 0,95 визначає значення параметра $t = 1{,}96$.

а) Якщо про частку деталей другого сорту в усій партії немає даних, тоді в якості добутку pq слід взяти його найбільше значення 0,25. Тому відповідно до формули (2.37) при $t = 1{,}96$, $\Delta = 0{,}03$ необхідний обсяг повторної вибірки

$$n = \frac{(1,96)^2 \cdot 0,25}{(0,03)^2} \approx 1068.$$

Якщо вибірка безповторна, тоді за формулою (2.38) при $n = 1068$, $N = 10000$ одержимо, що необхідний її обсяг

$$n' = \frac{1068 \cdot 10000}{1068 + 10000} \approx 965.$$

Зауваження. При визначенні n і n' заокруглення результатів до цілих чисел робимо з надлишком.

б) Необхідний обсяг повторної вибірки знайдемо за формулою (2.37) при попередніх значеннях t, Δ, а $p = 0,05$ (або 5%) і $q = 1 - p = 0,95$:

$$n = \frac{(1,96)^2 \cdot 0,05 \cdot 0,95}{(0,03)^3} \approx 203.$$

Якщо вибірка безповторна, то необхідний її обсяг знайдемо за формулою (2.38) при $n = 203$, $N = 10000$:

$$n' = \frac{203 \cdot 10000}{203 + 10000} \approx 199.$$

Висновок: обсяг повторної вибірки для випадку б) в 5,26 менший в порівнянні із випадком а). Це зумовлено тим, що у випадку б) число $pq = 0,0475$ і в 5,26 разів менше від максимального значення добутку pq. Тим самим визначається роль апріорної хоча б приблизної інформації про генеральну сукупність. ●

Задача 2.6. Знайти необхідні обсяги повторної і безповторної вибірок, щоб при визначенні середньої тривалості безперервної роботи блоків в партії із 6000 блоків з імовірністю 0,99 відхилення середньої генеральної від середньої вибіркової не перевищувало за абсолютною величиною 30 год. Середнє квадратичне відхилення генеральне вважати рівним 160 год.

○ Довірча імовірність 0,99 визначає $t = 2,58$ (за табл. 3 додатків з рівняння $\Phi(t) = 0,99/2 = 0,495$). За формулою (2.35) при $\Delta = 30$, $D_г = \sigma_г^2 = 160^2$ знайдемо необхідний обсяг повторної вибірки:

$$n = \frac{(2,58)^2 \cdot (160)^2}{(30)^2} \approx 189,34,$$

тобто вибірка повинна складатися з $n = 190$ електронних блоків.

Якщо вибірка безповторна, то обсяг її згідно формули (2.36) при $n = 190$, $N = 6000$ повинен складати

$$n' = \frac{190 \cdot 6000}{190 + 6000} \approx 185. \quad ●$$

Задача 2.7. Для визначення врожайності гречки зробили вибірку, до якої ввійшло вісім ділянок. Результати вибіркових спостережень за урожайністю (ц/га) наведені в таблиці:

Номер ділянки	1	2	3	4	5	6	7	8
Урожайність	18,2	15,1	16,9	17,8	19,1	15,4	20,5	16,3

Знайти довірчий інтервал, в якому з надійністю 0,95 перебуватиме середня врожайність $(\bar{x}_г)$ гречки всього поля.

○ Знайдемо точкові незміщені статистичні оцінки для $\bar{x}_г$ та $D_г$:

$$\bar{x}_в = \frac{\sum x_i}{n} = \frac{18{,}2+15{,}1+16{,}9+17{,}8+19{,}1+15{,}4+20{,}5+16{,}3}{8} = 17{,}4125;$$

$$\overline{x^2} = \sum x_i^2 / n = (331{,}24 + 228{,}01 + 285{,}61 + 316{,}84 + 364{,}81 + 237{,}16 + $$

$$+ 420{,}25 + 265{,}69)/8 = 306{,}20125;$$

$$D_в = \overline{x^2} - (\bar{x}_в)^2 = 306{,}21125 - 303{,}19515 = 3{,}0614;$$

$$S^2 = \frac{n}{n-1} D_в = \frac{8}{7} \cdot 3{,}0061 = 3{,}4355; \quad S = 1{,}8535.$$

Будемо вважати, що врожайність усього поля розподілена за нормальним законом. Тоді за даною надійністю $\gamma = 0{,}95$ та числом ступенів вільності $k = n - 1 = 8 - 1 = 7$, користуючись табл. 4 додатків, знайдемо значення $t(\gamma, n-1) = 2{,}365$. Обчислимо межі довірчого інтервалу (2.39):

$$\bar{x}_в - t(\gamma, n-1)\, S/\sqrt{n-1} = 17{,}4125 - 2{,}265 \cdot 1{,}8535/7 =$$

$$= 17{,}4125 - 0{,}6263 = 16{,}7862;$$

$$\bar{x}_в + t(\gamma, n-1)\, S/\sqrt{n-1} = 17{,}4125 + 2{,}265 \cdot 1{,}8535/7 =$$

$$= 17{,}4125 + 0{,}6263 = 18{,}0388;$$

Отже, довірчий інтервал для середньої врожайності гречки всього поля має такий вид:

$$16{,}7862 < \bar{x}_г < 18{,}0388. \quad ●$$

Задача 2.8. Кількісна ознака X об'єктів генеральної сукупності має нормальний закон розподілу. Із надійністю $\gamma = 0{,}98$ знайти довірчі інтервали для $D_г$ та $\sigma_г$, якщо відомо, що виправлена дисперсія вибіркова $S^2 = 4{,}6$, а обсяг вибірки $n = 26$.

○ Із рівнянь (2.43), (2.44) при $\alpha = 1 - \gamma = 1 - 0{,}98 = 0{,}02$ отримаємо:

$$P\left(\chi^2 > \chi_1^2\right) = \frac{1+\gamma}{2} = \frac{1+0{,}98}{2} = 0{,}99,$$

$$P\left(\chi^2 > \chi_2^2\right) = \frac{1-\gamma}{2} = \frac{1-0{,}98}{2} = 0{,}01.$$

Число ступенів вільності $k = n - 1 = 26 - 1 = 25$.

За табл. 6 додатків знаходимо

$$\chi_1^2(0{,}99; k = 25) = 11{,}52; \qquad \chi_2^2(0{,}01; k = 25) = 44{,}31.$$

Тоді згідно із (2.42) при $S^2 = 4{,}8$ отримаємо:

$$\frac{(n-1)S^2}{\chi_2^2} = \frac{(26-1)\cdot 4{,}6}{44{,}31} = 2{,}6; \qquad \frac{(n-1)S^2}{\chi_1^2} = \frac{25\cdot 4{,}6}{11{,}52} = 9{,}98.$$

Звідки шукані довірчі інтервали:

$$2{,}6 < D_\text{г} < 9{,}89;\ 1{,}61 < \sigma_\text{г} < 3{,}15. \quad \bullet$$

§ 3. СТАТИСТИЧНА ПЕРЕВІРКА СТАТИСТИЧНИХ ГІПОТЕЗ

Статистичні гіпотези та їх види. Статистичний критерій перевірки основної гіпотези. Параметричні статистичні гіпотези (Порівняння середньої вибіркової із гіпотетичною середньою генеральною нормальної сукупності. Порівняння виправленої дисперсії вибіркової з гіпотетичною дисперсією генеральною нормальної сукупності). Критерій узгодженості Пірсона $\left(\chi^2\right)$. Перевірка гіпотези про нормальний розподіл генеральної сукупності. Критерій узгодженості Колмогорова.

КОРОТКІ ТЕОРЕТИЧНІ ВІДОМОСТІ

Статистичною називається гіпотеза про вид невідомого розподілу випадкової величини (кількісної ознаки об'єктів генеральної сукупності) або про параметри відомого розподілу.

Поряд із висунутою гіпотезою розглядають і гіпотезу, яка суперечить їй. Тому надалі будемо припускати, що у нас є дві гіпотези: H_0 та H_1, які не перетинаються. Гіпотезу H_0 будемо називати **основною** або **нульовою**, а гіпотезу H_1 — **конкуруючою** або **альтернативною.**

Гіпотези розрізняються за числом припущень. **Простою** називається гіпотеза, яка містить тільки одне припущення. **Складною** називається гіпотеза, яка складається із скінченного або нескінченного числа простих гіпотез.

Висунута статистична гіпотеза може бути правильною або хибною. Для перевірки її правильності використовують статистичні дані і статистичні методи, тому перевірку називають **статистичною.**

Для перевірки основної гіпотези потрібно мати критерій правильності цієї гіпотези. Як вже зазначалося, в розпорядженні дослідника є тільки вибірка $X_1, X_2, \ldots, X_n$, де X_i — значення кількісної ознаки i-ого об'єкта вибірки $(i = \overline{1,n})$, або значення вибіркової частки ω у випадку вивчення якісної ознаки об'єктів генеральної сукупності.

Статистичним критерієм (або просто **критерієм**, чи **статистикою**) називається випадкова великина K, яка використовується для перевірки основної гіпотези і закон розподілу якої (точний або наближений) відомий. Для кожного конкретного випадку величина K спеціально підбирається і може позначатися різними літерами: U або Z, якщо вона нормально розподілена, F або v^2 — по закону Фішера-Снедокора, T — по закону Ст'юдента, χ^2 — по закону "хі-квадрат", K — по закону Колмогорова і т. д.

Можливі значення випадкової величини (критерію) K розбиваються на дві непорожні множини Q та $\overline{Q}$ $\left(Q \cap \overline{Q} = \varnothing \right)$ такі, що Q складається із значень критерію, при яких H_0 приймається, а $\overline{Q}$ — із тих значень критерію, при яких H_0 відхиляється (а отже, приймається H_1). Множину $\overline{Q}$ називають **критичною областю**, а множину Q — **областю прийняття гіпотези**, або **областю допустимих значень.**

Для конкретної вибірки обчислюється значення критерію як функції варіант. Отримане значення позначається $K_{сп}$ і називається **спостереженим значенням критерію.**

Сформулюємо **основний принцип статистичної перевірки статистичної гіпотези**: якщо спостережене значення критерію належить критичній області $\left(K_{сп} \in \overline{Q} \right)$, тоді гіпотезу H_0 відхиляють; якщо спостережене значення критерію належить області допустимих значень $\left(K_{сп} \in Q \right)$, тоді гіпотезу H_0 приймають.

Зауваження. Припустимо, що для конкретних гіпотез H_0 і H_1 множини Q і $\overline{Q}$ **визначені**. В ряді посібників по математичній статистиці під статистичним критерієм розуміється правило, яке реалізує основний принцип статистичної перевірки статистичної гіпотези.

Оскільки висновок про правильність гіпотези робиться за результатами скінченної вибірки, а вона може бути "невдалою", то завжди існує ризик прийняти хибне рішення. При цьому можуть бути допущені помилки двох родів.

Якщо буде відхилена гіпотеза H_0 (і прийнята H_1), в той час як насправді правильною є H_0, тоді це є **помилка першого роду**; її імовірність позначають α:

$$\alpha = P\left(K_{сп} \in \overline{Q} / H_0 \right) = P\left(H_1 / H_0 \right),$$

де $P\left(H_1 / H_0 \right)$ — імовірність того, що буде прийнята гіпотеза H_1, якщо насправді для генеральної сукупності правильною є гіпотеза H_0. Число α називають **рівнем значущості.**

Якщо буде прийнята гіпотеза H_0, в той час як насправді правильною є H_1, тоді буде допущена **помилка другого роду**, її імовірність позначають β:

$$\beta = P\left(K_{сп} \in Q / H_1 \right) = P\left(H_0 / H_1 \right).$$

В цьому параграфі розглянемо один із методів перевірки статистичних гіпотез, який передбачає виконання таких кроків:

1) формулювання основної і конкуруючої гіпотез;

2) вибір відповідного рівня значущості α;

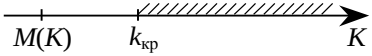
3) вибір статистичного критерія K перевірки гіпотези;

4) знаходження критичної області $\overline{Q}$ і області Q прийняття гіпотези;

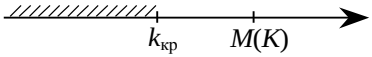
5) формулювання правила перевірки гіпотези: гіпотеза H_0 приймається при заданому рівні значущості α, якщо $K_{сп} \in Q$; гіпотеза H_0 відкидається, якщо $K_{сп} \in \overline{Q}$.

Розглянемо деякі особливості структури критичної області $\overline{Q}$. Якщо $\overline{Q}$ розташована зліва **і** справа від математичного сподівання випадкової величини K, то критична область називається **двосторонньою**, а критерій K — **двостороннім критерієм значущості**. Якщо ж $\overline{Q}$ розташована зліва **або** справа від математичного сподівання випадкової величини K, то критична область називається **односторонньою**, а критерій — **одностороннім**.

Реалізація основного принципу перевірки конкретної основної гіпотези H_0 передбачає знаходження множин Q та $\overline{Q}$. Оскільки ці множини числової осі не перетинаються, то існують точки, які їх розділяють. Такі точки називають **критичними точками** і позначаються $k_{кр}$. **Правосторонньою** називається критична область, яка визначається нерівністю $K > k_{кр}$, де $k_{кр} > M(K)$:



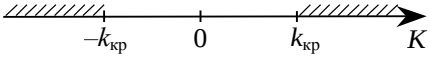
Лівосторонньою називається критична область, яка визначається нерівністю $K < k_{кр}$, де $k_{кр} < M(K)$:



Двосторонньою називається критична область, яка визначається сукупністю нерівностей

$$\left[\begin{array}{l} K < k_{кр}^{(1)}, \\ K > k_{кр}^{(2)}, \end{array}\right.$$

де $k_{кр}^{(1)} < M(K) < k_{кр}^{(2)}$. Зокрема, якщо критичні точки $k_{кр}^{(1)}$ та $k_{кр}^{(2)}$ відрізняються тільки знаком, тоді двостороння критична область визначається нерівністю $|K| > k_{кр}$:



Інтуїтивно зрозуміло, що множини Q та $\overline{Q}$ залежать від вибраного рівня значущості α, а тому слід очікувати залежності критичної точки $k_{кр}$ від α. Переконаємося в цьому, розглянувши підхід до використання основного принципу статистичної перевірки статистичної гіпотези у випадку правосторонньої критичної області. Аналіз решти випадків проводиться аналогічно. Для вибраного рівня значущості α $\overline{Q} = \{K > k_{кр}\}$. Критична точка $k_{кр}$ при умові правильності основної гіпотези H_0 задовольняє рівність

$$P(K > k_{кр}) = \alpha, \qquad (3.1)$$

яка трактується таким чином: при правильності H_0 малоімовірно (з врахуванням малості α), що спостережене значення критерію K виявиться більшим від $k_{кр}$.

Для кожного практично важливого критерію складено таблиці, за допомогою яких знаходиться $k_{кр}$, що задовольняє рівність (3.1). Наступний крок — обчислення спостереженого значення $K_{сп}$ критерію за даною вибіркою. А далі в силу вступає **правило**: якщо $K_{сп} > k_{кр}$, то гіпотезу H_0 відхиляють (вважається, що вона є хибною), якщо ж $K_{сп} < k_{кр}$, тоді H_0 приймають (вважається, що H_0 — правильна).

Якщо основна гіпотеза H_0 містить твердження про параметри розподілу, вид якого відомий, тоді статистичний критерій перевірки цієї гіпотези називається **параметричним**. Якщо ж у гіпотезі H_0 мова ведеться про невідомий розподіл, тоді відповідний критерій називається **критерієм узгодженості.**

Параметричні статистичні гіпотези

Розгляд статистичних гіпотез розпочнемо із параметричних гіпотез. Це зумовлено відносною простотою досліджень таких гіпотез в порівнянні із критеріями узгодженості.

Нижче будуть наведені тільки підсумки аналізу, який в повному обсязі можна знайти, зокрема, у посібнику [8].

Порівняння середньої вибіркової із гіпотетичною середньою генеральною нормальної сукупності.

Нехай кількісна ознака X генеральної сукупності розподілена нормально з параметрами a та σ. При цьому середня генеральна a хоча і невідома, але є підстави припускати, що вона дорівнює гіпотетичному значенню a_0. Для перевірки гіпотези $a = a_0$ потрібно знайти середню вибіркову $\bar{x}_в$ і з'ясувати, істотно чи ні $\bar{x}_в$ відхиляється від a_0.

Розв'язування окресленої задачі суттєво залежить від того, чи відомий параметр σ.

А. Дисперсія генеральної сукупності відома

Нехай із генеральної сукупності, кількісна ознака якої нормально розподілена, організована вибірка обсягом n і знайдена середня вибіркова $\bar{x}_в$, при цьому $D_г = \sigma^2$ — відома. Потрібно при заданому рівні значущості α перевірити основну гіпотезу про рівність середньої генеральної a гіпотетичному значенню a_0, тобто $H_0 : a = a_0$.

Оскільки середня вибіркова $\bar{X}$ (вибірка поки не фіксована) є незміщеною оцінкою середньої генеральної, тобто $M(\bar{X}) = a,$ то гіпотезу H_0 можна записати ще й таким чином: $M(\bar{X}) = a_0$.

Отже, потрібно перевірити, що математичне сподівання середньої вибіркової дорівнює гіпотетичній середній генеральній.

В якості критерія перевірки гіпотези H_0 використаємо випадкову величину:

$$U = (\bar{X} - a_0)/\sigma(\bar{X}) = (\bar{X} - a_0)\sqrt{n}/\sigma.$$

Ця величина розподілена нормально, причому якщо гіпотеза H_0 правильна, то $M(U) = 0,\ \sigma(U) = 1.$

Позначимо через $U_{спост}$ спостережене значення критерія для фіксованої вибірки:

$$U_{спост} = (\bar{x}_в - a_0)\sqrt{n}/\sigma. \qquad (3.2)$$

Критичну область побудуємо в залежності від виду конкуруючої гіпотези.

Правило 1. Гіпотеза $H_0 : a = a_0$ при конкуруючій гіпотезі $H_1 : a \neq a_0$ для рівня значущості α приймається, якщо $|U_{спост}| < u_{кр}$, де $u_{кр}$ — корінь рівняння

$$\Phi(u_{кр}) = (1-\alpha)/2\,. \qquad (3.3)$$

Якщо $|U_{спост}| > u_{кр},$ то гіпотеза H_0 відкидається.

Правило 2. При конкуруючій гіпотезі $H_1 : a > a_0$ критичну точку правосторонньої критичної області для рівня значущості α знаходять за рівністю (табл. 3 додатків):

$$\Phi(u_{кр}) = (1-2\alpha)/2. \qquad (3.4)$$

Якщо $U_{спост} < u_{кр},$ то основна гіпотеза H_0 приймається. Якщо ж $U_{спост} > u_{кр}$ — H_0 відкидається.

Правило 3. При конкуруючій гіпотезі $H_1 : a < a_0$ знаходять корінь $u_{кр}$ рівняння (3.4) (табл. 3 додатків).

Якщо $U_{\text{спост}} > -u_{\text{кр}}$, то нема підстав відкидати гіпотезу H_0. Якщо ж $U_{\text{спост}} < -u_{\text{кр}}$, то H_0 відкидається.

Б. Дисперсія генеральної сукупності невідома

В цьому випадку в якості критерія перевірки основної гіпотези береться випадкова величина

$$T = (\overline{X} - a_0)\sqrt{n}/S, \qquad (3.5)$$

де $\overline{X}$ та S — середня та "виправлене" середнє квадратичне відхилення для нефіксованої вибірки обсягу n.

Величина T розподілена за законом Ст'юдента із $k = n-1$ ступенями вільності.

Симетричність розподілу Ст'юдента дозволяє здійснювати побудову критичних областей (в залежності від виду конкуруючої гіпотези) аналогічно тому, як це робилося у випадку відомого параметра σ.

Правило 1*. Для того, щоб при заданому рівні значущості α перевірити основну гіпотезу $H_0 : a = a_0$ про рівність невідомої середньої генеральної a (нормально розподіленої генеральної сукупності з невідомою дисперсією) гіпотетичному значенню a_0 при конкуруючій гіпотезі $H_1 : a \neq a_0$, потрібно обчислити спостережене значення критерія

$$T_{\text{спост}} = (\overline{x}_в - a_0)\sqrt{n}/S$$

і за таблицею критичних точок розподілу Ст'юдента (табл. 5 додатків) по заданому рівню значущості α, розміщеному у верхньому рядку таблиці і числу ступенів вільності $k = n-1$ знайти критичну точку $t_{\text{двост.кр}}(\alpha, k)$.

Якщо $|T_{\text{спост}}| < t_{\text{двост.кр}}$, то нема підстав відкидати основну гіпотезу H_0.

Якщо ж $|T_{\text{спост}}| > t_{\text{двост.кр}}$, то гіпотеза H_0 відкидається на користь альтернативної гіпотези H_1.

Зауваження. Критична точка $t_{\text{двост.кр}} = t_\alpha$ є коренем рівняння

$$\int_0^{t_\alpha} g_k(t)dt = (1-\alpha)/2,$$

де $g_k(t)$ — густина розподілу Ст'юдента, $k = n-1$ — число ступенів вільності. Це рівняння є аналогом рівняння (3.3).

Правило 2*. При конкуруючій гіпотезі $H_1 : a > a_0$ по рівню значущості α, розміщеному в нижньому рядку табл. 5 додатків, і числу ступенів

вільності $k = n - 1$ знаходимо критичну точку $t_{\text{правост.кр}}(\alpha, k)$ правосторонньої критичної області.

Якщо $T_{\text{спост}} < t_{\text{правост.кр}}$, то нема підстав відкидати основну гіпотезу H_0.

Зауваження. Вибір критичної точки в табл. 5 додатків по нижньому рядку значень рівня значущості зумовлений аналогом рівності (3.4) для випадку правосторонньої критичної області, тобто $t_{\text{правост.кр}} = t_{2\alpha}$ є коренем рівняння

$$\int_0^{t_{2\alpha}} g_k(t) = (1 - 2\alpha)/2.$$

Правило 3*. При конкуруючій гіпотезі $H_1 : a < a_0$ спочатку знаходять "допоміжну" критичну точку $t_{\text{правост.кр}}(\alpha; k)$, а потім покладають межу лівосторонньої критичної області:

$$t_{\text{правост.кр}} = t_{\text{лівост.кр}},$$

Якщо $T_{\text{спост}} > -t_{\text{правост.кр}}$, то гіпотеза H_0 приймається.

Якщо $T_{\text{спост}} < -t_{\text{правост.кр}}$, то гіпотезу H_0 відкидають.

Порівняння виправленої дисперсії вибіркової з гіпотетичною дисперсією енеральною нормальної сукупності.

Нехай кількісна ознака генеральної сукупності розподілена за нормальним законом, причому дисперсія генеральна хоча і невідома, проте є підстави припускати, що вона дорівнює гіпотетичному значенню σ_0^2.

На підставі виправленої дисперсії вибіркової S^2, знайденої для вибірки обсягом n, при заданому рівні значущості α потрібно перевірити основну гіпотезу H_0, яка полягає в тому, що дисперсія генеральна дорівнює гіпотетичному значенню σ_0^2. Враховуючи незміщеність S^2 як оцінки дисперсії генеральної, основну гіпотезу можна записати таким чином:

$$H_0 : M(S^2) = \sigma_0^2. \qquad (3.6)$$

Іншими словами, потрібно з'ясувати, істотно чи неістотно відрізняються виправлена вибіркова і гіпотетична генеральна дисперсії.

В якості критерія перевірки гіпотези (3.6) візьмемо випадкову величину

$$\chi^2 = (n-1)S^2/\sigma_0^2, \qquad (3.7)$$

де права частина є випадковою величиною, розподіленою за законом χ^2 з $k = n - 1$ ступенями вільності.

Критична область будується в залежності від виду конкуруючої гіпотези.

Правило 1. Для того, щоб при заданому рівні значущості α перевірити основну гіпотезу $H_0 : \sigma^2 = \sigma_0^2$ при конкуруючій гіпотезі $H_1 : \sigma^2 > \sigma_0^2$, потрібно обчислити спостережене значення критерія $\chi^2_{\text{спост}}$ за формулою (3.7) і за табл. 6 додатків по заданому рівню значущості α і числу ступенів вільності $k = n - 1$ знайти критичну точку $\chi^2_{\text{кр}}(\alpha; k)$.

Якщо $\chi^2_{\text{спост}} < \chi^2_{\text{кр}}$, то нема підстав відкидати основну гіпотезу. Якщо $\chi^2_{\text{спост}} > \chi^2_{\text{кр}}$, то гіпотеза H_0 відкидається на користь гіпотези H_1.

Правило 2. Для того, щоб при заданому рівні значущості α перевірити гіпотезу $H_0 : \sigma^2 = \sigma_0^2$ при конкуруючій гіпотезі $H_1 : \sigma^2 \neq \sigma_0^2$, потрібно обчислити спостережене значення критерія $\chi^2_{\text{спост}}$ за формулою (3.7) і за табл. 6 додатків знайти ліву критичну точку $\chi^2_{\text{кр}}(1-\alpha/2; k)$ і праву критичну точку $\chi^2_{\text{кр}}(\alpha/2; k)$.

Якщо $\chi^2_{\text{лівост.кр}} < \chi^2_{\text{спост}} < \chi^2_{\text{правост.кр}}$, то нема підстав відкидати гіпотезу H_0.

Якщо $\chi^2_{\text{спост}} < \chi^2_{\text{лівост.кр}}$ або $\chi^2_{\text{спост}} > \chi^2_{\text{правост.кр}}$, то основна гіпотеза H_0 відкидається.

Правило 3. При конкуруючій гіпотезі $H_1 : \sigma^2 < \sigma_0^2$ знаходиться критична точка $\chi^2_{\text{кр}}(1-\alpha/2; k)$ лівосторонньої критичної області.

Якщо $\chi^2_{\text{спост}} > \chi^2_{\text{кр}}(1-\alpha/2; k)$, то нема підстав відкидати основну гіпотезу.

Якщо $\chi^2_{\text{спост}} < \chi^2_{\text{кр}}(1-\alpha/2; k)$, то гіпотеза H_0 відкидається.

Критерій узгодженості Пірсона (критерій χ^2).

Одна із найважливіших задач математичної статистики — знаходження невідомого закону розподілу випадкової величини X — кількісної ознаки об'єктів генеральної сукупності. В § 1 були висвітлені питання опрацювання вибірки з метою отримання інформації про вид емпіричного розподілу та його характеристик: середньої ознаки, величини розкиду, симетричності розподілу. Потім на основі цих даних підбирається той розподіл, який найкраще апроксимує дослідний розподіл випадкової величини.

Після вибору виду розподілу необхідно знайти (хоча б наближено) параметри того закону розподілу, який характеризує досліджувану випадкову величину.

Оскільки в більшості практично важливих випадків параметри теоретичного закону розподілу є або математичним сподіванням, або дисперсією випадкової великими, або виражаються через них (тобто через моменти перших порядків (§§4,5 Ч. 1)), то отримані вище висновки дають можливість знайти ці параметри за дослідними даними. Наприклад, для нормального розподілу параметри a та σ визначається рівностями $a = \bar{x}_в$, $\sigma = \sigma_в$ (або $\sigma = S$). Таким чином, можна говорити про "відомість" теоретичного розподілу досліджуваної ознаки X. При цьому лапки вказують на гіпотетичність такого знання.

Критерій узгодженості Пірсона $\left(\chi^2 \right)$ ґрунтується на виборі певної міри розбіжності між теоретичним і емпіричним (дослідним) розподілами. При цьому задачу перевірки узгодженості можна сформулювати таким чином: на основі вибірки спостережених значень деякої випадкової величини X потрібно визначити, що емпіричний розподіл належить певному розподілу (нормальному, показниковому, біноміальному і т. д.) із *визначеними* параметрами — гіпотеза H_0 проти альтернативної гіпотези H_1 — розподіл не належить вибраному розподілу.

Нехай у відповідності із гіпотезою H_0 **відома функція розподілу імовірностей $F(x)$ досліджуваної ознаки X**. Позначимо через $S_1, S_2, \ldots, S_m$ множини, на які розбивається вся область можливих значень випадкової величини X; ці множини — або інтервали для **неперервної** випадкової величини, або групи окремих значень **дискретної** випадкової величини, які не мають спільних точок. Тоді можна обчислити імовірності того, що при випробуванні випадкова величина X набере значення із множини S_i, тобто

$$p_i = P(X \in S_i), \quad i = \overline{1, m}. \qquad (3.8)$$

При цьому всі $p_i > 0, \ i = \overline{1, m}$, і

$$\sum_{i=1}^{m} p_i = 1.$$

Нехай $n_1, n_2, \ldots, n_m$ — відповідні емпіричні групові частоти, тобто суми частот тих варіант (значень випадкової величини X із вибірки), що потрапляють відповідно у множини $S_1, S_2, \ldots, S_m$.

Якщо гіпотеза H_0 правильна, то статистичний розподіл вибірки можна розглядати як емпіричний аналог для генерального розподілу, який визначається функцією $F(x)$. Це означає, що n_i є частотою (абсолютною) появи випадкової події $(X \in S_i), \ i = \overline{1, m}$, в послідовності із n спостережень. Отже, в першому розподілі кожній множині S_i ставляться у відповідність відносна частота n_i / n,

а в другому — імовірність p_i. Тоді за методом найменших квадратів в якості міри розходження між статистичним розподілом відносних частот вибірки і теоретичним розподілом імовірностей отримується міра розбіжності виду

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i} \qquad (3.9)$$

така, що при збільшенні обсягу вибірки розподіл величини χ^2 наближається до граничного розподілу χ^2 із $k = m - r - 1$ ступенями вільності, де m — число інтервалів або груп, на які розбита вся множина спостережених даних, r — число параметрів гіпотетичного розподілу імовірностей, що оцінюється за даними вибірки.

Зауваження. Числа

$$n_i^0 = np_i, \quad i = \overline{1, m}, \qquad (3.10)$$

називаються **теоретичними частотами** відбуття випадкових подій $(X \in S_i)$. Враховуючи рівності (3.10), із (3.9) отримаємо таку міру розбіжності між теоретичними і емпіричними частотами:

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - n_i^0)^2}{n_i^0}. \qquad (3.9^*)$$

Випадкова величина, що визначається рівностями (3.9) або (3.9*), називається **критерієм Пірсона**.

Для перевірки гіпотези H_0 задамо рівень значущості α і за табл. 6 додатків знайдемо критичну точку $\chi^2_{\text{кр}}(\alpha; k)$, де $p = \alpha$, $k = m - r - 1$. Таким чином, правостороння критична область визначається нерівністю $\chi^2 > \chi^2_{\text{кр}}(\alpha; k)$, а область прийняття нульової гіпотези — нерівністю $\chi^2 < \chi^2_{\text{кр}}(\alpha; k)$. За результатами вибірки обчислюємо $\chi^2_{\text{сп}}$. Якщо $\chi^2_{\text{сп}} < \chi^2_{\text{кр}}$, то гіпотезу H_0 приймаємо, а у випадку $\chi^2_{\text{сп}} > \chi^2_{\text{кр}}$ H_0 відкидаємо.

Перевірка гіпотези про нормальний розподіл генеральної сукупності

Використання критерія узгодженості Пірсона проілюструємо для випадку перевірки гіпотези H_0: кількісна ознака X об'єктів генеральної сукупності розподілена за нормальним законом.

Нехай статистичний розподіл має такий вид:

$[x_i; x_{i+1})$	$[x_1; x_2)$	$[x_2; x_3)$	...	$[x_m; x_{m+1}]$
n_i	n_1	n_2	...	n_m

(3.11)

де $x_{i+1} - x_i = h,\quad i = \overline{1,m},\quad \sum_{i=1}^{k} n_i = n,$ число частинних інтервалів визначається наближеною рівністю $m \approx \log_2 n.$ Якщо ж статистичний розподіл є дискретним, тоді потрібно перейти до інтервального (див. розв'язування задачі 1.2). Використовуючи метод добутків, знайдемо для розподілу (3.11) $\bar{x}_в$ та $D_в$. Тоді покладемо, що густина розподілу досліджуваної ознаки X має такий вид:

$$f(x) = \frac{1}{\sigma_в \sqrt{2\pi}} e^{-\frac{(x-\bar{x}_в)^2}{2\sigma_в^2}}.$$

Враховуючи те, що можливі значення нормально розподіленої випадкової величини заповнюють всю дійсну вісь, в якості множин $S_1, S_2, .. S_m$ візьмемо інтервали

$$(-\infty; x_2), (x_2; x_3), \ldots (x_m; \infty).$$

Для знаходження імовірностей $p_i = P(x_i < X < x_{i+1}),\quad i = \overline{1,m},$ де $x_1 = -\infty,$ $x_{m+1} = \infty,$ використаємо формулу

$$P(\alpha < X < \beta) = \Phi\left(\frac{\beta - a}{\sigma}\right) - \Phi\left(\frac{\alpha - a}{\sigma}\right),$$

в якій a та σ — параметри нормального розподілу.
Отже, за табл. 3 додатків обчислимо

$$p_i = \Phi\left(\frac{x_{i+1} - \bar{x}_в}{\sigma_в}\right) - \Phi\left(\frac{x_i - \bar{x}_в}{\sigma_в}\right),\quad i = \overline{1,m},$$

поклавши $x_1 = -\infty,\ x_{m+1} = \infty.$

В результаті за формулою (3.9) можна знайти для даної вибірки (розподілу (3.11)) $\chi^2_{\text{сп}}$. Для заданого рівня значущості α і ступеня вільності $k = m - 3$ (число параметрів нормального розподілу $r = 2$) за табл. 6 додатків знаходимо $\chi^2_{\text{кр}}(\alpha; k).$ Тоді гіпотеза H_0 приймається, якщо $\chi^2_{\text{сп}} < \chi^2_{\text{кр}}(\alpha; k),$ і відхиляється у випадку $\chi^2_{\text{сп}} > \chi^2_{\text{кр}}(\alpha; k).$

Зауваження. Вказане вище число ступенів вільності $k = m - 3$ відноситься тільки до того випадку, коли обидва параметри нормального закону розподілу знаходяться за даними вибірки, тобто коли замість точних значень a і σ використовуються їх емпіричні значення $\bar{x}_в$ та $\sigma_в$. Якщо значення a точно відоме (наприклад, у випадку знаходження відхилень від еталону), то число ступенів вільності дорівнює $k = m - 2.$ Якщо ж відомі обидва параметри, то число ступенів вільності $k = m - 1.$ На практиці така ситуація зустрічається рідко, а тому для отримання числа ступенів вільності не менше п'яти потрібно, щоб число частинних інтервалів було не меншим восьми.

Критерій узгодженості Колмогорова

Критерій Пірсона можна використовувати як для випадку неперервної ознаки X, так і для дискретної. Проте недоліком його є те, що випадкова величина χ^2, як правило, залежить від групування варіаційного ряду вибірки. Тому іноді доцільним є використання інших критеріїв узгодженості.

Нехай $F(x)$ — теоретична функція розподілу імовірностей **неперервної** випадкової величини X, що є кількісною ознакою об'єктів генеральної сукупності. При цьому згідно із гіпотезою H_0 вид цієї функції відомий. Нехай $F^*(x)$ — емпірична функція розподілу, отримана внаслідок опрацювання вибірки обсягом n. Згідно із теоремою Гливенко-Кантеллі $F^*(x)$ є спроможною оцінкою функції $F(x)$. Тому можна порівнювати емпіричну функцію розподілу $F^*(x)$ із гіпотетичною $F(x)$ і якщо міра розходженості між ними мала, то вважати правильною гіпотезу H_0. Найбільш природною і простою із таких мір є рівномірна віддаль

$$D_n = \max_{-\infty<x<\infty} \left| F^*(x) - F(x) \right|. \qquad (3.12)$$

Проте при побудові критерія Колмогорова більш зручно користуватись нормованою віддалю $D_n\sqrt{n}$.

Оскільки вибірка утворюється випадковим чином, то D_n є випадковою величиною, при цьому величина $D_n\sqrt{n}$ має граничний при $n\to\infty$ розподіл, обчислений в припущенні, що гіпотеза H_0 є правильною. При умові неперервності $F^*(x)$ цей розподіл має такий вид

$$P\left(D_n\sqrt{n} < x \right) \underset{n\to\infty}{\to} K(x) = \begin{cases} 0 & \text{при } x \le 0, \\ \sum\limits_{k=-\infty}^{\infty} (-1)^k e^{-2k^2x^2} & \text{при } x > 0. \end{cases} \qquad (3.13)$$

Функція розподілу імовірностей $K(x)$ називається **розподілом Колмогорова**. Суттєвим є те, що **вона не залежить від конкретного розподілу** $F(x)$.

Критерій Колмогорова приписує прийняти гіпотезу H_0, якщо $D_n < K_{n;\alpha}$ і відхиляти у випадку виконання нерівності $D(x) \ge K_{n;\alpha}$, де α — рівень значущості, $K_{n;\alpha}$ — критичне значення критерію, яке знаходиться за табл. 8 додатків. Відмітимо, що в деяких посібниках і довідниках наводяться таблиці критичних значень для статистики $D_n\sqrt{n}$.

Описання табл. 8 додатків

В табл. 8 наведені критичні значення $K_{n;\alpha}$ для $n=\overline{1,100}$; $\alpha=0{,}1$; 0,05; 0,02; 0,01.

Приклад для зчитування: при $n=26$, $\alpha=0{,}01$ знаходимо значення $K_{26;0{,}01}=0{,}311$. Для обсягів вибірки $n>100$ з урахуванням співвідношення (3.13) випливає асимптотичне співвідношення

$$K_{n;\alpha}\approx k_{1-\alpha}\big/\sqrt{n}, \qquad (3.14)$$

де $K(k_{1-\alpha})\approx 1-\alpha$. Значення $k_{1-\alpha}$ знаходяться за табл. 9 додатків. Наприклад, якщо $\alpha=0{,}05$, тоді $k_{1-0{,}05}=1{,}358$ і за формулою (3.14) отримаємо $K_{n;0{,}05}\approx 1{,}358\big/\sqrt{n}$. При $n=100$ наближене критичне значення дорівнює 0,1358, а відповідне точне значення за табл. 8. — 0,134.

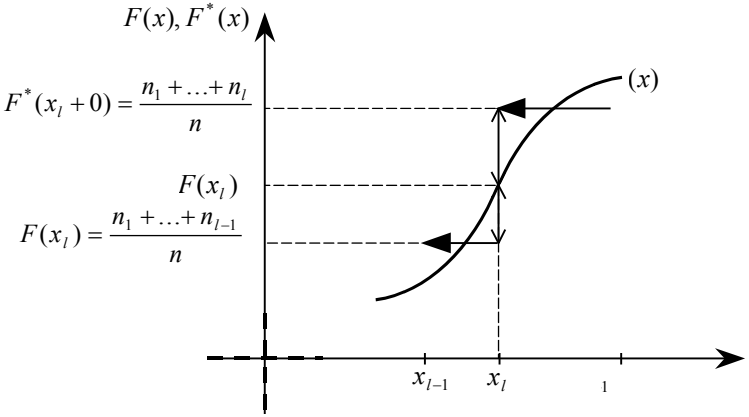
Розглянемо практичну реалізацію критерія Колмогорова. Нехай статистичний розподіл вибірки задається таблицею

x_i	x_1	x_2	…	x_k
n_i	n_1	n_2	…	n_k

де $\sum_{i=1}^{k} n_i = n$, $x_1 < x_2 < \ldots < x_k$. Тоді емпірична функція розподілу $F^*(x)$ має такий вид:

$$F^*(x)=\begin{cases} 0, & \text{якщо } x\le x_1,\\ n_1/n, & \text{якщо } x_1<x\le x_2,\\ (n_1+n_2)/n, & \text{якщо } x_2<x\le x_3,\\ \ldots & \ldots,\\ (n_1+n_2+\ldots+n_{k-1})/n & \text{якщо } x_{k-1}<x\le x_k,\\ 1, & \text{якщо } x>x_k. \end{cases} \qquad (3.15)$$

При знаходженні значення статистики D_n, що визначається за формулою (3.12), потрібно враховувати сходинковий характер графіка функції $F^*(x)$. Найбільша за абсолютною величиною різниця між $F^*(x)$ і $F(x)$ буде досягатися в одній із точок розподілу (мал. 3.1).



Мал. 3.1.

Тобто

$$D_n = \max\left\{ D_n^{(1)}; D_n^{(2)} \right\}, \qquad (3.16)$$

де

$$D_n^{(1)} = \max_{l=1,k}\left|\frac{n_1+n_2+\ldots+n_l}{n} - F(x_l)\right| = \max_{l=1,k}\left|F^*(x_l+0) - F(x_l)\right|,$$
$$D_n^{(2)} = \max_{l=1,k}\left|F(x_l) - \frac{n_1+n_2+\ldots+n_{l-1}}{n}\right| = \max_{l=1,k}\left|F(x_l) - F^*(x_l)\right|. \qquad (3.17)$$

Відмітимо, що другий алгебраїчний доданок в другому співвідношенні (3.17) для $l=1$ дорівнює 0, бо із (3.15) $F^*(x_1)=0,$ а перший доданок у перщому співвідношенні (3.17) при $l=k$ дорівнює 1.

Якщо у статистичному розподілі вибірки всі частоти дорівнюють одиниці, тоді D_n знаходиться з такого співвідношення

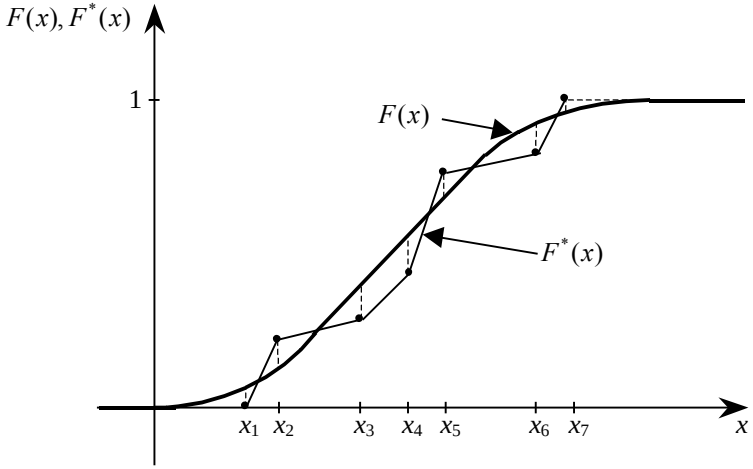
$$D_n = \max_{1\le l\le n}\left|F(x_l) - \frac{2l-1}{2n}\right| + \frac{1}{2n}. \qquad (3.18)$$

Формули (3.16), (3.17) або (3.18) дають практичну реалізацію **двостороннього** критерія Колмогорова для випадку **дискретного розподілу** частот або відносних частот вибірки. Розглянемо інтервальний статистичний розподіл вибірки

$[x_i; x_{i+1})$	$[x_1; x_2)$	$[x_2; x_3)$	…	$[x_k; x_{k+1})$
n_i	n_1	n_2	…	n_k

.

Ескізи графіків емпіричної та теоретичної функцій розподілу в цьому випадку (мал. 3.2) вказують на суттєву відмінність при знаходженні відхилень $\left|F^*(x)-F(x)\right|$ в порівнянні із дискретним розподілом (мал. 3.1).



Мал. 3.2.

В цьому випадку слід вибирати односторонню критичну область. Тоді статистика критерію Колмогорова задається формулою

$$D_n^+ = \max_{1\le i\le k+1}\left|F^*(x_i)-F(x_i)\right|. \qquad (3.19)$$

У випадку використання одностороннього критерія на рівні значущості α гіпотеза H_0 відкидається, якщо $D_n^+ > K_{n;\alpha}^+$, де $K_{n;\alpha}^+$ — критичне значення. Оскільки $K_{n;\alpha}^+ \approx K_{n;2\alpha}$, то $K_{n;\alpha}^+$ можна знаходити за табл. 8 для $n=\overline{1,100}$ і $\alpha =$ 0,05; 0,025; 0,01; 0,005.

При $n\to\infty$ має місце асимптотичне співвідношення ([17, с. 238])

$$K_{n;\alpha}^+ \approx \sqrt{\frac{-\ln\alpha}{2n}}. \qquad (3.20)$$

Наприклад, для $n=100$ і $\alpha=0{,}05$ приблизне значення, отримане з допомогою цього співвідношення, дорівнює 0,1224, а відповідне точне значення — 0,121 ($K_{100;0,05}^+ = K_{100;0,1}$).

РОЗВ'ЯЗУВАННЯ ТИПОВИХ ЗАДАЧ

Задача 3.1. Торгівельна фірма розглядає питання про відкриття в новому мікрорайоні міста філії. Відомо, що фірма буде працювати

прибутково, якщо щомісячний середній дохід мешканців мікрорайону перевищує 500 у. о. Відомо також, що дисперсія доходів $\sigma^2 = 400$ у. о.

Знайти умови прийняття рішення, з допомогою якого на підставі вибірки обсягом $n = 100$ і рівня значущості $\alpha = 0{,}05$ можна встановити, що робота філії буде прибутковою.

○ Вважаємо, що середній місячний дохід навмання вибраного мешканця є нормально розподіленою випадковою величиною.

Фірма не відкриє філію, якщо середній місячний дохід мешканців не перевищить 500 у. о. Тому будемо вважати, що $H_0 : a_0 = 500$, а $H_1 : a > 500$. Оскільки дисперсія відома, то згідно із правилом 2 гіпотеза H_1 приймається, якщо $U_{\text{спост}} > u_{\text{кр}}$, де $u_{\text{кр}}$ — корінь рівняння (3.4), тобто $\Phi(u_{\text{кр}}) = (1 - 2 \cdot 0{,}05)/2 = 0{,}45$. Звідки за табл. 3 додатків $u_{\text{кр}} = 1{,}65$. Із врахуванням (3.2) і умов задачі

$$U_{\text{спост}} = (\bar{x}_в - a_0)\sqrt{n}/\sigma = (\bar{x}_в - 500)\sqrt{100}/2.$$

Тому H_1 приймається і, отже, філію відкривають, якщо середній місячний дохід 100 мешканців

$$\bar{x}_в > 500 + 2 \cdot 1{,}65 = 503{,}3. \quad \bullet$$

Задача 3.2. Для уточнення норм виробки на підприємстві проведено 31 незалежне вимірювання продуктивності праці робітників, виконуючих однотипну операцію. Середня продуктивність праці склала $\bar{x}_в = 7{,}2$ (одиниць продукції за 1 год), а середнє квадратичне відхилення $\sigma_в = 0{,}5$ (одиниць продукції за год). Потрібно перевірити гіпотезу, що при масовому випуску цієї продукції середня продуктивність a_0 складе 7,5 одиниць продукції за 1 год при конкуруючій гіпотезі $a < 7{,}5$ одиниць продукції за 1 год при рівні значущості $\alpha = 0{,}01$.

○ Будемо вважати, що продуктивність праці навмання взятого робітника є нормально розподіленою випадковою величиною. Згідно із умовою задачі дисперсія цієї величини невідома, тому в якості статистичного критерія візьмемо випадкову величину (3.5), спостережене значення якої

$$T_{\text{спост}} = (\bar{x}_в - a_0)\sqrt{n}/S = (7{,}2 - 7{,}5)\sqrt{31}\Big/\left(0{,}5\sqrt{\frac{31}{31-1}}\right) = -3{,}286.$$

При обчисленні використовуємо такий зв'язок між $\sigma_в$ та S:

$$S = \sigma_в \sqrt{n/(n-1)}.$$

Оскільки $H_1 : a < a_0$, то потрібно побудувати лівосторонню критичну область. Згідно із правилом 3* знайдемо "допоміжну" критичну точку $t_{\text{правост.кр}}(0{,}01;30)$ по рівню значущості 0,01, розміщеному в нижньому рядку табл. 5 додатків: $t_{\text{правост.кр}}(0{,}01;30) = 2{,}457.$

Тоді $t_{\text{лівост.кр}} = -2{,}457.$

Так як $T_{\text{спост}} = -3{,}286 < -2{,}457 = t_{\text{лівост.кр}}$, то основну гіпотезу відкидаємо на користь альтернативної гіпотези $H_1 : a < 7{,}5.$ ●

Задача 3.3. Кількісна ознака генеральної сукупності розподілена за нормальним законом. За вибіркою обсягом $n = 16$ знайдена дисперсія вибіркова $D_в = 10{,}6.$ Для рівня значущості $\alpha = 0{,}02$ перевірити основну гіпотезу $H_0 : \sigma^2 = \sigma_0^2 = 12,$ якщо конкуруюча гіпотеза $H_0 : \sigma^2 \neq 12.$

○ Знайдемо спочатку виправлену дисперсію

$$S^2 = \left[\, n/(n-1)\right] D_в = (16/15)10{,}6 = 11{,}307,$$

а потім спостережене значення критерія

$$\chi^2_{\text{спост}} = (n-1)\, S^2 / \sigma_0^2 = 15 \cdot 11{,}307/12 = 14{,}133.$$

Оскільки $H_1 : \sigma^2 \neq 12,$ то критична область є двосторонньою. Згідно із правилом 2 за табл. 6 додатків знаходимо критичні точки: ліву — $\chi^2_{\text{кр}}(1-\alpha/2;k) = \chi^2_{\text{кр}}(1-0{,}02/2;15) = \chi^2_{\text{кр}}(0{,}99;15) = 5{,}23$ і праву — $\chi^2_{\text{кр}}(\alpha/2;k) = \chi^2_{\text{кр}}(0{,}01;15) = 30{,}58.$ Так як спостережене значення критерія належить області прийняття основної гіпотези (5,23 < 14,133 < 30,58), то нема підстав її відкидати. Іншими словами, виправлена дисперсія вибіркова (11,307) неістотно відрізняється від гіпотетичної дисперсії генеральної (12). ●

Задача 3.4. За даними табл. 1.2 про контрольні виміри товщини (в міліметрах) 200 вкладишів шатунних підшипників з допомогою критерія Пірсона перевірити гіпотезу H_0 про нормальний розподіл кількісної ознаки генеральної сукупності, якщо рівень значущості дорівнює 0,05.

○ Знайдемо число m частинних інтервалів однакової довжини, які отримуються при переході від дискретного ряду варіант з табл. 1.2 до інтервального розподілу частот, використовуючи формулу $m \approx \log_2 n.$ Оскільки обсяг

вибірки $n = 200$, то $m \approx \log_2 200 \approx 8$. Шуканий інтервальний розподіл наведений в табл. 1.3.

Для знаходження $\bar{x}_в$ та $\sigma_в$ від цього інтервального розподілу потрібно перейти до дискретного, "нові" варіанти якого є серединами частинних інтервалів. Для отриманого розподілу в задачі 1.2 методом добутків знайдено: $\bar{x}_в = 1{,}7417;\ \sigma_в = 0{,}0113269$.

Результати допоміжних розрахунків в процесі знаходження теоретичних частот $n_i^0 = p_i n$ розташуємо в табл. 3.1, де $x_1 = -\infty,\ x_9 = \infty$. Відмітимо, що згідно із властивостями функції Лапласа $\Phi(-x) = -\Phi(x),\ \Phi(-\infty) = -0{,}5,$ $\Phi(\infty) = 0{,}5$. Крім цього, при знаходженні значень $\Phi(x)$ за допомогою табл. 3 додатків для аргументів, що мають третій знак після коми, здійснимо лінійну інтерполяцію. Наприклад, при знаходженні $\Phi(1{,}916)$ використаємо табличні значення: $\Phi(1{,}91) = 0{,}47193,\ \Phi(1{,}92) = 0{,}47257$. Тоді

$$\Phi(1{,}916) = \Phi(1{,}91) + [\Phi(1{,}92) - \Phi(1{,}91)] \cdot 0{,}6 =$$
$$= 0{,}47193 + (0{,}47257 - 0{,}47193) \cdot 0{,}6 = 0{,}472314 \approx 0{,}47231.$$

Для знаходження $\chi^2_{\text{сп}}$ складемо розрахункову табл. 3.2, з якої отримаємо

$$\chi^2_{\text{сп}} = 2{,}5393\,.$$

Таблиця 3.1

Інтервал $(x_i; x_{i+1})$	n_i	$x_{i+1} - \bar{x}_в$	$x_i - \bar{x}_в$	$\frac{x_{i+1} - \bar{x}_в}{\sigma_в}$	$\frac{x_i - \bar{x}_в}{\sigma_в}$	$\Phi\left(\frac{x_{i+1} - \bar{x}_в}{\sigma_в}\right)$	$\Phi\left(\frac{x_i - \bar{x}_в}{\sigma_в}\right)$
(–∞; 1,720)	3	–0,0217	–∞	–1,916	–∞	–0,47231	–0,5
[1,720; 1,727)	13	–0,0147	–0,0217	–1,298	–1,916	–0,40285	–0,47231
[1,727; 1,734)	32	–0,0077	–0,0147	–0,680	–1,298	–0,25175	–0,40285
[1,734; 1,741)	49	–0,0007	–0,0077	–0,062	–0,680	–0,02472	–0,25175
[1,741; 1,748)	42	0,0063	–0,0007	0,556	–0,062	0,21089	–0,02472
[1,748; 1,755)	35	0,0133	0,0063	1,174	0,556	0,3798	0,21089
[1,755; 1,762)	19	0,0203	0,0133	1,792	1,174	0,46336	0,3798
(1,762; ∞)	7	∞	0,0203	∞	1,792	0,5	0,46336
Всього	200	—	—	—	—	—	—

Продовження таблиці 3.1

Інтервал $(x_i; x_{i+1})$	$p_i = \Phi\left(\frac{x_{i+1} - \bar{x}_в}{\sigma_в}\right) - \Phi\left(\frac{x_i - \bar{x}_в}{\sigma_в}\right)$	$n_i^0 = p_i n = 200 \cdot p_i$
(–∞; 1,720)	0,0277	5,54
[1,720; 1,727)	0,06946	13,89
[1,727; 1,734)	0,1511	30,22
[1,734; 1,741)	0,2270	45,41
[1,741; 1,748)	0,2356	47,12
[1,748; 1,755)	0,1689	33,78
[1,755; 1,762)	0,08356	16,71
(1,762; ∞)	0,03664	7,33
Всього	—	200

Таблиця 3.2

i	n_i	n_i^0	$n_i - n_i^0$	$(n_i - n_i^0)^2$	$(n_i - n_i^0)^2 / n_i^0$	n_i^2	n_i^2 / n_i^0
1	3	5,54	–2,54	6,4516	1,1646	9	1,6246
2	13	13,89	–0,89	0,7921	0,0570	169	12,1670
3	32	30,22	1,78	3,1684	0,1048	1024	33,8848
4	49	45,41	3,59	12,8881	0,2838	2401	52,8738
5	42	47,12	–5,12	26,2144	0,5563	1764	37,4363
6	35	33,78	1,22	1,4884	0,0441	1225	36,2641
7	19	16,71	2,29	5,2441	0,3138	361	21,6038
8	7	7,33	–0,33	0,1089	0,0149	49	6,6849
$\sum$	200	200			$\chi^2_{\text{сп}} = 2{,}5393$		202,5393

За таблицею критичних точок розподілу χ^2 (табл. 6 додатків) для рівня значущості $\alpha = 0{,}05$ і числа ступенів вільності $k = 8 - 3 = 5$ знайдемо $\chi^2_{\text{кр}}(0{,}05;5) = 11{,}07$.

Оскільки $\chi^2_{\text{сп}} < \chi^2_{\text{кр}}$, то нема підстав відкидати основну гіпотезу. Отже, дані спостережень узгоджуються із гіпотезою про нормальний розподіл кількісної ознаки генеральної сукупності. ●

Задача 3.5. На основі вибірки

0,6917; 0,1794; 0,7410; 0,3094; 0,1174;
0,5424; 0,0834; 0,6288; 0,9401; 0,6606

для рівня значущості $\alpha = 0{,}05$ з допомогою критерія Колмогорова перевірити гіпотезу H_0 про те, що дана генеральна сукупність на проміжку $[0; 1]$ має рівномірний розподіл, тобто $H_0 : F(x) = x$ $(0 \le x \le 1)$.

○ В даному випадку серед варіант відсутні рівні, тому спочатку утворимо варіаційний ряд, а потім використаємо співвідношення (3.18) для обчислення D_{10}. Результати проміжних обчислень занесені в табл. 3.3.

Таблиця 3.3

l	x_l	$F(x_l)$	$\frac{2l-1}{2n}$	$F(x_l) - \frac{2l-1}{2n}$	$\left\lvert F(x_l) - \frac{2l-1}{2n} \right\rvert$
1	0,0834	0,0834	0,05	0,0334	0,0334
2	0,1174	0,1174	0,15	–0,0326	0,0326
3	0,1794	0,1794	0,25	–0,0706	0,0706
4	0,3094	0,3094	0,35	–0,0406	0,0406
5	0,5424	0,5424	0,45	0,0924	0,0924
6	0,6288	0,6288	0,55	0,0788	0,0788
7	0,6606	0,6606	0,65	0,0106	0,0106
8	0,6917	0,6917	0,75	–0,0583	0,0583
9	0,7410	0,7410	0,85	–0,1090	0,1090
10	0,9401	0,9401	0,95	–0,0099	0,0099

Найбільше число 0,1090 в останньому стовпчику визначає перший доданок у формулі (3.18). Тому $D_{10} = 0{,}1090 + 0{,}05 = 0{,}1590$. За табл. 8 додатків при рівні значущості $\alpha = 0{,}05$ і $n = 10$ знаходимо критичне значення $K_{10;0,05} = 0{,}409$. Оскільки $D_{10} = 0{,}1590 < 0{,}409 = K_{10;0,05}$, то робимо висновок, що результати спостережень не суперечать гіпотезі H_0. ●

Задача 3.6. На основі інтервального статистичного розподілу вибірки, наведеного в табл. 1.3, з допомогою критерія Колмогорова здійснити перевірку гіпотези H_0 про нормальний розподіл кількісної ознаки генеральної сукупності, якщо рівень значущості $\alpha = 0{,}05$.

○ В задачі 1.2 для даного розподілу знайдені числові характеристики $\bar{x}_в = 1{,}7417$, $\sigma_в = 0{,}0113269$, які можна вважати параметрами нормального розподілу a та σ відповідно. Тоді теоретична функція розподілу $F(x)$ має такий вид:

$$F(x) = 0{,}5 + \Phi\left(\frac{x-a}{\sigma}\right) = 0{,}5 + \Phi\left(\frac{x-\bar{x}_в}{\sigma_в}\right),$$

де $\Phi(x)$ — функція Лапласа, значення якої наведені в табл. 3 додатків.

Враховуючи відмінність емпіричної функції, побудованої для інтервального розподілу, в порівнянні із випадком дискретного розподілу, виберемо статистику критерія Колмогорова (3.19) (для односторонньої критичної області).

Всі розрахунки для отримання D^+_{200} зведемо у табл. 3.4, при заповненні якої використаємо відповідні стовпці табл. 3.1. Зокрема, стовпці 1, 2 та 4 табл. 3.4 взяті із стовпців 1, 2 та 5 відповідно табл. 3.1. При заповненні стовпця 3 використовується означення емпіричної функції. Нарешті, при обчисленні даних стовпця 5 були використані результати стовпця 7 табл. 3.1.

Таблиця 3.4

$[x_i; x_{i+1})$	n_i	$F^*(x_{i+1})$	$\frac{x_{i+1}-\bar{x}_в}{\sigma_в}$	$F(x_{i+1}) = 0{,}5 + \Phi\left(\frac{x_{i+1}-\bar{x}_в}{\sigma_в}\right)$	$\lvert F^*(x) - F(x)\rvert$
(–∞; 1,713)	0	0	–2,534	0,00564	0,00564
[1,713; 1,720)	3	0,015	–1,916	0,02769	0,01269
[1,720; 1,727)	13	0,080	–1,298	0,09715	0,01715
[1,727; 1,734)	32	0,240	–0,680	0,24825	0,00825
[1,734; 1,741)	49	0,485	–0,062	0,47528	0,00972
[1,741; 1,748)	42	0,695	0,556	0,71089	0,01589
[1,748; 1,755)	35	0,870	1,174	0,8798	0,00980
[1,755; 1,762)	19	0,965	1,792	0,96336	0,00164
[1,762; 1,769)	7	1	2,410	0,99202	0,00798
[1,769; ∞)	1	1	∞	1	0

Найбільший модуль різниці між відповідними значеннями емпіричної і (гіпотетичної) теоретичної функції розподілу вибираємо із стовпця 6 табл. 3.4:

$$D^+_{200} = \max_{1 \le i \le k+1} \left| F(x_i) - F^*(x_i) \right| = 0{,}01715.$$

Критичне значення $K^+_{200;0,01}$ знайдемо за наближеною формулою (3.20), використавши рівність

$$\lg\alpha = 0{,}4343 \ln\alpha, \text{ з якої } \ln\alpha = 2{,}3026 \lg\alpha.$$

Отже,

$$\ln(0{,}1) = 2{,}3026 \lg 0{,}1 = -2{,}3026,$$

$$K^{+}_{200;0,1} \approx \sqrt{\frac{2{,}3026}{2 \cdot 200}} = 0{,}07587.$$

Оскільки $D^{+}_{200} = 0{,}01715 < 0{,}07587 = K^{+}_{200;0,1}$, то гіпотезу H_0 приймаємо.

Відмітимо, що в задачах 3.4 та 3.6 фактично досліджувався інтервальний статистичний розподіл з табл. 1.3. Цей розподіл породжений дискретним рядом варіант з табл. 1.2 і, природно, приводить до (хай незначних) похибок. Тому при перевірці гіпотези H_0 про нормальність розподілу генеральної сукупності, із якої отримано вибірку, задану табл. 1.2, з допомогою критерія Колмогорова бажано було б використати статистику Колмогорова (3.16) безпосередньо до варіант із табл. 1.2. Проте без використання обчислювальної техніки із відповідним програмним забезпеченням така задача є дуже працеємкою, якщо врахувати великий обсяг вибірки $(n = 200)$ і слабку згрупованість варіант — наявність 48 різних варіант у табл. 1.2, для кожної із яких потрібно знаходити значення функції Лапласа.

В зв'язку із цим цікаво було б дослідити таке питання: якщо інтервальний статистичний розподіл із табл. 1.3 замінити дискретним розподілом

x_i	1,7165	1,7235	1,7305	1,7375	1,7445	1,7515	1,7585	1,7655
n_i	3	13	32	49	42	35	19	7

, (3.21)

то чи буде для нього правильною гіпотеза H_0 згідно із критерієм узгодженості Колмогорова? Отримання відповіді на це питання передбачає використання статистики (3.16) двостороннього критерія Колмогорова для випадку, коли частоти n_i не дорівнюють 1, як це було для даних задачі 3.5, а остаточний висновок попередить про обережність стосовно подібних переходів. ●

Задача 3.7. На основі статистичного розподілу (3.21) з допомогою критерія Колмогорова здійснити перевірку гіпотези H_0 про нормальний розподіл генеральної сукупності при рівні значущості $\alpha = 0{,}05$.

○ В задачі 1.2 для розподілу (3.21) знайдено числові характеристики $\bar{x}_в = 1{,}7417$, $\sigma_в = 0{,}0113269$, які можна вважати параметрами нормального розподілу a та σ відповідно. Тоді, як і для задачі 3.6, теоретична функція розподілу $F(x)$ має такий вид:

$$F(x) = 0{,}5 + \Phi\left(\frac{x - \bar{x}_в}{\sigma_в}\right) = 0{,}5 + \Phi\left(\frac{x - 1{,}7417}{0{,}0113269}\right).$$

Всі розрахунки для обчислення D_{200} помістимо в табл. 3.5.

Таблиця 3.5

x_l	n_l	$F^*(x_l)$	$F^*(x_l+0)$	$\dfrac{x_l-\bar{x}_в}{\sigma_в}$
1,7165	3	0	0,015	–2,225
1,7235	13	0,015	0,080	–1,608
1,7305	32	0,080	0,240	–0,989
1,7375	49	0,240	0,485	–0,371
1,7445	42	0,485	0,695	0,247
1,7515	35	0,695	0,870	0,865
1,7585	19	0,870	0,965	1,483
1,7655	7	0,965	1	2,101

Продовження *таблиці 3.5*

x_l	$F(x_l)=0{,}5+$ $+\Phi\left(\dfrac{x_l-\bar{x}_в}{\sigma_в}\right)$	$F^*(x_l)-F(x_l)$	$F^*(x_l+0)-F(x_l)$
1,7165	0,01304	–0,01304	0,00196
1,7235	0,05392	–0,03892	0,02608
1,7305	0,16134	–0,08134	0,07866
1,7375	0,35606	–0,11606	0,12894
1,7445	0,59755	–0,11255	0,09745
1,7515	0,80648	–0,11148	0,06352
1,7585	0,93096	–0,06096	0,03404
1,7655	0,98218	–0,01718	0,01782

Найбільший модуль різниці між відповідними значеннями емпіричної і теоретичної функцій розподілу вибираємо із останнього стовпця табл. 3.5: $D_{200}=0{,}12894$. За формулою (3.14) критичне значення $K_{200;0,05}=k_{1-0,05}\big/\sqrt{200}$, де за табл. 9 додатків $k_{1-0,05}=1{,}358$. Остаточно $K_{200;0,05}=1{,}358\big/\sqrt{200}=0{,}09603$.

Оскільки $D_{200}=0{,}12894>0{,}09603=K_{200;0,05}$, то робимо висновок, що гіпотезу H_0 про нормальність розподілу генеральної сукупності на підставі дискретного розподілу вибірки (3.21) потрібно відкинути. ●

Таким чином, згідно із критеріями узгодженості Пірсона та Колмогорова, використаними до інтервального розподілу вибірки із табл. 1.3, гіпотеза H_0 не відхиляється (задачі 3.4 та 3.6), а використання критерію Колмогорова до дискретного розподілу, породженого цим самим інтервальним розподілом, вже

приводить до висновку про хибність цієї ж гіпотези. З'ясуємо причину такого неспівпадання висновків. Найбільші розходження між значеннями емпіричної та гіпотетичної теоретичної функцій розподілу, які перевищують критичне значення $K_{200;0,05}$, отримані для варіант 1,7375 та 1,7445 з найбільшими частотами 49 та 42 відповідно (див. табл. 3.5). Оскільки $\bar{x}_в = 1{,}7417$, то заміна модальних інтервалів їх серединами приводить до "віддалення" отриманих варіант від $\bar{x}_в$ із збереженням великих частот, що і приводить до різкого збільшення відхилення $F^*(x)$ від $F(x)$ в цих точках.

Нарешті, відмітимо, що критерій Колмогорова рекомендується використовувати при обсязі вибірки $n \geq 30$.

§ 4. ЕЛЕМЕНТИ КОРЕЛЯЦІЙНОГО ТА РЕГРЕСІЙНОГО АНАЛІЗУ

Поняття стохастичної та статистичної залежності, кореляції і регресії. Основні задачі кореляційного і регресійного аналізу. Лінійні емпіричні рівняння парної кореляції. Вибірковий коефіцієнт лінійної кореляції та його властивості. Оцінка достовірності емпіричних коефіцієнтів кореляції і регресії за даними вибірки. Нелінійна парна кореляція. Вибіркове кореляційне відношення та його властивості. Регресійний аналіз: парна лінійна регресія.

КОРОТКІ ТЕОРЕТИЧНІ ВІДОМОСТІ

Поняття стохастичної і статистичної залежності, кореляції і регресії.

Основні задачі кореляційного і регресійного аналізу

Дві випадкові величини можуть бути або незалежними, або пов'язані функціональною залежністю, або залежністю, яка називається стохастичною. Функціональна залежність, розглянута в §8 Ч. 1, досить рідко зустрічається в економічних дослідженнях. Дуже часто доводиться вивчати такі випадкові величини, для яких зміна можливих значень (при проведенні випробувань) приводить до зміни закону (умовного) розподілу іншої (див. §7 Ч. 1). Такий зв'язок між випадковими величинами називається **стохастичним**; він виникає тоді, коли на обидві величини впливають випадкові фактори, серед яких є спільні.

Найбільш важливим випадком стохастичного зв'язку є так званий **кореляційний** зв'язок, який встановлює залежність між значеннями випадкової величини X і умовним математичним сподіванням $M(Y \mid X = x)$ [*)] випадкової величини Y:

$$M(Y \mid X = x) = g(x), \qquad (4.1)$$

[*)] Умовні математичні сподівання дискретних і неперервних випадкових величин визначаються в §7 Ч. 1.

або між значенням випадкової величини Y і умовним математичним сподіванням $M(X \mid Y = y)$ [*)] випадкової величини X:

$$M(X \mid Y = y) = q(y). \tag{4.2}$$

Функції $g(x)$ і $q(y)$ називаються **функціями регресії** Y на X та X на Y відповідно, а їх графіки — лініями регресії Y на X та X на Y, рівняння (4.1) та (4.2) називаються **рівняннями регресії *Y* на *X* та *X* на *Y*** відповідно.

Використовуючи інформацію §§7,8 Ч. 1 можна отримати важливу для наступного викладу основну властивість регресії величини Y на величину X: якщо $g(x)$ є функцією регресії величини Y на X, то математичне сподівання квадрата відхилення величини Y від функції $g(x)$ менше, ніж від будь-якої функції $h(X) \neq g(X)$, тобто

$$M\left[\, Y - g(X)\right]^{\,2} < M\left[\, Y - h(X)\right]^{\,2}. \tag{4.3}$$

Аналогічною властивістю володіє і регресія величини X на величину Y.

Прикладами кореляційного зв'язку є стохастична взаємозалежність між: 1) окремими параметрами тіла людини або тварини; 2) обсягом виробництва підприємства та коефіцієнтом використання основних засобів.

На практиці сумісний закон розподілу випадкових величин X та Y (двовимірної випадкової величини $(X; Y)$ §7 Ч. 1) невідомий. В розпорядженні дослідника є двовимірна вибірка

$$(x^{(1)}; y^{(1)}), (x^{(2)}; y^{(2)}), \ldots, (x^{(n)}; y^{(n)}), \tag{4.4}$$

згрупувавши (для великих n) дані якої, можна отримати двовимірний статистичний розподіл, наведений в табл. 4.1.

Таблиця 4.1

X	Y						
	y_1	y_2	...	y_j	...	y_m	n_{x_i}
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1m}	n_{x_1}
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2m}	n_{x_2}
...	...	...	...	...	...	...	...
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{im}	n_{x_i}
...	...	...	...	...	...	...	...
x_k	n_{k1}	n_{k2}	...	n_{kj}	...	n_{km}	n_{x_k}
n_{y_j}	n_{y_1}	n_{y_2}	...	n_{y_j}	...	n_{y_m}	n

Тут n_{ij} — частота спільної появи ознак x_i, y_j (пари $(x_i; y_j)$);

$\sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij} = n$ — обсяг вибірки;

$$n_{x_i} = \sum_{j=1}^{m} n_{ij}, \ (i = \overline{1;k}); \quad n_{y_j} = \sum_{i=1}^{k} n_{ij}, \ \left(j = \overline{1;m}\right).$$

Кожному значенню X відповідає ряд значень Y, тобто зміна значень X приводить до зміни умовного статистичного розподілу $Y \mid x$. Аналогічно зміна значень Y веде до зміни умовного статистичного розподілу $X \mid y$.

Наприклад, при $X = x_1$ та $X = x_2$ відповідні умовні статистичні розподіли величини Y мають такі види:

$$\begin{array}{c|cccc} Y \mid x_1 & y_1 & y_2 & \dots & y_m \\ \hline n_i & n_{11} & n_{12} & \dots & n_{1m} \end{array}; \qquad \begin{array}{c|cccc} Y \mid x_2 & y_1 & y_2 & \dots & y_m \\ \hline n_i & n_{21} & n_{22} & \dots & n_{2m} \end{array}.$$

Статистичною називається така залежність між величинами X та Y, для якої зміна спостережених значень однієї із величин зумовлює зміну умовного статистичного розподілу іншої.

Умовною середньою $\bar{y}_x$ називається середнє арифметичне спостережених значень Y, які відповідають значенню $X = x$. Наприклад, згідно із табл. 4.1

$$\bar{y}_{x_2} = \frac{y_1 n_{21} + y_2 n_{22} + \dots + y_j n_{2j} + \dots + y_m n_{2m}}{n_{x_2}}.$$

Умовною середньою $\bar{x}_y$ називається середнє арифметичне спостережених значень X, які відповідають значенню $Y = y$. Наприклад, у відповідності із табл. 4.1

$$\bar{x}_{y_1} = \frac{x_1 n_{11} + x_2 n_{21} + \dots + x_k n_{k1}}{n_{y_1}}.$$

Можна довести, що умовні середні є незміщеними і спроможними точковими статистичними оцінками відповідних умовних математичних сподівань:

$$\bar{y}_x \approx M(Y \mid X = x); \quad \bar{x}_y \approx M(X \mid Y = y). \qquad (4.5)$$

Вкажемо підхід до знаходження статистичних наближень (емпіричних функцій) $\hat{g}(x) = \hat{g}(x; a_0, a_1, \dots, a_m)$ та $\hat{q}(y) = \hat{q}(y; b_0, b_1, \dots b_m)$ невідомих функцій регресій $g(x)$ та $q(y)$ відповідно, а отже, з урахуванням співвідношень (4.5) і емпіричних рівнянь Y на X

$$\bar{y}_x = \hat{g}(x, a_0, a_1, \dots, a_m) \qquad (4.6)$$

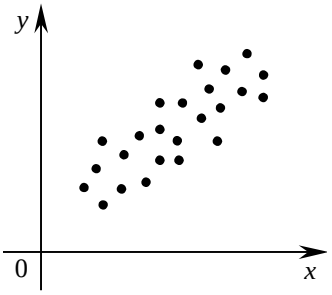
та X на Y

$$\bar{x}_y = \hat{q}(y, b_0, b_1, \dots b_m) \qquad (4.7)$$

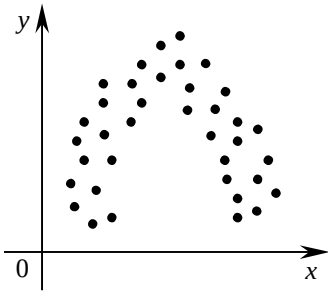
де $a_0, a_1, \dots, a_m$ та $b_0, b_1, \dots b_m$ — невідомі параметри.

На першому кроці в прямокутній системі координат xOy будується послідовність пар чисел (4.4). Отримана сукупність точок називається **кореляційним полем** або **діаграмою розсіювання**. Конфігурація

кореляційного поля дозволяє висунути гіпотезу про вид функції $\hat{g}(x)$. Наприклад, у випадку кореляційного поля, зображеного на мал. 4.1, $\hat{g}(x)$ доцільно шукати у вигляді лінійної функції a_1x+a_0, а у випадку мал. 4.2 у вигляді параболічної функції другого порядку $a_2x^2+a_1x+a_0$.



Мал. 4.1.



Мал. 4.2.

На другому кроці потрібно знайти невідомі параметри. Ця задача розв'язується з допомогою так званого **методу найменших квадратів** (МНК), суть якого полягає в наступному. Оскільки невідома функція регресії Y на X $g(X)$ згідно із (4.3) мінімізує величину $M\left[Y-h(X)\right]^2$, а оцінкою $M(Y\mid x)$ є умовна середня $\bar{y}_x$, що відповідає певним спостереженим значенням $X=x$, то емпірична лінія регресії (4.6) повинна задовольняти рівності

$$F(a_0,a_1,...,a_m)\equiv\sum_{i=1}^{n}\left[\ y^{(i)}-\hat{g}(x^{(i)};a_0,a_1,...,a_m)\right]^2\Rightarrow\min. \qquad (4.8)$$

Права частина цієї рівності є сума квадратів віддалей вздовж осі $0y$ від точок $(x^{(i)};y^{(i)})$ послідовності (4.4) до відповідних точок лінії регресії із тією ж абсцисою.

Для випадку згрупованих даних із табл. 4.1 рівність (4.8) набирає такого виду

$$F(a_0,a_1,\ldots,a_m)\equiv\sum_{i=1}^{k}\sum_{j=1}^{m}\left[\ y_j-\hat{g}(x_i;a_0,a_1,\ldots,a_m)\right]^2 n_{ij}\Rightarrow\min. \quad (4.8^*)$$

Таким чином, емпірична функція $\hat{g}(x;a_0,a_1,\ldots,a_m)$ повинна усереднити (згладити) спостережені дані $(x^{(i)};y^{(i)})$, $i=\overline{1,n}$, з послідовності (4.4) або (x_i,y_j), $i=\overline{1,k}$, $j=\overline{1,m}$, із табл. 4.1. При цьому невідомі параметри $a_0,a_1,\ldots,a_m$ повинні мінімізувати функцію $F(a_0,a_1,\ldots,a_m)$; необхідною умовою цього є виконання системи рівнянь

$$\begin{cases} \dfrac{\partial F(a_0, a_1, \ldots, a_m)}{\partial a_i} = 0, \quad i = \overline{0, m}, \end{cases} \tag{4.9}$$

яка називається **системою нормальних рівнянь**.

Метод найменших квадратів дозволяє **при встановленому виді емпіричної функції регресії** $\bar{g}(x; a_0, a_1, \ldots, a_m)$ так знайти невідомі параметри $a_0, a_1, \ldots, a_m$, що вона буде найкращою оцінкою функції регресії $g(x)$ в тому розумінні, що сума квадратів відхилень спостережених значень випадкової величини Y від відповідних ординат емпіричної функції буде найменшою.

Аналогічно знаходиться емпірична функція $\bar{q}(y; b_0, b_1, \ldots, b_m)$ регресії X на Y.

Зауваження. Враховуючи випадковий характер організації вибірки, можна зробити висновок, що для нефіксованої вибірки знайдені параметри є випадковими величинами.

Знаходження емпіричних рівнянь регресії — це тільки перший крок дослідження статистичних зв'язків між випадковими величинами. Наступним є встановлення сили або тісноти цих зв'язків. Відомо (§7 Ч. 1), що мірою лінійного зв'язку (стохастичного) між двома випадковими величинами X та Y є коефіцієнт кореляції

$$r = r_{XY} = \frac{K_{XY}}{\sigma_X \sigma_Y} = \frac{M(XY) - M(X)M(Y)}{\sigma_X \sigma_Y}, \tag{4.10}$$

де K_{XY} — кореляційний момент або коефіцієнт коваріації (сумісної варіації), який ще часто позначають $\operatorname{cov}(X, Y)$, $\sigma_X = \sqrt{D(X)}$, $\sigma_Y = \sqrt{D(Y)}$. Зокрема, якщо величини X та Y незалежні, то коефіцієнт кореляції $r = 0$; якщо $r = \pm 1$, то X та Y пов'язані **лінійною** функціональною залежністю. Звідси випливає, що коефіцієнт кореляції r вимірює силу (тісноту) **лінійного** зв'язку між X та Y.

Використовуючи метод моментів, тобто замінивши числові характеристики їх статистичними оцінками, можна отримати точкову статистичну оцінку коефіцієнта кореляції — вибірковий (емпіричний) коефіцієнт кореляції.

$$r_в = r_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y}, \tag{4.11}$$

де для незгрупованих даних (4.4)

$$\overline{xy} = \sum_{i=1}^{n} x^{(i)} y^{(i)} / n, \tag{4.12}$$

а для згрупованих даних із табл. 4.1

$$\overline{xy} = \sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij} x_i y_j / n = \sum n_{xy} xy / n, \tag{4.13}$$

σ_x та σ_y — середні квадратичні відхилення вибіркові для ознак X та Y відповідно.

Зауваження. Оскільки кореляційний момент рівносильно (4.10) визначається ще й таким чином: $K_{XY} = M\{ [X - M(X)] [Y - M(Y)] \}$, то формулу (4.11) можна подати і в такому вигляді

$$r_в = \frac{\sum_{i=1}^{n} (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{n\sigma_x \sigma_y}. \qquad (4.11^*)$$

Оскільки $r_в$ для нефіксованої вибірки є випадковою величиною, то знайдене значення $r_в$ для конкретної вибірки може суттєво відрізнятись від r. В зв'язку із цим необхідно побудувати довірчі інтервали для оцінки r, а також перевірити статистичну гіпотезу про некорельованість X та Y ($r = o$).

У випадку нелінійності хоча б однієї із функцій регресії $g(x)$ та $q(y)$ коефіцієнт кореляції r вже не дає інформації про силу зв'язку між X та Y. Натомість ступінь концентрації розподілу поблизу лінії регресії показує **кореляційне відношення** Y на X:

$$\eta_{Y/X} = 1 - \frac{\sigma^2_{Y/X}}{\sigma^2_Y}, \qquad (4.14)$$

де

$$\sigma^2_{Y/X} = D(Y \mid X = x) = M\{ [Y - g(x)]^2 \mid X = x \}. \qquad (4.15)$$

Із означення випливає, що кореляційне відношення змінюється в межах від 0 до 1 включно; воно дорівнює 1 тоді і тільки тоді, коли $\sigma^2_{Y/X} = 0$, тобто весь розподіл зосереджений на лінії регресії (має місце функціональна залежність). Це відношення дорівнює нулю тоді і тільки тоді, коли лінія регресій Y на X являє собою горизонтальну пряму, що проходить через центр розподілу, тобто якщо Y та X некорельовані.

Аналогічно вводиться означення кореляційного відношення X на Y $\eta_{X/Y}$.

Можна довести, що у всіх випадках виконуються нерівності

$$r^2 \le \eta^2_{X/Y}, \quad r^2 \le \eta^2_{Y/X}. \qquad (4.16)$$

За статистичними даними двовимірної вибірки можна знайти вибіркові кореляційні відношення $\hat{\eta}_{Y/X}$ та $\hat{\eta}_{X/Y}$, які є точковими статистичними оцінками $\eta_{Y/X}$ та $\eta_{X/Y}$ відповідно.

Сукупність методів оцінки кореляційних характеристик і перевірка статистичних гіпотез про них за даними вибірки називається **кореляційний аналізом**. В кореляційному аналізі використовуються такі основні методи:

1) побудова кореляційного поля і складання кореляційної таблиці;

2) знаходження вибіркового (емпіричного) коефіцієнта кореляції або кореляційного відношення;

3) перевірка статистичних гіпотез про значущість (істотність) зв'язку.

В практично важливих задачах одна із випадкових величин є "результуючою", а інша — від якої вона залежить (факторна величина). Наприклад, залежність урожайності сільськогосподарських культур Y від маси внесених добрив X. Це зразок односторонньої залежності між випадковими величинами, яка досліджується в регресійному аналізі.

Більш того, факторна величина може бути **детермінованою** (невипадковою), тобто значення якої не тільки відоме, але й ним можна керувати. При дослідженні лінійної залежності результуючої величини Y від факторної величини X загальний вираз емпіричного рівняння регресії має такий вид

$$\bar{y} = a_0 + a_1 x. \tag{4.17}$$

В загальному випадку емпіричне рівняння регресії визначається таким чином

$$\bar{y} = \hat{g}(x; a_0, a_1, \ldots, a_m), \tag{4.18}$$

де $a_0, a_1, \ldots, a_m$ — невідомі параметри.

Тепер, коли визначені об'єкти дослідження цього параграфа (рівняння регресії (4.6), (4.7), (4.17), (4.18)), можемо сформулювати основні задачі регресійного аналізу.

Регресійний аналіз — це дослідження односторонніх статистичних залежностей між випадковими величинами. При цьому деякі із факторних величин можуть бути невипадковими величинами.

Задачі регресійного аналізу:

1) визначення форми залежностей;

2) знаходження функції регресії;

3) побудова точкових та інтервальних оцінок параметрів функції регресії;

4) знаходження точкових та інтервальних оцінок умовних математичних сподівань, необхідних для визначення меж, в яких із заданою надійністю будуть міститися середні значення досліджуваної величини, якщо інші пов'язані з нею величини набирають певних значень;

5) перевірка узгодженості знайденої емпіричної функції регресії спостереженим даним.

Основна мета регресійного аналізу — **теоретично обґрунтований і статистично надійний точковий та інтервальний прогноз** значень залежної величини або умовного математичного сподівання цієї величини.

Зауваження. Якщо рівняння регресії описує об'єкт дослідження із економічної сфери і воно обґрунтоване в теоретично-економічному сенсі, то його називають **економетричним рівнянням**.

Найбільш простим випадком є той, коли **обидві функції регресії** $g(x)$ і $q(y)$ в рівняннях регресії (4.1) та (4.2) є лінійними, тобто обидві лінії регресії є прямими лініями; вони називаються **прямими регресії**. В цьому випадку будемо говорити про лінійну кореляцію між випадковими величинами X та Y.

Можна довести, що коли сумісний розподіл імовірностей величин X і Y є **нормальним розподілом** на площині (див. [7, п. 7.8]), тоді **кореляційний зв'язок завжди є лінійним**. В зв'язку із цим слід очікувати лінійного кореляційного зв'язку між статистично залежними випадковими величинами X та Y, якщо кожну із них можна розглядати як суму великого числа незалежних або майже незалежних випадкових доданків.

Нехай конфігурація кореляційного поля, отриманого внаслідок зображення у вигляді точки в прямокутній декартовій системі координат $x0y$ кожної із пар чисел вибіркової послідовності (4.4), дозволяє висунути припущення про лінійну кореляційну залежність між X та Y. Тобто рівняння регресії (4.1) та (4.2) в цьому випадку набирають відповідно такого виду

$$M(Y \mid X = x) = \alpha_1 x + \alpha_0, \tag{4.19}$$

$$M(X \mid Y = y) = \beta_1 y + \beta_0, \tag{4.20}$$

де α_i, β_i $(i = 1,0)$ — сталі.

Тоді емпіричні рівняння регресії Y на X та X на Y будемо шукати у вигляді

$$\bar{y}_x = a_1 x + a_0, \tag{4.21}$$

$$\bar{x}_y = b_1 y + b_0, \tag{4.22}$$

де a_0, a_1 та b_0, b_1 — невідомі параметри, які є **точковими статистичними оцінками** відповідних **чисел** рівнянь (4.19) і (4.20).

Використавши метод найменших квадратів як для незгрупованих даних(4.4), так і для згрупованих даних з табл.4.1, можна отримати **систему нормальних рівнянь**

$$\begin{cases} \overline{x^2} a_1 + \bar{x} a_0 = \overline{xy}, \\ \bar{x} a_1 + a_0 = \bar{y}, \end{cases} \tag{4.23}$$

для знаходження параметрів рівняння регресії (Y на X) (4.21).

Система (4.23) має єдиний розв'язок

$$a_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2}, \quad a_0 = \bar{y} - \bar{x} a_1. \tag{4.24}$$

Коефіцієнт a_1 рівняння (4.21) емпіричної прямої регресії Y на X називається **вибірковим (емпіричним або статистичним) коефіцієнтом регресії *Y* на *X*** і позначається $a_{Y/X}$. Оскільки $\overline{x^2} - (\bar{x})^2 = \sigma_x^2$, а

$$\overline{xy} - \bar{x} \cdot \bar{y} = K_{XY}^* (= \overline{\mathrm{cov}}(X, Y)) \tag{4.25}$$

— **вибірковий (емпіричний або статистичний) кореляційний момент** або **вибіркова коваріація**, то першу формулу (4.24) можна записати в такому виді

$$a_{Y/X} = a_1 = \frac{K^*_{XY}}{\sigma^2_X}\left(\equiv \frac{\overline{\mathrm{cov}}(X,Y)}{\sigma^2_X}\right), \tag{4.26}$$

а шукане рівняння (4.21) з урахуванням другої рівності (4.25) набуде вигляду

$$\bar{y}_x - \bar{y} = \frac{K^*_{XY}}{\sigma^2_X}(x - \bar{x}). \tag{4.27}$$

Це рівняння показує, що емпірична пряма регресії Y на X проходить через точку з координатами $(\bar{x}, \bar{y})$, яка називається **середньою точкою кореляційного поля**.

Аналогічно можна отримати емпіричне рівняння прямої X на Y, якщо мінімізувати сумарні квадрати відхилень точок $(\bar{x}_i, y_i)$, $i = \overline{1,m}$, від шуканої прямої, тобто прямої

$$\bar{x}_y - \bar{x} = b_{X/Y}(y - \bar{y}), \tag{4.28}$$

де вибірковий (емпіричний або статистичний) коефіцієнт регресії X на Y визначається за формулою

$$b_{X/Y} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\sigma^2_y} = \frac{K^*_{XY}}{\sigma^2_y}\left(\equiv \frac{\overline{\mathrm{cov}}(X,Y)}{\sigma^2_y}\right). \tag{4.29}$$

У формулах вибіркових коефіцієнтів регресії (4.26) і (4.29) чисельники співпадають, а знаменники завжди додатні, оскільки є дисперсіями вибірковими випадкових величин X та Y відповідно. Тому $a_{Y/X}$ і $b_{X/Y}$ мають однакові знаки. Відмітимо також, що вибірковий коефіцієнт регресії $a_{Y/X}(b_{X/Y})$ — це міра, яка на основі вибіркових даних в середньому вказує на вплив зміни змінної X (або Y) на змінну Y (або X).

Рівністю (4.11) **формально** був визначений вибірковий (емпіричний) коефіцієнт кореляції $r_в$ як точкова статистична оцінка коефіцієнта кореляції $r = r_{XY}$ — міри лінійного кореляційного зв'язку між випадковими величинами X та Y. Проте вид емпіричних коефіцієнтів регресії Y на X та X на Y вказує на природний зв'язок $r_в$ із $a_{Y/X}$ і $b_{X/Y}$ (за аналогією із зв'язком в теорії кореляції як розділу теорії імовірностей). А тому виникає необхідність детально вивчити властивості $r_в$ як **оцінки** сили лінійного кореляційного зв'язку між величинами X та Y. При цьому слід очікувати властивостей, аналогічних із властивостями r_{XY}.

Вибірковим (емпіричним або статистичним) коефіцієнтом кореляції $r_в = r_{XY}$ випадкових величини X та Y, між якими припускається лінійний кореляційний зв'язок, називається відношення емпіричного кореляційного

момента (коефіцієнта коваріації) $K^*(X,Y)(=\overline{\text{cov}}(X,Y))$ **до добутку середніх квадратичних відхилень вибіркових** σ_x **та** σ_y :

$$r_в = r_{xy} = \frac{K^*_{XY}}{\sigma_x\sigma_y}\left(=\frac{\overline{\text{cov}}(X,Y)}{\sigma_x\sigma_y}\right)=\frac{\overline{xy}-\bar{x}\cdot\bar{y}}{\sigma_x\sigma_y}, \tag{4.30}$$

де $\overline{xy}$ для незгрупованих даних (4.4) вибірки обчислюється за формулою (4.12), а для згрупованих даних із табл. 4.1 — за (4.13).

Зауваження. Для випадку, коли x_i та y_i $(i=\overline{1,k})$ є великими числами, а обсяг вибірки $n \geq 50$, то зручніше користуватися такою формулою

$$r_в = \frac{\sum\limits_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\left[\sum\limits_{i=1}^{n}(x_i-\bar{x})^2\right]\left[\sum\limits_{i=1}^{n}(y_i-\bar{y})^2\right]}}. \tag{4.30*}$$

Використовуючи формули (4.26) і (4.29), отримаємо вираз через емпіричні коефіцієнти регресій для вибіркового коефіцієнта кореляції:

$$r_в = \pm\sqrt{\frac{(K^*_{XY})^2}{\sigma_x^2\sigma_y^2}} = \pm\sqrt{\frac{K^*_{XY}}{\sigma_x^2}\cdot\frac{K^*_{XY}}{\sigma_y^2}} = \pm\sqrt{a_{Y/X}b_{X/Y}},$$

де знак перед коренем визначається знаком емпіричних коефіцієнтів регресій.

З другого боку, і емпіричні коефіцієнти регресій можна виразити через вибірковий коефіцієнт кореляції:

$$a_{Y/X} = \frac{K^*_{XY}}{\sigma_x^2} = \frac{K^*_{XY}}{\sigma_x\sigma_y}\cdot\frac{\sigma_y}{\sigma_x} = r_в\frac{\sigma_y}{\sigma_x},$$

$$b_{X/Y} = \frac{K^*_{XY}}{\sigma_y^2} = \frac{K^*_{XY}}{\sigma_x\sigma_y}\cdot\frac{\sigma_x}{\sigma_y} = r_в\frac{\sigma_x}{\sigma_y}.$$

Тоді емпіричне рівняння прямих регресій із врахуванням цих формул можна записати в такому вигляді:

$$\bar{y}_x - \bar{y} = r_в\frac{\sigma_y}{\sigma_x}(x-\bar{x}), \tag{4.31}$$

$$\bar{x}_y - \bar{x} = r_в\frac{\sigma_x}{\sigma_y}(y-\bar{y}). \tag{4.32}$$

Наведемо еквівалентні (4.30) вирази вибіркового коефіцієнта кореляції.

Нехай $\{(x_i,y_i), i=\overline{1,n}\}$ — незгруповані дані двовимірної вибірки обсягом *n*. Величину

$$D_{yx} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \bar{y} - r_в \frac{\sigma_y}{\sigma_x}(x_i - \bar{x}) \right]^2 \qquad (4.33)$$

природно називати дисперсією спостережених значень y_i навколо емпіричної прямої регресії (4.31) *Y* на *X*, врахувавши при цьому, що геометрично різниця $y_i - \widehat{y}_i$ означає відхилення по ординаті точки (x_i, y_i) від точки $(x_i, \widehat{y}_i)$ прямої (4.31).

Можна довести, що

$$D_{yx} = \sigma_y^2(1 - r_в^2), \qquad r_в = \pm\sqrt{1 - \frac{D_{yx}}{\sigma_y^2}}, \qquad (4.30^{**})$$

де знак перед коренем вибирається у відповідності із знаком коефіцієнта регресії.

Аналогічно ввівши дисперсію спостережених значень x_i навколо емпіричної прямої регресії (4.32) *X* на *Y* за формулою

$$D_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{x}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\left[x_i - \bar{x} - r_в \frac{\sigma_x}{\sigma_y}(y_i - \bar{y}) \right]^2,$$

можна отримати рівність

$$r_в = \pm\sqrt{1 - \frac{D_{xy}}{\sigma_x^2}}. \qquad (4.30^{***})$$

Властивості вибіркового коефіцієнта кореляції

Властивість 1. Величина $r_в$ є безрозмірною, тобто вона не залежить від вибору одиниць виміру випадкових величин *X* та *Y*.

Властивість 2. Вибірковий коефіцієнт кореляції $r_в$ не перевищує за абсолютною величиною одиницю, тобто $|r_в| \le 1$.

Властивість 3. Вибірковий коефіцієнт кореляції $r_в = \pm 1$ тоді і тільки тоді, коли між випадковими величинами *X* та *Y* існує лінійний функціональний зв'язок.

Властивість 4. Якщо між випадковими величинами *X* і *Y* відсутній хоча б один із кореляційних зв'язків, то вибірковий коефіцієнт кореляції $r_в$ дорівнює нулю.

Властивість 5. Рівність $r_в = \pm 1$ є необхідною і достатньою умовою співпадань регресій *Y* на *X* і *X* на *Y*.

Із розглянутих властивостей $r_в$ можна зробити висновок про те, що вибірковий коефіцієнт кореляції є мірою тісноти (сили) лінійного кореляційного зв'язку між випадковими величинами X і Y. Справді, якщо $|r_в| = 1,$ то між X і Y існує лінійний функціональний зв'язок, а якщо $r_в = 0,$ то лінійний кореляційний зв'язок відсутній. А із формул (4.30**) і (4.30***) випливає, що у випадку збільшення $|r_в|$ до одиниці сила кореляційного зв'язку зростає, оскільки сума квадратів відхилень спостережених значень від прямих регресій прямує до нуля. Якщо ж $|r_в| \to 0,$ то сила кореляційного зв'язку зменшується, бо сума квадратів відхилень зростає.

Спрощений метод обчислення параметрів прямих регресії за кореляційною таблицею

У випадку рівновіддаленості як варіант x_i ($i = \overline{1,k}$), так і варіант y_j ($j = \overline{1,m}$) кореляційної табл.4.1 параметри рівнянь (4.31) та (4.32) можуть бути обчислені спрощеним методом (за аналогією із обчисленням зведених характеристик вибірки методом добутків). З цією метою здійснюється перехід до умовних варіант u_i та v_j за формулами

$$u_i = \frac{x_i - C_1}{h_1} \; (i = \overline{1,k}), \quad v_j = \frac{y_j - C_2}{h_2} \; (j = \overline{1,m}), \tag{4.34}$$

де $h_1 = x_{i+1} - x_i$ ($i = \overline{1,k-1}$), $h_2 = y_{j+1} - y_j$ ($j = \overline{1,m-1}$), C_1 та C_2 – хибні нулі варіант x_i та y_j відповідно.Зв'язок між числовими характеристиками початкових та умовних варіант визначається такими формулами:

$$\bar{x} = \bar{u}h_1 + C_1, \; \bar{y} = \bar{v}h_2 + C_2, \; \sigma_x = \sigma_u h_1, \; \sigma_y = \sigma_v h_2, \tag{4.35}$$

$$r_в = \frac{\frac{1}{n}(\sum n_{uv} uv) - \bar{u} \cdot \bar{v}}{\sigma_u \sigma_v}.$$

Оцінка достовірності емпіричних коефіцієнтів кореляції і регресії за даними вибірки

Будемо вважати, що генеральна сукупність має нормальний розподіл. Тоді для вибірки великого обсягу ($n \geq 50$) довірчі інтервали з надійністю γ при оцінюванні теоретичного коефіцієнта кореляції r, а також теоретичних коефіцієнтів регресії α і β в рівняннях (4.19), (4.20) мають такий вид:

$$r_в - t\frac{1 - r_в^2}{\sqrt{n}} < r < r_в + t\frac{1 - r_в^2}{\sqrt{n}}, \tag{4.36}$$

$$a_{Y/X} - t\frac{\sigma_y}{\sigma_x} \cdot \frac{1 - r_в^2}{\sqrt{n}} < \alpha_1 < a_{Y/X} + t\frac{\sigma_y}{\sigma_x} \cdot \frac{1 - r_в^2}{\sqrt{n}}, \tag{4.37}$$

$$b_{X/Y} - t\frac{\sigma_x}{\sigma_y}\cdot\frac{1-r_в^2}{\sqrt{n}} < \beta_1 < b_{X/Y} + t\frac{\sigma_x}{\sigma_y}\cdot\frac{1-r_в^2}{\sqrt{n}}. \qquad (4.38)$$

де t – корінь рівняння $2\Phi(t)=\gamma$, $\Phi(t)$ – функція Лапласа.

Наведеними довірчими інтервалами можна користуватись лише при великих n $(n \geq 50)$. Якщо ж обсяг двовимірної вибірки невеликий $(n < 30)$, то розподіл вибіркового коефіцієнта кореляції суттєво відрізняється від нормального. В цьому випадку для оцінки теоретичного коефіцієнта кореляції використовується знайдене Фішером перетворення, при якому вибірковий коефіцієнт кореляції прирівнюється до гіперболічного тангенса*) деякої величини z: $r_в = \text{th}z$, звідки

$$z = \frac{1}{2}\ln\frac{1+r_в}{1-r_в}.$$

Відомо, що випадкова величина z ($r_в$ розглядається для нефіксованої вибірки) вже при невеликих значеннях n $(n > 10)$ розподілена приблизно нормально, причому

$$M(z) \approx \frac{1}{2}\ln\frac{1+r}{1-r} + \frac{r}{2(n-1)}, \qquad D(z) \approx \frac{1}{n-3}.$$

Довірчий інтервал з надійністю γ для теоретичного коефіцієнта кореляції r двовимірної нормально розподіленої генеральної сукупності для невеликих обсягів $(n < 30)$ має такий вид

$$\left(\ \text{th}(z - t/\sqrt{n-3});\ \text{th}(z + t/\sqrt{n-3})\right)\ , \qquad (4.39)$$

де $z = \frac{1}{2}\ln\frac{1+r_в}{1-r_в}$ і знаходиться за табл. 10 додатків перетворення Фішера (z-перетворення), t – корінь рівняння $2\Phi(t) = \gamma$,

th(...) знаходиться за табл. 11 додатків оберненого перетворення Фішера.

Для коефіцієнтів регресій при невеликих n $(n < 30)$ розглянемо статистики (випадкові величини)

$$T_1 = \frac{\sigma_x\sqrt{n-2}}{\sigma_y\sqrt{1-r_в^2}}(a_{Y/X} - \alpha_1), \qquad T_2 = \frac{\sigma_y\sqrt{n-2}}{\sigma_x\sqrt{1-r_в^2}}(b_{X/Y} - \beta_1).$$

*) Гіперболічний тангенс $y = \text{th}x$ визначається рівністю $y = \text{th}x = (e^x - e^{-x})/(e^x + e^{-x})$; відповідна обернена гіперболічна функція має вид $x = \text{arth}y = \frac{1}{2}\ln\frac{1+y}{1-y}$.

Можна довести, що у випадку нормально розподіленої вибірки кожна із них має розподіл Ст'юдента із $k = n - 2$ ступенями вільності. Довірчі інтервали для теоретичних коефіцієнтів регресії α_1 та β_1 будуються стандартним чином:

$$a_{Y/X} - t_\gamma \frac{\sigma_y \sqrt{1 - r_в^2}}{\sigma_x \sqrt{n-2}} < \alpha_1 < a_{Y/X} + t_\gamma \frac{\sigma_y \sqrt{1 - r_в^2}}{\sigma_x \sqrt{n-2}}, \tag{4.40}$$

$$b_{X/Y} - t_\gamma \frac{\sigma_x \sqrt{1 - r_в^2}}{\sigma_y \sqrt{n-2}} < \beta_1 < b_{X/Y} + t_\gamma \frac{\sigma_x \sqrt{1 - r_в^2}}{\sigma_y \sqrt{n-2}}, \tag{4.41}$$

де $t_\gamma = t(\gamma, k)$ знаходиться за табл. 4 додатків розподілу Ст'юдента з числом ступенів вільності $k = n - 2$.

Завершуючи розгляд статистичного оцінювання коефіцієнтів теоретичних прямих регресій, вкажемо довірчі інтервали для вільних членів цих прямих:

$$\bar{y} - t_\gamma \sigma_y \sqrt{\frac{(n-1)(1 - r_в^2)}{n(n-2)}} < \alpha_0 < \bar{y} + t_\gamma \sigma_y \sqrt{\frac{(n-1)(1 - r_в^2)}{n(n-2)}}, \tag{4.42}$$

$$\bar{x} - t_\gamma \sigma_x \sqrt{\frac{(n-1)(1 - r_в^2)}{n(n-2)}} < \beta_0 < \bar{x} + t_\gamma \sigma_x \sqrt{\frac{(n-1)(1 - r_в^2)}{n(n-2)}}, \tag{4.43}$$

де $t_\gamma = t(\gamma, k)$ знаходиться за табл. 4 додатків розподілу Ст'юдента з числом ступенів вільності $k = n - 2$.

Відмітимо, що довірчі інтервали (4.42), (4.43) дають добрі наближення як у випадку малих вибірок $(n < 30)$, так і великих, оскільки при зростанні n розподіл Ст'юдента наближається до нормального.

При кореляційному аналізі необхідно оцінити достовірність стохастичного зв'язку між випадковими величинами, тобто з'ясувати, чи не зумовлена величина отриманого вибіркового коефіцієнта кореляції випадковим характером утворення вибірки. Для цього оцінюється значущість або істотність $r_в$. Перевіряється гіпотеза $H_0 : r = 0$ (гіпотеза некорельованості величин) проти однієї із альтернативних гіпотез $H_1 : r \neq 0$, або $H_1 : r < 0$, або $H_1 : r > 0$.

Для великих n критерієм перевірки слугує вибірковий коефіцієнт кореляції (для нефіксованої вибірки!). При цьому критична область має вид

$$|r_в| > u_\alpha \frac{1 - r_в^2}{\sqrt{n}}, \tag{4.44}$$

де u_α – значення, знайдене за табл. 3 додатків нормального розподілу для рівня значущості α. Якщо спостережене значення $r_в$ (для конкретної вибірки) потрапить в цю область, то гіпотезу H_0 відкидають.

Для невеликих n $(n < 30)$ використовується статистика

$$T = \frac{r_в\sqrt{n-2}}{\sqrt{1-r_в^2}}, \tag{4.45}$$

яка має розподіл Ст'юдента із $k = n-2$ ступенів вільності. Якщо $T_{\text{спост}} > t_\alpha(n-2)$, де $t_\alpha(n-2) = t_{\text{двост.кр}}(\alpha; n-2)$ — критична точка розподілу Ст'юдента, що знаходиться за табл. 5 додатків для двосторонньої критичної області (рівень значущості α вибирається за верхнім рядком), тоді основну гіпотезу про відсутність кореляційної залежності випадкових величин слід відкинути.

Отже, **якщо основна гіпотеза $H_0 : r = 0$ відкидається, тоді для оцінки точності $r_в$ будуються довірчі інтервали**.

Для перевірки гіпотези про силу кореляційного зв'язку, тобто для перевірки основної гіпотези $H_0 : r = r_0$ при альтернативній гіпотезі $H_1 : r \neq r_0$, використовується статистика

$$U = (z - z_0)\sqrt{n-3}, \tag{4.46}$$

де

$$z_0 = \frac{1}{2}\ln\frac{1+r_0}{1-r_0} + \frac{r_0}{2(n-1)}, \quad z = \frac{1}{2}\ln\frac{1+r_в}{1-r_в}. \tag{4.47}$$

Якщо гіпотеза H_0 є правильною, тоді випадкова величина U розподілена асимптотично нормально з параметрами (0,1). Тому використовувати цю статистику можна для немалих обсягів вибірки.

При використанні (двостороннього) критерія на рівні значущості α гіпотеза H_0 відкидається, якщо

$$U_{\text{спост}} > u_{\text{кр}},$$

де $u_{\text{кр}}$ — корінь рівняння

$$\Phi(u_{кр}) = (1-\alpha)/2. \tag{4.48}$$

Зауваження 1. Відмітимо, що стосовно перевірки гіпотези $H_0 : r = 0$ **зберігається умова про нормальність розподілу двовимірної генеральної сукупності**.

Зауваження 2. Побудову довірчих інтервалів для теоретичних коефіцієнтів кореляції та регресії слід здійснювати **після** того, як статистичну гіпотезу про некорельованість випадкових величин відкинуто на підставі даних вибірки.

Нелінійна парна кореляція. Вибіркове кореляційне відношення та його властивості

Розглянемо тепер випадок, коли хоча б одна із двох послідовностей точок

$$(x_1, \bar{y}_{x_1}), (x_2, \bar{y}_{x_2}), \ldots, (x_k, \bar{y}_{x_k}), \tag{4.49}$$

$$(y_1, \bar{x}_{y_1}), (y_2, \bar{x}_{y_2}), \ldots, (y_m, \bar{x}_{y_m}) \tag{4.50}$$

дає підстави зробити висновок про існування нелінійного кореляційного зв'язку між випадковими величинами X та Y. На основі цих статистичних даних потрібно оцінити параметри нелінійної регресії та силу кореляційної залежності.

Нехай, наприклад, конфігурація точок (4.49) дозволяє зробити припущення про наявність параболічної кореляції другого порядку. В цьому випадку вибіркове рівняння регресії Y на X слід шукати в такому вигляді

$$\bar{y}_x = a_2 x^2 + a_1 x + a_0, \tag{4.51}$$

де $a_i (i = 0,1,2)$ — невідомі параметри.

Користуючись методом найменших квадратів, можна отримати систему нормальних рівнянь:

$$\begin{cases} \overline{x^4} a_2 + \overline{x^3} a_1 + \overline{x^2} a_0 = \overline{\bar{y}_x x^2}, \\ \overline{x^3} a_2 + \overline{x^2} a_1 + \bar{x} a_0 = \overline{\bar{y}_x x}, \\ \overline{x^2} a_2 + \bar{x} a_1 + a_0 = \overline{\bar{y}_x}, \end{cases} \tag{4.52}$$

де

$$\overline{x^l} = \left(\sum_{i=1}^{k} x_i^l n_{x_i} \right) \Big/ n, \quad l = \overline{1,4}, \tag{4.53}$$

$$\overline{\bar{y}_x x^l} = \left(\sum_{i=1}^{k} \bar{y}_{x_i} x_i^l n_{x_i} \right) \Big/ n, \quad l = \overline{0,2}.$$

Оскільки $\overline{\bar{y}_x} = \bar{y},$ то розв'язавши третє рівняння системи (4.52) відносно a_0 і підставивши в (4.51), після простих перетворень отримаємо

$$\bar{y}_x = \bar{y} + a_1 (x - \bar{x}) + a_2 (x^2 - \overline{x^2}). \tag{4.51*}$$

Знайдені із системи рівнянь (4.52) параметри a_1 та a_2 підставимо в (4.51*); в підсумку отримується шукане рівняння регресії Y на X.

У випадку параболічної регресії (другого порядку) X на Y

$$\bar{x}_y = b_2 y^2 + b_1 y + b_0$$

невідомі параметри b_2, b_1, b_0 знаходяться як розв'язок такої системи нормальних рівнянь

$$\begin{cases} \overline{y^4} b_2 + \overline{y^3} b_1 + \overline{y^2} b_0 = \overline{\bar{x}_y y^2}, \\ \overline{y^3} b_2 + \overline{y^2} b_1 + \bar{y} b_0 = \overline{\bar{x}_y y}, \\ \overline{y^2} b_2 + \bar{y} b_1 + b_0 = \bar{x}. \end{cases}$$

Вище було розглянуто вибірковий коефіцієнт кореляції $r_в$, який характеризує тісноту **лінійного** кореляційного зв'язку між випадковими величинами X та Y.

Якщо ж криві регресії відрізняються від прямолінійної форми, то в якості міри зв'язку за даними двовимірної вибірки використовується вибіркове (статистичне або емпіричне) кореляційне відношення, запропоноване Пірсоном.

Нехай дані спостережень над кількісними ознаками X та Y зведені в кореляційну табл. 4.1. Тим самим можна вважати спостережені значення Y розбитими на групи, при цьому кожна група містить ті значення Y, які відповідають конкретним значенням X.

Для регресії Y на X можна ввести відповідності

$$\begin{array}{lll} x_1 & \bar{y}_{x_1} & n_{x_1}, \\ x_2 & \bar{y}_{x_2} & n_{x_2}, \\ \dots & \dots & \dots \\ x_k & \bar{y}_{x_k} & n_{x_k}, \end{array}$$

а для регресії X на Y відповідності

$$\begin{array}{lll} y_1 & \bar{x}_{y_1} & n_{y_1}, \\ y_2 & \bar{x}_{y_2} & n_{y_2}, \\ \dots & \dots & \dots \\ y_m & \bar{x}_{y_m} & n_{y_m}, \end{array}$$

де $\sum_{i=1}^{k} n_{x_i} = n, \ \sum_{i=1}^{m} n_{y_j} = n.$

Оскільки всі значення ознаки $Y(X)$ розбиті на групи, то дисперсію (загальну) ознаки $Y(X)$ можна представити у вигляді суми внутрігрупової і міжгрупової дисперсій (див. теорему 1.2):

$$D_{y\ \text{заг}} = D_{y\ \text{внгр}} + D_{y\ \text{міжгр}}, \tag{4.54}$$

$$D_{x\ \text{заг}} = D_{x\ \text{внгр}} + D_{x\ \text{міжгр}}. \tag{4.55}$$

Вибірковим (статистичним чи емпіричним) кореляційним відношенням Y до X називається відношення міжгрупового середнього квадратичного відхилення до загального середнього квадратичного відхилення ознаки Y:

$$\eta_{yx} = \sigma_{y\ \text{міжгр}} \big/ \sigma_{y\ \text{заг}} \tag{4.56}$$

або

$$\eta_{yx} = \sigma_{\bar{y}_x} \big/ \sigma_y, \tag{4.56*}$$

де

$$\sigma_{\bar{y}_x} = \sqrt{D_{y\ \text{міжгр}}} = \sqrt{\sum_i n_{x_i} (\bar{y}_{x_i} - \bar{y})^2 / n}, \qquad (4.57)$$

$$\sigma_y = \sqrt{D_{y\ \text{заг}}} = \sqrt{\sum_j n_{y_j} (y_j - \bar{y})^2 / n},$$

де n —обсяг вибірки, n_{x_i} — частота значення x_i ознаки X, n_{y_i} — частота значення y_j ознаки Y, $\bar{y}$ — загальна середня ознаки Y, $\bar{y}_{x_i}$ — умовна середня ознаки Y, що відповідає значенню x_i ознаки X.

Аналогічно визначається вибіркове кореляційне відношення X до Y:

$$\eta_{xy} = \sigma_{\bar{x}_y} / \sigma_x .$$

Зміст вибіркових кореляційних відношень визначається такими властивостями.

Властивість 1. Вибіркові кореляційні відношення невід'ємні і не перевищують одиниці:

$$0 \le \eta_{yx} \le 1, \quad 0 \le \eta_{xy} \le 1.$$

Властивість 2. Рівність нулю вибіркового кореляційного відношення є необхідною і достатньою умовою відсутності кореляційного зв'язку між випадковими величинами *X* та *Y*.

Властивість 3. Рівність одиниці вибіркового кореляційного відношення є необхідною і достатньою умовою функціональної залежності між випадковими величинами *X* та *Y*.

Властивість 4. Для однієї і тієї ж двовимірної вибірки випадкові кореляційні відношення не менші від абсолютної величини вибіркового коефіцієнта кореляції:

$$\eta_{yx} \ge |r_в|, \quad \eta_{xy} \ge |r_в|.$$

Властивість 5. Для того, щоб регресія була точно лінійною, необхідно і достатньо, щоб вибіркові кореляційні відношення дорівнювали вибірковому коефіцієнту кореляції:

$$\eta_{yx} = |r_в|, \quad \eta_{xy} = |r_в|.$$

Розглянуті властивості розкривають зміст вибіркових кореляційних відношень, які є мірою тісноти (сили) відповідного кореляційного зв'язку між випадковими величинами X та Y, причому форма зв'язку може бути довільною. Справді, якщо $\eta_{yx} \to 1$ або $\eta_{xy} \to 1$, то відповідний зв'язок випадкових величин стає більш тісним, оскільки міжгрупова дисперсія (дисперсія умовних середніх)

прямує до загальної дисперсії ознаки Y або X, тобто точки $(x_i, \bar{y}_{x_i})$, $i=\overline{1,k}$, $\left((\bar{x}_{y_j}, y_j)_{j=\overline{1,m}}\right)$ наближаються до відповідної вибіркової кривої регресії. Відмітимо також, що **недоліком вибіркових кореляційних відношень** є те, що вони не дозволяють судити, наскільки близько розташовані спостережені точки до кривої певного виду, наприклад, до параболи, гіперболи тощо. Це пояснюється тим, що при визначенні кореляційних відношень **конкретна** форма зв'язку не бралася до уваги.

Зупинимося тепер детальніше на деяких практично важливих видах нелінійної кореляційної залежності між X та Y. При цьому обмежимося розглядом односторонньої залежності.

За методикою оцінок параметрів парні нелінійні регресії розподіляють на два види: 1) нелінійні за факторами, але лінійні за невідомими параметрами, які підлягають оцінці; 2) нелінійні за факторами і параметрами. Регресії, нелінійні за факторами, але лінійні за оцінюваними параметрами, називаються **квазілінійними.**

Парну квазілінійну регресію можна записати в такому загальному вигляді: $\bar{y}_x = a\varphi(x) + b.$ Заміною $z = \varphi(x)$ квазілійна парна регресія приводиться до лінійної парної регресії $\bar{y}_z = az + b.$ Формули для оцінок параметрів набувають вигляду

$$\hat{a} = \frac{n\sum_{i=1}^{n} z_i y_i - \sum_{i=1}^{n} z_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} z_i^2 - \left(\sum_{i=1}^{n} z_i\right)^2}, \qquad \hat{b} = \bar{y} - \hat{a}\bar{z}.$$

В табл. 4.2 наведено ряд парних квазілінійних регресій.

Таблиця 4.2

Квазілінійна регресія	Заміна змінної	Лінійна регресія	Формули для оцінок параметрів
1	2	3	4
$\bar{y}_x = a/x + b$	$1/x = z$	$\bar{y}_z = az + b$	$\hat{a} = \dfrac{n\sum_{i=1}^{n} y_i/x_i - \sum_{i=1}^{n} 1/x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} 1/x_i^2 - \left(\sum_{i=1}^{n} 1/x_i\right)^2},$ $\hat{b} = \bar{y} - \hat{a}\bar{z},$ де $\bar{z} = \left(\sum_{i=1}^{n} 1/x_i\right)\Big/ n.$

Продовження табл. 4.2

	2	3	4
$\bar{y}_x = ae^x + b$	$e^x = z$	$\bar{y}_z = az + b$	$\hat{a} = \dfrac{n\sum\limits_{i=1}^{n} y_i e^{x_i} - \sum\limits_{i=1}^{n} e^{x_i} \sum\limits_{i=1}^{n} y_i}{n\sum\limits_{i=1}^{n} e^{2x_i} - \left(\sum\limits_{i=1}^{n} e^{x_i}\right)^2}$, $\hat{b} = \bar{y} - \hat{a}\bar{z}$, де $\bar{z} = \left(\sum\limits_{i=1}^{n} e^{x_i}\right) \Big/ n$.
$\bar{y}_x = a\sqrt{x} + b$	$\sqrt{x} = z$	$\bar{y}_z = az + b$	$\hat{a} = \dfrac{n\sum\limits_{i=1}^{n} y_i \sqrt{x_i} - \sum\limits_{i=1}^{n} \sqrt{x_i} \sum\limits_{i=1}^{n} y_i}{n\sum\limits_{i=1}^{n} x_i - \left(\sum\limits_{i=1}^{n} \sqrt{x_i}\right)^2}$, $\hat{b} = \bar{y} - \hat{a}\left(\sum\limits_{i=1}^{n} \sqrt{x_i}\right) \Big/ n$.
$\bar{y}_x = ax^2 + b$	$x^2 = z$	$\bar{y}_z = az + b$	$\hat{a} = \dfrac{n\sum\limits_{i=1}^{n} x_i^2 y_i - \sum\limits_{i=1}^{n} x_i^2 \sum\limits_{i=1}^{n} y_i}{n\sum\limits_{i=1}^{n} x_i^4 - \left(\sum\limits_{i=1}^{n} x_i^2\right)^2}$,
			$\hat{b} = \bar{y} - \hat{a}\left(\sum\limits_{i=1}^{n} x_i^2\right) \Big/ n$.
$\bar{y}_x = ax^3 + b$	$x^3 = z$	$\bar{y}_z = az + b$	$\hat{a} = \dfrac{n\sum\limits_{i=1}^{n} x_i^3 y_i - \sum\limits_{i=1}^{n} x_i^3 \sum\limits_{i=1}^{n} y_i}{n\sum\limits_{i=1}^{n} x_i^6 - \left(\sum\limits_{i=1}^{n} x_i^3\right)^2}$, $\hat{b} = \bar{y} - \hat{a}\left(\sum\limits_{i=1}^{n} x_i^3\right) \Big/ n$.

Розглянемо тепер регресію, нелінійну за параметрами, виду $\bar{y}_x = 1/(ax + b)$. Заміною $y^{(1)} = 1/y$ вона приводиться до лінійної регресії $\bar{y}_x^{(1)} = ax + b$. Оцінки параметрів *a* та *b* знаходяться за формулами

$$\hat{a} = \frac{n\sum_{i=1}^{n} x_i / y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} 1/y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}, \qquad \hat{b} = \left(\sum_{i=1}^{n} 1/y_i\right)\Big/ n - \hat{a}\left(\sum_{i=1}^{n} x_i\right)\Big/ n.$$

Більш складною (з точки зору заміни змінних) є регресія виду $\bar{y}_x = 1/(ae^{-x} + b)$. Для приведення її до лінійної регресії потрібно зробити заміну обох величин: $y^{(1)} = 1/y$, $z = e^{-x}$. Тоді отримається лінійна регресія $\bar{y}_z^{(1)} = az + b$, оцінки параметрів якої знаходяться за формулами

$$\hat{a} = \frac{n\sum_{i=1}^{n} 1/y_i e^{x_i} - \sum_{i=1}^{n} 1/e^{x_i} \sum_{i=1}^{n} 1/y_i}{n\sum_{i=1}^{n} 1/e^{2x_i} - \left(\sum_{i=1}^{n} 1/e^{x_i}\right)^2}, \qquad \hat{b} = \left(\sum_{i=1}^{n} 1/y_i\right)\Big/ n - \hat{a}\left(\sum_{i=1}^{n} 1/e^{x_i}\right)\Big/ n.$$

Розглянемо регресію $\bar{y}_x = a^x b$. Якщо $a > 0$, $b > 0$, то прологарифмувавши ліву і праву частини рівняння, отримаємо

$$\ln \bar{y}_x = x \ln a + \ln b.$$

Після заміни $\bar{y}_x^{(1)} = \ln \bar{y}_x$, $a_1 = \ln a$, $b_1 = \ln b$ одержується лінійна парна регресія $\bar{y}_x^{(1)} = a_1 x + b_1$. Для неї знаходяться оцінки параметрів a_1 і b_1 :

$$\hat{a}_1 = \frac{n\sum_{i=1}^{n} x_i \ln y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} \ln y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}, \qquad \hat{b}_1 = \left(\sum_{i=1}^{n} \ln y_i\right)\Big/ n - \hat{a}\left(\sum_{i=1}^{n} x_i\right)\Big/ n.$$

Тоді оцінки параметрів вихідної регресії матимуть такий вид: $\hat{a} = e^{\hat{a}_1}$, $\hat{b} = e^{\hat{b}_1}$.

Для регресії $\bar{y}_x = e^{a/x} b$ після логарифмування отримаємо

$$\ln \bar{y}_x = \frac{a}{x} + \ln b.$$

Внаслідок виконання заміни $y^{(1)} = \ln y$, $z = 1/x$, $b_1 = \ln b$ приходимо до парної лінійної регресії $\bar{y}_z^{(1)} = az + b_1$, оцінки параметрів якої знаходяться за формулами

$$\hat{a} = \frac{n\sum_{i=1}^{n} \frac{\ln y_i}{x_i} - \sum_{i=1}^{n} \frac{1}{x_i} \sum_{i=1}^{n} \ln y_i}{n\sum_{i=1}^{n} \frac{1}{x_i^2} - \left(\sum_{i=1}^{n} \frac{1}{x_i}\right)^2}, \qquad \hat{b}_1 = \frac{\sum_{i=1}^{n} \ln y_i}{n} - \hat{a}\frac{\sum_{i=1}^{n} \frac{1}{x_i}}{n}.$$

Тоді $\widehat{b} = e^{\widehat{b}_1}$.

Розглянемо регресію $\bar{y}_x = x^a b, \ \ b > 0$. Після логарифмування рівняння і заміни отримаємо лінійну парну регресію

$$\bar{y}_z^{(1)} = az + b_1, \text{ де } y^{(1)} = \ln y, \ \ z = \ln x, \ \ b_1 = \ln b.$$

Оцінки параметрів a і b_1 знаходяться за формулами

$$\widehat{a} = \frac{n\sum\limits_{i=1}^{n} \ln x_i \ln y_i - \sum\limits_{i=1}^{n} \ln x_i \sum\limits_{i=1}^{n} \ln y_i}{n\sum\limits_{i=1}^{n} \ln^2 x_i - \left(\sum\limits_{i=1}^{n} \ln x_i\right)^2},$$

$$\widehat{b}_1 = \left(\sum_{i=1}^{n} \ln y_i\right) \Big/ n - \widehat{a} \left(\sum_{i=1}^{n} \ln x_i\right) \Big/ n, \quad x_i > 0, \quad y_i > 0 \quad (i = \overline{1,n}).$$

Оцінка параметра b знаходиться таким чином

$$\widehat{b} = e^{\widehat{b}_1}, \text{ де } \widehat{b}_1 > 0.$$

Для показниково-степеневої парної регресії $\bar{y}_x = x^{ax} b \quad (x > 0, b > 0)$ після логарифмування і заміни отримаємо $\bar{y}_z^{(1)} = az + b_1$, де $y^{(1)} = \ln y$, $z = x \ln x, \ \ b_1 = \ln b.$ Параметри a і b_1 оцінюються за формулами

$$\widehat{a} = \frac{n\sum\limits_{i=1}^{n} x_i \ln x_i \ln y_i - \sum\limits_{i=1}^{n} x_i \ln x_i \sum\limits_{i=1}^{n} \ln y_i}{n\sum\limits_{i=1}^{n} x_i^2 \ln^2 x_i - \left(\sum\limits_{i=1}^{n} x_i \ln x_i\right)^2},$$

$$\widehat{b}_1 = \left(\sum_{i=1}^{n} \ln y_i\right) \Big/ n - \widehat{a} \left(\sum_{i=1}^{n} x_i \ln x_i\right) \Big/ n, \quad x_i > 0, \quad y_i > 0 \quad (i = \overline{1,n}).$$

Тоді оцінка параметра b визначається за формулою $\widehat{b} = e^{\widehat{b}_1}$.

Нарешті розглянемо нелінійну регресію $\bar{y}_x = \dfrac{x}{ax + b}$. Запишемо це рівняння в такому вигляді: $\dfrac{x}{\bar{y}_x} = ax + b.$ Заміною $y^{(1)} = \dfrac{x}{y}$ приводимо парну нелінійну регресію до парної лінійної $\bar{y}_x^{(1)} = ax + b.$ Параметри a і b оцінюються за такими формулами:

$$\widehat{a} = \frac{n\sum\limits_{i=1}^{n} x_i^2 / y_i - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} x_i / y_i}{n\sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2},$$

$$\widehat{b} = \left(\sum_{i=1}^{n} x_i / y_i \right) \Big/ n - \widehat{a} \left(\sum_{i=1}^{n} x_i \right) \Big/ n, \quad y_i > 0 \quad (i = \overline{1,n}).$$

Регресійний аналіз: парна лінійна регресія

Для описання, аналізу і прогнозування явищ і процесів в економіці використовують математичні моделі у вигляді рівнянь або функцій. Модель економічного об'єкта або виробничого процесу, відображаючи основні його властивості і абстрагуючись від другорядних, дозволяє передбачати його поведінку в певних умовах.

При використанні регресійних моделей результат дії економічної системи або об'єкта у вигляді одного або кількох результуючих показників представляється як функція факторів, що впливають на нього. Як правило, суттєвих факторів небагато, а несуттєвих — достатньо велике число, а тому останніми повністю нехтувати не можна.

В процесі виробництва і розподілу продукції спостерігають масові події, що багаторазово повторюються. Наприклад, серійне виготовлення деталей, вузлів, виробів, багаторазово повторювані акти продажу в продовольчих та промтоварних магазинах. Незважаючи на розвиток економіки і її окремих частин, протягом відносно невеликих часових відрізків і в межах окремих економічних підсистем спостерігається стабільність в умовах відбуття масових подій. По крайній мірі, особливо при прогнозуванні, передбачається можливість багаторазового повторення виробничої ситуації.

Об'єктом дослідження даного пункту є найпростіша регресійна модель — парна лінійна регресія, яку схематично можна зобразити таким чином:

$$Y = \alpha_1 x + \alpha_0 + \Delta, \tag{4.58}$$

де x — детермінований (невипадковий) фактор, перші два доданки — детермінована складова; α_1 — теоретичний коефіцієнт регресії, який показує, наскільки в середньому зміниться детермінована складова, якщо фактор зміниться на одиницю; Δ — випадкова складова, що уособлює сумарну дію різних випадкових факторів. Складова Δ називається **похибкою** (залишком, флуктуацією або збуренням).

Нехай для значень $x_1, x_2, \ldots, x_n$ (нефіксованих) спостерігаються або вимірюються відповідні моделі (4.58) значення Y: $y_1, y_2, \ldots, y_n$. В результаті отримуються n рівнянь

$$y_i = \alpha_1 x_i + \alpha_0 + \Delta_i, \quad i = \overline{1,n}. \tag{4.59}$$

Відносно випадкових величин $\Delta_1, \Delta_2, \ldots, \Delta_n$ введемо такі припущення (в літературі по економетрії вони називаються **специфікацією** моделі):

1) похибка $\Delta_i (i = \overline{1,n})$ є нормально розподілена випадкова величин з параметрами

$$M(\Delta_i) = 0, \quad D(\Delta_i) = \sigma^2, \quad i = \overline{1,n};$$

2) похибки Δ_i та Δ_j є незалежними для $i \neq j$, $i = \overline{1,n}$, $j = \overline{1,n}$.

Мета цього пункту полягає в розв'язуванні таких задач:

1) знаходження оцінок невідомих параметрів α_0 і α_1, вивчення їх властивостей і побудова для них довірчих інтервалів;

2) знаходження оцінки прогнозного значення $y^*(x)$ регресії $y(x)$ для довільного значення x і побудова для $y^*(x)$ довірчих інтервалів;

3) вказати критерій того, наскільки добре регресія описує дану систему спостережень.

Статистичні оцінки $\widehat{\alpha}_0, \widehat{\alpha}_1$ параметрів регресії α_0, α_1 вибираються таким чином, щоб емпіричні значення детермінованої складової $\widehat{y}_i = \widehat{a}_0 + \widehat{a}_1 x_i$ якомога ближче наближались до фактичних значень результуючої ознаки y_i. Оскільки випадкова складова моделі має нормальний розподіл, то оцінки, що володіють найкращими властивостями, отримуються з допомогою МНК.

Нехай δ_i — реалізація (можливе значення конкретної вибірки) випадкової величини $\Delta_i, i = \overline{1,n}$. Згідно із МНК мінімізуємо функцію

$$Q(\widehat{\alpha}_0, \widehat{\alpha}_1) = \sum_{i=1}^{n} \delta_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \widehat{\alpha}_1 x_i - \widehat{\alpha}_0)^2.$$

В результаті отримаємо систему нормальних рівнянь відносно невідомих $\widehat{\alpha}_0, \widehat{\alpha}_1$:

$$\begin{cases} \left(\sum_{i=1}^{n} x_i^2 \right) \widehat{\alpha}_1 + \left(\sum_{i=1}^{n} x_i \right) \widehat{\alpha}_0 = \sum_{i=1}^{n} x_i y_i, \\ \left(\sum_{i=1}^{n} x_i \right) \widehat{\alpha}_1 + n \widehat{\alpha}_0 = \sum_{i=1}^{n} y_i. \end{cases}$$

За формулою Крамера

$$\widehat{\alpha}_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}, \tag{4.60}$$

а з другого рівняння системи

$$\widehat{\alpha}_0 = \left(\sum_{i=1}^{n} y_i - \widehat{a}_1 \sum_{i=1}^{n} x_i \right) \Big/ n$$

або

$$\widehat{\alpha}_0 = \overline{y} - \widehat{\alpha}_1 \overline{x}. \tag{4.61}$$

Незміщена оцінка дисперсії випадкової регресійної моделі (4.58) має такий вид:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\delta_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \qquad (4.62)$$

де

$\hat{y}_i = \bar{y} + \hat{\alpha}_1(x_i - \bar{x})$.

Емпіричні оцінки $\sigma_{\hat{\alpha}_1}$ та $\sigma_{\hat{\alpha}_0}$ визначаються формулами

$$\hat{\sigma}_{\hat{\alpha}_1} = \hat{\sigma} \Big/ \sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}\,, \qquad (4.63)$$

$$\hat{\sigma}_{\hat{\alpha}_0} = \hat{\sigma}\sqrt{\sum_{i=1}^{n}x_i^2 \Big/ \left[\, n\sum_{i=1}^{n}(x_i - \bar{x})^2 \right]}\,, \qquad (4.64)$$

а довірчі інтервали для оцінок невідомих параметрів α_1 та α_0:

$$\hat{\alpha}_1 - t(\gamma; n-2)\hat{\sigma}_{\hat{\alpha}_1} < \alpha_1 < \hat{\alpha}_1 + t(\gamma; n-2)\hat{\sigma}_{\hat{\alpha}_1}\,, \qquad (4.65)$$

$$\hat{\alpha}_0 - t(\gamma; n-2)\hat{\sigma}_{\hat{\alpha}_0} < \alpha_0 < \hat{\alpha}_0 + t(\gamma; n-2)\hat{\sigma}_{\hat{\alpha}_0} \qquad (4.66)$$

де $t(\gamma; k)$ знаходиться за табл. 4 додатків.

В дійсності може виявитися, що фактор x не впливає на результуючу ознаку Y, що еквівалентно рівності $\alpha_1 = 0$, однак при цьому вибірковий коефіцієнт $\hat{\alpha}_1$, взагалі кажучи, відмінний від нуля. Для перевірки істотності відхилення $\hat{\alpha}_1$ від нуля використовується критерій значущості. Розглядається статистична гіпотеза $H_0 : \alpha_1 = 0$ і конкуруюча гіпотеза $H_1 : \alpha_1 \neq 0$. В якості статистичного критерію береться випадкова величина

$$t = \frac{\hat{\alpha}_1}{\hat{\sigma}_{\hat{\alpha}_1}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\hat{\sigma}\Big/\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad (4.67)$$

яка має розподіл Ст'юдента із $k = n-2$ ступенями вільності. Для фіксованої вибірки спостережене значення критерію $t_{\text{спост}}$, знайдене за формулою (4.67), порівнюється із критичною точкою $t_\alpha = t_{\text{двост.кр}}(\alpha; k)$, знайденою за табл. 5 додатків (рівень значущості α вибирається за верхнім рядком, тобто для двосторонньої критичної області). Якщо $|t_{\text{спост}}| < t_\alpha$, то з імовірністю помилки α приймається гіпотеза H_0, в протилежному випадку, тобто при $|t_{\text{спост}}| > t_\alpha$, приймається гіпотеза H_1. У випадку прийняття гіпотези H_0 фактор x виключається із моделі і тим самим приймається, що найкраще описання

системи спостережень здійснюється з допомогою середньої арифметичної $\hat{y} = \bar{y}$.

Знайдемо **довірчу зону рівняння регресії**. Теоретичне (підібране за емпіричними даними) лінійне рівняння регресії можна записати у вигляді $\hat{y} = \bar{y} + \hat{\alpha}_1(x - \bar{x})$, де $\bar{y}$ і $\hat{\alpha}_1$ зазнають впливу помилок, тобто є випадковими величинами для нефіксованої вибірки. Виявляється, що $\bar{y}$ і $\hat{\alpha}_1$ є некорельованими випадковими величинами.

Теоретичне значення $\hat{y}_k$ при заданому $x_k (k = \overline{1,n})$ із врахуванням рівності (4.61) має такий вид:

$$\hat{y}_k = \hat{\alpha}_0 + \hat{\alpha}_1 x_k = \bar{y} + \hat{\alpha}_1 (x_k - \bar{x}).$$

При цьому

$$\hat{\sigma}_{\hat{y}_k} = \hat{\sigma}\left\{ 1/n + (x_k - \bar{x})^2 \Big/ \sum_{i=1}^{n} (x_i - \bar{x})^2 \right\}^{1/2}, \quad k = \overline{1,n}, \tag{4.68}$$

де $\hat{\sigma}^2$ — точкова незміщена оцінка σ^2, визначена формулою (4.62).

Випадкова величина

$$t = \hat{y}_k / \hat{\sigma}_{\hat{y}_k}$$

має розподіл Ст'юдента із $k = n - 2$ ступенями вільності, тому за даною надійністю γ

$$P\left(-\hat{\sigma}_{\hat{y}_k} t(\gamma; k) < \hat{y}_k < \hat{\sigma}_{\hat{y}_k} t(\gamma; k) \right) = \gamma.$$

Знаючи надійність γ та число ступенів вільності $k = n - 2$, за табл. 4 додатків відшукаємо значення $t(\gamma; k)$.

Отже, довірчий інтервал для так званих **базисних даних** має такий вигляд

$$\left(\hat{y}_k - \hat{\sigma}_{\hat{y}_k} t(\gamma; k); \hat{y}_k + \hat{\sigma}_{\hat{y}_k} t(\gamma; k) \right). \tag{4.69}$$

Якщо для всіх $k = \overline{1,n}$, знайти значення лівих і правих кінців інтервала (4.69), а потім прямолінійними відрізками сполучити кожні дві сусідні тчоки, то в результаті отримається **довірча зона для базисних точок**. При цьому вважається, що послідовність x_i розташована в порядку зростання.

Розрахунки і перевірка достовірності отриманих оцінок коефіцієнтів регресії не є самоціллю, це тільки необхідний проміжний етап. Основне — це використання моделі для аналізу і прогнозу поведінки досліджуваного економічного явища. Прогноз здійснюється підстановкою фактора x в оцінку детермінованої складової:

$$\hat{y}_{\text{п}} = \bar{y} + \alpha_1 (x_{\text{п}} - \bar{x}). \tag{4.70}$$

Це точкова оцінка детермінованої складової в прогнозній точці $x_{\text{п}}$. Можна довести, що вираз (4.70) є найбільш точним (в середньоквадратичному) прогнозом. Для того, щоб виміряти точність цієї оцінки і побудувати довірчий інтервал для $\hat{y}_{\text{п}}$, необхідно знайти дисперсію точкової оцінки $\hat{y}_{\text{п}}(x_{\text{п}})$. При цьому потрібно врахувати ще одну невизначеність — розсіювання навколо прямої регресії. Оскільки $D(\Delta_i)=\sigma^2$, то рівнянню $y_{\text{п}}=\hat{\alpha}_1 x_{\text{п}}+\hat{\alpha}_0+\Delta_{\text{п}}$ відповідає дисперсія

$$D(y_{\text{п}_i})=\sigma^2(x_{\text{п}}-\bar{x})^2\Big/\sum_{i=1}^{n}(x_i-\bar{x})^2+\sigma^2\Big/n+\sigma^2.$$

Замінюючи σ його точковою оцінкою $\hat{\sigma}$, визначеною формулою (4.62), отримаємо за аналогією із вище розглянутим випадком (для базисних даних) такий довірчий інтервал

$$\left(\hat{y}_{\text{п}}-t(\gamma;k)\ \hat{\sigma}_{\hat{y}_{\text{п}}};\hat{y}_{\text{п}}+t(\gamma;k)\ \hat{\sigma}_{\hat{y}_{\text{п}}}\right), \tag{4.71}$$

де $t(\gamma;k)$ знаходиться за табл. 4 додатків для заданої надійності γ і числа ступенів вільності $k=n-2$,

$$\hat{\sigma}_{\hat{y}_{\text{п}}}=\hat{\sigma}\left[(x_{\text{п}}-\bar{x})^2\Big/\sum_{i=1}^{n}(x_i-\bar{x})^2+1/n+1\right]^{1/2}. \tag{4.72}$$

Залишається ще розглянути питання: наскільки добре рівняння регресії описує дану систему спостережень $\left\{(x_i,y_i),i=\overline{1,n}\right\}$. В якості міри якості такого узгодження служить **коефіцієнт детермінації**; при цьому за базу порівняння береться описання з допомогою середнього арифметичного.

Складемо наступні суми квадратів відхилень:

$S^2=\sum_{i=1}^{n}(y_i-\bar{y})^2$ — спостережених значень від їх середньої арифметичної;

$\hat{S}^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ — розрахункових (за моделлю) від середньої арифметичної спостережених значень;

$S_R^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ — спостережених від розрахункових значень.

Співвідношення між ними має такий вид:

$$S^2=\hat{S}^2+S_R^2. \tag{4.73}$$

Суму $\hat{S}^2$ природно назвати поясненою частиною варіації, оскільки значення $\hat{y}_i$ знаходиться за рівнянням регресії, а $\hat{S}_R$ — непоясненою (згідно із моделлю) частиною. Тому відповідно до рівності (4.73) чим "менше" значення S_R^2, тим точніше рівняння регресії описує спостережені дані.

Вибірковим коефіцієнтом детермінації називається відношення поясненої частини варіації до усієї варіації в цілому: $d = \widehat{S}^2 / S^2$. Враховуючи рівність (4.77), отримаємо:

$$d = \widehat{S}^2 / S^2 = (S^2 - S_R^2)/S^2 = 1 - S_R^2 / S^2 . \qquad (4.74)$$

Отже, чим "ближчий" коефіцієнт d до одиниці, тим кращим є опис рівнянням регресії емпіричних даних, якщо при цьому, зрозуміло, модель є методично правильною.

Статистичну значущість вибіркового коефіцієнта детермінації можна перевірити за допомогою F-статистики:

$$F = \frac{1-d}{d} \cdot \frac{m-1}{n-m}, \qquad (4.75)$$

де d — вибірковий коефіцієнт детермінації, n — число спостережень, m — число параметрів рівняння регресії.

Випадкова величина F розподілена за законом Фішера-Снедокора із параметрами $k_1 = m-1,\ k_2 = n-m.$

Оскільки в даному випадку $m = 2,$ то спостережене значення критерію обчислюється за формулою

$$F_{\text{спост}} = \frac{1-d}{d} \cdot \frac{1}{n-2}. \qquad (4.76)$$

Для заданої надійності γ знаходиться рівень значущості $\alpha = 1-\gamma$ і за табл. 7 додатків шукається критична точка $F_{\text{кр}}(\alpha; k_1, k_2).$ Якщо $F_{\text{спост}} < F_{\text{кр}}(\alpha; k_1, k_2),$ тоді з надійністю γ можна вважати, що вибірковий коефіцієнт детермінації статистично значущий і включений у регресію фактор достатньо пояснює стохастичну залежність показника.

ПРИКЛАДИ РОЗВ'ЯЗУВАННЯ ЗАДАЧ

Задача 4.1. Отримані статистичні дані десяти однотипних підприємств стосовно кофіцієнта використання основних засобів X і добового обсягу виробництва Y(тис. грн.):

x_i	0,4	0,45	0,5	0,55	0,6	0,65	0,7	0,75	0,8	0,9
y_i	2,3	2,4	3,2	3,8	4,5	4,7	5,1	5,6	7,2	8,4

1) скласти систему нормальних рівнянь і знайти коефіцієнти рівняння $\bar{y}_x = \alpha_1 x + a_0$ прямої регресії Y на X;

2) обчислити вибірковий коефіцієнт кореляції $r_{\text{в}}$;

3) перевірити правильність статистичної гіпотези $H_0 : r = 0$ для рівня значущості $\alpha = 0{,}05$;

4) якщо гіпотеза з 3) відхилена, тоді для рівня значущості $\alpha = 0{,}05$ перевірити гіпотезу $H_0 : r = r_{\text{в}}$ при конкуруючій гіпотезі $H_1 : r \neq r_{\text{в}}$;

5) побудувати довірчі інтервали для α_1, α_0, r для надійності $\gamma = 0{,}95$.

○ 1) Для знаходження коефіцієнтів системи нормальних рівнянь (4.23)

$$\begin{cases} \overline{x^2} a_1 + \bar{x} a_0 = \overline{xy}, \\ \bar{x} a_1 + a_0 = \bar{y}, \end{cases}$$

складемо розрахункову табл. 4.3,

Таблиця 4.3

i	x_i	y_i	x_i^2	$x_i y_i$	y_i^2
1	0,4	2,3	0,16	0,92	5,29
2	0,45	2,4	0,2025	1,08	5,76
3	0,5	3,2	0,25	1,6	10,24
4	0,55	3,8	0,3025	2,09	14,44
5	0,6	4,5	0,36	2,7	20,25
6	0,65	4,7	0,4225	3,055	22,09
7	0,7	5,1	0,49	3,57	26,01
8	0,75	5,6	0,5625	4,2	31,36
9	0,8	7,2	0,64	5,76	54,76
10	0,9	8,4	0,81	7,56	70,56
Σ	6,3	47,2	4,2	32,535	260,76

останній стовпець якої потрібний для обчислення σ_y. Використовуючи нижній рядок табл. 4.3, отримаємо (обсяг вибірки n=10):

$$\bar{x} = \sum_{i=1}^{10} x_i / n = 6{,}3/10 = 0{,}63\,; \quad \bar{y} = \sum_{i=1}^{10} y_i / n = 47{,}2/10 = 4{,}72\,;$$

$$\overline{x^2} = \sum_{i=1}^{10} x_i^2 / n = 4{,}2/10 = 0{,}42\,; \quad \overline{xy} = \sum_{i=1}^{10} x_i y_i / n = 32{,}535/10 = 3{,}2535\,;$$

$$\overline{y^2} = \sum_{i=1}^{10} y_i^2 / n = 260{,}76/10 = 26{,}076;$$

$$\begin{cases} 0{,}42 a_1 + 0{,}63 a_0 = 3{,}2535, \\ 0{,}63 a_1 + a_0 = 4{,}72. \end{cases}$$

Знайдемо єдиний розв'язок системи нормальних рівнянь згідно із формулами (4.24):

$$a_1 = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\overline{x^2} - (\bar{x})^2} = \frac{3{,}2535 - 0{,}63\cdot 4{,}72}{0{,}42 - (0{,}63)^2} = \frac{0{,}2799}{0{,}0231} \approx 12{,}117,$$

$$a_0 = \bar{y} - \bar{x}a_1 = 4{,}72 - 0{,}63\cdot 12{,}117 = -2{,}914.$$

Отже, емпіричне рівняння прямої регресії Y на X має такий вид:

$$\bar{y}_x = 12{,}117x - 2{,}914.$$

2) Вибірковий коефіцієнт кореляції $r_в$ знайдемо за формулою (4.30):

$$r_в = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\sigma_x\sigma_y} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\sqrt{\overline{x^2} - (\bar{x})^2}\sqrt{\overline{y^2} - (\bar{y})^2}} =$$

$$= \frac{3{,}2535 - 0{,}63\cdot 4{,}72}{\sqrt{0{,}42 - (0{,}63)^2}\sqrt{26{,}076 - (4{,}72)^2}} \approx 0{,}945.$$

3) Оскільки обсяг вибірки n=10 – малий, то використаємо критерій (статистику) (4.45). Для даної вибірки

$$T_{спост.} = \frac{r_в\sqrt{n-2}}{\sqrt{1-r_в^2}} = \frac{0{,}945\sqrt{8}}{\sqrt{1-(0{,}945)^2}} = 8{,}172.$$

В даному випадку критична область двостороння. Тому критичну точку $t_{двост.кр.}(\alpha; n\text{-}2) = t_\alpha(n\text{-}2)$ знайдемо за верхнім рядком табл. 5 додатків при α=0,05, k=n-2=8: $t_{0,05}(8)$=2,306.

Так як $t_{спост.}$=8,172> $t_{0,05}(8)$=2,306, то гіпотезу H_0 слід відкинути на користь конкуруючої гіпотези $H_1 : r \neq 0$, тобто можна зробити висновок про існування кореляційного зв'язку між X та Y.

4) Перевіримо гіпотезу $H_0 : r = r_0 = 0{,}94$ при конкуруючій гіпотезі $H_1 : r \neq r_0 = 0{,}94$.

Для обчислення спостереженого значення критерія (4.46) знайдемо спочатку значення z_0 та z за формулами (4.47), використавши табл. 10 додатків (перетворення Фішера) і здійснивши лінійну інтерполяцію:

$$z = \frac{1}{2}\ln\frac{1+r_в}{1-r_в} = \frac{1}{2}\ln\frac{1+0{,}945}{1-0{,}945} = (1{,}7380 + 1{,}8318)/2 = 1{,}7849,$$

$$z_0 = \frac{1}{2}\ln\frac{1+r_0}{1-r_0} + \frac{r_0}{2(n-1)} = \frac{1}{2}\ln\frac{1+0{,}94}{1-0{,}95} + \frac{0{,}94}{2(10-1)} =$$

$$= 1{,}7380 + 0{,}0522 = 1{,}7902.$$

Тоді

$$U_{спост.} = (z - z_0)\sqrt{n-3} = (1{,}7849 - 1{,}7902)\sqrt{7} = -0{,}0141.$$

Коренем рівняння (4.48) $\Phi(u_{кр}) = (1-0{,}05)/2 = 0{,}475$ (за табл.3 додатків) є $u_{кр.}$=1,96.

Оскільки $|U_{спост.}| = |-0{,}0141| < u_{кр.} = 1{,}96$, то нема підстав відкидати основну гіпотезу $H_0 : r = 0{,}94$.

5) Знову врахуємо малість обсягу вибірки. Тоді довірчі інтервали з надійністю γ=0,95 для оцінок теоретичного коефіцієнта регресії α_1 та вільного члена α_0 використаємо (4.40) та (4.42):

$$a_1 - t_\gamma \frac{\sigma_y\sqrt{1-r_в^2}}{\sigma_x\sqrt{n-2}} < \alpha_1 < a_1 + t_\gamma \frac{\sigma_y\sqrt{1-r_в^2}}{\sigma_x\sqrt{n-2}},$$

$$\bar{y} - t_\gamma\sigma_y\sqrt{\frac{(n-1)(1-r_в^2)}{n(n-2)}} < \alpha_0 < \bar{y} + t_\gamma\sigma_y\sqrt{\frac{(n-1)(1-r_в^2)}{n(n-2)}},$$

де $t_\gamma = t(\gamma, k)$ знаходиться за табл. 4 додатків розподілу Ст'юдента з числом ступенів вільності $k = n-2$.

Згідно із 1) та 2)

$$\sigma_x = \sqrt{\overline{x^2} - (\bar{x})^2} = \sqrt{0{,}42 - (0{,}63)^2} = 0{,}152,$$

$$\sigma_y = \sqrt{\overline{y^2} - (\bar{y})^2} = \sqrt{26{,}076 - (4{,}72)^2} = 1{,}949,$$

а за табл. 4 додатків $t_{0{,}95}(0{,}95;\ 8)$=2,306.

Підставивши дані, отримаємо

$$12{,}117 - 2{,}306\frac{1{,}949\sqrt{1-(0{,}945)^2}}{0{,}152\sqrt{8}} < \alpha_1 < 12{,}117 + 2{,}306\frac{1{,}949\sqrt{1-(0{,}945)^2}}{0{,}152\sqrt{8}},$$

$$4{,}72 - 2{,}306\cdot 1{,}949\sqrt{\frac{9[1-(0{,}945)^2]}{8}} < \alpha_0 < 4{,}72+{,}306\cdot 1{,}949\sqrt{\frac{9[1-(0{,}945)^2]}{8}},$$ або

остаточно

$$8{,}705 < \alpha_1 < 15{,}529,$$

$$3{,}161 < \alpha_0 < 6{,}279.$$

Для знаходження довірчого інтервалу з надійністю γ=0,95 для оцінки теоретичного коефіцієнта кореляції r використаємо (4.39):

$$\left(\ \text{th}(z - t/\sqrt{n-3});\ \text{th}(z + t/\sqrt{n-3})\right),$$

де $z = \frac{1}{2}\ln\frac{1+r_в}{1-r_в}$ і знаходиться за табл. 10 додатків перетворення Фішера (z-перетворення), t — корінь рівняння $2\Phi(t) = \gamma$, th(...) знаходиться за табл. 11 додатків оберненого перетворення Фішера.

Для даного випадку t=1,96, за табл. 10 додатків

$$z=\frac{1}{2}\ln\left(\frac{1+0{,}945}{1-0{,}945}\right)=(1{,}7380+1{,}8318)/2=1{,}7849,$$

$$z-t/\sqrt{n-3}=1{,}7849-1{,}96/\sqrt{7}=1{,}044,$$

$$z+t/\sqrt{n-3}=1{,}7849+1{,}96/\sqrt{7}=2{,}526,$$

за табл. 11 додатків з використанням інтерполяції

$$\text{th}(1{,}044)=0{,}7616+(0{,}7658-0{,}7616)\cdot 0{,}44=0{,}7635,$$

$$\text{th}(2{,}526)=0{,}9874.$$

Отже, шуканий довірчий інтервал має вид:

$$0{,}7635<r<0{,}9874 \quad \bullet$$

Зробимо зауваження відносно кореляційної табл. 4.1. Досі вважалося, що вона отримується внаслідок групування спостережених пар чисел (x_i, y_i) таких, для яких є однаковими або x_i, або y_i, або і x_i і y_i. Така ситуація є характерною для великих обсягів вибірки і дискретних величин *X* та *Y*. Проте для неперервних величин навіть для дуже великих обсягів вибірки вище згадане співпадання значень є рідкісним. Тим не менше, формування кореляційної таблиці і в таких випадках є корисним, оскільки внаслідок цього зменшується обсяг обчислень, а геометрична інтерпретація кореляційного поля стає більш наочною. Проте, як і у випадку одновимірної вибірки, "платою" за такі зручності є похибка числових характеристик вибірки, значення якої стають достатньо малими для великих *n* і вдалому виборі "нових" варіант.

Ідея побудови кореляційної таблиці у випадку великого обсягу вибірки досліджуваних **неперервних** ознак полягає в розбитті основного прямокутника, в якому знаходиться все кореляційне поле, на сукупність частинних прямокутників з однаковими довжинами сторін із наступним обчисленням частот попадання спостережених точок (пар чисел) в кожний із них. Ілюструє цю ідею, зокрема, наступна задача.

Задача 4.2. В табл. 4.4 наведені значення ваги (кг) x_i, $i=\overline{1,50}$, та росту (см) y_i, $i=\overline{1,50}$, випадково відібраних 50 чоловіків.

Потрібно: 1) скласти кореляційну таблицю спостережених даних; 2) знайти вибіркові рівняння регресії *Y* на *X* та *X* на *Y* за отриманою кореляційною таблицею.

Таблиця 4.4

x_i	69,80	71,46	82,41	77,08	63,77	66,78	88,15	73,60	77,54	56,84
y_i	165,8	175,7	176,0	170,5	163,0	170,4	172,1	181,0	180,2	157,4
x_i	72,00	73,03	73,98	73,70	69,90	72,33	78,63	83,46	80,46	82,73
y_i	171,8	171,7	169,2	166,2	161,7	162,7	167,7	170,5	171,7	184,5

x_i	81,85	71,01	76,69	68,97	65,41	75,28	83,33	75,73	70,43	73,01
y_i	173,8	175,5	179,8	167,8	163,7	177,4	176,4	178,1	170,2	179,2
x_i	77,24	64,48	77,93	72,51	72,08	70,02	79,91	67,22	68,42	76,20
y_i	168,1	159,1	171,3	179,6	166,9	162,9	182,8	166,3	174,4	177,8
x_i	68,41	69,92	58,18	79,16	73,34	71,66	84,04	77,41	73,36	72,94
y_i	171,1	170,6	153,9	175,3	170,6	173,8	181,4	174,2	170,4	163,0

○ 1) Спостережені ознаки x та y знаходяться в проміжках [56,84; 88,15] та [153,9; 184,5]. Основним прямокутником, в якому знаходиться все кореляційне поле, будемо вважати прямокутник $\{56 \le x \le 89; 153 \le y \le 185]$. Проміжок [56; 89] довжиною 33 розіб'ємо на 11 частинних інтервалів, а проміжок [153; 185] довжиною 32 — на 8 інтервалів. В результаті отримується 88 частинних прямокутників. Побудову кореляційної таблиці здійнимо в два етапи: на першому кроці підрахуємо число попадань спостереженої пари чисел у частинний прямокутник, помічаючи кожне попадання штрихом всередині відповідного прямокутника кореляційної таблиці; на другому кроці здійснимо перехід до "нових" варіант — середин кожного із частинних інтервалів. Результатом першого етапу є табл. 4.5, а другого — табл. 4.6.

Таблиця 4.5

X	Y								
	[153; 157)	[157; 161)	[161; 165)	[165; 169)	[169; 173)	[173; 177)	[177; 181)	[181; 185]	n_{x_i}
[56;59)	′	′							2
[59;62)									0
[62;65)		′	′						2
[65;68)			′	′	′				3
[68;71)			″	″	‴	′			8
[71;74)			″	″	⁗	‴	‴		15
[74;77)							⁗		4
[77;80)				″	″	″	′	′	8
[80;83)					′	″		′	4
[83;86)					′	′		′	3
[86;89]					′				1
n_{y_j}	1	2	6	7	14	9	7	4	50

Таблиця 4.6

X	Y								
	155	159	163	167	171	175	179	183	n_{x_i}
57,5	1	1							2
60,5									0
63,5		1	1						2

66,5			1	1	1				3
69,5			2	2	3	1			8
72,5			2	2	5	3	3		15
75,5							4		4
78,5				2	2	2	1	1	8
81,5					1	2		1	4
84,5					1	1		1	3
87,5					1				1
n_{y_i}	1	2	6	7	14	9	8	3	50

2) Для знаходження вибіркових рівнянь регресії Y на X та X на Y використаємо спрощений метод обчислення параметрів за кореляційною таблицею. В якості хибних нулів візьмемо моди варіант x та y: C_1=72,5, C_2=171. Варіанти x_i та y_j є рівновіддаленими із кроками h_1=3 та h_2=4 відповідно. За формулами (4.34) перейдемо до умовних варіант. В результаті отримаємо кореляційну табл. 4.7

Таблиця 4.7

	V								
U	-4	-3	-2	-1	0	1	2	3	n_u
-5	1	1							2
-4									0
-3		1	1						2
-2			1	1	1				3
-1			2	2	3	1			8
0			2	2	5	3	3		15
1							4		4
2				2	2	2	1	1	8
3					1	2		1	4
4					1	1		1	3
5					1				1
n_v	1	2	6	7	14	9	8	3	50

Величини $\bar{u}$, $\bar{v}$, σ_u, σ_v можна обчислити методом добутків, проте оскільки числа u_i, v_j малі, то знайдемо їх за означенням:

$$\bar{u} = \left(\sum n_u u\right)\Big/ n = [(-5)\cdot 2 + (-4)\cdot 0 + (-3)\cdot 2 + (-2)\cdot 3 + (-1)\cdot 8 + 1\cdot 4 + 2\cdot 8 + 3\cdot 4 + 4\cdot 3 + 5\cdot 1]/50 = 0{,}38;$$

$$\overline{u^2} = \left(\sum n_u u^2\right)\Big/ n = [(-5)^2\cdot 2 + (-4)^2\cdot 0 + (-3)^2\cdot 2 + (-2)^2\cdot 3 + (-1)^2\cdot 8 + 1^2\cdot 4 + 2^2\cdot 8 + 3^2\cdot 4 + 4^2\cdot 3 + 5^2\cdot 1]/50 = 4{,}68;$$

$$\sigma_u = \sqrt{\overline{u^2} - (\bar{u})^2} = \sqrt{4{,}68 - (0{,}38)^2} = 2{,}130;$$

$$\bar{v} = \left(\sum n_v v\right)\Big/ n = [(-4)\cdot 1 + (-3)\cdot 2 + (-2)\cdot 6 + (-1)\cdot 7 + 1\cdot 9 + 2\cdot 8 +$$
$$+ 3\cdot 3]/50 = 0{,}1;$$
$$\overline{v^2} = \left(\sum n_v v^2\right)\Big/ n = [(-4)^2\cdot 1 + (-3)^2\cdot 2 + (-2)^2\cdot 6 + (-1)^2\cdot 7 + 1^2\cdot 9 +$$
$$+ 2^2\cdot 8 + 3^2\cdot 3]/50 = 2{,}66;$$
$$\sigma_v = \sqrt{\overline{v^2} - (\bar{v})^2} = \sqrt{2{,}66 - (0{,}1)^2} = 1{,}628\,.$$

Для знаходження $r_в$ обчислимо $\left(\sum n_{uv} uv\right)\Big/ n$, врахувавши нульові значення варіант u та v:

$$\left(\sum n_{uv} uv\right)\Big/ n = [(-5)\cdot(-4)\cdot 1 + (-5)\cdot(-3)\cdot 1 + (-3)\cdot(-3)\cdot 1 +$$
$$+ (-3)\cdot(-2)\cdot 1 + (-2)\cdot(-2)\cdot 1 + (-2)\cdot(-1)\cdot 1 + (-1)\cdot(-2)\cdot 2 +$$
$$+ (-1)\cdot(-1)\cdot 2 + (-1)\cdot 1\cdot 1 + 1\cdot 2\cdot 4 + 2\cdot(-1)\cdot 2 + 2\cdot 1\cdot 2 + 2\cdot 2\cdot 1 +$$
$$+ 2\cdot 3\cdot 1 + 3\cdot 1\cdot 2 + 3\cdot 3\cdot 1 + 4\cdot 1\cdot 1 + 4\cdot 3\cdot 1]/50 = 2{,}2.$$

Тоді за формулами (4.35)

$$\bar{x} = \bar{u}h_1 + C_1 = 0{,}38\cdot 3 + 72{,}5 = 73{,}64;$$
$$\bar{y} = \bar{v}h_2 + C_2 = 0{,}1\cdot 4 + 171 = 171{,}4;$$
$$\sigma_x = \sigma_u \cdot h_1 = 2{,}13\cdot 3 = 6{,}39;$$
$$\sigma_y = \sigma_v \cdot h_2 = 1{,}628\cdot 4 = 6{,}512;$$
$$r_в = \frac{\frac{1}{n}\left(\sum n_{uv} uv\right) - \bar{u}\cdot\bar{v}}{\sigma_u \sigma_v} = \frac{2{,}2 - 0{,}38\cdot 0{,}1}{2{,}13\cdot 1{,}628} = 0{,}624\,.$$

Емпіричні рівняння (4.31) та (4.32) прямих регресій із врахуванням отриманих даних матимуть такий вид

$$\bar{y}_x - 171{,}4 = 0{,}624\frac{6{,}512}{6{,}39}(x - 73{,}64),$$
$$\bar{x}_y - 73{,}64 = 0{,}624\frac{6{,}39}{6{,}512}(y - 171{,}4)$$

або остаточно

$$\bar{y}_x = 0{,}636x + 124{,}57,$$
$$\bar{x}_y = 0{,}612y - 31{,}31.\ \bullet$$

Задача 4.3. 1) За даними кореляційної табл. 4.6 задачі 4.2 з надійністю $\gamma = 0{,}95$ знайти довірчі інтервали для теоретичних коефіцієнтів кореляції і регресії Y на X та X на Y, а також вільних членів прямих регресій. 2) Ігноруючи немалість обсягу вибірки, знайти довірчий інтервал для коефіцієнта кореляції, використавши формули для малих вибірок. Порівняти результати.

○ 1) Для знаходження інтервальних оцінок теоретичного коефіцієнта кореляції та параметрів теоретичних прямих регресії використаємо числові характеристики двовимірної вибірки, знайдені в задачі 4.2:

$$r_в = 0{,}624,\ \ \sigma_x = 6{,}39,\ \ \sigma_y = 6{,}512,$$

$$a_{Y/X} = 0{,}636,\ \ b_{X/Y} = 0{,}612,\ \ \bar{x} = 73{,}64,\ \ \bar{y} = 171{,}4.$$

Оскільки обсяг даної вибірки $n = 50$ достатньо великий, то 95%-ові довірчі інтервали для r, α_1 та β_1 будемо шукати за формулами (4.36)-(4.38), де $t = 1{,}96$ — корінь рівняння $2\Phi(t) = 0{,}95,$ знайдений за табл. 3 додатків.

Тобто

$$0{,}624 - 1{,}96 \cdot \frac{1-(0{,}624)^2}{\sqrt{50}} < r < 0{,}624 + 1{,}96 \cdot \frac{1-(0{,}624)^2}{\sqrt{50}},$$

$$0{,}636 - 1{,}96 \cdot \frac{6{,}512}{6{,}31} \cdot \frac{1-(0{,}624)^2}{\sqrt{50}} < \alpha_1 < 0{,}636 + 1{,}96 \cdot \frac{6{,}512}{6{,}39} \cdot \frac{1-(0{,}624)^2}{\sqrt{50}},$$

$$0{,}612 - 1{,}96 \cdot \frac{6{,}39}{6{,}512} \cdot \frac{1-(0{,}624)^2}{\sqrt{50}} < \beta_1 < 0{,}612 + 1{,}96 \cdot \frac{6{,}39}{6{,}512} \cdot \frac{1-(0{,}624)^2}{\sqrt{50}}.$$

Виконавши обчислення, отримаємо

$$0{,}455 < r < 0{,}793; \quad 0{,}464 < \alpha_1 < 0{,}808; \quad 0{,}446 < \beta_1 < 0{,}778.$$

Для знаходження довірчих інтервалів для α_0 та β_0 використаємо формули (4.42), (4.43), знайшовши попередньо $t_{0{,}95} = t(0{,}95;47)$ за табл. 4 додатків розподілу Ст'юдента: $t = 2{,}025$ (при цьому здійснюємо лінійну інтерполяцію для значень $k = 40$ та $k = 50$).

Тоді

$$171{,}4 - 2{,}025 \cdot 6{,}512\sqrt{\frac{49(1-0{,}624^2)}{50 \cdot 48}} < \alpha_0 < 171{,}4 + 2{,}025 \cdot 6{,}512\sqrt{\frac{49(1-0{,}624^2)}{50 \cdot 48}},$$

$$73{,}64 - 2{,}025 \cdot 6{,}39\sqrt{\frac{49(1-0{,}624^2)}{50 \cdot 48}} < \beta_0 < 73{,}64 + 2{,}025 \cdot 6{,}39\sqrt{\frac{49(1-0{,}624^2)}{50 \cdot 48}},$$

або остаточно $169{,}928 < \alpha_0 < 172{,}872, \quad 72{,}195 < \beta_0 < 75{,}085.$

2) Для інтервального оцінювання теоретичного коефіцієнта кореляції r використаємо тепер (4.39). Для $r_в = 0{,}624$ за табл. 8 додатків перетворення Фішера знайдемо $z = 0{,}725 + (0{,}74/4 - 0{,}725) \cdot 0{,}4 = 0{,}7316$, а для $\gamma = 0{,}95$ за табл. 3 додатків знайдемо корінь рівняння $2\Phi(t) = 0{,}95,$ тобто $t = 1{,}96.$ Тоді інтервал (4.39) набере такого виду

$$\left(\text{th}(0{,}7316 - 1{,}96/\sqrt{47}); \text{th}(0{,}7341 + 1{,}96/\sqrt{47})\right)$$

або $(\text{th}0{,}45;\text{th}1{,}02)$. Знайшовши за табл. 11 додатків межі останнього інтервалу, остаточно отримаємо шуканий довірчий інтервал $0{,}422 < r < 0{,}770$. Порівняння цього інтервалу із довірчим інтервалом, знайденим в першій частині, дозволяє зробити висновок про незначну похибку, яка виникає внаслідок використання інтервалу (4.39) при оцінюванні r для немалих вибірок. ●

Задача 4.4. Перевірити основну гіпотезу $H_0 : r = 0$ при конкуруючій гіпотезі $H_1 : r \neq 0$ за даними кореляційної табл.4.6 задачі 4.2 при рівні значущості $\alpha = 0{,}05$. При цьому використати критерії як для великих n, так і для малих n.

○ Оскільки альтернативна гіпотеза $H_1 : r \neq 0,$ то критична область — двостороння. Тому u_α в (4.44) є коренем рівняння (4.48), тобто $u_\alpha = 1{,}96$ (за табл. 3 додатків). Враховуючи, що $r_в = 0{,}624,\ \ n = 50,$ обчислимо в даному випадку праву частину нерівності (4.48):

$$u_\alpha \frac{1 - r_в^2}{\sqrt{n}} = 1{,}96 \cdot \frac{1 - (0{,}624)^2}{\sqrt{50}} = 0{,}169.$$

Але $r_в = 0{,}624 > 0{,}169,$ тому основну гіпотезу про некорельованість випадкових величин X та Y відкидаємо, допускаючи помилку лише у 5% випадків.

Здійснимо тепер перевірку H_0, використавши статистику (4.45) для невеликих обсягів вибірки. Для даної вибірки

$$T_{\text{спост}} = \frac{r_в \sqrt{n-2}}{\sqrt{1 - r_в^2}} = \frac{0{,}624\sqrt{48}}{\sqrt{1 - (0{,}624)^2}} = 5{,}533.$$

В даному випадку критична область також двостороння. Тому критичну точку $t_\alpha(n-2)$ знайдемо за верхнім рядком табл. 5 додатків при $\alpha = 0{,}05,$ $k = n - 2 = 48,$ здійснивши лінійну інтерполяцію:

$$t_{0{,}05}(48) = t_{0{,}05}(40) + [t_{0{,}05}(60) - t_{0{,}05}(40)] \cdot \frac{8}{20} =$$

$$= 2{,}021 + (2{,}000 - 2{,}021) \cdot \frac{8}{20} = 2{,}021 - 0{,}0084 = 2{,}0126.$$

Оскільки $T_{\text{спост}} = 5{,}533 > t_{0{,}05}(48) = 2{,}0126,$ то робимо висновок про існування кореляційного зв'язку між X та Y, тобто приймаємо конкуруючу гіпотезу. ●

Задача 4.5. Для рівня значущості $\alpha = 0{,}05$ перевірити гіпотезу $H_0 : r = 0{,}62$ при конкуруючій гіпотезі $H_1 : r \neq 0{,}62$ за даними кореляційної табл. 4.6 задачі 4.2.

○ Для обчислення спостереженого значення критерія (4.46) знайдемо спочатку значення z_0 та z (формули (4.47)), використавши табл. 10 додатків перетворення Фішера і лінійну інтерполяцію:

$$z = \frac{1}{2}\ln\frac{1+r_в}{1-r_в} = \frac{1}{2}\ln\frac{1+0{,}624}{1-0{,}624} = 0{,}725 + (0{,}74/4 - 0{,}725)\cdot 0{,}4 = 0{,}7316,$$

$$z_0 = \frac{1}{2}\ln\frac{1+r_0}{1-r_0} + \frac{r_0}{2(n-1)} = \frac{1}{2}\ln\frac{1+0{,}62}{1-0{,}62} + \frac{0{,}62}{2\cdot 49} = 0{,}725 + 0{,}0063 = 0{,}7313.$$

Тоді

$$U_{\text{спост}} = (z - z_0)\sqrt{n-3} = (0{,}7313 - 0{,}7313)\sqrt{47} = 0{,}0021.$$

Коренем рівняння $\Phi(u_{\text{кр}}) = (1 - 0{,}05)/2$ є $u_{\text{кр}} = 1.96$.

Оскільки $U_{\text{спост}} = 0{,}0021 < u_{\text{кр}} = 1{,}96$, то нема підстав відкидати основну гіпотезу про те, що теоретичний коефіцієнт кореляції $r = 0{,}62$. ●

Задача 4.6. Знайти емпіричне рівняння регресії Y на X за даними кореляційної табл. 4.8.

Таблиця 4.8

X	Y				
	[5; 7)	[7; 9)	[9; 11)	[11; 13]	n_x
[0,9; 1,1)	—	—	1	8	9
[1,1; 1,3)	2	—	6	—	8
[1,3; 1,5)	6	1	—	—	7
[1,5; 1,7)	1	9	1	—	11
[1,7; 1,9)	—	2	5	—	7
[1,9; 2,1]	—	—	1	7	8
n_y	9	12	14	15	n = 50

○ Для знаходження послідовності точок виду (4.49) перейдемо від кореляційної табл. 4.8 до табл. 4.9, для якої "новими" варіантами є середини відповідних частинних інтервалів, а останній стовпчик заповнимо, використовуючи означення умовної середньої:

$$\bar{y}_{x_i} = \left(\sum y_j n_{ij}\right) / n_{x_i}$$

(Наприклад, $\bar{y}_{x_1} = (10\cdot 1 + 12\cdot 8)/9 \approx 11{,}78$; $\bar{y}_{x_2} = (6\cdot 2 + 10\cdot 6)/8 = 9$).

Таблиця 4.9

X	Y					
	6	8	10	12	n_x	$\bar{y}_x$

1	—	—	1	8	9	11,78
1,2	2	—	6	—	8	9
1,4	6	1	—	—	7	6,29
1,6	1	9	1	—	11	8
1,8	—	2	5	—	7	9,43
2	—	—	1	7	8	11,75
n_y	9	12	14	15	$n = 50$	

Із табл. 4.9 отримуємо послідовність точок

(1; 11,78), (1,2; 9), (1,4; 6,29), (1,6; 8), (1,8; 9,43), (2; 11,75),

розташування яких в прямокутній декартовій системі координат дозволяє висунути припущення про наявність параболічної кореляції другого порядку.

Для знаходження коефіцієнтів системи нормальних рівнянь (4.52) складемо розрахункову табл. 4.10.

Таблиця 4.10

x_i	n_{x_i}	$\bar{y}_{x_i}$	$n_{x_i}x_i$	$n_{x_i}x_i^2$	$n_{x_i}x_i^3$	$n_{x_i}x_i^4$	$n_{x_i}\bar{y}_{x_i}$	$n_{x_i}\bar{y}_{x_i}x_i$	$n_{x_i}\bar{y}_{x_i}x_i^2$
1	9	11,78	9	9	9	9	106,02	106,02	106,02
1,2	8	9	9,6	11,52	13,824	16,589	72	86,4	103,68
1,4	7	6,29	9,8	13,72	19,208	26,891	44,03	61,642	86,3
1,6	11	8	17,6	28,16	45,056	72,090	88	140,8	225,28
1,8	7	9,43	12,6	22,68	40,824	73,483	66,01	118,818	213,872
2	8	11,75	16	32	64	128	94	188	376
$\sum$	50	—	74,6	117,08	191,912	326,053	470,06	701,68	1111,152

За формулами (4.53) із останнього рядка табл. 4.10 отримаємо:

$\bar{x} = 74{,}6/50 = 1{,}492;\ \overline{x^2} = 117{,}08/50 = 2{,}342;\ \overline{x^3} = 119{,}912/50 = 3{,}838;$

$\overline{x^4} = 326{,}053/50 = 6{,}521;\ \bar{y} = \overline{\bar{y}_x} = 470{,}06/50 = 9{,}401;\quad \overline{\bar{y}_x x} = 701{,}68/50 = 14{,}034;$

$\overline{\bar{y}_x x^2} = 1111{,}152/50 = 22{,}223.$

Тоді система рівнянь (4.52) набере такого виду

$$\begin{cases} 6{,}521a_2 + 3{,}838a_1 + 2{,}342a_0 = 22{,}223, \\ 3{,}838a_2 + 2{,}342a_1 + 1{,}492a_0 = 14{,}034, \\ 2{,}342a_2 + 1{,}492a_1 + a_0 = 9{,}401. \end{cases} \qquad (4.77)$$

Розв’язавши її, отримаємо

$$a_2 = 11{,}492,\quad a_1 = -34{,}010,\quad a_0 = 33{,}230.$$

Отже, шукане емпіричне рівняння регресії Y на X має такий вид

$$\bar{y}_x = 11{,}492x^2 - 34{,}01x + 33{,}23. \qquad (4.78)\ \bullet$$

Зауваження. При розв’язуванні системи нормальних рівнянь з’ясовується, що визначник матриці, складеної із коефіцієнтів при невідомих, близький до

нуля. Для таких систем характерним є те, що навіть дуже малі зміни коефіцієнтів приводять до **значних** змін розв'язків. Для ілюстрації цього положення розглянемо систему рівнянь

$$\begin{cases} 6{,}52a_2 + 3{,}84a_1 + 2{,}34a_0 = 22{,}22, \\ 3{,}84a_2 + 2{,}34a_1 + 1{,}49a_0 = 14{,}03, \\ 2{,}34a_2 + 1{,}49a_1 + a_0 = 9{,}4, \end{cases}$$

яка отримується із (4.77) внаслідок заокруглення коефіцієнтів до сотих. Її розв'язок

$$a_2 = 55{,}359, \quad a_1 = -162{,}958, \quad a_0 = 122{,}666$$

суттєво відрізняється від розв'язку системи (4.77) і визначає емпіричне рівняння регресії в такому вигляді:

$$\bar{y}_x = 55{,}359x^2 - 162{,}958x + 122{,}666. \qquad (4.79)$$

Табл. 4.11 характеризує "якість" отриманих емпіричних рівнянь регресії в плані порівняння розрахункових значень $y_x^{(4.78)}$ і $y_x^{(4.79)}$, отриманих із формул (4.78) та (4.79) відповідно, із спостереженими значеннями $\bar{y}_x^{(\text{спост})}$ (дані останнього стовпця табл. 4.9).

Таблиця 4.11

x_i	1	1,2	1,4	1,6	1,8	2
$y_x^{(4.78)}$	10,71	8,97	8,14	8,24	9,25	11,18
$y_x^{(4.79)}$	15,07	6,84	3,03	3,66	8,71	18,19
$\bar{y}_x^{(\text{спост})}$	11,78	9	6,29	8	9,43	11,75

На підставі цієї таблиці можна зробити висновок, що рівняння регресії (4.78) є кращим в порівнянні із (4.79). Разом з тим слід очікувати покращення результатів внаслідок збільшення числа знаків після коми у коефіцієнтів нормальної системи рівнянь. Іншими словами, **досягнення потрібної точності результатів без використання обчислювальної техніки є дуже трудоємкою задачею.**

Задача 4.7. Обчислити вибіркове кореляційне відношення за даними задачі 4.6.

○ За означенням (4.56*) $\eta_{yx} = \sigma_{\bar{y}_x} / \sigma_y$, тому потрібно знайти

$$\sigma_{\bar{y}_x} = \sqrt{\sum_i n_{x_i} (\bar{y}_{x_i} - \bar{y})^2 / n}, \quad \sigma_y = \sqrt{\sum_j (y_j - \bar{y})^2 n_{y_j} / n},$$

де за даними розв'язування задачі 4.6 $\bar{y} = 9{,}401$, $n = 50$, n_{x_i} та $\bar{y}_{x_i}$ наведені в двох останніх стовпцях табл. 4.9, а n_{y_j} — в останньому рядку цієї ж таблиці.

Підставивши вказані значення, отримаємо:

$$\sigma_{\bar{y}_x} = \left\{ \left[9(11{,}87-9{,}401)^2 + 8(9-9{,}401)^2 + 7(6{,}29-9{,}401)^2 + 11(8-9{,}401)^2 + \right.\right.$$

$$\left.\left. + 7(9{,}43-9{,}401)^2 + 8(11{,}75-9{,}401)^2 \right] /50 \right\}^{1/2} = 1{,}927;$$

$$\sigma_y = \left\{ \left[9(6-9{,}401)^2 + 12(8-9{,}401)^2 + 14(10-9{,}401)^2 + \right.\right.$$

$$\left.\left. + 15(12-9{,}401)^2 \right] /50 \right\}^{1/2} = 2{,}169.$$

Тоді $\eta_{yx} = 1{,}927/2{,}169 = 0{,}888$ і, отже, між випадковими величинами X та Y існує достатньо тісна кореляційна залежність. ●

Задача 4.8. Торгівельне підприємств має велику кількість філій, і керівництво цього підприємства вивчає питання про залежність y (річний товарообіг однієї філії, млн. грн.) від x (торгівельної площі, тис. м^2). Для дванадцяти філій за певний рік зафіксовані такі значення показників y та x:

i	1	2	3	4	5	6	7	8	9	10	11	12
y_i	2,93	5,27	6,85	7,01	7,02	8,35	4,33	5,77	7,68	3,16	1,52	3,15
x_i	0,31	0,98	1,21	1,29	1,12	1,49	0,78	0,94	1,29	0,48	0,24	0,55

На обсяг товарообігу впливають такі фактори: середньоденна інтенсивність потоку покупців, об'єм основних фондів, їх структура, середньосписочна чисельність працівників, площа підсобних приміщень тощо. Припускається, що в досліджуваній групі філій значення цих факторів приблизно однакові, тому вплив відмінностей їх значень на зміну обсягу товарообігу є незначним.

Потрібно:

1) знайти статистичні оцінки параметрів лінійного рівняння регресії;
2) обчислити вибірковий коефіцієнт детермінації;
3) для рівня значущості $\alpha = 0{,}05$ перевірити адекватність лінійної моделі емпіричним даним (з'ясувати статистичну значущість вибіркового коефіцієнта детермінації);
4) перевірити правильність статистичної гіпотези $H_0 : \widehat{\alpha}_1 = 0$ для рівня значущості $\alpha = 0{,}05$;
5) побудувати довірчі інтервали для параметрів рівняння регресії з надійністю $\gamma = 0{,}95$;
6) побудувати довірчу зону для базисних даних із надійністю $\gamma = 0{,}95$;
7) знайти прогнозне значення річного товарообігу для нової філії, торгівельна площа якої складає 1,82 тис. м^2, а також із надійністю

$\gamma = 0{,}95$ побудувати довірчий інтервал для цього прогнозного значення.

○ Для спрощення розрахунків при розв'язуванні задачі складемо табл. 4.12, заповнивши спочатку перших 5 стовпців.

№ п/п філії	x_i	y_i	x_i^2	$x_i y_i$	$x_i - \bar{x}$
1	2	3	4	5	6
1	0,31	2,93	0,0961	0,9083	-0,58
2	0,98	5,27	0,9604	5,1646	0,09
3	1,21	6,85	1,4641	8,2885	0,32
4	1,29	7,01	1,6641	9,0429	0,40
5	1,12	7,02	1,2544	7,8624	0,23
6	1,49	8,35	2,2201	12,4415	0,60
7	0,78	4,33	0,6084	3,3774	-0,11
8	0,94	5,77	0,8836	5,4238	0,05
9	1,29	7,68	1,6641	9,9072	0,40
10	0,48	3,16	0,2304	1,5168	-0,41
11	0,24	1,52	0,0576	0,3648	-0,65
12	0,55	3,15	0,3025	1,7325	-0,34
$\sum$	10,68	63,04	11,4058	66,0307	

1) Обсяг вибірки $n = 12$. Сума другого-п'ятого стовпців визначають коефіцієнти системи нормальних рівнянь. Тому згідно із формулою (4.60)

Таблиця 4.12

Продовження *таблиці 4.12*

№ п/п філії	$(x_i - \bar{x})^2$	$(x_i - \bar{x})y_i$	$(y_i - \bar{y})^2$	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	7	8	9	10	11
1	0,3364	-1,6994	5,3977	2,2245	0,4977
2	0,0081	0,4743	0,0003	5,7233	0,2055
3	0,1024	2,1920	2,5495	6,9243	0,0743

4	0,1600	2,8040	3,0860	7,3421	0,1103
5	0,0529	1,6146	3,1212	6,4544	0,3199
6	0,3600	5,0100	9,5896	8,3865	0,0013
7	0,0121	-0,4763	0,8525	4,6779	0,1217
8	0,0025	0,2885	0,2670	5,5144	0,0653
9	0,1600	3,0720	5,8889	7,3421	0,1142
10	0,1681	-1,2956	4,3819	3,1122	0,0023
11	0,4225	-0,9880	13,9375	1,8589	0,1149
12	0,1156	-1,071	4,4239	3,4778	0,1074
$\sum$	1,9006	9,9251	53,496		1,7348

$$\widehat{a}_1 = \frac{n\sum\limits_{i=1}^{n} x_i y_i - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i}{n\sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2} = \frac{12 \cdot 66{,}0307 - 10{,}68 \cdot 63{,}04}{12 \cdot 11{,}4058 - (10{,}68)^2} = 5{,}2221.$$

Із останнього рядка табл. 4.12 (2-го і 3-го стовпців) знаходимо

$$\bar{x} = \sum_{i=1}^{n} x_i / n = 10{,}68/12 = 0{,}89; \quad \bar{y} = \sum_{i=1}^{n} y_i / n = 63{,}04/12 = 5{,}2533,$$

тоді за формулою (4.61)

$$\widehat{\alpha}_0 = \bar{y} - \widehat{\alpha}_1 \bar{x} = 5{,}2533 - 5{,}2221 \cdot 0{,}89 = 0{,}6056.$$

Використавши дані 10-го стовпця табл. 4.12, послідовно отримаємо довірчі інтервали:

$k = 1$ $(2{,}2245 - 0{,}2125 \cdot 2{,}228; 2{,}2245 + 0{,}2125 \cdot 2{,}228) \leftrightarrow (1{,}7511; 2{,}2730)$,

$k = 2$ — (5,4486; 5,9980), $k = 3$ — (6,5805; 7,2681),
$k = 4$ — (6,9623; 7,7219), $k = 5$ — (6,1449; 6,7639),
$k = 6$ — (7,9019; 8,8711), $k = 7$ — (4,4001; 4,9557),
$k = 8$ — (5,2443; 5,7844), $k = 9$ — (6,9622; 7,7229),
$k = 10$ — (2,7277; 3,468), $k = 11$ — (1,3458; 2,3720),
$k = 12$ — (3,1256; 3,8300).

Для графічної побудови довірчої зони для базисних даних попередньо потрібно розсортувати значення $x_i (i = \overline{1, n})$ в порядку зростання і кожному значенню x_i поставити у відповідність дві точки з координатами $(x_i; y_i^*)$, $(x_i; y_i^{**})$, де y_i^* та y_i^{**} — лівий та правий кінець i-ого побудованого довірчого інтервалу (для базисних даних). Кожні дві сусідні точки $\left((x_i; y_i^*), (x_{i+1}; y_{i+1}^*) \right)$ та $\left((x_i; y_i^{**}), (x_{i+1}; y_{i+1}^{**}) \right)$ сполучаються прямолінійними відрізками.

7) Знайдемо прогноз значення річного товарообігу для нової філії, торгівельна площа якої складає 1,82 тис. м^2. Для цього в рівняння (4.80)

підставимо $x_{\text{п}} = 1{,}82$, тоді $\hat{y}_{\text{п}} = 5{,}2221 \cdot 1{,}82 + 0{,}6056 = 10{,}1098$. За формулою (4.72)

$$\hat{\sigma}_{\hat{y}_{\text{п}}} = \hat{\sigma}\left[(x_{\text{п}} - \bar{x})^2 \Big/ \sum_{i=1}^{n} (x_i - \bar{x})^2 + 1/n + 1 \right]^{1/2} =$$

$$= 0{,}4165\left[(1{,}82 - 0{,}89)^2 / 1{,}9006 + 1/12 + 1 \right]^{1/2} = 0{,}5166,$$

а за співвідношенням (4.71) довірчий інтервал для цього прогнозного значення з надійністю $\gamma = 0{,}95$ має вид

$$(10{,}1098 - 2{,}228 \cdot 0{,}5166; 10{,}1098 + 2{,}228 \cdot 0{,}5166)$$

або остаточно

$$8{,}9588 < y_{\text{п}} < 11{,}2608. \quad \bullet$$

ДОДАТКИ

Значення функції Лапласа $\Phi(x) = \dfrac{1}{\sqrt{2\pi}} \int_0^x e^{\frac{t^2}{2}} dt$

Таблиця 3

x	Соті долі x									
	0	1	2	3	4	5	6	7	8	9
0,0	0,00000	00399	00798	01197	01595	01994	02392	02790	03188	03586
0,1	03983	04380	04776	05172	05567	05962	06356	06749	07142	07535
0,2	07926	08317	08706	09095	09483	09871	10257	10642	11026	11409
0,3	11791	12172	12552	12930	13307	13683	14058	14431	14803	15173
0,4	15542	15910	16276	16640	17003	17364	17724	18082	18439	18793
0,5	19146	19497	19847	20194	20540	20884	21226	21566	21904	22240
0,6	22575	22907	23237	23565	23891	24215	24537	24857	25175	25490
0,7	25804	26115	26424	26730	27035	27337	27637	27935	28230	28524
0,8	28814	29103	29389	29673	29955	30234	30511	30785	31057	31327
0,9	31594	31859	32121	32381	32639	32894	33147	33398	33646	33891
1,0	34134	34375	34614	34850	35083	35314	35543	35769	35993	36214
1,1	36433	36650	36864	37076	37286	37493	37698	37900	38100	38298
1,2	38493	38686	38877	39065	39251	39435	39617	39796	39973	40147
1,3	40320	40490	40658	40824	40988	41149	41308	41466	41621	41774
1,4	41924	42073	42220	42364	42507	42647	42786	42922	43056	43189
1,5	43319	43448	43574	43699	43822	43943	44062	44179	44295	44408
1,6	44520	44630	44738	44845	44950	45053	45154	45254	45352	45449
1,7	45543	45637	45728	45818	45907	45994	46080	46164	46246	46327
1,8	46407	46485	46562	46638	46712	46784	46856	46926	46995	47062
1,9	47128	47193	47257	47320	47381	47441	47500	47558	47615	47670
2,0	47725	47778	47831	47882	47932	47982	48030	48077	48124	48169
2,1	48214	48257	48300	48341	48382	48422	48461	48500	48537	48574
2,2	48610	48645	48679	48713	48745	48778	48809	48840	48870	48899
2,3	48928	48956	48983	49010	49036	49061	49086	49111	49134	49158
2,4	49180	49202	49224	49245	49266	49286	49305	49324	49343	49361
2,5	49379	49396	49413	49430	49446	49461	49477	49492	49506	49520
2,6	49534	49547	49560	49573	49585	49598	49609	49621	49632	49643
2,7	49653	49664	49674	49683	49693	49702	49711	49720	49728	49736
2,8	49744	49752	49760	49767	49774	49781	49788	49795	49801	49807

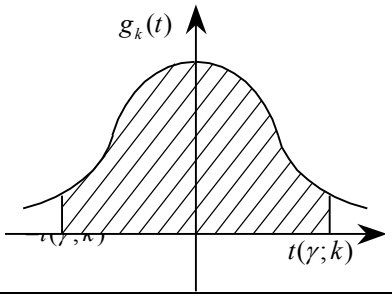
2,9	49813	49819	49825	49831	49836	49841	49846	49851	49856	49861
3,0	49865	49869	49874	49878	49882	49886	49889	49893	49897	49900
3,1	49903	49906	49910	49913	49916	49918	49921	49924	49926	49929
3,2	49931	49934	49936	49938	49940	49942	49944	49946	49948	49950
3,3	49952	49953	49955	49957	49958	49960	49961	49962	49964	49965
3,4	49966	49968	49969	49970	49971	49972	49973	49974	49975	49976
3,5	49977	49978	49978	49979	49980	49981	49981	49982	49983	49983
3,6	49984	49985	49985	49986	49986	49987	49987	49988	49988	49989
3,7	49989	49990	49990	49990	49991	49991	49992	49992	49992	49992
3,8	49993	49993	49993	49994	49994	49994	49994	49995	49995	49995
3,9	49995	49995	49996	49996	49996	49996	49996	49996	49997	49997
x	Десяті долі x									
	0		2		4		6		8	
4,	0,4999683		4999867		4999946		4999979		4999992	
5,	4999997									

Таблиця 4

Значення $t = t(\gamma; k)$,
що задовільняють рівнянню

$$P\left(|T| < t\right) = 2\int_0^t g_k(t)dt = \gamma,$$

де $g_k(t)$ — густина **розподілу Ст'юдента** (***t*-розподілу**), $k = n-1$ — число ступенів вільності



$k = n-1$	γ				
	0,9	0,95	0,98	0,99	0,999
1	6,314	12,706	31,821	63,657	636,619
2	2,920	4,303	6,965	9,925	31,599
3	2,353	3,182	4,541	5,841	12,924
4	2, 132	2,776	3,747	4,604	8,610
5	2,015	2,571	3,365	4,032	6,869
6	1,943	2,447	3,143	3,707	5,969
7	1,895	2,365	2,998	3,499	5,408
8	1,860	2,306	2,896	3,355	5,041
9	1,833	2,262	2,821	3,250	4,781
10	1,812	2,228	2,764	3,169	4,587
11	1,796	2,201	2,718	3,106	4,437
12	1,782	2,179	2,681	3,055	4,318
13	1,771	2,160	2,650	3,012	4,221
14	1,761	2,145	2,624	2,977	4,140
15	1,753	2,131	2,602	2,947	4,073
16	1,746	2,120	2,583	2,921	4,015
17	1,740	2,110	2,567	2,898	3,965
18	1,734	2,101	2,552	2,878	3,922

19	1,729	2,093	2,539	2,861	3,883
20	1,725	2,086	2,528	2,845	3,850
25	1,708	2,060	2,485	2,785	3,725
30	1,697	2,042	2,457	2,750	3,646
40	1,684	2,021	2,423	2,704	3,551
50	1,676	2,009	2,403	2,678	3,496
60	1,671	2,000	2,390	2,660	3,460
70	1,667	1,994	2,381	2,648	3,435
80	1,664	1,990	2,374	2,639	3,416
90	1,662	1,987	2,368	2,632	3,402
100	1,660	1,984	2,364	2,626	3,390
120	1,658	1,980	2,358	2,617	3,373
∞	1,645	1,960	2,326	2,576	3,291

Таблиця 5

Критичні точки розподілу Ст'юдента (*t*-розподілу)

Для двосторонньої критичної області критична точка $t_{\text{двост.кр}}(\alpha;k)=t_{\alpha}$ є коренем рівняння $\int\limits_{0}^{t_{\alpha}} g_k(t)dt=(1-\alpha)/2$; для односторонньої (правосторонньої) критичної області точка $t_{\text{правост.кр}}(\alpha;k)=t_{2\alpha}$ є коренем рівняння $\int\limits_{0}^{\infty} g_k(t)dt=(1-2\alpha)/2$, де $g_k(t)$ — густина розподілу Ст'юдента, $k=n-1$ — число ступенів вільності. Для лівосторонньої критичної області $t_{\text{лівост.кр}}(\alpha;k)=-t_{2\alpha}$.

Число ступенів вільності $k=n-1$	Рівень значущості α (**двостороння** критична область)					
	0,10	0,05	0,02	0,01	0,002	0,001
1	6,314	12,71	31,82	63,66	318,3	637,0
2	2,920	4,303	6,965	9,925	22,33	31,60
3	2,353	3,182	4,541	5,841	10,22	12,94
4	2,132	2,776	3,747	4,604	7,173	8,610
5	2,015	2,571	3,365	4,032	5,893	6,859
6	1,943	2,447	3,143	3,707	5,208	5,959
7	1,895	2,365	2,998	3,499	4,785	5,405
8	1,860	2,306	2,896	3,355	4,501	5,041
9	1,833	2,262	2,821	3,250	4,297	4,781
10	1,812	2,228	2,764	3,169	4,144	4,587
11	1,796	2,201	2,718	3,106	4,025	4,437
12	1,782	2,179	2,681	3,055	3,930	4,318
13	1,771	2,160	2,650	3,012	3,852	4,221
14	1,761	2,145	2,624	2,977	3,787	4,140
15	1,753	2,131	2,602	2,947	3,733	4,073
16	1,746	2,120	2,583	2,921	3,686	4,015
17	1,740	2,110	2,567	2,898	3,646	3,965
18	1,734	2,101	2,552	2,878	3,611	3,922
19	1,729	2,093	2,539	2,861	3,579	3,883
20	1,725	2,086	2,528	2,845	3,562	3,850
21	1,721	2,080	2,518	2,831	3,527	3,819
22	1,717	2,074	2,508	2,819	3,505	3,792

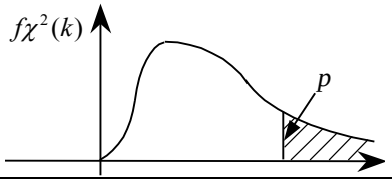
23	1,714	2,069	2,500	2,807	3,485	3,767
24	1,711	2,064	2,492	2,797	3,467	3,745
25	1,708	2,060	2,485	2,787	3,450	3,725
26	1,706	2,056	2,479	2,779	3,435	3,707
27	1,703	2,052	2,473	2,771	3,421	3,690
28	1,701	2,048	2,467	2,763	3,408	3,674
29	1,699	2,045	2,462	2,756	3,396	3,659
30	1,697	2,042	2,457	2,750	3,385	3,646
40	1,684	2,021	2,423	2,704	3,307	3,551
60	1,671	2,000	2,390	2,660	3,232	3,460
120	1,658	1,981	2,362	2,624	3,172	3,374
∞	1,645	1,960	2,326	2,576	3,090	3,291
Число ступенів вільності $k = n - 1$	0,05	0,025	0,01	0,005	0,001	0,0005
	Рівень значущості α (**одностороння** критична область)					

Таблиця 6

Значення

$$P\left(\chi^2(k) > \chi^2(p;k) \right) = p,$$

де k — число ступенів вільності



k	p							
	0,999	0,99	0,95	0,90	0,10	0,05	0,01	0,001
1	$0{,}157 \cdot 10^{-5}$	0,0002	0,004	0,02	2,71	3,84	6,63	10,83
2	0,002	0,02	0,10	0,21	4,61	5,99	9,21	13,82
3	0,02	0,12	0,35	0,58	6,25	7,82	11,34	16,27
4	0,09	0,30	0,71	1,06	7,78	9,49	13,28	18,47
5	0,21	0,55	1,15	1,61	9,24	11,07	15,08	20,51
6	0,38	0,87	1,64	2,20	10,65	12,59	16,81	22,46
7	0,60	1,24	2,17	2,83	12,02	14,06	18,48	24,32
8	0,86	1,65	2,73	3,49	13,36	15,51	20,09	26,12
9	1,15	2,09	3,33	4,17	14,68	16,92	21,67	27,88
10	1,48	2,56	3,94	4,87	15,99	18,31	23,21	29,59
11	1,83	3,05	4,58	5,58	17,28	19,68	24,72	31,26
12	2,21	3,57	5,23	6,30	18,55	21,03	26,22	32,91
13	2,62	4,11	5,89	7,04	19,81	22,36	27,69	34,53
14	3,04	4,66	6,57	7,79	21,06	23,69	29,14	36,12
15	3,48	5,23	7,26	8,55	22,31	25,00	30,58	37,70
16	3,94	5,81	7,96	9,31	23,54	26,30	32,00	39,25
17	4,42	6,41	8,67	10,09	24,77	27,59	33,41	40,79
18	4,90	7,02	9,39	10,86	25,99	28,87	34,81	42,31
19	5,41	7,63	10,12	11,65	27,20	30,14	36,19	43,82
20	5,92	8,26	10,85	12,44	28,41	31,41	37,57	45,31

21	6,45	8,90	11,59	13,24	29,62	32,67	38,93	46,80
22	6,98	9,54	12,34	14,04	30,81	33,92	40,29	48,27
23	7,53	10,20	13,20	14,85	32,01	35,17	41,64	49,73
24	8,08	10,86	13,85	15,66	33,19	36,42	43,98	51,18
25	8,65	11,52	14,61	16,47	34,38	37,65	44,31	52,62
26	9,22	12,20	15,37	17,29	35,56	38,89	45,64	54,05
27	9,80	12,88	16,15	18,11	36,74	40,11	46,96	55,48
28	10,39	13,56	16,93	18,94	37,92	41,34	48,28	56,89
29	10,99	14,26	17,71	19,77	39,09	42,56	49,59	58,30
30	11,59	14,95	18,49	20,60	40,26	43,77	50,89	59,70
40	17,92	22,16	26,51	29,05	51,81	55,76	63,69	73,40
50	24,67	29,71	34,76	37,69	63,17	67,51	76,15	86,66
100	61,92	70,07	77,93	82,36	118,50	124,34	135,81	149,45

Таблиця 7

Критичні точки $F_{\text{кр}}(\alpha;k_1,k_2)$ розподілу Фішера-Снедокора,
що задовільняють рівнянню $P\left[\ F > F_{\text{кр}}(\alpha;k_1,k_2)\right] = \alpha$ при $\alpha = 0{,}05$

k_2	k_1									
	1	2	3	4	5	6	8	12	24	∞
1	161,45	199,50	215,71	224,58	230,16	233,99	238,88	243,91	249,05	254,32
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19	4,38	3.52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
25	4,24	3,38	2,99	2,76	2,60	2.49	2,34	2.16	1,96	1,71
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62

40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,51
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,70	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,02	1,83	1,61	1,25
∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	1,00

Таблиця 8

Критичні значення $K_{n;\alpha}$ для статистики критерія Колмогорова

n	α				n	α			
	0,10	0,05	0,02	0,01		0,10	0,05	0,02	0,01
1	0,950	0,975	0,990	0,995	51	0,168	0,187	0,208	0,224
2	776	842	900	929	52	166	185	207	222
3	636	708	789	829	53	165	183	205	220
4	565	624	689	734	54	163	181	203	218
5	510	563	627	669	55	162	180	201	216
6	468	519	577	617	56	160	178	199	214
7	436	483	538	576	57	159	177	198	212
8	410	454	507	542	58	158	175	196	210
9	388	430	480	513	59	156	174	194	208
10	369	409	457	489	60	155	172	193	207
11	352	391	437	468	61	154	171	191	205
12	338	375	419	449	62	153	170	190	203
13	326	361	404	433	63	151	168	188	202
14	314	349	390	418	64	150	167	187	200
15	304	338	377	404	65	149	166	185	199
16	295	327	366	392	66	148	164	184	197
17	286	318	355	381	67	147	163	183	196
18	279	309	346	371	68	146	162	180	194
19	271	301	337	361	69	145	161	181	193
20	265	294	329	352	70	144	160	179	192
21	259	287	321	344	71	143	159	177	190
22	253	281	314	337	72	142	158	176	189
23	248	275	307	330	73	141	157	175	188

24	242	269	301	323	74	140	155	174	187
25	238	264	295	317	75	139	154	173	185
26	233	259	290	311	76	138	153	172	184
27	229	254	284	305	77	137	152	171	183
28	225	250	279	300	78	136	152	169	182
29	221	246	275	295	79	136	151	168	181
30	218	242	270	290	80	135	150	167	180
31	215	238	266	285	81	134	149	166	178
32	211	234	262	281	82	133	148	165	177
33	208	231	258	277	83	132	147	164	176
34	205	227	254	273	84	132	146	163	175
35	202	224	251	269	85	131	145	162	174
36	199	221	247	265	86	130	144	161	173
37	196	218	244	262	87	129	144	161	172
38	194	215	241	258	88	129	143	160	171
39	192	213	238	255	89	128	142	159	170
40	189	210	235	252	90	127	141	158	169
41	187	208	232	249	91	126	140	157	169
42	185	205	229	246	92	126	140	156	168
43	183	203	227	243	93	125	139	155	167
44	181	201	224	241	94	124	138	155	166
45	179	198	222	238	95	124	138	154	165
46	177	196	219	235	96	123	137	153	164
47	175	194	217	233	97	123	136	152	163
48	173	192	215	231	98	122	135	151	162
49	171	190	213	228	99	121	135	151	162
50	170	188	211	226	100	121	134	150	161

Таблиця 9

Асимптотичні критичні значення $n_{1-\alpha}$
для статистики Колмогорова і Смирнова

α	0,20	0,10	0,05	0,02	0,01
$n_{1-\alpha}$	1,073	1,224	1,358	1,517	1,628

Таблиця 10

Перетворення Фішера (*Z*-перетворення)

$$Z = \frac{1}{2}\ln\frac{1+r_в}{1-r_в}.$$

	Соті долі $r_в$

$r_в$	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0100	0,0200	0,0300	0,0400	0,0501	0,0601	0,0701	0,0802	0,0902
0,1	0,1003	0,1105	0,1206	0,1308	0,1409	0,1511	0,1614	0,1717	0,1820	0,1923
0,2	0,2027	0,2132	0,2237	0,2342	0,2448	0,2561	0,2661	0,2769	9,2877	0,2986
0,3	0,3095	0,3206	0,3317	0,3428	0,3541	0,3654	0,3769	0,3884	0,4001	0,4118
0,4	0,4236	0,4356	0,4477	0,4599	0,4722	0,4847	0,4973	0,5101	0,5230	0,5361
0,5	0,5493	0,5627	0,5763	0,5901	0,6042	0,6184	0,6328	0,6475	0,6625	0,6777
0,6	0,6931	0,7089	0,7250	0,7414	0,7582	0,7753	0,7928	0,8107	0,8291	0,8480
0,7	0,8673	0,8872	0,9076	0,9287	0,9505	0,9730	0,9962	1,0203	1,0454	1,0714
0,8	1,0986	1,1270	1,1568	1,1881	1,2212	1,2562	1,2933	1,3331	1,3758	1,4219
0,9	1,4722	1,5275	1,5890	1,6584	1,7380	1,8318	1,9459	2,0923	2,2976	2,6467

Таблиця 11

Обернене перетворення Фішера $r_в = \text{th}Z$.

	Соті долі Z									
Z	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0100	0,0200	0,0300	0,0400	0,0500	0,0599	0,0699	0,0799	0,0898
0,1	0997	1096	1194	1293	1391	1489	1586	1684	1781	1877
0,2	1974	2070	2165	2260	2355	2449	2543	2636	2729	2821
0,3	2913	3004	3095	3185	3275	3364	3452	3540	3627	3714
0,4	3800	3885	3969	4053	4136	4219	4301	4382	4462	4542
0,5	4621	4699	4777	4854	4930	5005	5080	5154	5227	5299
0,6	5370	5441	5511	5580	5649	5717	5784	5850	5915	5980
0,7	6044	6107	6169	6231	6291	6351	6411	6469	6527	6584
0,8	6640	6696	6751	6805	6858	6911	6963	7014	7064	7114
0,9	7163	7211	7259	7306	7352	7398	7443	7487	7531	7574
1,0	7616	7658	7699	7739	7779	7818	7857	7895	7932	7969
1,1	8005	8041	8076	8110	8144	8178	8210	8243	8275	8306
1,2	8337	8367	8397	8426	8455	8483	8511	8538	8565	8591
1,3	8617	8643	8668	8692	8717	8741	8764	8787	8810	8832
1,4	8854	8875	8896	8917	8937	8957	8977	8996	9015	9033
1,5	9051	9069	9087	9104	9121	9138	9154	9170	9186	9201
1,6	9217	9232	9246	9261	9275	9289	9302	9316	9329	9341
1,7	9354	9366	9379	9391	9402	9414	9425	9436	9447	9458
1,8	9468	9478	9488	9498	9508	9518	9527	9536	9545	9554
1,9	9562	9571	9579	9587	9595	9603	9611	9618	9626	9633
2,0	9640	9647	9654	9661	9668	9674	9680	9686	9693	9699
2,1	9704	9710	9716	9722	9727	9732	9738	9743	9748	9753
2,2	9757	9762	9767	9771	9776	9780	9785	9789	9793	9797

2,3	9801	9805	9809	9812	9816	9820	9823	9827	9830	9834
2,4	9837	9840	9843	9846	9849	9852	9855	9858	9861	9864
2,5	9866	9869	9871	9874	9876	9879	9881	9884	9886	9888
2,6	9890	9892	9894	9897	9899	9901	9903	9904	9906	9908
2,7	9910	9912	9914	9915	9917	9919	9920	9922	9923	9925
2,8	9926	9928	9929	9931	9932	9933	9935	9936	9937	9938
2,9	9940	9941	9942	9943	9944	9945	9946	9947	9948	9949

ЛІТЕРАТУРА

1. *Бочаров П. П., Печенкин А. В.* **Теория вероятностей. Математическая статистика.** — М.: Гардарика, 1998. — 328 с.
2. *Бугір М. К.* **Практикум з теорії імовірностей та математичної статистики.** — Тернопіль: ЦМДС, 1998. — 171 с.
3. *Булдик Г. М.* **Теория вероятностей и математичская статистика.** — Минск: Вышэйшая школа, 1989. — 285 с.
4. *Гмурман В. Е.* **Руководство к решению задач по теории вероятностей и математической статистике.** — М.: Высш. шк., 1979. — 400 с.
5. *Гмурман В. Е.* **Теория вероятностей и математическая статистика.** — М.: Высш. шк., 2004. — 479 с.
6. *Грубер Й.* **Эконометрия.** — Том 1. **Введение в эконометрию.** — К.: Астарта, 1996. — 397 с.
7. *Єрьоменко В. О., Шинкарик М. І.* **Теорія імовірностей.** — Тернопіль: Економічна думка, 2000. — 176 с.
8. *Єрьоменко В. О., Шинкарик М. І.* **Математична статистика.** — Тернопіль: Економічна думка, 2002. — 247 с.
9. *Єрьоменко В. О., Шинкарик М. І., Бабій Р. М., Процик А. І.* **Практикум з теорії імовірностей та математичної статистики /** Навчальний посібник для студентів економічних спеціальностей. — Тернопіль: Економічна думка, 2006. — 320 с.
10. *Жлуктенко В. І., Наконечний С. І.* **Теорія ймовірностей і математична статистика:** Навч. метод. посібник. У 2ч. — ч.І. Теорія ймовірностей. — К.: КНЕУ, 2000. — 304 с.
11. *Карасев А. И.* **Теория вероятностей и математическая статистика.** — М.: Статистика, 1977. — 279 с.
12. *Кибзун А.И., Горяинов Е.Р. Наумов А.В., Сиротин А.Н.* **Теория вероятностей и математическая статистика. Базовый курс с примерами и задачами** / Учебн. пособие. М.: ФИЗМАТЛИТ, 2002. — 224с.
13. *Коваленко И. Н., Гнеденко Б. В.* **Теория вероятностей.** — К.: Вища школа, 1990. — 328 с.
14. *Колемаев В. А., Калинина В. Н.* **Теория вероятностей и математическая статистика:** Учебник / Под ред. В. А. Колемаева. — М.: ИНФРА-М, 2000. — 302 с.
15. *Колемаев В. А., Староверов О. В., Турундаевский В. Б.* **Теория вероятностей и математическая статистика.** — М.: Высш. шк., 1991. — 400 с.

16. *Кремер М.Ш.* **Теория вероятностей и математическая статистика.** — М.: ЮНИТИ-ДАНА, 2002. — 543 с.

17. *Мармоза А.Т.* **Практикум з математичної статистики:** Навч. посібник. —К.: Кондор, 2004. —264с.

18. *Мюллер П., Найман П., Шторм Р.* **Таблицы по математической статистике.** — М.: Финансы и статистика, 1982. — 278 с.

19. *Румшиский Л. З.* **Математическая обработка результатов эксперимента.** — М.: Наука, 1971. — 192 с.

20. **Статистика підприємництва:** Навч. посібник. — Вашків П. Г., Пастер П. І., Сторожук В. П., Ткач Є. І. — К.: Слобожанщина, 1999. — 600 с.

21. *Черняк І.О., Обушна О.М., Ставицький А.В.* **Теорія ймовірностей та математична статистика:** Збірник задач: Навч. посібник. — К.: Т-во "Знання", КОО, 2001. — 199 с. — (Вища освіта XXI століття).

ЗМІСТ

МАТЕМАТИЧНА СТАТИСТИКА