

Київський національний університет імені Тараса Шевченка
механіко-математичний факультет
кафедра загальної математики

Данілов В. Я.

**СТАТИСТИЧНА
ОБРОБКА ДАНИХ**

навчальний посібник

Київ – 2019

Рецензенти
Бідюк П.І., доктор технічних наук, професор,
Кушніренко С.В., кандидат фізико-математичних наук, доцент

*Затверджено Вченою радою
механіко-математичного факультету
Київського національного університету імені Тараса Шевченка
(протокол № 1 від 16 вересня 2019 року)*

Данілов Володимир Якович

Статистична обробка даних: навчальний посібник. 2019. – 156 с.

Пропонований навчальний посібник містить методичні рекомендації для вивчення дисципліни «Математична статистика» і складається з двох розділів: «Основи математичної статистики» та «Статистичний аналіз даних з пакетом STATISTICA», що викладені в 7 главах. У посібнику наряду з теоретичним матеріалом наведено приклади, які ілюструють методи прикладної статистики. В кінці кожної глави дається перелік контрольних питань для самоперевірки. В додатку 1 наведено таблиці математичної статистики, а в додатку 2- розрахункові завдання з використанням програми STATISTICA.

Для студентів кваліфікаційного рівня “бакалавр”.

ПЕРЕДМОВА

Математична статистика являє собою галузь знань, спрямована на збір, групування, обробку та інтерпретацію статистичних даних, одержаних в результаті спостережень або в результаті спеціально поставлених експериментів. Її завдання полягає в проведенні спостереження, обробці та використанні одержаних статистичних даних для встановлення статистичних закономірностей ознаки чи ряду ознак певної сукупності елементів. Існують дві основні форми спостережень – статистична звітність (статистичний реєстр) та спеціально організоване спостереження (опитування, обстеження, облік, перепис), які використовують при проведенні статистичного дослідження.

На практиці значення числових характеристик випадкових величин, як правило, невідомі й підлягають визначенню на основі результатів спостережень. Знаючи дані спостережень, можна оцінити невідомі ймовірнісні характеристики (ймовірність випадкової події, функцію розподілу випадкової величини, математичне сподівання, дисперсію, середнє квадратичне відхилення, кореляцію та ін.), а потім висунути гіпотезу про те, що випадкове явище, яке спостерігається, може бути описане конкретною математичною моделлю – певним законом розподілу. Ця гіпотеза підлягає перевірці за допомогою математичних методів. І та математична модель, яка не суперечить фактичним статистичним даним, може описати математичні закономірності досліджуваного випадкового явища.

Дослідження закономірності розподілу включає в себе встановлення загального характеру розподілу, вирівнювання емпіричного розподілу за теоретичною кривою розподілу та встановлення відповідності емпіричного розподілу теоретичному за певними критеріями.

Для вивчення наявності зв'язку між досліджуваними ознаками, встановлення виду функціональної залежності (регресійної моделі), оцінки параметрів функції зв'язку та достовірності отриманих результатів використовують дисперсійний, кореляційний та регресійний аналіз.

Важливим питанням, яке постає при плануванні вибіркового обстеження, як забезпечити необхідну точність результатів, уникаючи при цьому зайвих витрат. Сьомий розділ посібника присвячено проблемі визначення оптимальної кількості вибіркового даних.

Наряду з вивченням теоретичних основ і методів математичної статистики важливими є навички практичного використання статистичного програмного забезпечення. Адже сучасна статистична обробка даних практично неможлива без певних комп'ютерних програм. У даному посібнику описуються деякі найпростіші прийоми обробки емпіричних даних в пакеті STATISTICA. В розділах 4-6 розглянуто алгоритми побудови лінійних регресійних моделей, елементи кластерного аналізу та аналізу часових рядів.

РОЗДІЛ 1

ОСНОВИ МАТЕМАТИЧНОЇ СТАТИСТИКИ

Глава 1. Статистичні розподіли вибірок та їх числові характеристики

1.1. Поняття вибіркового обстеження

Для задоволення інформаційних потреб у різних сферах людської діяльності часто проводять спланований, систематичний і науково організований збір певних даних про різноманітні суспільні, економічні, наукові явища, процеси і величини. Такі дослідження називаються **статистичними спостереженнями**, а їх метою є збір статистичних даних або статистичної інформації. Статистичне спостереження, яке охоплює всі елементи певної множини, називається **суцільним**, а коли досліджується лише її частина, то – **вибірковим**.

Під терміном **генеральна сукупність** будемо розуміти множину однорідних елементів, яким властиві деякі, як правило, **кількісні** ознаки, які виражаються числами – заробітна плата працівників підприємства, обсяг виготовленої чи реалізованої продукції, температура, вага, маса та ін. Кількість елементів генеральної сукупності становить **об'єм генеральної сукупності**. Кількісні ознаки елементів генеральної сукупності можуть бути дискретними або неперервними.

На практиці, з метою економії засобів та часу, дослідження генеральної сукупності проводять не повністю, а застосовують підбір характерних „ключів” або точок, просторових або часових обмежень, які називають **вибіркою** із генеральної сукупності. Іншими словами, вибірка – це деяка непорожня підмножина елементів, відібраних для вивчення із генеральної сукупності. Залежно від способу відбору елементів, які потрапили до вибірки, розрізняють **випадкову** і **невипадкову** вибірки. При випадковому відборі кожен елемент генеральної сукупності має певну, заздалегідь задану ймовірність бути відібраним. Це дає можливість досліднику забезпечити **репрезентативність** вибірки, тобто повторити властивості всієї досліджуваної множини у вибраній її частині та найбільш точно і повно характеризувати генеральну сукупність, з якої виділена вибірка.

Кількість елементів вибірки називають **об'ємом вибірки**. Якщо досліджувана ознака змінюється швидко (наприклад, середні добові температури певного місяця), то кількість вимірювань потрібно збільшити, і навпаки, при незначній зміні ознаки, надійний результат може бути одержаний при малому об'ємі вибірки. Задачу визначення об'єму

вибіркової сукупності можна розв'язати, використовуючи таблицю великих чисел (див. таблицю 8.1), або деякими розрахунковими методами, які будуть обговорюватися далі. В обох випадках кількість спостережень (вимірювань) визначається з урахуванням величини допустимої ймовірності, з якою планується зробити висновок, та величини точності проведення дослідю. Так, при рівні ймовірності $P=0,99$ і точності дослідю $\alpha=7\%$ число спостережень по таблиці великих чисел дорівнює 338. У випадку збільшення точності дослідю до $\alpha=2\%$ відповідний об'єм вибірки зростає до 4146.

Орієнтовний об'єм вибірки при $P=0,68$ можна знайти за формулами:

$$n = \sigma^2 / m^2, \quad (1)$$

$$n = V^2 / p^2, \quad (2)$$

тут σ – середнє квадратичне відхилення досліджуваної ознаки, m – похибка середнього арифметичного, V – коефіцієнт варіації, p – точність дослідю.

Так якщо варіація ознаки (скажімо, коливання температури) складає 8°C і точність дослідю $p = \pm 0,4^{\circ}\text{C}$, отримаємо такий орієнтовний об'єм вибірки $n = 8^2 / 0,4^2 = 400$.

Для знаходження орієнтовного об'єму вибірки при допустимому рівні ймовірності $P=0,95$ у формулах (1) та (2) величини σ та V потрібно домножити на коефіцієнт 1,96.

Математична статистика дозволяє за результатами аналізу вибірки характеризувати всю генеральну сукупність з відомим ступенем достовірності. Тому поняття генеральної сукупності трактується як сукупність усіх можливих спостережень, що можуть з'явитись при певному комплексі умов, які відповідають досліджуваній реальній ситуації чи конкретному стохастичному експерименту, де спостерігається випадкова величина ξ із функцією розподілу $F(x)$. Набір n незалежних спостережень, в результаті яких випадкова величина ξ набуває значень x_1, x_2, \dots, x_n називають вибіркою об'єму n із генеральної сукупності випадкової величини ξ із функцією розподілу $F(x)$. Відмітимо, що за вибіркою оцінюється розподіл випадкової величини, з генеральної сукупності якої отримали вибірку. При цьому визначаються не параметри генеральної сукупності, а межі, в яких вони знаходяться.

Для отримання випадкової вибірки можна йти двома шляхами: після того, як об'єкт спостереження випадково відібраний і над ним проведено спостереження, він може бути повернений або не повернений в генеральну сукупність. Стосовно типу відбору (з поверненням та без повернення), вибірки поділяються на **вибірки з поверненням**, коли відібраний об'єкт

перед відбором наступного повертається в генеральну сукупність та **вибірки без повернення**, коли відібраний об'єкт не повертається до генеральної сукупності. Частіше на практиці використовують випадковий відбір без повернення.

Варіантою кількісної ознаки X при реалізації вибірки будемо називати конкретні числові значення $X = x_i$, які вона набуває. Варіанта вибірки може бути спостереженою n_i число раз ($n_i \geq 1$), яке називається **частотою варіанти**.

Якщо k – кількість варіант x_i , що відрізняються своїм числовим значенням, а n_i відповідно їх частоти, то об'єм вибірки дорівнює

$$n = \sum_{i=1}^k n_i . \quad (3)$$

Одержаний при відборі зростаючий (спадаючий) числовий ряд варіант називають **варіаційним**. Дію впорядкування за величиною елементів вибірки називають **ранжуванням** статистичних даних.

Відносною частотою варіанти x_i називається відношення частоти n_i варіанти x_i до об'єму вибірки n і позначають її через W_i :

$$W_i = \frac{n_i}{n} . \quad (4)$$

Очевидно, що для кожної вибірки виконується співвідношення

$$\sum_{i=1}^k W_i = 1 . \quad (5)$$

У випадку, коли досліджувана ознака генеральної сукупності X є неперервною, то варіант буде багато. В цьому випадку варіаційний ряд являє собою певну кількість рівних (або нерівних) частинних інтервалів (або груп) варіант зі своїми частотами. Розміщенні у зростаючій послідовності такі частинні інтервали варіант, називають **інтервальним варіаційним рядом**. Для групування за кількісною ознакою слід встановити число відокремлених груп (інтервалів) та розмір інтервалу.

Число інтервалів k можна встановити аналітично за формулою Стерджеса:

$$k = 1 + 3,322 \lg n . \quad (6)$$

де n – кількість одиниць сукупності. Нижче наведено таблицю, яка визначає число інтервалів за об'ємом вибірки n :

об'єм вибірки	15-22	23-45	46-90	91-180	181-360	361-719
число інтервалів	5	6	7	8	9	10

При цьому, величина інтервалу h визначається за формулою

$$h = \frac{(x_{\max} - x_{\min})}{k} = \frac{R}{k} , \quad (7)$$

де x_{\max} , x_{\min} – відповідно найбільше та найменше значення ознаки X ; R – розмах вибірки.

Для зручності на практиці, як правило, розглядають інтервальні варіаційні ряди, в яких інтервали є рівними між собою. За початок першого інтервалу рекомендують брати величину $x_0 = x_{\min} - h/2$.

Зауважимо, що ті значення варіант, які різко відрізняються від варіант вибіркової сукупності і викликають сумнів дослідника, (вони, зазвичай, представляють собою крайні значення статистичної сукупності) можуть бути виключені із обробки, але при цьому відбраковка повинна бути статистично доведена. Існуючі критерії відбраковки обґрунтовані, як правило, на припущенні, що вибірка розподілена за нормальним або близьким до нього законом. Таким критерієм може бути τ критерій (див. таблицю 8. 2).

Позначимо через $\tau_{\text{розрах}} -$ розрахункове значення випадкової величини, $\tau_T -$ табличне критичне значення. Розрахункові значення для величин, які викликають сумнів і підлягають перевірці обчислюються за формулами:

для найменшого значення змінної величини варіаційного ряду ($\tilde{\alpha}_1$)

$$\tau_1 = \frac{(x_2 - x_1)}{(x_{n-1} - x_1)}; \quad (8)$$

для найбільшого значення змінної величини варіаційного ряду (x_n)

$$\tau_n = \frac{(x_n - x_{n-1})}{(x_n - x_2)}. \quad (9)$$

Якщо $\tau_{\text{розрах}} \geq \tau_T$, при об'ємі вибірки n , та рівні значимості α , то відповідне значення варіант x_1 або x_n слід відкинути; якщо $\tau_{\text{розрах}} < \tau_T$, то варіанта вважається репрезентативною і використовується для подальшої статистичної обробки.

Приклад 1.1. При спостереженні мікроклімату певної території в травні місяці було одержано ряд значень температур: $10,6^{\circ}\text{C}$; $12,7^{\circ}\text{C}$; $13,1^{\circ}\text{C}$; $13,5^{\circ}\text{C}$; $26,4^{\circ}\text{C}$. За допомогою τ критерію потрібно обґрунтувати відбраковку максимального значення температури $x_5 = 26,4^{\circ}\text{C}$.

Розв'язання. Скористаємося формулою (9):

$$\tau_5 = \frac{(x_n - x_{n-1})}{(x_n - x_2)} = \frac{(26,4 - 13,5)}{(26,4 - 12,7)} = \frac{12,9}{13,7} \approx 0,942.$$

За таблицею 8.2 при $n = 5$, $\alpha = 0,05$ і $\alpha = 0,01$ табличні (критичні) значення τ_T дорівнюють відповідно 0,807 та 0,916, що менше $\tau_{\text{розрах}} = \tau_5 = 0,942$. Отже, варіанта $x_5 = 26,4^{\circ}\text{C}$ виключається із статистичної вибірки.

Перейдемо до розгляду дискретних та інтервальних статистичних розподілів вибірки.

1.2. Дискретний статистичний розподіл вибірки та його числові характеристики

Дискретним статистичним розподілом вибірки будемо називати перелік варіант варіаційного ряду і відповідних їм частот (або відносних частот).

Даний розподіл зручно подати у табличній формі:

$X = x_i$	x_1	x_2	\dots	x_k
n_i	n_1	n_2	\dots	n_k
W_i	W_1	W_2	\dots	W_k

Такий дискретний статистичний розподіл вибірки можна представити емпіричною функцією розподілу (її називають ще функцією нагромадження відносних частот) $F^*(x)$.

Емпірична функція розподілу $F^*(x)$ та її властивості.

Функція $F^*(x)$ аргументу x , що визначає відносну частоту події $\{X < x\}$, тобто

$$F^*(x) = W(X < x) = \frac{n_x}{n} \quad (10)$$

називається **емпіричною функцією розподілу**, або функцією розподілу вибірки.

У формулі (10) n – об'єм вибірки; n_x – кількість варіант статистичного розподілу вибірки, значення яких менше за x ;

Наведемо **властивості функції розподілу**:

1. $0 \leq F^*(x) \leq 1$;
2. $F(x_{min}) = 0$, тут x_{min} - найменша варіанта варіаційного ряду;
3. $F(x) \mid$ при $x > x_{max} = 1$, тут x_{max} - найбільша варіанта варіаційного ряду;
4. $F(x)$ - неспадна функція, тобто, якщо $x_2 \geq x_1$ то $F(x_2) \geq F(x_1)$

Дискретний статистичний розподіл вибірки можна зобразити графічно у вигляді ламаної лінії, відрізки якої послідовно сполучають координати точок $(x_i; n_i)$ чи $(x_i; W_i)$.

Коли вздовж осі Ox відкладають частоти, то таку ламану лінію будемо називати **полігоном частот**, у другому, коли вздовж осі Ox відкладають відносні частоти, – **полігоном відносних частот**.

Приклад 1.2. В результаті спостережень за ранковою температурою повітря на певній території протягом перших двох декад грудня було одержано ряд значень температур: $7^{\circ}C$; $3^{\circ}C$; $3^{\circ}C$; $-1^{\circ}C$; $-1^{\circ}C$; $-2^{\circ}C$; $-5^{\circ}C$; $-3^{\circ}C$; $2^{\circ}C$; $0^{\circ}C$; $-2^{\circ}C$; $0^{\circ}C$; $2^{\circ}C$; $3^{\circ}C$; $0^{\circ}C$; $2^{\circ}C$; $0^{\circ}C$; $-2^{\circ}C$; $-3^{\circ}C$; $-2^{\circ}C$.

1. Знайти дискретний статистичний розподіл вибірки.

2. Побудувати $F^*(x)$ і зобразити її графічно.
3. Накреслити полігони частот і відносних частот.

Розв'язання. 1. Випишемо відповідний варіаційний ряд: $-5^{\circ}\text{C}; -3^{\circ}\text{C}; -3^{\circ}\text{C}; -2^{\circ}\text{C}; -2^{\circ}\text{C}; -2^{\circ}\text{C}; -2^{\circ}\text{C}; -1^{\circ}\text{C}; -1^{\circ}\text{C}; 0^{\circ}\text{C}; 0^{\circ}\text{C}; 0^{\circ}\text{C}; 0^{\circ}\text{C}; 2^{\circ}\text{C}; 2^{\circ}\text{C}; 2^{\circ}\text{C}; 3^{\circ}\text{C}; 3^{\circ}\text{C}; 3^{\circ}\text{C}; 7^{\circ}\text{C}$.

Бачимо, що температура -5°C спостерігалась 1 раз, тобто частота варіанти $x_1 = -5$ буде $n_1 = 1$, для варіанти $x_2 = -3$ частота $n_2 = 2$ і так далі. Отримаємо таку таблицю частот:

$X = x_i$ ($^{\circ}\text{C}$)	-5	-3	-2	-1	0	2	3	7
n_i	1	2	4	2	4	3	3	1

Переконаємось, що для даного розподілу має місце формула (3):

$$\sum_{i=1}^9 n_i = 1 + 2 + 4 + 2 + 4 + 3 + 3 + 1 = 20.$$

За формулою (4) знайдемо відносну частоту для кожної варіанти і отримаємо дискретний статистичний розподіл вибірки у вигляді таблиці:

$X = x_i$ ($^{\circ}\text{C}$)	-5	-3	-2	-1	0	2	3	7
W_i	0,05	0,1	0,2	0,1	0,2	0,15	0,15	0,05

2. За означенням емпіричної функції $F^*(x)$ знаходимо:

$$F^*(x) = W(X < x) = \begin{cases} 0, & x \leq -5, \\ 0.05, & -5 < x \leq -3, \\ 0.15, & -3 < x \leq -2, \\ 0.35, & -2 < x \leq -1, \\ 0.45, & -1 < x \leq 0, \\ 0.65, & 0 < x \leq 2, \\ 0.8, & 2 < x \leq 3, \\ 0.95, & 3 < x \leq 7, \\ 1, & x > 7. \end{cases}$$

Графічне зображення $F^*(x)$ представлено на рис. 1.1.

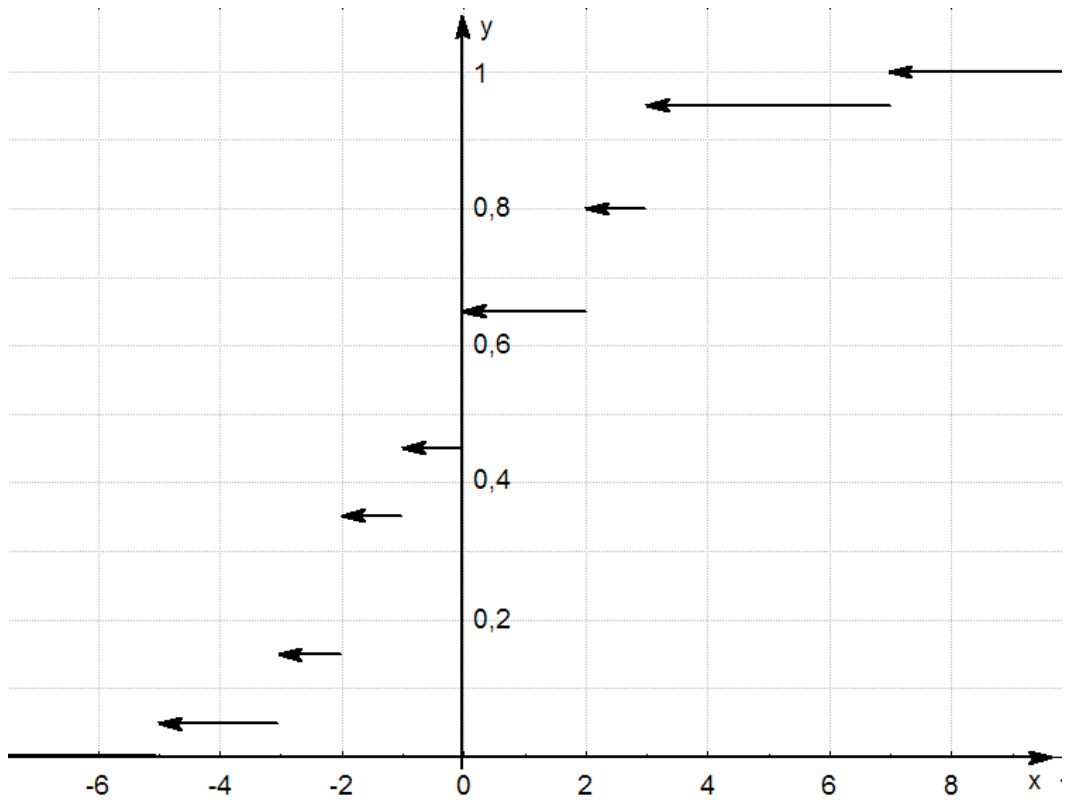


Рис. 1.1

Полігони частот та відносних частот зображено відповідно на рис. 1.2 та рис. 1.3.

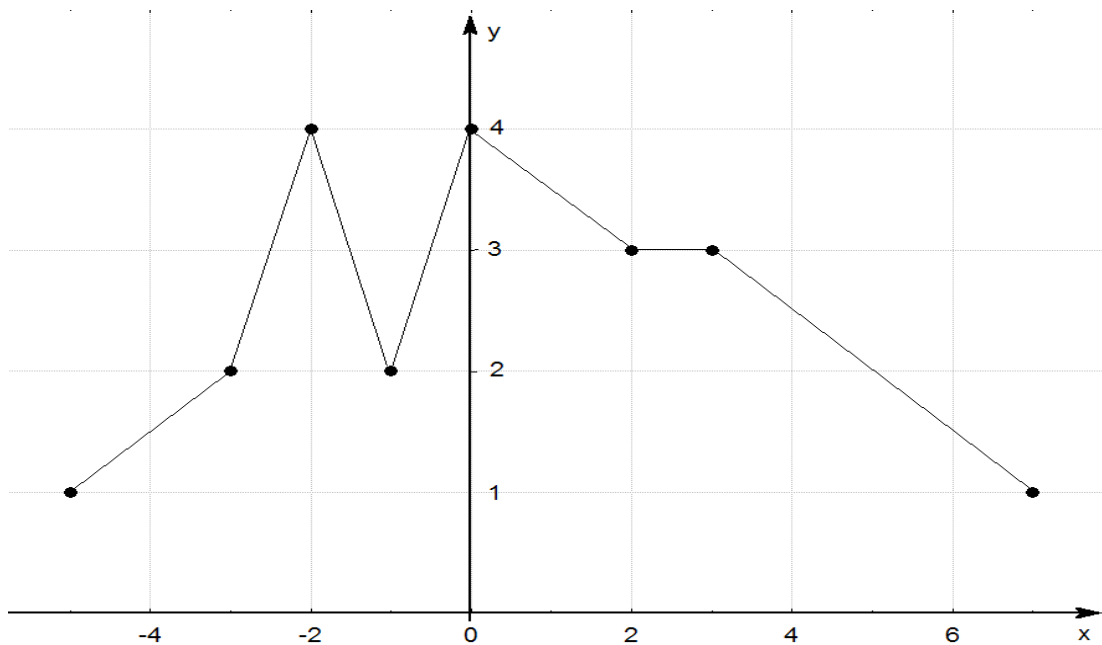


Рис. 1.2

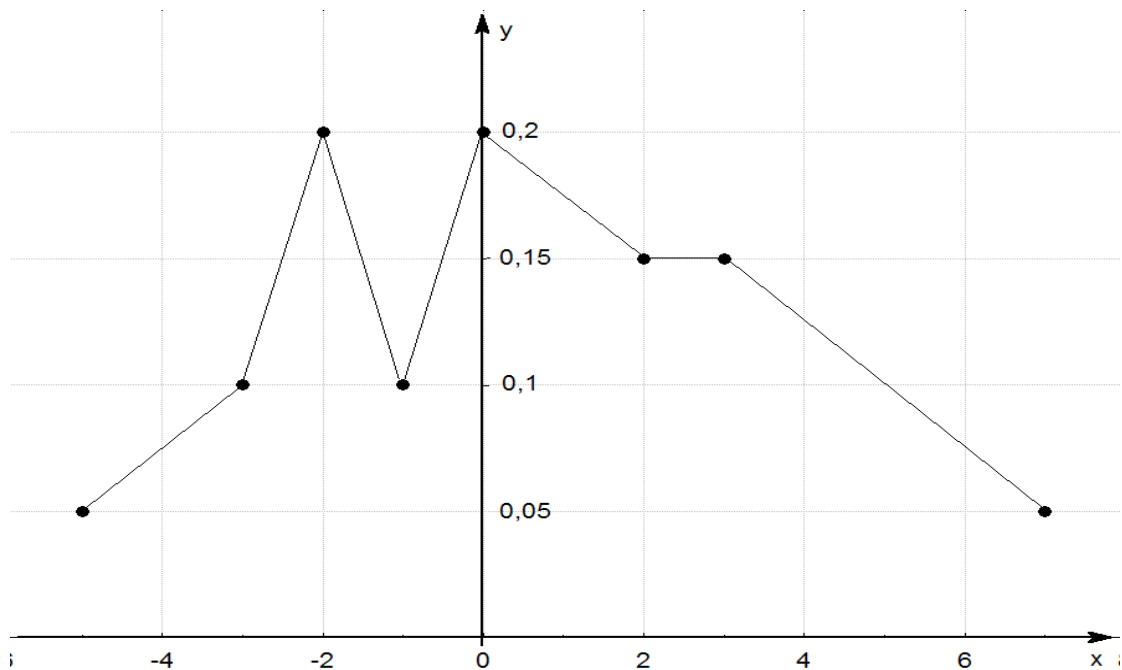


Рис. 1.3

Аналіз варіаційного ряду розподілу полягає у виявленні закономірностей зміни частот залежно від зміни кількісної ознаки, яка покладена в основу групування. Характеристику генеральної сукупності можна давати за значеннями числових параметрів, одержаних на основі вибірових даних. При аналізі варіаційних рядів вивчають такі групи показників:

- характеристики центру розподілу;
- характеристики розміру варіації;
- характеристики форми розподілу.

До статистичних характеристик центру розподілу відносяться такі вибірові характеристики: **вибірове середнє, вибірова мода, вибірова медіана.**

1. Вибірове середнє (\bar{x}_B).

Величину, яка визначається співвідношенням

$$\bar{x}_B = \frac{\sum_{i=1}^k x_i n_i}{n}, \quad (11)$$

називають **вибіровим середнім** дискретного статистичного розподілу вибірки.

У формулі (11): x_i – варіанта варіаційного ряду вибірки; n_i – частота цієї варіанти; n – об'єм вибірки ($n = \sum_{i=1}^k n_i$).

Очевидно, що формулу (11) можна переписати у вигляді $\bar{x}_B = \sum_{i=1}^k x_i W_i$, де W_i – відносна частота варіанти x_i . Остання формула нагадує формулу для математичного сподівання дискретної випадкової величини, що приймає значення x_i з ймовірностями W_i . Тому властивості вибіркового середнього аналогічні властивостям математичного сподівання випадкової величини. Наприклад, якщо всі варіанти збільшити (зменшити) в k разів, то вибіркоче середнє збільшиться (зменшиться) в k разів. Якщо всі варіанти збільшити (зменшити) на число c , то вибіркоче середнє збільшиться (зменшиться) на число c .

У випадку, коли всі варіанти варіаційного ряду з'являються у вибірці лише по одному разу ($n_i = 1$), то вибіркоче середнє є звичайним (незваженим) середнім арифметичним варіант і обчислюється згідно з формулою

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

Зауважимо, що вибіркоче середнє, яке обчислюється за формулою (11), є зваженим середнім арифметичним варіант з ваговими множниками – частотами відповідних варіант. Формулу (11) можна отримати при $p = 1$ із узагальненої формули для степеневих середніх

$$\bar{x}_p = \left(\frac{\sum_{i=1}^k x_i^p n_i}{n} \right)^{\frac{1}{p}}. \quad (12)$$

При вивченні закономірностей розподілу, як правило, користуються формулою (11). Якщо відомі не самі варіанти, а обернені до них величини, то для обчислення вибіркового середнього використовують формулу (12) при $p = -1$ (середнє гармонічне). Якщо досліджуються закономірності інтенсивності розвитку, то використовують формулу (12) при $p = 0$, яка після розкриття невизначеності при обчисленні границі, коли $p \rightarrow 0$, набуває вигляду $\bar{x}_0 = \sqrt[n]{x_1^{n_1} x_2^{n_2} \dots x_k^{n_k}}$ (середнє геометричне). При вивченні варіації користуються середнім квадратичним, що отримується при $p = 2$ з формули (12). Для степеневих середніх має місце властивість мажорантності $\bar{x}_{-1} < \bar{x}_0 < \bar{x}_1 < \bar{x}_2 < \bar{x}_3 < \dots$

Поряд з розглянутими аналітичними середніми в статистичному аналізі часто використовують порядкові середні: вибіркочі моду та медіану.

2. **Вибіркова мода (Mo^*)**. Модою дискретного статистичного розподілу вибірки називають ту її варіанту, що має найбільшу частоту (відносну частоту) появи. На графіку вибіркова мода відповідає максимальній ординаті й знаходиться на вершині варіаційної кривої (полігону частот або відносних частот).

Очевидно, мод може бути кілька. Якщо дискретний статистичний розподіл має одну моду, то він називається одномодальним. Якщо він має дві моди – двомодальним і т. д. Якщо мод більше однієї, то кажуть, що розподіл вибірки полімодальний.

3. **Вибіркова медіана (Me^*)**. Медіаною дискретного статистичного розподілу вибірки називають варіанту, яка ділить варіаційний ряд на дві (рівні за кількістю варіант) частини. Якщо варіаційний ряд складено за парною кількістю спостережень, то медіаною є півсума двох середніх варіант.

До статистичних показників розміру варіації відносяться такі числові характеристики вибірки: *відхилення варіант, вибіркова дисперсія, середнє квадратичне відхилення вибірки, розмах та вибірковий коефіцієнт варіації*.

4. **Відхилення варіант**. Різницю $x_i - \bar{x}_B$ називають відхиленням варіант.

Оскільки

$$\sum_{i=1}^k (x_i - \bar{x}_B) n_i = \sum_{i=1}^k x_i n_i - \sum_{i=1}^k \bar{x}_B n_i = n \cdot \bar{x}_B - n \cdot \bar{x}_B = 0,$$

то сума відхилень усіх варіант варіаційного ряду вибірки завжди дорівнює нулеві.

5. **Вибіркова дисперсія (D_B)**. Мірою розсіювання варіант вибірки відносно вибіркової середньої величини \bar{x}_B є вибіркова дисперсія.

Дисперсія вибірки – це середнє арифметичне квадратів відхилень варіант відносно вибіркової середньої величини \bar{x}_B , яке визначається формулою

$$D_B = \frac{\sum_{i=1}^k (x_i - \bar{x}_B)^2 n_i}{n}. \quad (13)$$

На практиці дисперсію вибірки обчислюють за спрощеною формулою

$$D_B = \frac{\sum_{i=1}^k x_i^2 n_i}{n} - (\bar{x}_B)^2. \quad (14)$$

6. **Середнє квадратичне відхилення вибірки (σ_B).** При знаходженні вибіркової дисперсії D_B відхилення підноситься до квадрату, а отже, змінюється одиниця виміру ознаки X . Тому на основі дисперсії вводиться **середнє квадратичне відхилення** (стандартне відхилення)

$$\sigma_B = \sqrt{D_B}, \quad (15)$$

яке дозволяє вимірювати розсіювання варіант вибірки відносно \bar{x}_B , але вже в тих самих одиницях, в яких вимірюється досліджувана ознака X .

7. **Розмах вибірки (R).** Величина, яка дорівнює різниці між найбільшою x_{\max} і найменшою x_{\min} варіантами варіаційного ряду називається розмахом вибірки:

$$R = x_{\max} - x_{\min}. \quad (16)$$

Розмах вибірки використовується для грубого оцінювання розсіювання варіант відносно свого середнього \bar{x}_B .

8. **Вибірковий коефіцієнт варіації (V).** Порівнюючи варіацію різних ознак в одній сукупності або варіацію однієї ознаки в різних сукупностях, недостатньо виявити абсолютну величину варіації. Тому для порівняння оцінок варіацій статистичних рядів із різними значеннями \bar{x}_A , які не дорівнюють нулеві, використовують відносний показник варіації, який називається вибірковий коефіцієнт варіації і визначається формулою

$$V = \frac{\sigma_B}{\bar{x}_B} \cdot 100\%. \quad (17)$$

Він використовується як показник однорідності вибірки. Вважається, що при $V \leq 33\%$ сукупність є однорідною, отже, вибіркоче середнє є типовою й надійною характеристикою сукупності. Однорідність сукупності – передумова використання інших статистичних методів – середніх величин, регресійного аналізу та ін. Однорідними вважаються такі сукупності, елементи яких мають спільні властивості й належать до одного типу (класу). При цьому однорідність означає не повну тотожність властивостей елементів, а лише наявність у них спільного в істотному (головному).

Приклад 1.3. За заданим в попередньому прикладі статистичним розподілом вибірки

- 1) обчислити розмах вибірки R , вибіркоче середнє \bar{x}_B , вибіркочу дисперсію D_B та середнє квадратичне відхилення вибірки σ_B ;
- 2) знайти вибіркочі моду Mo^* та медіану Me^* ;
- 3) зробити висновок про однорідність вибірки за обчисленим коефіцієнтом варіації V .

Розв'язання. 1) За формулою (16):

$$R = x_{\max} - x_{\min} = 7 - (-5) = 12.$$

Користуючись формулою (11), запишемо:

$$\bar{x}_B = \frac{\sum_{i=1}^8 x_i n_i}{n} = \frac{(-5) \cdot 1 + (-3) \cdot 2 + (-2) \cdot 4 + (-1) \cdot 2 + 0 \cdot 4 + 2 \cdot 3 + 3 \cdot 3 + 7 \cdot 1}{20} = 0,05.$$

Для обчислення D_B визначимо спочатку вираз:

$$\frac{\sum_{i=1}^8 x_i^2 n_i}{n} = \frac{(-5)^2 \cdot 1 + (-3)^2 \cdot 2 + (-2)^2 \cdot 4 + (-1)^2 \cdot 2 + 0 + 2^2 \cdot 3 + 3^2 \cdot 3 + 7^2 \cdot 1}{20} = 7,45.$$

Тоді

$$D_B = \frac{\sum_{i=1}^8 x_i^2 n_i}{n} - (\bar{x}_B)^2 = 7,45 - (0,05)^2 = 7,4475.$$

Отже, $D_B \approx 7,45$.

$$\sigma_B = \sqrt{D_B} \approx 2,73.$$

2) Оскільки $Mo_1^* = -2$, $Mo_2^* = 0$, то даний дискретний статистичний розподіл вибірки є двомодальним.

Варіаційний ряд складено за парною кількістю спостережень, тому $Me^* = \frac{0+0}{2} = 0$ (півсума x_{10} та x_{11}).

3) Далі, оскільки, $V = \frac{\sigma_B}{\bar{x}_B} \cdot 100\% = \frac{2,73}{0,05} \cdot 100\% = 5460$, то сукупність не є

однорідною і середнє арифметичне не буде служити надійною характеристикою сукупності, тобто не матиме практичного значення.

Аналіз закономірностей розподілу передбачає оцінювання ступеня однорідності сукупності, **асиметрії** та **ексцесу** розподілу.

В однорідних сукупностях розподіли одновершинні (одноmodalні). Багатовершинність свідчить про неоднорідний склад сукупності, про різнотиповість окремих складових. У такому разі слід перегрупувати дані, виокремити однорідні групи.

Серед одновершинних розподілів є симетричні та асиметричні (скошені), гостро та плосковершинні. У симетричному розподілі рівновіддалені від центра значення ознаки мають однакові частоти, в асиметричному – вершина розподілу зміщена. Напрямо асиметрії протилежний напряму зміщення вершини. Якщо вершина зміщена ліворуч, маємо правосторонню асиметрію, і навпаки. Зазначимо, що асиметрія

виникає внаслідок обмеженої варіації в одному напрямі або під впливом домінуючої причини розвитку, яка призводить до зміщення центра розподілу. Ступінь асиметрії різний – від помірної до значної.

У вибірці з генеральної сукупності з симетричним розподілом характеристики центра – вибіркові середнє, мода, медіана – мають приблизно однакові значення, в асиметричному розподілі між ними існують певні розбіжності. Для правосторонньої асиметрії $\bar{x}_B > Me > Mo$, а в разі лівосторонньої – навпаки: $\bar{x}_B < Me < Mo$. Чим більша асиметрія, тим більше відхилення $\bar{x}_B - Mo^*$.

Найпростішою мірою асиметрії є відносне відхилення

$$A = \frac{\bar{x}_B - Mo^*}{\sigma_B}, \quad (18)$$

яке характеризує напрям і міру скошеності в середині розподілу; при правосторонній асиметрії $A > 0$, при лівосторонній – $A < 0$.

Теоретично коефіцієнт асиметрії не має меж, проте на практиці його значення не буває надто великим і в помірно скісних розподілах не перевищує одиниці.

Іншою властивістю одновершинних розподілів є ступінь зосередженості елементів сукупності навколо центра розподілу. Цю властивість називають **ексцесом** розподілу.

Асиметрія та ексцес – дві пов'язані з варіацією властивості форми розподілу. Комплексне їх оцінювання виконується на базі центральних моментів розподілу. Алгебраїчно **центральный момент розподілу** – це середнє арифметичне k -го степеня відхилення індивідуальних значень ознаки від середнього:

$$\mu_k^* = \frac{\sum_i (x_i - \bar{x}_B)^k n_i}{n}. \quad (19)$$

Очевидно, що центральний момент 2-го порядку є дисперсією, яка характеризує варіацію. Центральні моменти 3-го і 4-го порядків характеризують відповідно асиметрію та ексцес. У симетричному розподілі $\mu_3=0$, а, отже, $A=0$. Чим більша скошеність ряду, тим більше значення μ_3 . Для того, щоб характеристика скошеності не залежала від масштабу вимірювання ознаки, для порівняння ступеня асиметрії різних розподілів використовується стандартизований момент:

$$A_S^* = \frac{\mu_3^*}{\sigma_B^3}, \quad (20)$$

який, на відміну від коефіцієнта скошеності A , залежить від крайніх значень ознаки.

При правосторонній асиметрії вибірковий коефіцієнт $A_S^* > 0$, при лівосторонній – $A_S^* < 0$. Звідси правостороння асиметрія називається

додатною, а лівостороння – від'ємною. Кажуть, що при $A_S^* < 0,25$ асиметрія низька, якщо A_S^* не перевищує 0,5 – середня, при $A_S^* > 0,5$ – висока.

Для вимірювання ексцесу використовується стандартизований момент четвертого порядку:

$$E_S^* = \frac{\mu_4^*}{\sigma_B^4} - 3. \quad (21)$$

У симетричному, близькому до нормального розподілу E_S^* близький до нуля. Очевидно, для **гостровершинного розподілу** $E_S^* > 0$, для **плосковершинного** – $E_S^* < 0$.

Похибку коефіцієнта асиметрії можна знайти за формулою

$$\delta_A = \sqrt{\frac{6(n-1)}{(n+1)(n+3)}}, \quad (22)$$

похибка коефіцієнта ексцесу обчислюється за формулою

$$\delta_E = \sqrt{\frac{24n(n-2)(n-3)}{(n-1)^2(n+3)(n+5)}}. \quad (23)$$

Приклад 1.4. За заданим в прикладі 1.2 статистичним розподілом вибірки

- 1) обчислити показники асиметрії та ексцесу;
- 2) знайти похибки коефіцієнтів асиметрії та ексцесу.

Розв'язання. Скористаємось результатами, що були отримані в попередньому прикладі 1.3. У формулу (19) підставимо знайдені раніше вибірконе середнє $\bar{x}_B = 0,05$ та середнє квадратичне відхилення вибірки $\sigma_B \approx 2,73$. Обчислимо спочатку

$$\begin{aligned} \mu_3^* &= \frac{\sum_{i=1}^8 (x_i - \bar{x}_B)^3 n_i}{n} = \frac{(-5 - 0,05)^3 \cdot 1 + (-3 - 0,05)^3 \cdot 2 + (-2 - 0,05)^3 \cdot 4 + (-1 - 0,05)^3 \cdot 2}{20} + \\ &+ \frac{(0 - 0,05)^3 \cdot 4 + (2 - 0,05)^3 \cdot 3 + (3 - 0,05)^3 \cdot 3 + (7 - 0,05)^3 \cdot 1}{20} \approx 8,41. \end{aligned}$$

$$A_S^* = \frac{\mu_3^*}{\sigma_B^3} = \frac{8,41}{(2,73)^3} \approx 0,41.$$

Отже, статистичний розподіл вибірки має середню правосторонню асиметрію. Для знаходження вибіркового коефіцієнта ексцесу знайдемо вибірковий центральний момент четвертого порядку

$$\begin{aligned} \mu_4^* &= \frac{\sum_{i=1}^8 (x_i - \bar{x}_B)^4 n_i}{n} = \\ &= \frac{(-5 - 0,05)^4 \cdot 1 + (-3 - 0,05)^4 \cdot 2 + (-2 - 0,05)^4 \cdot 4 + (-1 - 0,05)^4 \cdot 2}{20} + \\ &+ \frac{(0 - 0,05)^4 \cdot 4 + (2 - 0,05)^4 \cdot 3 + (3 - 0,05)^4 \cdot 3 + (7 - 0,05)^4 \cdot 1}{20} \approx 175,01. \end{aligned}$$

За формулою (21)

$$E_S^* = \frac{\mu_4^*}{\sigma_B^4} - 3 = \frac{175,01}{(2,73)^4} - 3 \approx 0,15.$$

Обчислимо похибку коефіцієнта асиметрії за формулою (22)

$$\delta_A = \sqrt{\frac{6(n-1)}{(n+1)(n+3)}} = \sqrt{\frac{6 \cdot 19}{21 \cdot 23}} \approx 0,49$$

та похибку коефіцієнта ексцесу за формулою (23)

$$\delta_E = \sqrt{\frac{24n(n-2)(n-3)}{(n-1)^2(n+3)(n+5)}} = \sqrt{\frac{24 \cdot 20 \cdot 18 \cdot 17}{19^2 \cdot 23 \cdot 25}} \approx 0,84.$$

1.3. Інтервальний статистичний розподіл вибірки та його числові характеристики

Інтервальним статистичним розподілом вибірки називається перелік часткових інтервалів і відповідних їм частот (або відносних частот).

Такий розподіл у табличній формі має вигляд:

h	$x_0 - x_1$	$x_1 - x_2$...	$x_{k-1} - x_k$
n_i	n_1	n_2	...	n_k
W_i	W_1	W_2	...	W_k

Величина $h = x_i - x_{i-1}$ є довжиною часткового i -того інтервалу. На практиці, як правило, ці інтервали вибираються однаковими. Вибір кількості інтервалів та визначення їх меж обговорювалося у підрозділі 1.1.

Такий інтервальний розподіл вибірки можна подати графічно у вигляді **гістограми частот** або **гістограми відносних частот**, а також (як і для дискретного статистичного розподілу), емпіричною функцією розподілу.

Гістограма частот та відносних частот

Гістограма частот є ступінчата фігурою, що складається з прямокутників, кожний з яких має основу h і висоту $\frac{n_i}{h}$. Очевидно, площа такої гістограми дорівнює об'єму вибірки:

$$S = \sum_{i=1}^k h \frac{n_i}{h} = \sum_{i=1}^k n_i = n.$$

Гістограма відносних частот є ступінчатою фігурою, що складається з прямокутників, кожний з яких має основу h і висоту $\frac{W_i}{h}$.

Оскільки,

$$S = \sum_{i=1}^k h \frac{W_i}{h} = \sum_{i=1}^k W_i = 1,$$

то площа гістограми відносних частот дорівнює 1.

Емпірична функція розподілу та кумулята

Аналогічно, як і для дискретного варіаційного ряду, визначається емпірична функція розподілу

$$F_n^*(x) = W(X < x) = \frac{n_x}{n}. \quad (24)$$

За значеннями цієї функції для інтервального статистичного розподілу вибірки за припущення рівномірного розподілу ознаки на кожному частинному інтервалі будується кумулятивна крива або **кумулята** $F^*(x)$. Вона має вигляд ламаної лінії, яка зростає на кожному частковому інтервалі і з'єднує точки $(x_i, F_n^*(x_i))$; тут x_i визначають межі часткових інтервалів.

Приклад 1.5. За заданим інтервальним статистичним розподілом вибірки

$h = 10$	$-10 - 0$	$0 - 10$	$10 - 20$	$20 - 30$	$30 - 40$
n_i	5	10	15	12	8
W_i	0,1	0,2	0,3	0,24	0,16

побудувати гістограму частот і відносних частот, знайти емпіричну функцію розподілу і побудувати кумулятивну криву.

Розв'язання. Для побудови гістограми частот знайдемо для кожного часткового інтервалу відношення $\frac{n_i}{h}$. Для заданого інтервального статистичного розподілу вибірки ці відношення будуть відповідно такими: 0,5; 1; 1,5; 1,2; 0,8. Отже, гістограма частот має наступний вигляд:

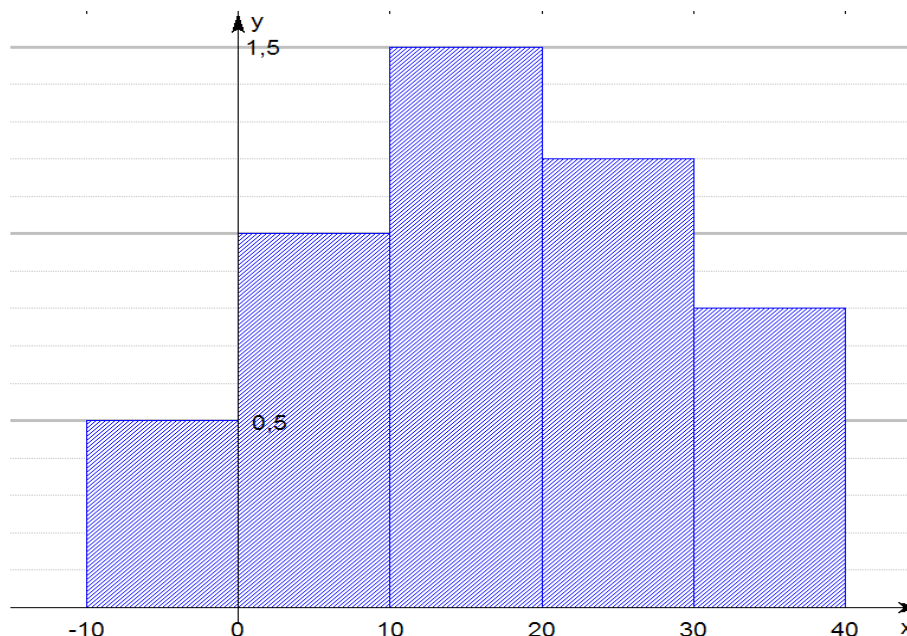


Рис. 1.4

Для гістограми відносних частот стовпчики матимуть висоти 0,01; 0,02; 0,03; 0,024; 0,016 відповідно. Гістограма відносних частот має такий вигляд:

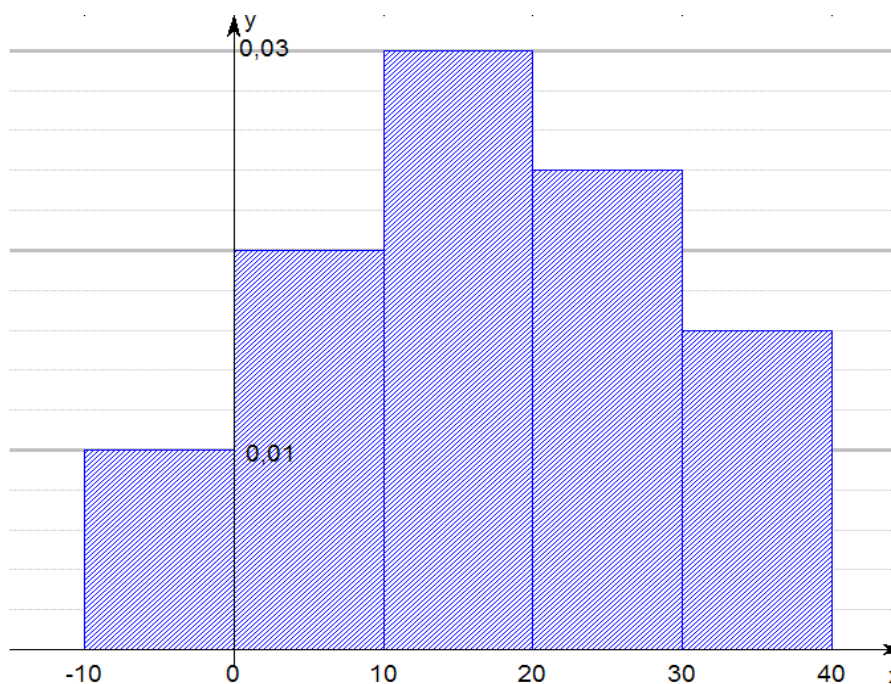


Рис. 1.5

Для побудови кумулятивної кривої знайдемо емпіричну функцію розподілу. Скористаємось формулою (24). Послідовно обчислимо $F^*(x)$ на кожному частинному інтервалі. Остаточоно отримаємо:

$$F_n^*(x) = W(X < x) = \begin{cases} 0, & x \leq -10; \\ 0.1, & -10 < x \leq 0; \\ 0.3, & 0 < x \leq 10; \\ 0.6, & 10 < x \leq 20; \\ 0.84, & 20 < x \leq 30; \\ 1, & 30 < x \leq 40. \end{cases}$$

Кумулята представлена на рис.1.6 .

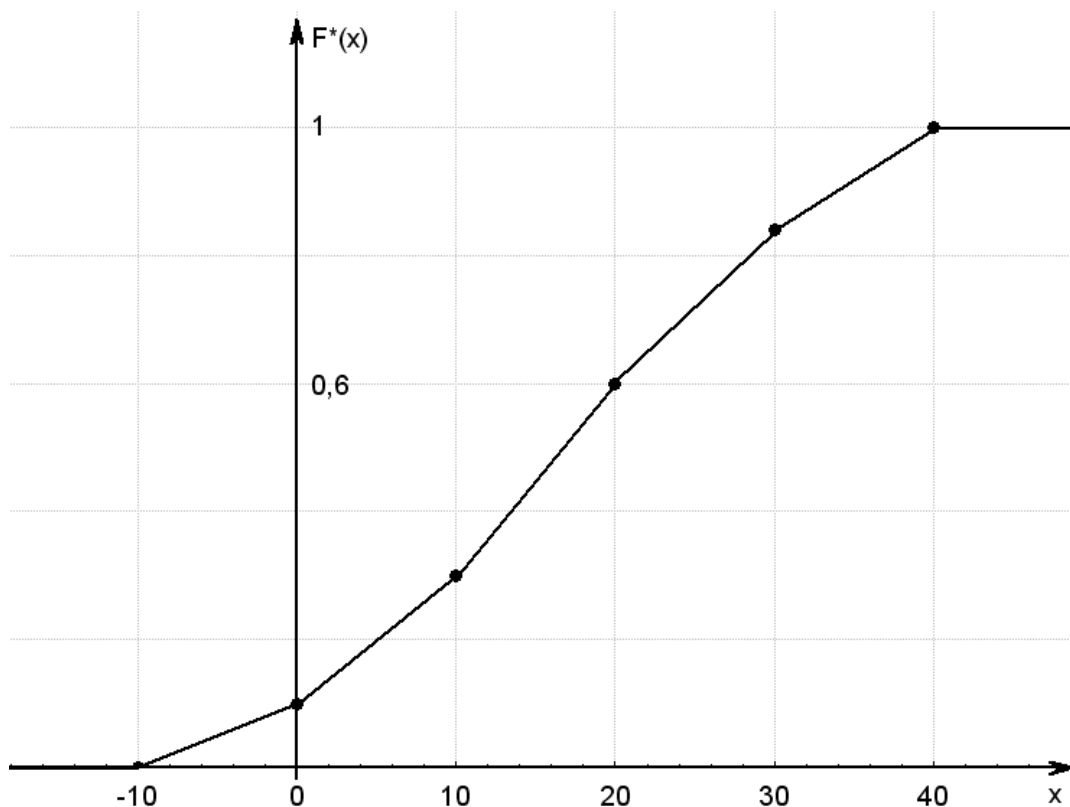


Рис. 1.6

Відмітимо, що аналогом емпіричної функції $F^*(x)$ у теорії ймовірностей є функція розподілу випадкової величини $F(x) = P\{X < x\}$.

Медіана

Для відшукування медіани інтервального статистичного розподілу вибірки слід визначити медіанний частковий інтервал. Так, якщо на

інтервалі $[x_{i-1} - x_i]$ $F^*(x_{i-1}) < 0,5$ і $F^*(x_i) > 0,5$, то всередині цього інтервалу існує $x = Me^*$, така, що $F^*(Me^*) = 0,5$. Для знаходження вибіркової медіани користуються формулою

$$Me^* = x_{i-1} + \frac{0,5 - F^*(x_{i-1})}{F^*(x_i) - F^*(x_{i-1})} h, \quad (25)$$

де $h = x_i - x_{i-1}$ є довжина часткового інтервалу.

Мода

Для відшукування моди інтервального статистичного розподілу вибірки потрібно визначити модальний інтервал, тобто такий частковий інтервал, що має найбільшу частоту появи досліджуваної ознаки. Вибіркову моду інтервального статистичного розподілу вибірки можна обчислити за формулою

$$Mo^* = x_{i-1} + \frac{n_{Mo} - n_{Mo-1}}{2n_{Mo} - n_{Mo-1} - n_{Mo+1}} h, \quad (26)$$

тут h – довжина часткового інтервалу; x_{i-1} – початок модального інтервалу; n_{Mo} – частота модального інтервалу; n_{Mo-1} – частота домодального інтервалу; n_{Mo+1} – частота післямодального інтервалу.

Розглянемо приклад.

Приклад 1.6. За заданим в попередньому прикладі інтервальним статистичним розподілом вибірки знайти Me^* , Mo^* .

Розв'язання. Для визначення медіанного часткового інтервалу скористаємось рис. 1.6. Маємо $F^*(10) = 0,3$, $F^*(20) = 0,6$, $h = 10$, тому

$$Me^* = 10 + \frac{0,5 - 0,3}{0,6 - 0,3} \cdot 10 \approx 16,7.$$

Для визначення модального часткового інтервалу скористаємось заданим інтервальним статистичним розподілом вибірки. Шуканий інтервал $[10, 20]$, тому

$$Mo^* = 10 + \frac{15 - 10}{2 \cdot 15 - 10 - 12} \cdot 10 = 16,25.$$

Вибіркове середнє

Для знаходження вибіркового \bar{x}_B перейдемо від інтервального статистичного розподілу до дискретного, варіантами якого є середини

часткових інтервалів $x_i^* = x_i - \frac{h}{2} = x_{i-1} + \frac{h}{2}$, а частотами – частоти інтервального розподілу:

$x_i^* = x_i - \frac{h}{2} = x_{i-1} + \frac{h}{2}$	x_1^*	x_2^*	...	x_k^*
n_i	n_1	n_2	...	n_k

Тоді, очевидно, можемо записати:

$$\bar{x}_B = \frac{\sum_{i=1}^k x_i^* n_i}{n}. \quad (27)$$

$$D_B = \frac{\sum_{i=1}^k (x_i^*)^2 n_i}{n} - (\bar{x}_B)^2, \quad (28)$$

$$\sigma_B = \sqrt{D_B}. \quad (29)$$

Приклад 1.7. За заданим в прикладі 1.5 інтервальним статистичним розподілом вибірки обчислити \bar{x}_B, D_B, σ_B .

Розв'язання. Побудуємо дискретний статистичний розподіл за заданим інтервальним. Оскільки $h = 10$, то дістанемо:

$x_i^* = x_i - \frac{h}{2} = x_{i-1} + \frac{h}{2}$	-5	5	15	25	35
n_i	5	10	15	12	8

За формулами (27), (28), (29) при $n = 50$ дістанемо:

$$\bar{x}_B = \frac{\sum_{i=1}^5 x_i^* n_i}{n} = \frac{-25 + 50 + 225 + 300 + 280}{50} = 16,6.$$

$$\frac{\sum_{i=1}^5 (x_i^*)^2 n_i}{n} = \frac{125 + 250 + 3375 + 7500 + 9800}{50} = 421.$$

$$D_B = \frac{\sum_{i=1}^5 (x_i^*)^2 n_i}{n} - (\bar{x}_B)^2 = 421 - 275,56 = 145,44.$$

$$\sigma_B = \sqrt{D_B} = \sqrt{145,44} \approx 12,06.$$

1.4. Двовимірний статистичний розподіл вибірки та його числові характеристики

Перелік варіант $Y = y_i$, $X = x_j$ та відповідних їм частот n_{ij} спільної їх появи утворюють **двовимірний статистичний розподіл вибірки**. Вона реалізована з генеральної сукупності, елементам якої притаманні кількісні ознаки X та Y . У табличній формі двовимірний розподіл вибірки має вигляд:

$Y = y_i$	$X = x_j$					
	x_1	x_2	x_3	...	x_m	x_{yi}
y_1	n_{11}	n_{12}	n_{13}	...	n_{1m}	n_{y1}
y_2	n_{21}	n_{22}	n_{23}	...	n_{2m}	n_{y2}
y_3	n_{31}	n_{32}	n_{33}	...	n_{3m}	n_{y3}
...
y_k	n_{k1}	n_{k2}	n_{k3}	...	n_{km}	n_{yk}
n_{xj}	n_{x1}	n_{x2}	n_{x3}	...	n_{xm}	

Тут n_{ij} – частота спільної появи варіант

$$Y = y_i, \quad X = x_j;$$

$$n_{y_i} = \sum_{j=1}^m n_{ij}, \quad n_{x_j} = \sum_{i=1}^k n_{ij};$$

$$n = \sum_{i=1}^k \sum_{j=1}^m n_{ij} = \sum_{i=1}^k n_{y_i} = \sum_{j=1}^m n_{x_j}.$$

Загальні числові характеристики ознаки X :

загальне середнє ознаки X

$$\bar{x} = \frac{\sum_{i=1}^k \sum_{j=1}^m x_j n_{ij}}{n} = \frac{\sum_{j=1}^m x_j n_{x_j}}{n}; \quad (30)$$

загальна дисперсія ознаки X

$$D_x = \frac{\sum_{i=1}^k \sum_{j=1}^m x_j^2 n_{ij}}{n} - (\bar{x})^2 = \frac{\sum_{j=1}^m x_j^2 n_{x_j}}{n} - (\bar{x})^2; \quad (31)$$

загальне середнє квадратичне відхилення ознаки X

$$\sigma_x = \sqrt{D_x}. \quad (32)$$

Загальні числові характеристики ознаки Y :

загальне середнє ознаки Y

$$\bar{y} = \frac{\sum_{i=1}^k \sum_{j=1}^m y_i n_{ij}}{n} = \frac{\sum_{i=1}^k y_i n_{y_i}}{n}; \quad (33)$$

загальна дисперсія ознаки Y

$$D_y = \frac{\sum_{i=1}^k \sum_{j=1}^m y_j^2 n_{ij}}{n} - (\bar{y})^2 = \frac{\sum_{i=1}^k y_i^2 n_{y_i}}{n} - (\bar{y})^2; \quad (34)$$

загальне середнє квадратичне відхилення ознаки Y

$$\sigma_y = \sqrt{D_y}. \quad (35)$$

Умовні статистичні розподіли та їх числові характеристики

Умовним статистичним розподілом ознаки Y при фіксованому значенні $X = x_j$ називають перелік варіант ознаки Y та відповідних їм частот, узятих при фіксованому значенні X .

$$Y / X = x_j.$$

$Y = y_i$	y_1	y_2	y_3	...	y_k
n_{ij}	n_{1j}	n_{2j}	n_{3j}	...	n_{kj}

Тут $\sum_{i=1}^k n_{ij} = n_{x_j}$.

Числові характеристики для такого статистичного розподілу називають **умовними**. До них належать:
умовне середнє ознаки Y

$$\bar{y}_{X=x_j} = \frac{\sum_{i=1}^k y_i n_{ij}}{\sum_{i=1}^k n_{ij}} = \frac{\sum_{i=1}^k y_i n_{ij}}{n_{x_j}}; \quad (36)$$

умовна дисперсія ознаки Y

$$D(Y / X = x_j) = \frac{\sum_{i=1}^k y_i^2 n_{ij}}{n_{x_j}} - (\bar{y}_{X=x_j})^2; \quad (37)$$

умовне середнє квадратичне відхилення ознаки Y

$$\sigma(Y / X = x_j) = \sqrt{D(Y / X = x_j)}. \quad (38)$$

$D(Y / X = x_j), \sigma(Y / X = x_j)$ вимірюють розсіювання варіант ознаки Y щодо умовного середнього ознаки $Y - \bar{y}_{X=x_j}$.

Аналогічно, умовним статистичним розподілом ознаки X при фіксованому значенні $Y = y_i$ називають перелік варіант $X=x_j$ та відповідних їм частот, узятих при фіксованому значенні ознаки $Y = y_i$.

$$X / Y = y_i.$$

$X = x_j$	x_1	x_2	x_3	...	x_m
n_{ij}	n_{i1}	n_{i2}	n_{i3}	...	n_{im}

$$\text{Тут } \sum_{j=1}^m n_{ij} = n_{y_i}.$$

Умовні числові характеристики для цього розподілу:
умовне середнє ознаки X

$$\bar{x}_{y=y_i} = \frac{\sum_{j=1}^m x_j n_{ij}}{\sum_{j=1}^m n_{ij}} = \frac{\sum_{j=1}^m x_j n_{ij}}{n_{y_i}}; \quad (39)$$

умовна дисперсія ознаки X

$$D(X / Y = y_i) = \frac{\sum_{j=1}^m x_j^2 n_{ij}}{n_{y_i}} - (\bar{x}_{Y=y_i})^2; \quad (40)$$

умовне середнє квадратичне відхилення ознаки X

$$\sigma((X / Y = y_i)) = \sqrt{D((X / Y = y_i))}. \quad (41)$$

При відомих значеннях умовних середніх $\bar{y}_{x_j}, \bar{x}_{y_i}$ загальні середні ознаки X та Y можна обчислити за формулами:

$$\bar{y} = \frac{\sum_{j=1}^m \bar{y}_{x_j} n_{x_j}}{n}; \quad (42)$$

$$\bar{x} = \frac{\sum_{i=1}^k \bar{x}_{y_i} n_{y_i}}{n}. \quad (43)$$

Кореляційний момент, вибірковий коефіцієнт кореляції

При дослідженні двомірного статистичного розподілу вибірки виникає питання з'ясувати наявність зв'язку між ознаками X і Y . Такий зв'язок (якщо він існує) у статистиці називають кореляційним. Спочатку визначається емпіричний **кореляційний момент** K_{xy}^* згідно формули:

$$K_{xy}^* = \frac{\sum_{i=1}^k \sum_{j=1}^m y_i x_j n_{ij}}{n} - \bar{x} \cdot \bar{y}. \quad (44)$$

У випадку, коли $K_{xy}^* = 0$, то між ознаками X і Y немає лінійної кореляційної залежності і такі ознаки називаються некорельованими. Якщо кореляційний момент $K_{xy}^* \neq 0$, то цей зв'язок існує. Ясно, що кореляційний момент дає лише відповідь на запитання: чи є лінійний зв'язок між ознаками X і Y , чи він відсутній.

Тіснота кореляційного зв'язку визначення з використанням вибіркового коефіцієнту кореляції r_B :

$$r_B = \frac{K_{xy}^*}{\sigma_x \sigma_y}. \quad (45)$$

По аналогії з теорією ймовірностей, коефіцієнт кореляції, задовольняє умові $-1 \leq r_B \leq 1$. В залежності від того, на скільки модуль r_B близький до одиниці розрізняють слабкий, помірний, помітний, досить тісний та чітко виражений лінійний зв'язок між ознаками. Чим ближче $|r_B|$ до 1, тим тісніший зв'язок. Розглянемо приклад наведений в [7] ст.20.

Приклад 1.8. За заданим двомірним статистичним розподілом вибірки ознак X і Y

$Y=y_i$	$X=x_j$				n_{y_i}
	10	20	30	40	
2	-	2	4	4	10
4	10	8	6	6	30
6	5	10	5	-	20
8	15	-	15	10	40
n_{x_j}	30	20	30	20	-

потрібно:

- 1) обчислити K_{xy}^* , r_B ;
- 2) побудувати умовні статистичні розподіли $Y / X=30$, $X / Y=4$ й обчислити умовні числові характеристики.

Роз'язання.1) Для обчислення K_{xy}^* , r_B визначимо \bar{x} , σ_x , \bar{y} , σ_y .

Оскільки $n = \sum_{i=1}^k \sum_{j=1}^m n_{ij} = 100$, то

$$\begin{aligned}\bar{x} &= \frac{\sum_{j=1}^m x_j n_{x_j}}{n} = \frac{10 \cdot 30 + 20 \cdot 20 + 30 \cdot 30 + 40 \cdot 20}{100} = \\ &= \frac{300 + 400 + 900 + 800}{100} = \frac{2400}{100} = 24 \\ \bar{x} &= 24.\end{aligned}$$

$$\begin{aligned}\frac{\sum_{j=1}^m x_j^2 n_{x_j}}{n} &= \frac{(10)^2 \cdot 30 + (20)^2 \cdot 20 + (30)^2 \cdot 30 + (40)^2 \cdot 20}{100} = \\ &= \frac{3000 + 8000 + 27000 + 32000}{100} = \frac{70000}{100} = 700.\end{aligned}$$

$$D_x = \frac{\sum_{j=1}^m x_j^2 n_{x_j}}{n} - (\bar{x})^2 = 700 - (24)^2 = 700 - 576 = 124.$$

$$\sigma_x = \sqrt{D_x} = \sqrt{124} \approx 11,14.$$

Звідки, $\sigma_x = 11,14$.

$$\bar{y} = \frac{\sum_{i=1}^k y_i n_{y_i}}{n} = \frac{2 \cdot 10 + 4 \cdot 30 + 6 \cdot 20 + 8 \cdot 40}{100} = \frac{20 + 120 + 120 + 320}{100} = 5,8.$$

Отже, $\bar{y} = 5,8$.

$$\begin{aligned}\frac{\sum_{i=1}^k y_i^2 n_{y_i}}{n} &= \frac{(2)^2 \cdot 10 + (4)^2 \cdot 30 + (6)^2 \cdot 20 + (8)^2 \cdot 40}{100} = \\ &= \frac{40 + 480 + 720 + 2560}{100} = \frac{3800}{100} = 38.\end{aligned}$$

$$D_y = \frac{\sum_{i=1}^k y_i^2 n_{y_i}}{n} - (\bar{y})^2 = 38 - (5,8)^2 = 38 - 33,64 = 4,36,$$

$$\sigma_y = \sqrt{D_y} = \sqrt{4,36} \approx 2,1.$$

Для визначення кореляційного моменту K_{xy}^* обчислюють

$$\begin{aligned} \sum_{i=1}^k \sum_{j=1}^m y_i x_j n_{ij} &= 2 \cdot 10 \cdot 0 + 2 \cdot 20 \cdot 2 + 2 \cdot 30 \cdot 4 + 2 \cdot 40 \cdot 4 + 4 \cdot 10 \cdot 10 + 4 \cdot 20 \cdot 8 + \\ &+ 4 \cdot 30 \cdot 6 + 4 \cdot 40 \cdot 6 + 6 \cdot 10 \cdot 5 + 6 \cdot 20 \cdot 10 + 6 \cdot 30 \cdot 5 + 6 \cdot 40 \cdot 0 + 8 \cdot 10 \cdot 15 + \\ &+ 8 \cdot 20 \cdot 0 + 8 \cdot 30 \cdot 15 + 8 \cdot 40 \cdot 10 = 0 + 80 + 240 + 320 + 400 + 640 + 720 + \\ &+ 960 + 300 + 1200 + 900 + 0 + 1200 + 0 + 3600 + 3200 = 13760. \end{aligned}$$

Тоді

$$K_{xy}^* = \frac{\sum_{i=1}^k \sum_{j=1}^m y_i x_j n_{ij}}{n} - \bar{x} \cdot \bar{y} = \frac{13760}{100} = 24 \cdot 5,8 = 137,6 - 139,2 = -1,6.$$

Це свідчить про те, що між ознаками X і Y існує від'ємний кореляційний зв'язок.

Для вимірювання тісноти цього зв'язку обчислимо вибірковий коефіцієнт кореляції.

$$r_B = \frac{K_{xy}^*}{\sigma_x \sigma_y} = \frac{-1,6}{11,14 \cdot 2,1} = \frac{-1,6}{23,394} \approx -0,068.$$

Отже, $r_B = -0,068$, тобто тіснота кореляційного зв'язку між знаками X і Y є слабкою.

3) Умовний статистичний розподіл $X/Y = 30$ має такий вигляд:

$Y = y_i$	2	4	6	8
n_{i3}	4	6	5	15

Обчислимо умовні числові характеристики для цього розподілу:

Умовне середнє

$$\bar{y}_{X=30} = \frac{\sum_{i=1}^k y_i n_{i3}}{\sum_{i=1}^k n_{i3}} = \frac{2 \cdot 4 + 4 \cdot 6 + 6 \cdot 5 + 8 \cdot 15}{30} = \frac{8 + 24 + 30 + 120}{30} = \frac{182}{30} = 6,07.$$

Умовна дисперсія та середнє квадратичне відхилення

$$\frac{\sum_{i=1}^k y_i^2 n_{i3}}{\sum_{i=1}^k n_{i3}} = \frac{(2)^2 \cdot 4 + (4)^2 \cdot 6 + (6)^2 \cdot 5 + (8)^2 \cdot 15}{30} =$$

$$= \frac{16 + 96 + 180 + 960}{30} = \frac{1252}{30} = 41,73$$

$$D(Y / X = 30) = \frac{\sum_{i=1}^k y_i^2 n_{i3}}{\sum_{i=1}^k n_{i3}} - (\bar{y}_{X=30})^2 = 41,73 - 36,8449 \approx 4,89;$$

$$\sigma(Y / X = 30) = \sqrt{D_{(Y/X=30)}} = \sqrt{4,89} \approx 2,21.$$

Отримали, що $\sigma(Y / X = 30) \approx 2,21$.

Умовний статистичний розподіл $X / Y = 4$ має вигляд:

$X = x_j$	10	20	30	40
n_{2j}	10	8	6	6

Обчислимо умовні числові характеристики.

Умовне середнє

$$\bar{x}_{Y=4} = \frac{\sum_{j=1}^m x_j n_{2j}}{\sum_{j=1}^m n_{2j}} = \frac{10 \cdot 10 + 20 \cdot 8 + 30 \cdot 6 + 40 \cdot 6}{30} =$$

$$= \frac{100 + 160 + 180 + 240}{30} = \frac{680}{30} \approx 22,7.$$

Маємо: $\bar{x}_{Y=4} \approx 22,7$.

Умовна дисперсія та середнє квадратичне відхилення мають вигляд:

$$\frac{\sum_{j=1}^m x_j^2 n_{2j}}{\sum_{j=1}^m n_{2j}} = \frac{(10)^2 \cdot 10 + (20)^2 \cdot 8 + (30)^2 \cdot 6 + (40)^2 \cdot 6}{30} =$$

$$= \frac{1000 + 3200 + 5400 + 9600}{30} = 640.$$

$$D(X / Y = 4) = \frac{\sum_{j=1}^m x_j n_{2j}}{\sum_{j=4}^m n_{2j}} - (\bar{x}_{Y=4})^2 = 640 - (22,7)^2 = 640 - 515,29 = 124,71.$$

$$\sigma(X / Y = 4) = \sqrt{D_{(X/Y=4)}} = \sqrt{124,71} \approx 11,17.$$

$$\sigma(X / Y = 4) \approx 11,17.$$

Отримали, що $\sigma(X / Y = 4) \approx 11,17$.

1.5. Парний статистичний розподіл вибірки та його числові характеристики

У випадку, коли частота спільної появи ознак X і Y $n_{ij} = 1$ для всіх варіант, то двовимірний статистичний розподіл набуває такого вигляду:

$Y = y_i$	y_1	y_2	y_3	y_4	...	y_n
$X = x_i$	x_1	x_2	x_3	x_4	...	x_n

Його називають **парним статистичним розподілом вибірки**. В ньому кожна пара значень ознак X і Y з'являється лише один раз.

Об'єм вибірки у цьому випадку дорівнює кількості пар, тобто n .

Числові характеристики ознаки X :

вибіркове середнє

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}; \quad (46)$$

вибіркова дисперсія

$$D_x = \frac{\sum_{i=1}^n x_i^2}{n} - (\bar{x})^2; \quad (47)$$

вибіркове середнє квадратичне відхилення

$$\sigma_x = \sqrt{D_x}. \quad (48)$$

Числові характеристики ознаки Y :

вибіркове середнє

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}; \quad (49)$$

вибіркова дисперсія

$$D_y = \frac{\sum_{i=1}^n y_i^2}{n} - (\bar{y})^2; \quad (50)$$

вибіркове середнє квадратичне відхилення

$$\sigma_y = \sqrt{D_y}; \quad (51)$$

емпіричний кореляційний момент

$$K_{xy}^* = \frac{\sum_{i=1}^n x_i y_i}{n} - \bar{x} \cdot \bar{y}; \quad (52)$$

вибірковий коефіцієнт кореляції обчислюється за формулою (45).

Попереднє уявлення про двовимірну генеральну сукупність можна отримати, якщо зображувати елементи вибірки (x_i, y_i) точками на площині з вибраною прямокутною системою координат. Таке зображення вибірки називається **діаграмою розсіювання**.

У випадку повної кореляції всі точки (x_i, y_i) , $i = \overline{1, n}$, будуть розміщені на одній прямій. Якщо $r_B = 1$, то між вибірковими даними існує прямий лінійний зв'язок: із збільшенням значень однієї вибірки відповідні значення другої вибірки також збільшуються. Якщо $r_B = -1$, то між вибірковими даними є обернений лінійний зв'язок: із збільшенням значень однієї вибірки відповідні значення другої вибірки зменшуються. Якщо $r_B = 0$, то говорять, що дві вибірки є некорельовані, при цьому точки (x_i, y_i) розміщені на площині хаотично.

Якщо $0 < r_B < 1$, то можна знайти таку пряму, від якої точки (x_i, y_i) відхиляються найменше у тому сенсі, що сума квадратів відстаней від точок (x_i, y_i) до цієї прямої буде мінімальною. Вказана пряма називається **прямою вибіркової лінійної регресії** у на x . Вона визначається

рівнянням $y = ax + b$, де $a = r_B \frac{\sigma_y}{\sigma_x}$, $b = \bar{y} - a\bar{x}$. Кутовий коефіцієнт даної

прямої a називається **вибірковим коефіцієнтом регресії** у на x , він показує, на скільки одиниць в середньому змінюється змінна y при збільшенні x на одну одиницю.

Невідомі параметри a і b в рівнянні вибіркової лінійної регресії у на x можуть бути знайдені і як розв'язки нормальної системи методу найменших квадратів:

$$\begin{cases} a \sum_{i=1}^n x_i^2 + b \sum_{i=1}^n x_i = \sum_{i=1}^n x_i y_i, \\ a \sum_{i=1}^n x_i + b \cdot n = \sum_{i=1}^n y_i. \end{cases} \quad (53)$$

Приклад 1.9. Обчислити вибіркового коефіцієнта кореляції, знайти рівняння вибіркової лінійної регресії y на x , та побудувати діаграму розсіювання за вибілковими даними:

x_i	7	8	5	3	7
y_i	1	2	3	1	3

Роз'язання. 1-й спосіб. Знаходимо числові характеристики

$$\bar{x} = \frac{1}{5}(7 + 8 + 5 + 3 + 7) = 6;$$

$$\bar{y} = \frac{1}{5}(1 + 2 + 3 + 1 + 3) = 2;$$

$$D_x = \frac{1}{5}(7^2 + 8^2 + 5^2 + 3^2 + 7^2) - 6^2 = 3,2;$$

$$D_y = \frac{1}{5}(1^2 + 2^2 + 3^2 + 1^2 + 3^2) - 2^2 = 0,8;$$

$$K_{xy} = \frac{1}{5}(7 + 16 + 15 + 3 + 21) - 12 = 0,4;$$

$$r_B = \frac{0,4}{\sqrt{3,2} \cdot \sqrt{0,8}} = 0,25; a = 0,25 \cdot \frac{\sqrt{0,8}}{\sqrt{3,2}} = 0,125; \quad b = 2 - 0,125 \cdot 6 = 1,25.$$

Отже, $y = 0,125x + 1,25$ – рівняння вибіркової лінійної регресії y на x .

2-й спосіб. Запишемо нормальну систему методу найменших квадратів (53). Для цього знайдемо суми:

$$\sum_{i=1}^n x_i = 30, \quad \sum_{i=1}^n x_i^2 = 196, \quad \sum_{i=1}^n y_i = 10, \quad \sum_{i=1}^n x_i y_i = 62.$$

Маємо:

$$\begin{cases} 196a + 30b = 62 \\ 30a + 5b = 10 \end{cases} \Leftrightarrow \begin{cases} b = 2 - 6a \\ 196a + 60 - 180a = 62 \end{cases} \Leftrightarrow$$

$$\begin{cases} b = 2 - 6a \\ 16a = 2 \end{cases} \Leftrightarrow \begin{cases} a = 0,125 \\ b = 1,25 \end{cases}.$$

Отже, $y = 0,125x + 1,25$.

Діаграма розсіювання має вигляд:

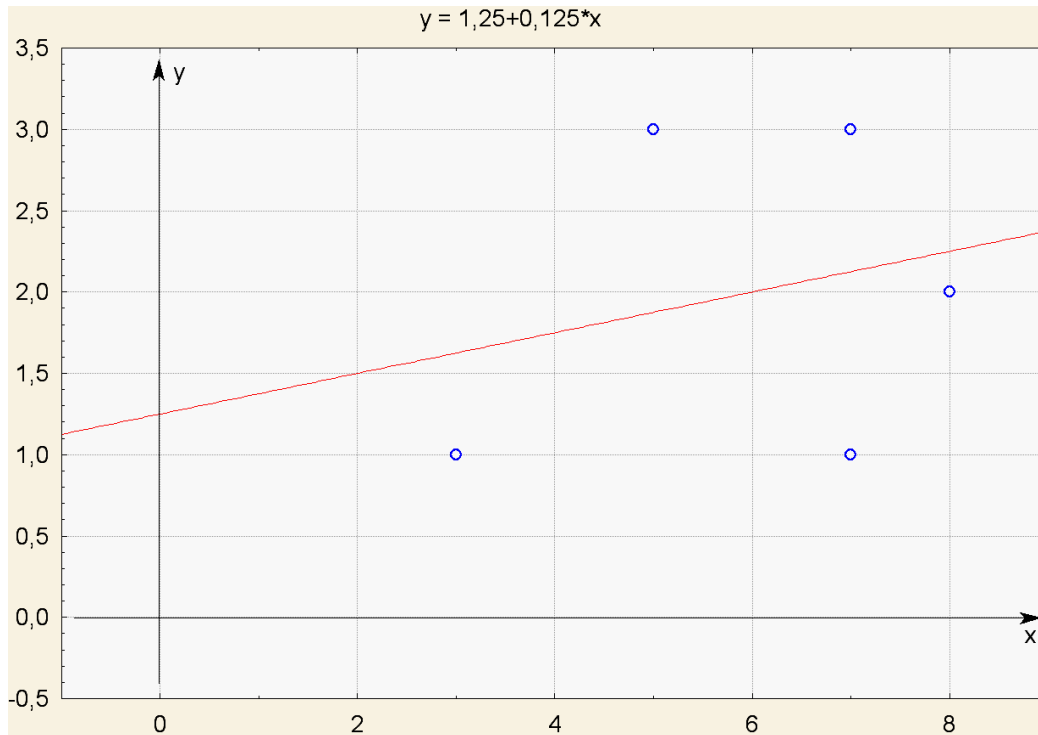


Рис. 1.7

**Контрольні питання для самоперевірки до теми
„Статистичні розподіли вибірки та їх числові характеристики”**

1. Що є предметом математичної статистики?
2. Які основні задачі розв’язує математична статистика?
3. Наведіть визначення генеральної сукупності та вибірки.
4. Що називається варіантою, варіаційним рядом?
5. Як визначається частота та відносна частота варіант?
6. Наведіть визначення дискретного статистичного розподілу вибірки.
7. З якою метою вводиться розмах та коефіцієнт варіації вибірки?
8. Як визначається мода, медіана дискретного статистичного розподілу?
9. Дайте означення емпіричної функції розподілу. Які властивості вона має?
10. Як побудувати полігон частот і полігон відносних частот?
11. Як визначаються вибіркове середнє, вибіркова дисперсія, вибіркове середнє квадратичне відхилення для дискретного статистичного розподілу вибірки?
12. Як визначається інтервальний статистичний розподіл вибірки?
13. Як визначаються вибіркове середнє, вибіркова дисперсія, вибіркове середнє квадратичне відхилення для інтервального статистичного розподілу?
14. Як визначається емпірична функція розподілу для інтервального статистичного розподілу вибірки?
15. Як визначається медіана для інтервального статистичного розподілу?
16. Як визначається мода для інтервального статистичного розподілу?
17. Як побудувати гістограму частот і відносних частот?

18. Що називається початковим емпіричним моментом k -го порядку?
19. Що називається центральним емпіричним моментом k -го порядку?
20. Як визначається асиметрія та ексцес статистичного розподілу вибірки?
21. Що називається двовимірним статистичним розподілом вибірки?
22. Запишіть формули для обчислення основних числових характеристик ознак X і Y для двовимірного статистичного розподілу вибірки.
23. Наведіть визначення емпіричного кореляційного моменту та сформулюйте його властивості.
24. Як визначається і які властивості вибіркового коефіцієнта кореляції?
25. Що називається умовним статистичним розподілом $Y/X = x_i$?
26. Як визначаються умовні числові характеристики для умовного статистичного розподілу $Y/X = x_i$?
27. Що називається умовним статистичним розподілом $X/Y = y_i$ та як визначаються умовні числові характеристики цього розподілу?
28. Що називають діаграмою розсіювання?
29. Яким рівнянням визначається пряма вибіркової лінійної регресії y на x ? Який зміст її кутового коефіцієнта?
30. Який вигляд має нормальна система методу найменших квадратів? Що є розв'язком цієї системи?

Вправи до теми „Статистичні розподіли вибірки та їх числові характеристики”

1.1. На телефонній станції проводились спостереження за випадковою величиною ξ – кількість неправильних з'єднань за хвилину. Спостереження протягом години дали такі результати: 4, 0, 2, 1, 4, 3, 2, 4, 1, 3, 0, 2, 2, 0, 2, 1, 3, 3, 3, 1, 4, 2, 2, 1, 1, 2, 1, 0, 3, 5, 1, 3, 2, 7, 2, 0, 0, 1, 3, 3, 1, 2, 4, 2, 0, 2, 3, 1, 2, 5, 1, 1, 0, 1, 1, 2, 2, 1, 1, 4.

Провести первинну статистичну обробку даних.

1.2. Із генеральної сукупності випадковим чином одержано вибірку з наступними значеннями x :

10, 1, 7, 2, 5, 6, 5, 9, 6, 2, 8, 7, 6, 3, 3, 8, 8, 10, 7, 6, 4, 9, 3, 6, 5.

Потрібно:

- 1) побудувати дискретний статистичний розподіл вибірки;
- 2) побудувати емпіричну функцію розподілу $F^*(x)$ і зобразити її графічно;
- 3) накреслити полігони частот і відносних частот.

1.3. За одержаним в прикладі 1.2 статистичним розподілом вибірки

4) обчислити розмах вибірки R , вибіркове середнє \bar{x}_B , вибіркочну дисперсію D_B та середнє квадратичне відхилення вибірки σ_B ;

5) знайти вибіркові моду Mo^* та медіану Me^* ;

б) зробити висновок про однорідність вибірки за обчисленим коефіцієнтом варіації V .

1.4. Одержано дані вибірки із 50 спостережень випадкової величини ξ (таблиця 1.1):

Таблиця 1.1

23.76	37.50	16.5	32.05	33.47	20.55	32.63	34.71	16.04	44.10
34.51	56.60	48.2	25.22	14.47	27.40	28.64	21.97	20.27	41.02
34.00	26.70	36.6	37.34	35.19	40.67	38.31	30.74	33.99	18.52
31.81	54.80	37.9	43.21	21.73	24.15	29.81	12.95	49.87	23.06
32.39	35.3	20.2	22.65	25.14	26.81	31.50	42.77	22.27	25.72

Побудувати інтервальний варіаційний ряд за даною вибіркою.

1.5. Для інтервального статистичного розподілу

інтервал	1.6-1.8	1.8-2.0	2.0-2.2	2.2-2.4	2.4-2.6	2.6-2.8	2.8-3.0
Частота, n_i	6	20	15	24	8	17	10

потрібно:

- 1) обчислити розмах вибірки R , вибіркове середнє \bar{x}_B , вибіркочу дисперсію D_B та середнє квадратичне відхилення вибірки σ_B ;
- 2) знайти моду Mo^* та медіану Me^* ;
- 3) коефіцієнт варіації V , коефіцієнт асиметрії A_S^* та ексцесу E_S^* .

1.6. Побудувати полігон, гістограму відносних частот та графік емпіричної функції розподілу, користуючись даними таблиці 1.2.

Таблиця 1.2

Номер інтервалу, i	Межі інтервалу		Частота, n_i	Накопичена частота	Відносна частота, n_i/n	Накопичена відносна частота
	нижня	верхня				
1	12	19	5	5	0.1	0.1
2	19	26	13	18	0.26	0.36
3	26	33	11	29	0.22	0.58
4	33	40	12	41	0.24	0.82
5	40	47	5	46	0.1	0.92
6	47	54	2	48	0.04	0.96
7	54	61	2	50	0.04	1

1.7. Користуючись даними таблиці 1.2

- 1) обчислити розмах вибірки R , вибіркоче середнє \bar{x}_A , вибіркочу дисперсію D_B та середнє квадратичне відхилення вибірки σ_B ;
- 2) знайти моду Mo^* та медіану Me^* ;
- 3) обчислити коефіцієнт варіації V , коефіцієнт асиметрії A_S^* та ексцесу E_S^* . Знайти похибку коефіцієнтів асиметрії та ексцесу.

1.8. За заданим двовимірним статистичним розподілом вибірки ознак X і Y

$Y=y_i$	$X=x_j$					n_{y_i}
	5	10	15	20		
100	-	2	4	8		14
120	8	-	6	4		18
140	4	6	-	5		15
160	8	-	15	-		23
n_{x_j}	20	8	25	17		$n=70$

потрібно:

- 1) обчислити кореляційний момент та вибіркочий коефіцієнт кореляції K_{xy}^*, r_B ;
- 2) побудувати умовні статистичні розподіли $Y / X=15, X / Y=120$ й обчислити умовні числові характеристики.

1.9. За заданим двовимірним статистичним розподілом вибірки ознак X і Y

Y	X								n_y
	18	23	28	33	38	43	48		
12.5	-	1	-	-	-	-	-	-	1
15.0	1	2	5	-	-	-	-	-	8
17.5	-	3	2	12	-	-	-	-	8
20.0	-	-	1	8	7	-	-	-	16
22.5	-	-	-	-	3	3	-	-	6
25.0	-	-	-	-	-	1	1	-	2
n_x	1	6	8	20	10	4	1		$n = 50$

потрібно:

1) обчислити кореляційний момент та вибірковий коефіцієнт кореляції K_{xy}^* , r_B ;

2) побудувати умовні статистичні розподіли $Y/X=33$, $X/Y=20$ й обчислити умовні числові характеристики.

1.10. Обчислити вибірковий коефіцієнт кореляції, знайти рівняння вибіркової лінійної регресії y на x , та побудувати діаграму розсіювання за вибірковими даними:

x_i	0	1	2	4	7	10
y_i	3.5	2.7	2.2	-2.0	-1.3	0.6

1.11. Побудувати рівняння регресії залежності заробітної плати y (грн.) працівників закладів освіти від віку x (роки):

Вік	40	42	44	46	48	50	52	54	56
Заробітна плата	2330	2385	2353	2447	2459	2542	2570	2619	2690

За одержаним рівнянням регресії спрогнозувати рівень середньої заробітної плати працівника, який досягне віку 60 та 65 років.

Глава 2. Статистичні оцінки параметрів розподілу

2.1. Постановка задачі оцінювання параметрів розподілу

У багатьох задачах вид теоретичного розподілу досліджуваної ознаки генеральної сукупності можна вважати відомим. Але у кожному конкретному випадку виникає потреба оцінити (наближено обчислити) значення параметрів цього розподілу за наявними вибірковими даними.

Наприклад, якщо є підстави вважати, що генеральна сукупність має нормальний розподіл, то треба оцінити математичне сподівання a та дисперсію σ^2 цього розподілу, оскільки вказані два параметри повністю визначають нормальний розподіл.

Одержана на основі обробки вибірки інформація про конкретну ознаку генеральної сукупності завжди містить певні похибки, оскільки об'єм вибірки набагато менший від об'єму генеральної сукупності. Як ми уже наголошували вище, вибірку слід організувати так, щоб вона була репрезентативною і забезпечувала можливість якомога точніше оцінити параметри розподілу генеральної сукупності.

Нехай x_1, x_2, \dots, x_n – вибірка об'єму n із генеральної сукупності з відомою функцією розподілу $F(x, \theta)$, яка залежить від невідомого параметра θ . Наприклад, для нормального розподілу $\theta = (a, \sigma^2)$. Задача

знаходження параметра θ полягає в побудові наближених формул $\theta \approx h(x_1, x_2, \dots, x_n)$, де h – функція від вибірки.

2.2. Точкові статистичні оцінки параметрів розподілу

Точковою оцінкою параметра θ називають довільну функцію від вибірки $\theta^* = h(x_1, x_2, \dots, x_n)$. Зрозуміло, що така статистична оцінка θ^* визначається одним числом, тобто точкою.

Зауважимо, що оцінюваний параметр генеральної сукупності $\theta = \text{const}$, а його статистична оцінка $\theta^* = h(x_1, x_2, \dots, x_n)$, яку називають **статистикою**, випадкова величина. Вважають, що до реалізації вибірки кожна її варіанта є випадковою величиною, що має закон розподілу ймовірностей досліджуваної ознаки генеральної сукупності з відповідними *теоретичними* числовими характеристиками:

$$M(x_i) = \bar{x}_T = M(x), \quad D(x_i) = D_T, \quad \sigma(x_i) = \sigma_T.$$

Оскільки θ^* є випадковою величиною, точкову статистичну оцінку називають **незмщеною оцінкою невідомого параметра θ** , коли математичне сподівання цієї оцінки точно дорівнює оцінюваному параметру розподілу генеральної сукупності θ ,

$$M(\theta^*) = \theta, \tag{54}$$

і **зміщеною** відносно параметра генеральної сукупності θ , коли

$$M(\theta^*) \neq \theta. \tag{55}$$

Різниця $\theta^* - \theta = \delta$ називається **зміщенням статистичної оцінки θ^*** .

Вимога незміщеності оцінки гарантує відсутність систематичних помилок при оцінюванні.

Зауважимо, що оцінюваний параметр θ може мати кілька точкових незміщених статистичних оцінок. Але помилково вважати, що незміщена оцінка дає завжди добре наближення параметра θ . Серед незміщених оцінок значення θ^* можуть бути сильно розсіяні навколо свого середнього значення, тобто дисперсія $D(\theta^*)$ може бути великою.

Міру розсіювання θ^* відносно θ описує величина $M|\theta^* - \theta|^2$, яка називається **середнім квадратом похибки**, або **середньоквадратичною похибкою оцінки θ^*** .

Точкова незміщена оцінка θ^* параметра θ називається **ефективною**, коли при заданому обсязі вибірки вона має мінімальну дисперсію серед усіх інших незміщених оцінок параметра θ . Отже, ефективна оцінка – це оцінка з найменшим середнім квадратичним відхиленням.

Ефективність оцінки $\tilde{\theta}$ визначають відношенням

$$e = \frac{D(\theta^*)}{D(\tilde{\theta})}, \quad (56)$$

де $D(\theta^*)$ і $D(\tilde{\theta})$ – дисперсії відповідно ефективної і даної оцінок. Чим ближче e до 1, тим ефективнішою є дана оцінка. Якщо $e \rightarrow 1$ при $n \rightarrow \infty$, то така оцінка називається асимптотично ефективною.

Точкова статистична оцінка називається **спроможною**, якщо у разі необмеженого збільшення об'єму вибірки θ^* наближається до оцінюваного параметра θ , а саме:

$$\lim_{n \rightarrow \infty} P(|\theta^* - \theta| < \delta) = 1, \quad (57)$$

тобто, “точність” оцінки збільшується зі збільшенням об'єму вибірки.

2.3. Методи визначення точкових статистичних оцінок параметрів генеральної сукупності

Існують три основні методи визначення точкових статистичних оцінок для параметрів генеральної сукупності.

Метод аналогій. Даний метод базується на тому, що для параметрів генеральної сукупності вибирають такі самі параметри вибірки, тобто для оцінювання $M(x) = \bar{x}_T$, D_T вибирають аналогічні статистики – \bar{x}_A, D_B .

Метод найменших квадратів. Статистичні оцінки цим методом визначаються з умови мінімізації суми квадратів відхилень варіант вибірки від статистичної оцінки θ^* .

Використовуючи метод найменших квадратів, можна, наприклад, визначити точкову статистичну оцінку для $M(x) = \bar{x}_T$. Для цього скористаємося функцією $u = \sum_{i=1}^n (x_i - \theta^*)^2 n_i$. Згідно з необхідною умовою екстремуму, дістанемо:

$$\frac{\partial u}{\partial \theta^*} = -2 \sum_{i=1}^n (x_i - \theta^*)^2 n_i = 0;$$

$$\sum_{i=1}^n x_i n_i - \sum_{i=1}^n \theta^* n_i = 0, \quad \text{звідки } \theta^* = \frac{\sum_{i=1}^n x_i n_i}{n} = \bar{x}_B.$$

Отже, для $\theta = \bar{x}_T$ точковою статистичною оцінкою буде вибіркове середнє, тобто $\theta^* = \bar{x}_B$.

Метод максимальної вірогідності. Нехай ознака генеральної сукупності X визначається параметром θ і має щільність розподілу ймовірностей $f(x, \theta)$. При реалізації вибірки з варіантами x_1, x_2, \dots, x_n в силу того, що x_1, x_2, \dots, x_n – незалежні, однаково розподілені випадкові величини, щільність розподілу вибірки буде

$$f(x_1, x_2, \dots, x_n, \theta^*) = f(x_1, \theta^*) \cdot f(x_2, \theta^*) \cdot \dots \cdot f(x_n, \theta^*). \quad (58)$$

Вона називається **функцією вірогідності** і позначається $L = L(\theta^*)$. При цьому варіанти розглядаються як незалежні випадкові величини, котрі мають один і той самий закон розподілу, що й досліджувана ознака генеральної сукупності X .

Суть цього методу полягає в тому, що фіксуючи значення варіант x_1, x_2, \dots, x_n , визначають таке значення параметра θ^* , при якому функція (58) максимізується. **Оцінкою максимальної вірогідності параметра θ** називають те значення параметра θ , при якому функція вірогідності досягає найбільшого значення при заданих x_1, x_2, \dots, x_n , тобто є розв'язком рівняння $L(\theta^*) = \max L(\theta)$.

Так, наприклад, якщо ознака генеральної сукупності X має нормальний закон розподілу, то функція вірогідності набере вигляду:

$$f(x_1, x_2, \dots, x_n, \theta_1^*, \theta_2^*) = \frac{1}{(2\pi\theta_2^*)^{\frac{n}{2}}} \cdot e^{-\frac{\sum_{i=1}^n (x_i - \theta_1^*)^2}{2\theta_2^*}}. \quad (59)$$

При цьому за статистичні оцінки θ_1^* і θ_2^* вибирають ті їх значення, за яких задана вибірка буде найімовірнішою, тобто функція (59) досягає максимуму.

Оскільки $L(\theta^*)$ і $\ln L(\theta^*)$ досягають найбільшого значення у одних і тих самих точках, то зручно від функції (59) перейти до її логарифма:

$$\begin{aligned} \ln f(x_1, x_2, \dots, x_n, \theta_1^*, \theta_2^*) &= \\ &= L(x_1, x_2, \dots, x_n, \theta_1^*, \theta_2^*) = -\frac{n}{2} (\ln \pi + \ln \theta_2^*) - \frac{\sum_{i=1}^n (x_i - \theta_1^*)^2}{2\theta_2^*}. \end{aligned}$$

Використовуючи необхідні умови екстремуму для цієї функції, маємо:

$$\begin{cases} \frac{\partial L}{\partial \theta_1^*} = -\frac{1}{\theta_2^*} \sum_{i=1}^n (x_i - \theta_1^*) = 0 \\ \frac{\partial L}{\partial \theta_2^*} = -\frac{n}{2\theta_2^*} + \frac{1}{2(\theta_2^*)^2} \sum_{i=1}^n (x_i - \theta_1^*)^2 = 0 \end{cases} \quad (60)$$

З першого рівняння системи (7) дістанемо:

$$\begin{aligned} \sum_{i=1}^n x_i - n\theta_1^* &= 0, \\ \theta_1^* &= \frac{1}{n} \cdot \sum_{i=1}^n x_i = \bar{x}_B, \end{aligned} \quad (61)$$

а з другого рівняння системи (60), враховуючи (61):

$$\begin{aligned}\theta_2^* n &= \sum_{i=1}^n (x_i - \bar{x}_B)^2, \\ \theta_2^* &= \frac{1}{n} \cdot \sum_{i=1}^n (x_i - \bar{x}_B)^2 = D_B.\end{aligned}\quad (62)$$

Таким чином, для середнього генеральної сукупності $M(x) = \bar{x}_T$ точковою статистичною оцінкою є вибіркве середнє \bar{x}_B , а для теоретичної дисперсії D_T – вибірква дисперсія D_B .

2.4. Властивості точкових оцінок для середнього та дисперсії генеральної сукупності \bar{x}_B , D_B . виправлена дисперсія, виправлене середнє квадратичне відхилення

Точковою незміщеною статистичною оцінкою для середнього генеральної сукупності $\bar{x}_T = M(X)$ є вибіркве середнє \bar{x}_B .

Дійсно,

$$M(\bar{x}_B) = M\left(\frac{\sum_{i=1}^n x_i}{n}\right) = \frac{\sum_{i=1}^n M(x_i)}{n} = \left| M(x_i) = \bar{x}_T = a \right| = \frac{\sum_{i=1}^n a}{n} = \frac{na}{n} = a.$$

Отже, $M(\bar{x}_B) = \bar{x}_T$.

Перевіримо на незміщеність статистичну оцінку D_B .

$$\begin{aligned}M(D_B) &= M\left(\frac{\sum_{i=1}^n (x_i - \bar{x}_B)^2}{n}\right) = M\left(\frac{\sum_{i=1}^n ((x_i - a) - (\bar{x}_B - a))^2}{n}\right) = \\ &= M \frac{\sum_{i=1}^n ((x_i - a)^2 - 2(x_i - a)(\bar{x}_B - a) + (\bar{x}_B - a)^2)}{n} = \\ &= M \frac{\sum_{i=1}^n (x_i - a)^2 - 2\sum_{i=1}^n (x_i - a)(\bar{x}_B - a) + \sum_{i=1}^n (\bar{x}_B - a)^2}{n} = \\ &= M \frac{\sum_{i=1}^n (x_i - a)^2 - 2(\bar{x}_B - a)\sum_{i=1}^n (x_i - a) + (\bar{x}_B - a)^2 n}{n} =\end{aligned}$$

$$\begin{aligned}
&= M \frac{\sum_{i=1}^n (x_i - a)^2 - 2(\bar{x}_B - a) \left(\sum_{i=1}^n x_i - \sum_{i=1}^n a \right) + (\bar{x}_B - a)^2 n}{n} = \\
&= M \frac{\sum_{i=1}^n (x_i - a)^2 - 2(\bar{x}_B - a)(n\bar{x}_B - na) + (\bar{x}_B - a)^2 n}{n} = \\
&= M \frac{\sum_{i=1}^n (x_i - a)^2 - 2n(\bar{x}_B - a)^2 + (\bar{x}_B - a)^2 n}{n} = \\
&= M \frac{\sum_{i=1}^n (x_i - a)^2}{n} - M(\bar{x}_B - a)^2 =
\end{aligned}$$

$$= \left[\begin{array}{l} M(x_i - a)^2 = D_T, \\ M(\bar{x}_B - a)^2 = D(\bar{x}_B) = D\left(\frac{\sum_{i=1}^n x_i}{n}\right) = \frac{\sum_{i=1}^n D(x_i)}{n^2} = \frac{D_T}{n}, \end{array} \right] =$$

$$\begin{aligned}
&= M \frac{\sum_{i=1}^n (x_i - a)^2}{n} - M(\bar{x}_B - a)^2 = \frac{\sum_{i=1}^n D_T}{n} - \frac{D_T}{n} = \\
&= \frac{nD_T}{n} - \frac{D_T}{n} = D_T - \frac{1}{n} D_T = \left(1 - \frac{1}{n}\right) D_T = \frac{n-1}{n} D_T.
\end{aligned}$$

Остаточно маємо:

$$M(D_B) = \frac{n-1}{n} D_T.$$

Таким чином, вибіркова дисперсія D_B є точковою зміщеною статистичною оцінкою для дисперсії генеральної сукупності D_T , де $\frac{n-1}{n}$ - коефіцієнт зміщення, який зменшується зі збільшенням обсягу вибірки n .

Помножимо D_B на $\frac{n}{n-1}$ і знайдемо математичне сподівання:

$$M\left(\frac{n}{n-1} D_B\right) = \frac{n}{n-1} M(D_B) = \frac{n}{n-1} \cdot \frac{n-1}{n} D_T = D_T.$$

Отже, $\frac{n}{n-1}D_B$ буде точковою незміщеною статистичною оцінкою для теоретичної дисперсії D_T . Вона називається **виправленою дисперсією** і позначається S^2 .

Маємо, що точковою незміщеною статистичною оцінкою для D_T є виправлена дисперсія $S^2 = \frac{n}{n-1}D_B$, або

$$S^2 = \frac{n}{n-1} \cdot \frac{\sum_{i=1}^n (x_i - \bar{x}_B)^2}{n} = \frac{\sum_{i=1}^n (x_i - \bar{x}_B)^2}{n-1}. \quad (63)$$

Величину

$$S = \sqrt{\frac{n}{n-1}D_B} \quad (64)$$

називають **виправленим середнім квадратичним відхиленням**.

Виправлене середнє квадратичне відхилення буде зміщеною точковою статистичною оцінкою для теоретичного середнього квадратичного відхилення σ_T , оскільки

$$M(S) = \sqrt{\frac{2}{k}} \cdot \frac{\tilde{A}(\frac{k+1}{2})}{\tilde{A}(\frac{k}{2})} \sigma_T, \quad (65)$$

де $k = n - 1$ – число ступенів свободи (k дорівнює числу незалежних змінних мінус число зв'язків між ними);

$$\sqrt{\frac{2}{k}} \cdot \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})} - \text{коефіцієнт зміщення, } \Gamma(y) = \int_0^{\infty} e^{-t} t^{y-1} dt.$$

Приклад 2.1. За інтервальним статистичним розподілом вибірки з прикладу 1.5 знайти точкові незміщені оцінки для математичного сподівання та дисперсії досліджуваної ознаки генеральної сукупності.

Розв'язання. Оскільки точковою незміщеною оцінкою для $\bar{x}_T = M(X)$ є \bar{x}_B , то можемо використати результати прикладу 1.7.

$$\bar{x}_B = \frac{\sum_{i=1}^5 x_i^* n_i}{n} = 16,6.$$

Нагадаємо, що вибіркове середнє обчислено для відповідного дискретного статистичного розподілу вибірки і варіанти x_i^* є серединами часткових інтервалів. Тобто, точковою незміщеною оцінкою для математичного сподівання є $\bar{x}_B = 16,6$.

Для визначення точкової незміщеної статистичної оцінки для D_T скористаємося обчисленою в прикладі 1.7 вибірковою дисперсією $D_B=145,44$. Тоді точкова незміщена статистична оцінка для D_T має вигляд:

$$S^2 = \frac{n}{n-1} D_B = \frac{50}{50-1} \cdot 145,44 \approx 148,41.$$

2.5. Закони розподілу ймовірностей для вибіркового середнього \bar{x}_B , виправленої дисперсії S^2 , виправленого середнього квадратичного відхилення S

Числові характеристики вибірки є випадковими величинами, що мають певні закони розподілу ймовірностей. На підставі центральної граничної теореми теорії ймовірностей вибіркоче середнє \bar{x}_B матиме нормальний закон розподілу з такими числовими характеристиками:

математичне сподівання

$$M(\bar{x}_B) = M\left(\frac{\sum_{i=1}^k x_i n_i}{n}\right) = \frac{1}{n} \sum_{i=1}^k M(x_i) \cdot n_i = \frac{1}{n} \sum_{i=1}^k a \cdot n_i = a \frac{\sum_{i=1}^k n_i}{n} = a,$$

$$(a = M(x) = \bar{x}_T);$$

дисперсія

$$D(\bar{x}_B) = D\left(\frac{\sum_{i=1}^k x_i n_i}{n}\right) = \frac{D_T}{n};$$

середньоквадратичне відхилення

$$\sigma(\bar{x}_B) = \frac{\sigma_T}{\sqrt{n}}.$$

Отже, випадкова величина \bar{x}_B має закон розподілу $N\left(a; \frac{\sigma_T}{\sqrt{n}}\right)$.

Нехай ознака генеральної сукупності X має нормальний закон розподілу $N(a; \sigma)$, тобто $X \sim N(a; \sigma)$. При реалізації вибірки кожному з варіант $X = x_i$ розглядають як випадкову величину, що також має закон розподілу $N(a; \sigma)$. При цьому варіанти вибірки є незалежними, тобто

$K_{ij} = 0$, а випадкова величина $z = \frac{x_i - a}{\sigma}$ відповідно матиме закон розподілу $N(0;1)$.

Можна показати, що випадкова величина

$$\frac{(n-1)}{\sigma^2} S^2 = \sum_{i=1}^{n-1} \left(\frac{y_i}{\sigma} \right)^2$$

матиме розподіл χ^2 із $k = n - 1$ ступенями свободи.

Звідси випливає, що випадкова величина $\frac{\sqrt{n-1}}{\sigma} S$ матиме розподіл χ із $k = n - 1$ ступенями свободи. Отже, випадкова величина S^2 має закон розподілу $\frac{\chi^2(n-1)}{n-1} \sigma^2$, випадкова величина S має закон розподілу $\frac{\chi(n-1)}{\sqrt{n-1}} \sigma$.

2.6. Інтервальне оцінювання

При заміні істинного значення параметра розподілу генеральної сукупності θ його точковою оцінкою θ^* потрібно знати можливу похибку, яка виникає при використанні такої оцінки. Навіть така “гарна” точкова оцінка θ^* , яка є незміщеною (в середньому співпадає з θ), спроможною (наближається до θ із збільшенням об’єму вибірки) і ефективною (має мінімальну міру відхилення від θ) може істотно відрізнятись від справжнього значення параметра θ . Для того, щоб мати уявлення про точність і надійність оцінки θ^* параметра θ використовують інтервальні статистичні оцінки.

При інтервальному оцінюванні вказують такий інтервал, що покриває оцінюваний параметр θ генеральної сукупності з заданою ймовірністю γ . Величину γ вибирають заздалегідь.

Нехай для заданого $\gamma \in (0;1)$ існує таке $\delta > 0$, що

$$P(|\theta^* - \theta| < \delta) = \gamma \quad (P(\theta^* - \delta < \theta < \theta^* + \delta) = \gamma). \quad (66)$$

Тоді випадковий інтервал $(\theta^* - \delta; \theta^* + \delta)$ називають **надійним** (або **довірчим**) **інтервалом**, число γ – **надійністю** (або **надійним рівнем**), δ – **точністю оцінки**, значення $\theta^* - \delta$, $\theta^* + \delta$ – відповідно **нижньою** і **верхньою надійними межами**.

2.7. Надійні інтервали для параметрів нормального закону

Нехай ознака X генеральної сукупності має нормальний закон розподілу $N(a; \sigma)$. Побудуємо довірчий інтервал для математичного сподівання, знаючи числове значення середнього квадратичного відхилення генеральної сукупності σ_T , із заданою надійністю γ . Оскільки випадкова величина \bar{x}_B ,

як точкова незміщена статистична оцінка для $\bar{x}_T = M(X)$, має нормальний закон розподілу $N\left(a; \frac{\sigma_T}{\sqrt{n}}\right)$, то дістанемо

$$P(|\bar{x}_B - a| < \delta) = \gamma, \quad (67)$$

де випадкова величина $\bar{x}_B - a$ має нормальний закон розподілу з числовими характеристиками

$$M(\bar{x}_B - a) = M(\bar{x}_B) - a = a - a = 0;$$

$$D(\bar{x}_B - a) = D(\bar{x}_B) = \frac{D_T}{n};$$

$$\sigma(\bar{x}_B) = \frac{\sigma_T}{\sqrt{n}}.$$

Тому випадкова величина $\frac{\bar{x}_B - a}{\frac{\sigma_T}{\sqrt{n}}}$ матиме стандартний нормальний

закон розподілу $N(0;1)$.

Позначимо $\frac{\delta}{\frac{\sigma_T}{\sqrt{n}}} = x$ і перепишемо (67) у вигляді

$$P\left(\left|\frac{\bar{x}_B - a}{\frac{\sigma_T}{\sqrt{n}}}\right| < x\right) = \gamma \quad (68)$$

або

$$P\left(\bar{x}_B - \frac{x \cdot \sigma_T}{\sqrt{n}} < a < \bar{x}_B + \frac{x \cdot \sigma_T}{\sqrt{n}}\right) = \gamma.$$

Тобто,

$$P\left(\left|\frac{\bar{x}_B - a}{\frac{\sigma_T}{\sqrt{n}}}\right| < x\right) = 2\Phi(x) - 1 = \gamma, \quad (69)$$

де $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left\{-\frac{z^2}{2}\right\} dz$ - функція розподілу стандартної нормальної випадкової величини (див. таблицю 8.3).

З рівності (69) знаходимо

$$\Phi(x) = \frac{1 + \gamma}{2}.$$

Останнє рівняння розв'язуємо відносно x за допомогою таблиць функції розподілу стандартної нормальної випадкової величини.

Таким чином, шуканий довірчий інтервал матиме вигляд:

$$\bar{x}_B - \frac{x \cdot \sigma_T}{\sqrt{n}} < a < \bar{x}_B + \frac{x \cdot \sigma_T}{\sqrt{n}} . \quad (70)$$

Величина $\frac{x \cdot \sigma_T}{\sqrt{n}}$ називається **точністю оцінки**, або **похибкою вибірки**.

Зауважимо, коли є таблиці функції Лапласа $\Phi_1(x) = \frac{1}{\sqrt{2\pi}} \int_0^x \exp\left\{-\frac{z^2}{2}\right\} dz$, то x буде розв'язком рівняння $\Phi_1(x) = \frac{\gamma}{2}$.

Приклад 2.1. Знайти інтервальну оцінку з надійністю $\gamma=0,95$ для математичного сподівання a нормально розподіленої генеральної сукупності, якщо відомо $\sigma_T=5$, а за вибіркою об'єму $n=25$ знайшли вибіркове середнє $\bar{x}_B=14$.

Розв'язання. З умови задачі маємо: $\bar{x}_B=14$, $\sigma_T=5$, $n=25$. Величину x знаходимо за таблицею 8.3 як корінь рівняння $\Phi(x) = \frac{1+\gamma}{2}$, тобто $\Phi(x) = 0,975$. Маємо $x=1,96$.

Знайдемо числові значення нижньої і верхньої надійних меж:

$$\bar{x}_B - \frac{x \cdot \sigma_T}{\sqrt{n}} = 14 - \frac{1,96 \cdot 5}{\sqrt{25}} = 12,04 .$$

$$\bar{x}_B + \frac{x \cdot \sigma_T}{\sqrt{n}} = 14 + \frac{1,96 \cdot 5}{\sqrt{25}} = 15,96 .$$

Підставимо у формулу (70): $12,04 < a < 15,96$.

Отже, з надійністю 0,95 (95% гарантії) математичне сподівання a нормально розподіленої генеральної сукупності покривається інтервалом (12,04;15,96).

Приклад 2.2. Знайти мінімальний об'єм вибірки при якому з надійністю 0,975 точність оцінки математичного сподівання a нормально розподіленої генеральної сукупності $\delta=0,3$, якщо відомо $\sigma_T=1,2$.

Розв'язання. Похибка вибірки знаходиться за формулою

$$\delta = \frac{x \cdot \sigma_T}{\sqrt{n}} ,$$

звідки

$$n = \frac{x^2 \sigma_T^2}{\delta^2} .$$

Знайдемо корінь рівняння $\Phi(x) = \frac{1+\gamma}{2}$, де $\gamma=0,975$. Звідки $\Phi(x) = 0,9875$. Отже, $x = 2,24$.

Обчислюємо мінімальний об'єм вибірки

$$n = \frac{(2,24)^2 \cdot (1,2)^2}{(0,3)^2} = 81.$$

Для побудови довірчого інтервалу, який оцінює математичне сподівання a нормально розподіленої генеральної сукупності при невідомому середньоквадратичному відхиленні із заданою надійністю γ , застосовується випадкова величина

$$t = \frac{\bar{x}_B - a}{\frac{S}{\sqrt{n}}}, \quad (71)$$

що має розподіл Стьюдента з $k = n - 1$ ступенями свободи.

При цьому

$$P\left(\left|\frac{\bar{x}_B - a}{\frac{S}{\sqrt{n}}}\right| < t_\gamma\right) = P\left(\bar{x}_B - \frac{t_\gamma \cdot S}{\sqrt{n}} < a < \bar{x}_B + \frac{t_\gamma \cdot S}{\sqrt{n}}\right) = \gamma.$$

Для побудови довірчого інтервалу обчислюємо за даним статистичним розподілом вибіркове середнє \bar{x}_B , виправлене середнє квадратичне відхилення S і визначаємо за таблицею розподілу Стьюдента (таблиця 8.6) значення $t_\gamma = t(\gamma, n)$ при заданому γ і n . Шуканий надійний інтервал має вигляд:

$$\bar{x}_B - \frac{t_\gamma \cdot S}{\sqrt{n}} < a < \bar{x}_B + \frac{t_\gamma \cdot S}{\sqrt{n}}. \quad (72)$$

Приклад 2.3. З надійністю $\gamma = 0,95$ побудувати довірчий інтервал вигляду (72) за вибіркою 4; 2; 1; -2; 3; -2; 2; 5; 4; 3.

Розв'язання. Знайдемо вибіркове середнє і виправлене середнє квадратичне відхилення. Для спрощення обчислень побудуємо дискретний статистичний розподіл вибірки

x_i	-2	1	2	3	4	5
n_i	2	1	2	2	2	1

Обчислимо \bar{x}_B :

$$\bar{x}_B = \frac{\sum_{i=1}^6 x_i n_i}{n} = \frac{(-2) \cdot 2 + 1 \cdot 1 + 2 \cdot 2 + 3 \cdot 2 + 4 \cdot 2 + 5 \cdot 1}{10} = \frac{20}{10} = 2.$$

Для визначення D_B спочатку знайдемо

$$\frac{\sum_{i=1}^6 x_i^2 n_i}{n} = \frac{(-2)^2 \cdot 2 + 1^2 \cdot 1 + 2^2 \cdot 2 + 3^2 \cdot 2 + 4^2 \cdot 2 + 5^2 \cdot 1}{10} = 9,2.$$

$$D_B = \frac{\sum_{i=1}^6 x_i^2 n_i}{n} - (\bar{x}_B)^2 = 9,2 - (2)^2 = 5,2.$$

Виправлене середнє квадратичне відхилення

$$S = \sqrt{\frac{n}{n-1} D_B} = \sqrt{\frac{10}{10-1} \cdot 5,2} \approx 2,4.$$

За таблицю значень розподілу Стьюдента (таблиця 8.6) за заданою надійністю $\gamma = 0,95$ і $n = 10$ знаходимо значення $t(\gamma = 0,95; n = 10) = 2,26$.

Знайдемо числові значення нижньої і верхньої надійних меж:

$$\bar{x}_B - \frac{t_\gamma S}{\sqrt{n}} = 2 - \frac{2,26 \cdot 2,4}{\sqrt{10}} \approx 0,3.$$

$$\bar{x}_B + \frac{t_\gamma S}{\sqrt{n}} = 2 + \frac{2,26 \cdot 2,4}{\sqrt{10}} \approx 3,7.$$

Отже, з надійністю $\gamma = 0,95$ можна стверджувати, що $0,3 < a < 3,7$.

Зауважимо, що при великих обсягах вибірки ($n > 30$) на підставі центральної граничної теореми розподіл Стьюдента наближається до нормального закону розподілу і значення t_γ знаходиться за таблицею значень функції Лапласа $\Phi_1(x)$. При цьому $\Phi_1(x) = \Phi(x) - \frac{1}{2}$, де значення $\Phi(x)$ містяться в таблиці 8.3 даного посібника.

2.8. Побудова довірчих інтервалів із заданою надійністю γ для дисперсії D_T та середнього квадратичного відхилення σ_T

Якщо досліджувана ознака X має нормальний закон розподілу, то для побудови довірчого інтервалу із заданою надійністю γ для D_T , σ_T використаємо випадкову величину

$$\chi^2 = \frac{n-1}{\sigma_T^2} S^2, \quad (73)$$

що має розподіл χ^2 із $k = n - 1$ ступенями свободи.

В рівність $P(\chi_1^2 < \chi^2 < \chi_2^2) = P\left(\frac{1}{\chi_2^2} < \frac{1}{\chi^2} < \frac{1}{\chi_1^2}\right)$ для рівноймовірних випадкових подій $(\chi_1^2 < \chi^2 < \chi_2^2)$ і $\left(\frac{1}{\chi_2^2} < \frac{1}{\chi^2} < \frac{1}{\chi_1^2}\right)$ підставимо $\chi^2 = \frac{n-1}{\sigma_T^2} S^2$:

$$\begin{aligned} P\left(\frac{1}{\chi_2^2} < \frac{1}{\chi^2} < \frac{1}{\chi_1^2}\right) &= P\left(\frac{1}{\chi_2^2} < \frac{1}{\frac{n-1}{\sigma_T^2} S^2} < \frac{1}{\chi_1^2}\right) = \\ &= P\left(\frac{1}{\chi_2^2} < \frac{\sigma_T^2}{(n-1)S^2} < \frac{1}{\chi_1^2}\right) = P\left(\frac{(n-1)S^2}{\chi_2^2} < \sigma_T^2 < \frac{(n-1)S^2}{\chi_1^2}\right) = \gamma. \end{aligned}$$

Тобто, довірчий інтервал для D_T можна записати у вигляді:

$$\frac{n-1}{\chi_2^2} S^2 < D_T < \frac{n-1}{\chi_1^2} S^2, \quad (74)$$

звідки довірчий інтервал для σ_T :

$$\frac{\sqrt{n-1}}{\chi_2} S < \sigma_T < \frac{\sqrt{n-1}}{\chi_1} S, \quad (75)$$

де значення χ_1^2, χ_2^2 знаходимо за таблицею 8.4 ймовірностей $P(\chi^2 > \chi_{\alpha;k}^2)$. Величини χ_1^2, χ_2^2 вибирають так, щоб

$$P(\chi^2 < \chi_1^2) = P(\chi^2 > \chi_2^2) = \frac{1-\gamma}{2}.$$

Врахуємо, що $P(\chi^2 < \chi_1^2) = 1 - P(\chi^2 > \chi_1^2)$, тобто умова $P(\chi^2 < \chi_1^2) = \frac{1-\gamma}{2}$

рівносильна умові $P(\chi^2 > \chi_1^2) = 1 - \frac{1-\gamma}{2} = \frac{1+\gamma}{2}$. Отже, значення χ_1^2, χ_2^2

знаходимо за таблицею 8.4 з використанням рівностей:

$$P(\chi^2 > \chi_1^2) = \frac{1+\gamma}{2}; \quad (76)$$

$$P(\chi^2 > \chi_2^2) = \frac{1-\gamma}{2}. \quad (77)$$

Приклад 2.4. За вибіркою 4; 2; 1; -2; 3; -2; 2; 5; 4; 3 з нормально розподіленої генеральної сукупності побудувати довірчі інтервали з надійністю $\gamma = 0,9$ для дисперсії D_T та середнього квадратичного відхилення σ_T .

Розв'язання. Для побудови довірчих інтервалів скористаємось формулами (74) і (75). виправлене середнє квадратичне відхилення було знайдено в прикладі 2.3: $S \approx 2,4$. виправлена дисперсія $S^2 = (2,4)^2 \approx 5,8$.

Кількість ступенів свободи $k = n - 1 = 9$. Для $\gamma = 0,9$ маємо $\frac{1-\gamma}{2} = 0,05$. $\frac{1+\gamma}{2} = 0,95$. За таблицею 8.4 для ймовірностей 0,95 і 0,05 при $k = 9$ знаходимо $\chi_1^2 = 3,3$ та $\chi_2^2 = 16,9$. Тоді довірчий інтервал з надійністю $\gamma = 0,9$ для дисперсії D_T за (74) має вигляд:

$$\frac{9}{16,9} 5,8 < D_T < \frac{9}{3,3} 5,8 \text{ або } 3,1 < D_T < 15,8.$$

Довірчий інтервал з надійністю $\gamma = 0,9$ для σ_T :

$$\sqrt{3,1} < \sigma_T < \sqrt{15,8} \text{ звідки } 1,8 < \sigma_T < 4,0.$$

Довірчий інтервал для σ_T із заданою надійністю γ можна побудувати з використанням розподілу χ . Маємо:

$$P(S - \delta < \sigma_T < S + \delta) = \gamma,$$

звідки

$$P\left(S\left(1 - \frac{\delta}{S}\right) < \sigma_T < S\left(1 + \frac{\delta}{S}\right)\right) = \gamma,$$

або

$$P(S(1 - q) < \sigma_T < S(1 + q)) = \gamma, \quad \text{де } q = \frac{\delta}{S}.$$

Для знаходження q розглянемо випадкову величину

$$\chi = \frac{S}{\sigma_T} \sqrt{n-1},$$

що має розподіл χ (хі-розподіл).

Розглянемо рівність

$$P(S(1 - q) < \sigma_T < S(1 + q)) = P\left(\frac{1}{S(1 + q)} < \frac{1}{\sigma_T} < \frac{1}{S(1 - q)}\right),$$

яка є вірною при $q < 1$.

Якщо помножити всі члени подвійної нерівності

$$\frac{1}{S(1+q)} < \frac{1}{\sigma_T} < \frac{1}{S(1-q)}$$

на $S\sqrt{n-1}$, то отримаємо:

$$\begin{aligned} P(S(1-q) < \sigma_T < S(1+q)) &= P\left(\frac{\sqrt{n-1}}{(1+q)} < \frac{S\sqrt{n-1}}{\sigma_T} < \frac{\sqrt{n-1}}{(1-q)}\right) = \\ &= P\left(\frac{\sqrt{n-1}}{(1+q)} < \chi < \frac{\sqrt{n-1}}{(1-q)}\right) = \gamma. \end{aligned} \quad (78)$$

Отже, за заданою надійністю γ і обсягом вибірки n знаходимо за таблицею 8.6 значення величини $q_\gamma = q(\gamma, n)$ і при $q < 1$ записуємо довірчий інтервал для σ_T у вигляді:

$$S(1 - q(\gamma, n)) < \sigma_T < S(1 + q(\gamma, n)). \quad (79)$$

Якщо $q \geq 1$, то відповідний довірчий інтервал для σ_T має вигляд:

$$0 < \sigma_T < S(1 + q(\gamma, n)). \quad (80)$$

Приклад 2.5. Побудувати довірчі інтервали для σ_T з надійностями $\gamma_1 = 0,95$ та $\gamma_2 = 0,99$ за вибіркою об'єму $n = 10$, якщо знайдено $S = 2,4$.

Розв'язання. Для побудови потрібних довірчих інтервалів знайдемо за таблицею 8.6 значення $q_1 = q(\gamma_1, n)$ та $q_2 = q(\gamma_2, n)$.

Маємо $q_1 = 0,65$ та $q_2 = 1,08$, тому довірчий інтервал з надійністю $\gamma_1 = 0,95$ запишемо у вигляді (79), а з надійністю $\gamma_2 = 0,99$ – у вигляді (80). Визначимо величини:

$$S(1 - q_1) = 2,4(1 - 0,65) = 2,4 \cdot 0,35 = 0,84;$$

$$S(1 + q_1) = 2,4(1 + 0,65) = 2,4 \cdot 1,65 = 3,96.$$

$$S(1 + q_2) = 2,4(1 + 1,08) = 2,4 \cdot 2,08 = 4,99.$$

Отже, довірчий інтервал для σ_T з надійністю $\gamma_1 = 0,95$ має вигляд:

$$0,84 < \sigma_T < 3,96;$$

з надійністю $\gamma_2 = 0,99$:

$$0 < \sigma_T < 4,99.$$

2.9. Побудова довірчого інтервалу для коефіцієнта кореляції r_{xy} генеральної сукупності із заданою надійністю γ

Точковою незміщеною статистичною оцінкою для теоретичного коефіцієнта кореляції r_{xy} є вибірковий коефіцієнт кореляції r_B з виправленим середнім квадратичним відхиленням $S = \frac{1-r_B^2}{\sqrt{n}}$.

Якщо центрувати і нормувати випадкову величину r_B , то отримаємо величину

$$x_\gamma = \frac{r_B - r_{xy}}{\sigma(r_B)} = \frac{r_B - r_{xy}}{\frac{1-r_B^2}{\sqrt{n}}},$$

яка має стандартний нормальний закон розподілу $N(0;1)$, тому

$$P\left(\left|\frac{r_B - r_{xy}}{\frac{1-r_B^2}{\sqrt{n}}}\right| < x_\gamma\right) = P\left(r_B - x_\gamma \frac{1-r_B^2}{\sqrt{n}} < r_{xy} < r_B + x_\gamma \frac{1-r_B^2}{\sqrt{n}}\right) = \gamma = 2\Phi(x_\gamma) - 1$$

Отже, довірчий інтервал для r_{xy} із заданою надійністю γ має вигляд:

$$r_B - x_\gamma \frac{1-r_B^2}{\sqrt{n}} < r_{xy} < r_B + x_\gamma \frac{1-r_B^2}{\sqrt{n}}, \quad (81)$$

де x_γ знаходимо за таблицею 8.3 з рівняння

$$\Phi(x_\gamma) = \frac{1+\gamma}{2}. \quad (82)$$

Приклад 2.6. Побудувати довірчий інтервал з надійністю $\gamma = 0,99$ для коефіцієнта кореляції r_{xy} за двовимірним статистичним розподілом вибірки з прикладу 1.8.

Розв'язання. Шуканий довірчий інтервал запишемо у вигляді (81). Скористаємось знайденим у прикладі 1.8 вибірковим коефіцієнтом кореляції $r_B = -0,068$. Запишемо рівняння для знаходження x_γ :

$$\Phi(x_\gamma) = \frac{1+\gamma}{2} = \frac{1,99}{2} = 0,995.$$

За таблицею 8.3 знаходимо $x_\gamma = 2,58$. Підставимо r_B , x_γ та $\sqrt{n} = \sqrt{100} = 10$ в формулу (81):

$$\begin{aligned} -0,068 - 2,58 \cdot \frac{1 - (-0,068)^2}{10} < r_{xy} < -0,068 + 2,58 \cdot \frac{1 - (-0,068)^2}{10}, \\ -0,068 - 0,257 < r_{xy} < -0,068 + 0,257, \\ -0,325 < r_{xy} < 0,189. \end{aligned}$$

Отже, з надійністю $\gamma = 0,99$ отримали, що $r_{xy} \in (-0,325; 0,189)$.

2.10. Побудова довірчого інтервалу для математичного сподівання за допомогою нерівності Чебишова із заданою надійністю γ

Якщо немає впевненості, що досліджувана ознака генеральної сукупності має нормальний розподіл, тоді для побудови довірчого інтервалу для математичного сподівання a із заданою надійністю γ використовують нерівність Чебишова:

$$P(|\bar{x}_B - a| < \delta) \geq 1 - \frac{\sigma_T^2}{n\delta^2} = \gamma. \quad (83)$$

Останню рівність розв'яжемо відносно точності оцінки δ :

$$\delta = \frac{\sigma_T}{\sqrt{(1-\gamma)n}}.$$

Отже, шуканий довірчий інтервал має вигляд:

$$\bar{x}_B - \frac{\sigma_T}{\sqrt{(1-\gamma)n}} < a < \bar{x}_B + \frac{\sigma_T}{\sqrt{(1-\gamma)n}}. \quad (84)$$

Якщо теоретичне середнє квадратичне відхилення невідоме, то його замінюють виправленим вибірковою середнім квадратичним відхиленням S і довірчий інтервал для математичного сподівання a із заданою надійністю γ шукають у вигляді

$$\bar{x}_B - \frac{S}{\sqrt{(1-\gamma)n}} < a < \bar{x}_B + \frac{S}{\sqrt{(1-\gamma)n}}. \quad (85)$$

Приклад 2.7. З надійністю $\gamma = 0,95$ побудувати довірчий інтервал для математичного сподівання a за допомогою нерівності Чебишова, якщо за вибіркою об'єму $n = 10$ знайдено $\bar{x}_B = 2$ і $S = 2,4$.

Розв'язання. Скористаємось формулою (85). Обчислимо

$$\frac{S}{\sqrt{(1-\gamma)n}} = \frac{2,4}{\sqrt{(1-0,95) \cdot 10}} = \frac{2,4}{0,7} \approx 3,4.$$

Шуканий довірчий інтервал має вигляд

$$2 - 3,4 < a < 2 + 3,4 \quad \text{або} \quad -1,4 < a < 5,4.$$

Отже, з надійністю $\gamma = 0,95$ математичне сподівання $a \in (-1,4; 5,4)$.

Зауважимо, що при наявності інформації про нормальний розподіл генеральної сукупності довірчий інтервал для математичного сподівання a шукають за формулою (72). При цьому з такою самою надійністю $\gamma = 0,95$ математичне сподівання a покривається меншим довірчим інтервалом (див. приклад 2.3).

Контрольні питання для самоперевірки до теми „Статистичні оцінки параметрів розподілу”

1. Що таке точкова оцінка невідомого параметра розподілу?
2. Яка точкова статистична оцінка називається незміщеною, зміщеною, ефективною?
3. У чому полягає суть методу найменших квадратів?
4. У чому полягає суть методу максимальної вірогідності?
5. Що є точковою незміщеною статистичною оцінкою для математичного сподівання генеральної сукупності?
6. Що означає точкова незміщена статистична оцінка для дисперсії генеральної сукупності?
7. Що називається виправленою дисперсією, виправленим середнім квадратичним відхиленням?
8. Як знаходять інтервальні оцінки невідомих параметрів розподілу?
9. Що таке надійний (довірчий) інтервал? Як визначається рівень надійності, верхня та нижня межі надійного інтервалу?
10. Як визначається надійний інтервал для математичного сподівання нормально розподіленої генеральної сукупності при відомому значенні середнього квадратичного відхилення із заданою надійністю?
11. Як визначається надійний інтервал для математичного сподівання нормального закону розподілу при невідомому значенні середнього квадратичного відхилення із заданою надійністю?
12. Як визначаються надійні інтервали для дисперсії та середнього квадратичного відхилення нормального закону розподілу із заданою надійністю?
13. Як побудувати надійний інтервал для коефіцієнта кореляції генеральної сукупності із заданою надійністю?
14. Як побудувати надійний інтервал для математичного сподівання за допомогою нерівності Чебишова із заданою надійністю?

Вправи до теми „Статистичні оцінки параметрів розподілу”

2.1. При відомій дисперсії нормально розподіленої випадкової величини ξ рівній 0,25; та об’ємі вибірки $n = 25$ і вибіркового середньому $\bar{x}_B = 52$ знайти для невідомого математичного сподівання a довірчий інтервал. Надійний рівень $\mathcal{U} = 0,95$.

2.2. Побудувати 96% довірчий інтервал для оцінки математичного сподівання нормально розподіленої випадкової величини ξ , якщо дисперсія $D\xi = 5,78$, об’єм вибірки $n = 10$, а вибіркоче середнє, знайдене за вибіркою, дорівнює $\bar{x}_B = 2$.

2.3. Одержано дані із навмання вибраних підприємств щодо зростання виробітку на одного працівника (у % відносно попереднього року), які мають наступний інтервальний статистичний розподіл:

інтервал	80-90	90-100	100-110	110-120	120-130	130-140
n_i	4	7	55	23	5	6

Побудувати довірчий інтервал для математичного сподівання з надійністю 0.95, якщо відоме значення середнього квадратичного відхилення 5%.

2.4. Для нормально розподіленої випадкової величини ξ при об’ємі вибірки $n=25$, виправленій вибіркової дисперсії рівній 0,25 та вибіркового середньому $\bar{x}_B = 52$ знайти довірчий інтервал для невідомого математичного сподівання a . Довірча ймовірність α рівна 0,95.

2.5 Побудувати 96% довірчий інтервал для оцінки математичного сподівання $M\xi$ нормально розподіленої випадкової величини ξ , якщо об’єм вибірки $n=10$, виправлена вибіркоче дисперсія $S^2 = 5,78$, а вибіркоче середнє, знайдене за вибіркою, $\bar{x}_B = 2$.

2.6 Дано вибірку вимірювання відстані до об’єкту в км: 129, 125, 130, 122, 135, 125, 120, 130, 127. Вважаючи, що вимірювання віддалі є нормально розподіленою випадковою величиною, знайти довірчий інтервал для дисперсії цієї величини з надійністю $\mathcal{U} = 0,8$.

Розрахункові завдання до теми „Статистичні оцінки параметрів розподілу”

Знайти інтервальну оцінку з надійністю 0,95 для математичного сподівання a нормального розподілу, знаючи вибіркоче середнє \bar{x} , об’єм вибірки n і середнє квадратичне відхилення σ .

№ варіанту	об’єм вибірки n	вибіркоче середнє \bar{x} ,	середнє квадратичне відхилення σ
1.	16	45.34	7.12
2.	25	45.30	7.48
3.	36	45.26	8.23
4.	49	45.22	8.68
5.	64	45.18	9.24
6.	81	45.14	9.66
7.	100	45.10	9.94
8.	121	45.06	10.05
9.	144	45.02	10.56
10.	169	44.98	10.95
11.	196	44.96	11.42
12.	225	44.92	11.66
13.	256	44.88	11.89
14.	289	44.84	12.43
15.	324	44.82	12.68
16.	361	44.78	13.56
17.	400	44.74	13.98
18.	441	44.70	14.32
19.	484	44.66	14.85
20.	529	44.62	15.43

Глава 3. Перевірка статистичних гіпотез

3.1. Постановка задачі перевірки статистичних гіпотез

З теорією оцінювання параметрів генеральної сукупності пов’язана перевірка статистичних гіпотез.

Статистичною гіпотезою називають довільне припущення, яке перевіряється за вибіркою, про вид чи параметри розподілів величин, що спостерігаються в експерименті.

Статистичні гіпотези бувають **прості** та **складні**. Проста гіпотеза полягає в однозначному припущенні і повністю визначає функцію розподілу досліджуваної випадкової величини (теоретичну функцію розподілу). Гіпотеза “досліджувана випадкова величина має стандартний нормальний розподіл” є простою, а гіпотези “закон розподілу випадкової

величини нормальний з додатнім середнім” та “розподіл випадкової величини не є нормальним” – складні.

Розрізняють **параметричні** та **непараметричні** статистичні гіпотези. Параметричними гіпотезами є припущення про параметри розподілів досліджуваних випадкових величин. Гіпотеза “середнє випадкової величини рівне нулю” – параметрична, гіпотеза про рівномірний розподіл досліджуваної випадкової величини - непараметрична.

Гіпотезу, яку потрібно перевірити, називають **основною** або **нульовою** гіпотезою. Її позначають H_0 . Всі інші припущення, які суперечать гіпотезі H_0 , називаються **альтернативними** або **конкурентними** гіпотезами і позначаються H_1 . Основна і альтернативна гіпотези є двома можливими варіантами вибору в задачі перевірки статистичних гіпотез.

Наприклад, коли висувається основна гіпотеза про те, що значення параметра θ досліджуваного розподілу рівне θ_0 , тобто $H_0: \theta = \theta_0$, то альтернативною гіпотезі H_0 можна висунути гіпотезу H_1 про те, що $\theta = \theta_1$, або гіпотезу $H_1: \theta \neq \theta_0$, або гіпотезу $H_1: \theta > \theta_0$.

Випадкова величина K із відомим законом розподілу, яка використовується для перевірки основної гіпотези, називається **статистичним критерієм**. На основі вибірки, керуючись значеннями критерію K , відхиляють або не відхиляють нульову гіпотезу. При цьому можуть виникати помилки першого та другого роду.

Помилку, яка виникає тоді, коли гіпотеза H_0 відхиляється, якщо вона є вірною, називають **помилкою першого роду**. Ймовірність зробити помилку першого роду називають **рівнем значущості статистичного критерію** і позначають α . Помилку, яка виникає тоді, коли гіпотеза H_0 не відхиляється, якщо вона не має місця, називають **помилкою другого роду**, її ймовірність позначають β . Ймовірність не допустити помилку другого роду ($1 - \beta$) називають **потужністю критерію**. Це є ймовірність відхилити основну гіпотезу при справедливості альтернативи.

Якщо зафіксувати об’єм вибірки, то при зменшенні ймовірності однієї з помилок першого чи другого роду буде збільшуватись ймовірність іншої. На практиці можна (при фіксованому об’ємі вибірки) зробити як завгодно малим або α , або β . При перевірці статистичної гіпотези ймовірність помилки першого роду вибирають заздалегідь. Як правило, розглядають $\alpha = 0,05$; $\alpha = 0,01$; $\alpha = 0,001$. Останнє α означає, що при такому рівні значущості у одному випадку із тисячі ми ризикуємо відхилити правильну нульову гіпотезу. Зрозуміло, що чим серйозніші наслідки помилки першого роду, тим меншим потрібно вибирати рівень значущості. Основну гіпотезу серед сукупності всіх гіпотез вибирають такою, для якої важливіше уникнути помилки першого роду.

Множина усіх можливих значень статистичного критерію K (одновимірної випадкової величини) розбивається **критичними**

точками k на область прийняття гіпотези S (сукупність значень критерію, при яких основну гіпотезу не відхиляють) та **критичну область \bar{S}** (сукупність значень критерію, при яких основну гіпотезу відхиляють). При вибраному рівні значущості α критичну область будують так, щоб потужність критерію була максимальною, при цьому область прийняття гіпотези визначається автоматично. Критичні точки k визначаються за таблицями розподілу статистичного критерію K .

Якщо при $K < k$ нульова гіпотеза відхиляється, то таку критичну область називають **лівосторонньою**, при $K > k$ – **правосторонньою**, при $K < k_1, K > k_2$ – **двосторонньою**.

В силу практичної неможливості малоймовірних подій при вибраному малому рівні значущості в одному експерименті подія $K \in \bar{S}$ практично не відбувається. Якщо ж вона відбулась, то це можна пояснити тим, що основна гіпотеза H_0 несправедлива і її потрібно відхилити. В іншому випадку вважається, що спостереження узгоджені з гіпотезою H_0 .

Зазначимо, що узгодження даних спостережень із гіпотезою H_0 не означає, що ця гіпотеза є вірною: цілком можливо, що дані спостережень узгоджуються також і з іншою гіпотезою. У цьому випадку можна тільки стверджувати, що дані не суперечать гіпотезі H_0 при певному рівні значущості α .

3.2. Схема перевірки статистичної гіпотези

1. Визначити основну H_0 і альтернативну H_1 гіпотези.
2. Підібрати потрібний статистичний критерій K .
3. Обрати вид критичної області. Якщо $H_0: \theta = \theta_0$, то при $H_1: \theta > \theta_0$ вибирається правостороння критична область, при $H_1: \theta < \theta_0$ – лівостороння, при $H_1: \theta \neq \theta_0$ – двостороння критична область.
4. Зафіксувати рівень значущості α залежно від змісту основної гіпотези H_0 .
5. Побудувати критичну область. Для цього за таблицями розподілу вибраного критерію потрібно знайти критичні точки. Нехай $k = s_\alpha$ – критична точка, що відповідає вибраному рівню значущості α . Тоді s_α знаходиться з умови $P(K > s_\alpha) = \alpha$ для правосторонньої критичної області, $P(K < s_\alpha) = \alpha$ для лівосторонньої критичної області та $P(K < s_\alpha^1) + P(K > s_\alpha^2) = \alpha$ (або $P(K > s_\alpha) = \alpha/2$, якщо розподіл критерію симетричний відносно нуля) для двосторонньої критичної області.
6. За вибіркою обчислити емпіричне значення критерію \hat{K} .

7. Зробити висновок про справедливість основної гіпотези. Якщо \hat{K} належить до критичної області, то гіпотезу H_0 відхиляють, якщо до області прийняття гіпотези, то роблять висновок, що дані не суперечать гіпотезі H_0 при вибраному рівні значущості α .

3.3. Перевірка гіпотези про рівність середніх значень двох нормальних генеральних сукупностей у випадку відомих стандартних відхилень

Нехай є дві незалежні генеральні сукупності, які мають нормальний розподіл із невідомими генеральними середніми a_1 та a_2 і відомими стандартними відхиленнями σ_1 та σ_2 . Потрібно перевірити гіпотезу H_0 про те, що $a_1 = a_2$ проти альтернативи $H_1 : a_1 \neq a_2$.

Для перевірки основної гіпотези розглянемо дві незалежні вибірки об'ємів n_1 і n_2 із даних генеральних сукупностей. Відомо, що оцінками теоретичних середніх a_1 і a_2 є вибіркові середні \bar{x} та \bar{y} . При цьому \bar{x} і \bar{y} також незалежні та мають приблизно нормальний розподіл із параметрами $(a_1, \sigma_1/\sqrt{n_1})$ і $(a_2, \sigma_2/\sqrt{n_2})$, їх різниця $(\bar{x} - \bar{y})$ також має нормальний

розподіл із параметрами $\left(a_1 - a_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$.

Отже, величина $z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ при гіпотезі H_0 має стандартний

нормальний розподіл $N(0,1)$.

Якщо вибрали рівень значущості α , то за таблицями значень функції стандартного нормального розподілу знайдемо таке значення z_α , що $P\{|z| \leq z_\alpha\} = 1 - \alpha$. Ті значення z , для яких $|z| \leq z_\alpha$, утворюють область прийняття гіпотези; значення z , для яких $|z| > z_\alpha$, утворюють критичну область, яку заштриховано на рис. 3.1.

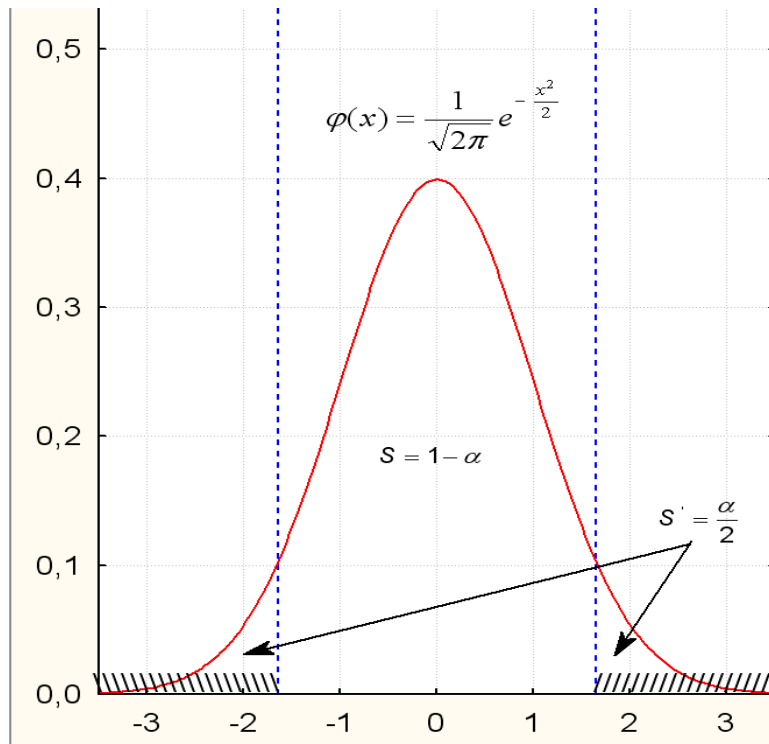


Рис. 3.1
Двостороння критична область

Нехай $d_{1-\frac{\alpha}{2}}$ – квантиль рівня $1-\frac{\alpha}{2}$ стандартного нормального розподілу $N(0,1)$. Тоді при рівні значущості α гіпотезу $H_0: a_1 = a_2$ при альтернативі $H_1: a_1 \neq a_2$ будемо відхиляти, якщо

$$\hat{K} = \frac{|\bar{x} - \bar{y}|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq d_{1-\frac{\alpha}{2}}. \text{ Якщо } \hat{K} = \frac{|\bar{x} - \bar{y}|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < d_{1-\frac{\alpha}{2}}, \text{ то вибіркові дані не}$$

суперечать гіпотезі H_0 при рівні значущості α .

При односторонній альтернативній гіпотезі $a_1 > a_2$ гіпотезу $H_0: a_1 = a_2$ будемо відхиляти, якщо $\hat{K} > d_{1-\alpha}$, $d_{1-\alpha}$ – квантиль рівня $1-\alpha$ стандартного нормального розподілу $N(0,1)$.

Відповідну критичну область заштриховано на рис. 3.2.

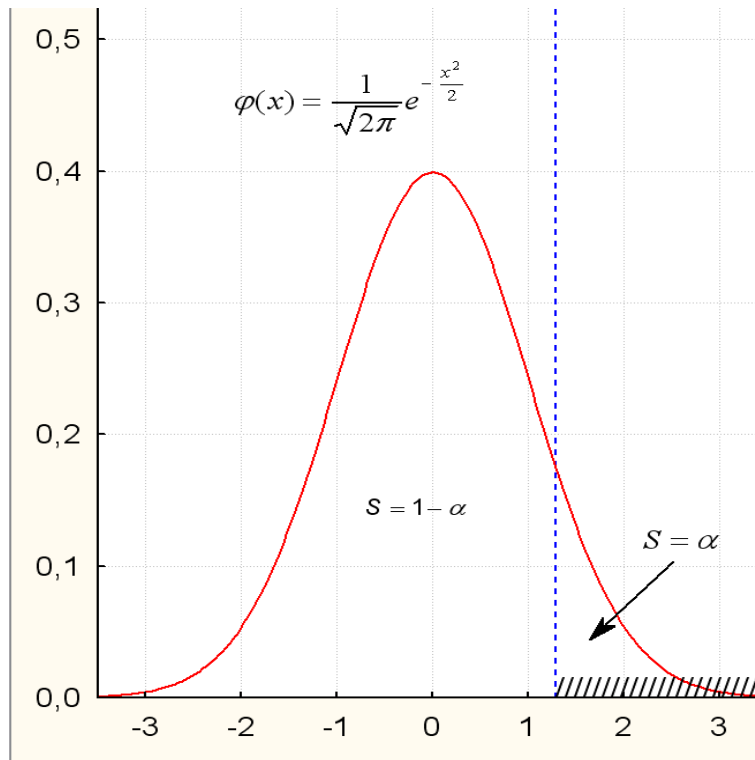


Рис. 3.2

Правостороння критична область

Приклад 3.1. Для перевірки ефективності нової технології відібрали дві групи робітників: для першої групи чисельністю $n_1 = 50$ чол., де застосовувалась нова технологія, знайшли вибіркоче середнє вироблених деталей $\bar{x} = 85$, для другої групи з $n_2 = 70$ чол., де не застосовувалась нова технологія, вибіркоче середнє $\bar{y} = 78$ деталей. Дисперсії вироблених деталей у цих групах $\sigma_1^2 = 100$, $\sigma_2^2 = 74$. При рівні значущості $\alpha = 0,05$ перевірити ефективність нової технології.

Розв'язання. Будемо перевіряти гіпотезу H_0 про те, що $a_1 = a_2$ (середнє вироблених деталей при застосуванні нової технології та без її застосування співпадають) проти альтернативи $H_1 : a_1 > a_2$. Остання гіпотеза означає ефективність нової технології. Обчислимо емпіричне значення критерію

$$\hat{K} = \frac{|\bar{x} - \bar{y}|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{|85 - 78|}{\sqrt{\frac{100}{50} + \frac{74}{70}}} = 4.$$

Знайдемо за таблицею 8.3 квантиль стандартного нормального розподілу $d_{1-\alpha} = d_{0,95}$, тобто таке t , для якого $\Phi(t) = 0,95$, звідки $d_{0,95} = 1,64$. Маємо, що $\hat{K} > d_{0,95}$, тому гіпотезу $H_0 : a_1 = a_2$ будемо відхиляти і

при рівні значущості $\alpha = 0,05$ можна стверджувати, що нова технологія підвищує середнє вироблених деталей.

Зауважимо, якщо б ми не були впевнені в позитивності нової технології, то варто було б розглянути альтернативу $H_1 : a_1 \neq a_2$, для якої $d_{1-\frac{\alpha}{2}} = d_{0,975} = 1,96$, тобто $\hat{K} > d_{0,975}$ і при рівні значущості $\alpha = 0,05$ гіпотезу

$H_0 : a_1 = a_2$ будемо відхиляти.

3.4. Перевірка гіпотези про рівність середніх значень двох нормальних генеральних сукупностей у випадку невідомого стандартного відхилення

Нехай є дві незалежні генеральні сукупності, які мають нормальні розподіли $N(a_1, \sigma^2)$ та $N(a_2, \sigma^2)$ відповідно із невідомими параметрами a_1 , a_2 і σ . Потрібно перевірити гіпотезу H_0 про те, що $a_1 = a_2$ проти альтернативи $H_1 : a_1 \neq a_2$.

Для перевірки основної гіпотези розглянемо дві незалежні вибірки об'ємів n_1 і n_2 із даних генеральних сукупностей. Різниця вибірових середніх $\bar{x} - \bar{y}$ оцінює різницю теоретичних середніх $a_1 - a_2$, яка при справедливості основної гіпотези перетворюється на 0. Отже, гіпотезу H_0 будемо відхиляти, якщо $\bar{x} - \bar{y}$ не близьке до нуля.

Зауважимо, що при $a_1 = a_2$ випадкова величина

$$\frac{|\bar{x} - \bar{y}|}{s \cdot \sqrt{\frac{n_1 + n_2}{n_1 n_2}}}, \text{ де } s = \sqrt{\frac{1}{n_1 + n_2 - 2} \left(\sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{j=1}^{n_2} (y_j - \bar{y})^2 \right)},$$

має розподіл Стьюдента з $(n_1 + n_2 - 2)$ ступенями свободи.

Якщо вибрали рівень значущості α та за таблицею 8.5 розподілу Стьюдента знайшли $t_{1-\frac{\alpha}{2}; n_1+n_2-2}$, тоді при

$$\hat{K} = \frac{|\bar{x} - \bar{y}|}{s \cdot \sqrt{\frac{n_1 + n_2}{n_1 n_2}}} \geq t_{\frac{\alpha}{2}; n_1+n_2-2}$$

гіпотезу H_0 будемо відхиляти. При $\hat{K} < t_{\frac{\alpha}{2}; n_1+n_2-2}$ вибірові дані не суперечать гіпотезі H_0 при рівні значущості α .

Якщо розглядати односторонню альтернативу $H_1 : a_1 > a_2$, то гіпотезу H_0 будемо відхиляти при умові

$$\hat{K} = \frac{|\bar{x} - \bar{y}|}{s \cdot \sqrt{\frac{n_1 + n_2}{n_1 n_2}}} \geq t_{\alpha; n_1 + n_2 - 2}.$$

Приклад 3.2. За даними, отриманими при дослідженні однієї й тієї ж поверхні двома вимірвальними приладами, зробити висновок про наявність систематичної розбіжності між показами цих приладів при рівні значущості $\alpha = 0,05$.

Прилад I: 0,8; 1,9; 3,0; 3,5; 3,8; 2,5; 1,7; 0,9; 1,0; 2,3; 3,3; 3,4.

Прилад II: 1,4; 2,1; 3,1; 3,6; 2,7; 1,7; 1,1; 0,2; 1,6; 2,8; 4,0; 4,7.

Розв'язання. Потрібно перевірити гіпотезу H_0 про те, що $a_1 = a_2$ (покази приладів не мають систематичної розбіжності) проти альтернативи $H_1: a_1 \neq a_2$ (наявність систематичної розбіжності між показами цих приладів).

Маємо $n_1 = n_2 = 12$. Обчислимо $\bar{x} = 2,34$. $\bar{y} = 2,42$. $s^2 = 1,44$.

$$\hat{K} = \frac{|\bar{x} - \bar{y}|}{s \cdot \sqrt{\frac{n_1 + n_2}{n_1 n_2}}} = \frac{|2,34 - 2,42|}{\sqrt{\frac{1,44(12 + 12)}{12 \cdot 12}}} = 0,16.$$

За таблицею 8.5 знайдемо $t_{\frac{\alpha}{2}; n_1 + n_2 - 2} = t_{0,025; 22} = 2,074$.

Отже, $\hat{K} = 0,16 < 2,074$, тому гіпотеза H_0 при рівні значущості $\alpha = 0,05$ не відхиляється і припущення про те, що покази приладів не мають систематичної розбіжності, не суперечить експериментальним даним.

Розглянутий критерій Стьюдента можна використовувати для виключення помилок спостережень. Нехай маємо вибірку $x_1, x_2, \dots, x_n, x^*$, де значення x^* суттєво відрізняється від решти спостережень (набагато менше чи більше) і викликає сумнів. Таке аномальне значення у статистиці називають **викидом**. Для спостережень x_1, x_2, \dots, x_n обчислюють вибіркове середнє \bar{x} та виправлене середнє квадратичне відхилення S . Якщо справедлива гіпотеза $H_0: \bar{x} = x^*$ про належність x^* до решти спостережень, то статистика $\hat{t} = \frac{|\bar{x} - x^*|}{S}$, що отримується з \hat{K} при $\bar{y} = x^*$, $n_2 = 1$, має розподіл Стьюдента з $n - 1$ ступенями свободи. Альтернативна гіпотеза H_1 має вигляд $\bar{x} > x^*$ чи $\bar{x} < x^*$ залежно від того менше чи більше x^* від решти спостережень.

Приклад 3.3. При спостереженні мікроклімату певної території в травні місяці було одержано ряд значень температур: $10,6^{\circ}\text{C}$; $12,7^{\circ}\text{C}$; $13,1^{\circ}\text{C}$; $13,5^{\circ}\text{C}$; $26,4^{\circ}\text{C}$. Перевірити належність максимального значення температури $x^* = 26,4^{\circ}\text{C}$ до даних спостережень при рівні значущості $\alpha = 0,05$.

Розв'язання. Обчислимо $\bar{x} = \frac{1}{4}(10,6 + 12,7 + 13,1 + 13,5) = 12,475$.

$$S^2 = \frac{1}{3}[(10,6 - 12,475)^2 + (12,7 - 12,475)^2 + (13,1 - 12,475)^2 + (13,5 - 12,475)^2],$$

звідки $S = \sqrt{1,669} \approx 1,29$.

Будемо перевіряти справедливість гіпотези $H_0: \bar{x} = x^*$ при альтернативі $H_1: \bar{x} < x^*$.

Знайдемо $\hat{t} = \frac{|\bar{x} - x^*|}{S} = \frac{|12,475 - 26,4|}{1,29} \approx 10,8$ та порівняємо з табличним

$t_{\alpha; n-1} = t_{0,05; 3} = 2,353$. Маємо, що $\hat{t} \approx 10,8 > 2,353 = t_{0,05; 3}$, тому при рівні значущості $\alpha = 0,05$ гіпотезу H_0 будемо відхиляти, чим обґрунтовано відбраковку максимального значення температури $x^* = 26,4^{\circ}\text{C}$, що узгоджується з результатами прикладу 1.1.

3.5. Перевірка гіпотези про рівність стандартних відхилень двох нормальних генеральних сукупностей у випадку невідомих параметрів розподілів

Нехай є дві незалежні генеральні сукупності, які мають нормальний розподіл із невідомими генеральними середніми a_1 та a_2 і невідомими стандартними відхиленнями σ_1 та σ_2 . Потрібно перевірити гіпотезу H_0 про те, що $\sigma_1 = \sigma_2$.

Для перевірки основної гіпотези розглянемо дві незалежні вибірки об'ємів n_1 і n_2 із даних генеральних сукупностей. За цими вибірками знайдемо вибіркові середні \bar{x} та \bar{y} та виправлені вибіркові стандартні відхилення S_1 та S_2 . При $\sigma_1^2 / \sigma_2^2 = 1$ (гіпотеза H_0 справджується) випадкова величина S_1^2 / S_2^2 має розподіл Фішера з $(n_1 - 1)$, $(n_2 - 1)$ ступенями вільності.

У випадку двобічної альтернативи $H_1: \sigma_1^2 / \sigma_2^2 \neq 1$ гіпотезу H_0 будемо відхиляти при

$$\frac{S_1^2}{S_2^2} \notin \left(\frac{1}{F_{\alpha; (n_2-1); (n_1-1)}}, F_{\alpha; (n_1-1); (n_2-1)} \right)$$

і не відхиляти в інших випадках. При цьому рівень значущості такого критерію 2α .

Якщо альтернатива одностороння, наприклад, $\sigma_1^2 / \sigma_2^2 > 1$, то при вибраному рівні значущості α гіпотезу H_0 будемо відхиляти при $\frac{S_1^2}{S_2^2} > F_{\alpha; (n_1-1); (n_2-1)}$. Критичні значення такого критерію знаходимо за таблицею 8.7.

Приклад 3.4. За вибірками результатів незалежних вимірювань двома приладами зробити висновок про точність вимірювань даними приладами при рівні значущості $\alpha = 0,02$.

Прилад I: 1,32; 1,35; 1,32; 1,35; 1,30; 1,30; 1,37; 1,31; 1,39; 1,39.

Прилад II: 1,35; 1,31; 1,31; 1,41; 1,39; 1,37; 1,32; 1,34.

Розв'язання. Будемо вважати, що є дві незалежні вибірки об'ємів $n_1 = 10$ і $n_2 = 8$ із нормальних розподілів $N(a_1, \sigma_1^2)$ та $N(a_2, \sigma_2^2)$. Відносно невідомих параметрів σ_1 та σ_2 висунемо гіпотезу $H_0: \sigma_1^2 / \sigma_2^2 = 1$ (однакова точність вимірювань приладами). Розглядатимемо двобічну альтернативу $H_1: \sigma_1^2 / \sigma_2^2 \neq 1$ (точність вимірювань приладами різна).

Обчислимо $\bar{x} = 1,34$. $\bar{y} = 1,35$. $S_1^2 = 0,00122$. $S_2^2 = 0,0014$. $S_1^2 / S_2^2 = 0,87$. За таблицею 8.7.2 знайдемо $F_{\alpha; (n_1-1); (n_2-1)} = F_{0,01; 9; 7} = 6,72$.

$$F_{\alpha; (n_2-1); (n_1-1)} = F_{0,01; 7; 9} = 5,61. \frac{1}{F_{\alpha; (n_2-1); (n_1-1)}} = \frac{1}{5,61} \approx 0,18.$$

Отже,

$$\frac{S_1^2}{S_2^2} = 0,87 \in (0,18; 6,72)$$

і при рівні значущості $\alpha = 0,02$ гіпотеза H_0 не суперечить наведеним в умові вибіровим даним, тобто можна вважати, що прилади I та II мають однакову точність.

3.6. Перевірка гіпотези про закон розподілу за допомогою критерію χ^2

Нехай маємо вибірку об'єму n з генеральної сукупності дискретної випадкової величини. Висувається гіпотеза H_0 про те, що закон розподілу даної випадкової величини ξ має вигляд:

ξ	x_1	x_2	...	x_k
p	p_1	p_2	...	p_k

Нехай варіантам x_1, x_2, \dots, x_k вибірки відповідають частоти n_1, n_2, \dots, n_k , при цьому $\sum_{i=1}^k n_i = n$. Побудуємо функцію χ^2 , яка характеризує відхилення спостережуваних частот від теоретичних ймовірностей:

$$\hat{\chi}^2 = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i}.$$

При $n \rightarrow \infty$ функція $\hat{\chi}^2$ має розподіл χ^2 з $(k-1)$ ступенем свободи. Якщо рівень значущості α є заданим, то за таблицею 8.4 шукаємо таке значення χ_{α}^2 , що $P\{\hat{\chi}^2 > \chi_{\alpha}^2\} = \alpha$. Тоді, значення $\hat{\chi}^2$, для яких $\hat{\chi}^2 \leq \chi_{\alpha}^2$, є областю допустимих значень; значення $\hat{\chi}^2$, для яких $\hat{\chi}^2 > \chi_{\alpha}^2$, утворюють критичну область.

Розглянутим критерієм χ^2 можна користуватися, коли об'єм вибірки такий, що виконуються умови

$$np_i \geq 10, i = 1, 2, \dots, k. \quad (86)$$

Приклад 3.5. При $n = 4000$ незалежних випробуваннях події A_1, A_2, A_3 , які утворюють повну групу подій, відбулися відповідно 1905, 1015 та 1080 раз. Перевірити узгодженість наведених даних при рівні значущості $\alpha = 0,05$ з гіпотезою H_0 про те, що $P(A_1) = 1/2$, $P(A_2) = P(A_3) = 1/4$ у кожному випробуванні.

Розв'язання. Гіпотеза H_0 полягає в тому, що $p_1 = 0,5$; $p_2 = p_3 = 0,25$. Очевидно, що умови (86) виконуються. Обчислимо емпіричне значення критерію

$$\begin{aligned} \hat{\chi}^2 &= \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i} = \\ &= \frac{(1905 - 4000 \cdot 0,5)^2}{4000 \cdot 0,5} + \frac{(1015 - 4000 \cdot 0,25)^2}{4000 \cdot 0,25} + \frac{(1080 - 4000 \cdot 0,25)^2}{4000 \cdot 0,25} = 11,13. \end{aligned}$$

За таблицею 8.4 знайдемо $\chi_{\alpha;n}^2 = \chi_{0,05;3-1}^2 = 5,99$.

Отже, $\hat{\chi}^2 > \chi_{\alpha}^2$ і при рівні значущості $\alpha = 0,05$ гіпотезу H_0 будемо відхиляти.

Критерій узгодження χ^2 використовується також і для перевірки гіпотез про неперервну функцію розподілу $F(x)$. Для цього, числова вісь розбивається на інтервали, що не перетинаються, підраховуються

теоретичні та спостережувані частоти для кожного інтервалу, а потім вони порівнюються за допомогою функції χ^2 . Критерієм χ^2 можна користуватися, коли частоти для кожного інтервалу $n_i \geq 5$. Якщо це не так, то перед застосуванням критерію об'єднують сусідні інтервали з малими частотами.

Якщо при перевірці гіпотези про закон розподілу гіпотетичний розподіл характеризують s невідомих параметрів, то за вибіркою шукають їх відповідні оцінки, а критичне значення χ^2_α при вибраному α шукають за таблицею 8.4 для кількості ступенів свободи $(k - s - 1)$.

Приклад 3.6. За наведеним нижче інтервальним розподілом середньодобової температури повітря при рівні значущості $\alpha = 0,05$ перевірити гіпотезу H_0 про нормальний розподіл середньодобової температури повітря.

Інтервали	(8,1 – 8,4)	(8,4 – 8,7)	(8,7 – 9,0)	(9,0 – 9,3)	(9,3 – 9,6)	(9,6 – 9,9)
n_i	7	10	9	4	6	4

Розв'язання. Розіб'ємо числову вісь на інтервали, що не перетинаються, при цьому об'єднаємо сусідні інтервали з малими частотами. Утворений інтервальний розподіл має вигляд

Інтервали	$(-\infty; -8,4)$	(8,4 – 8,7)	(8,7 – 9,3)	$(9,3; +\infty)$
n_i	7	10	13	10

Маємо:

$$n = \sum_{i=1}^k n_i = 40.$$

Гіпотетичний розподіл має два невідомих параметри ($s = 2$) a та σ , точкові незміщені оцінки яких обчислимо за даним в умові задані розподілом: $\bar{x} = 8,88$, $S = 0,484$. Ймовірності p_i потрапляння нормальної випадкової величини з параметрами a та σ в інтервал (α, β) обчислимо за формулою

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

де значення функції $\Phi(x)$ стандартного нормального розподілу наведено у таблиці 8.3:

$$p_1 = \Phi\left(\frac{8,4 - 8,88}{0,484}\right) - \Phi(-\infty) = \Phi(-0,99) - 0 = 0,1611;$$

$$p_2 = \Phi\left(\frac{8,7 - 8,88}{0,484}\right) - \Phi\left(\frac{8,4 - 8,88}{0,484}\right) = \Phi(-0,37) - \Phi(-0,99) = \\ = 0,3557 - 0,1611 = 0,1946;$$

$$p_3 = \Phi\left(\frac{9,3 - 8,88}{0,484}\right) - \Phi\left(\frac{8,7 - 8,88}{0,484}\right) = \Phi(0,87) - \Phi(-0,37) = \\ = 0,8078 - 0,3557 = 0,4521;$$

$$p_4 = \Phi(+\infty) - \Phi\left(\frac{9,3 - 8,88}{0,484}\right) = 1 - \Phi(0,87) = 1 - 0,8078 = 0,1922.$$

Обчислимо

$$\hat{\chi}^2 = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i} = \frac{(7 - 40 \cdot 0,1611)^2}{40 \cdot 0,1611} + \frac{(10 - 40 \cdot 0,1946)^2}{40 \cdot 0,1946} + \\ + \frac{(13 - 40 \cdot 0,4521)^2}{40 \cdot 0,4521} + \frac{(10 - 40 \cdot 0,1922)^2}{40 \cdot 0,1922} \approx 2,8.$$

За таблицею 8.4 для кількості ступенів свободи $(k - s - 1) = (4 - 2 - 1) = 1$ при $\alpha = 0,05$ знайдемо $\chi_{\alpha;n}^2 = \chi_{0,05;1}^2 = 3,84$.

Отже, $\hat{\chi}^2 \approx 2,8 < 3,84 = \chi_{\alpha}^2$ і при рівні значущості $\alpha = 0,05$ будемо стверджувати, що гіпотеза H_0 про нормальний розподіл середньодобової температури повітря узгоджується з вибірковими даними.

3.7. Критерій Колмогорова для перевірки гіпотези про закон розподілу

Нехай маємо вибірку об'єму n з певного невідомого неперервного розподілу $F(x)$. Необхідно перевірити гіпотезу $H_0: F(x) = F_0(x)$, де $F_0(x)$ – повністю визначений неперервний розподіл. Оцінкою невідомої функції розподілу $F(x)$ є емпірична функція розподілу $\hat{F}_n(x)$, побудована за даною вибіркою. Мірою відхилення емпіричного розподілу $\hat{F}_n(x)$ від теоретичного $F_0(x)$ є величина $D_n = \sup_x |F_0(x) - \hat{F}_n(x)|$, яка при справедливості гіпотези H_0 набуває значень $D_n = \sup_x |F_0(x) - \hat{F}_n(x)|$. При досить великих n нормоване відхилення $\sqrt{n} \cdot D_n$ має розподіл Колмогорова, функція розподілу якого

$$K(\lambda) = \sum_{k=-\infty}^{+\infty} (-1)^k \exp\{-2k^2\lambda^2\}.$$

Якщо гіпотеза H_0 має місце, то відхилення D_n мало відрізняється від нуля, в іншому випадку відхилення D_n велике. Межі $\varepsilon_{\alpha;n}$, що відокремлюють великі значення D_n (критичну область) від малих (область прийняття гіпотези) при вибраному рівні значущості α визначають як найменше ε , для якого

$$P\left\{\sup_x |F(x) - \hat{F}_n(x)| \geq \varepsilon\right\} \leq \alpha.$$

Ці критичні значення наведено у таблиці 8.8.

При $n > 100$ користуються асимптотичними границями і критичні значення вибирають так:

$$\varepsilon_{0,05;n} = \frac{1,36}{\sqrt{n}}; \quad \varepsilon_{0,01;n} = \frac{1,63}{\sqrt{n}},$$

для них справжні коефіцієнти надійності навіть трохи більші від 0,95 і 0,99 відповідно.

Зауважимо, що критерієм Колмогорова можна користуватися лише у випадку, коли гіпотетичний розподіл неперервний і не залежить від невідомих параметрів. Якщо для гіпотетичного розподілу виникає потреба оцінити його параметри за вибіркою, то треба скористатися іншим критерієм, наприклад, розглянутим вище критерієм χ^2 .

Приклад 3.7. Перевірити при рівні значущості $\alpha = 0,01$ гіпотезу H_0 про рівномірний на проміжку $[0; 12]$ розподіл вибірки 8,00; 5,88; 7,13; 11,27; 4,47; 7,78; 10,20; 4,36; 7,83; 3,10; 11,50; 10,75; 8,58; 9,30.

Розв'язання. Теоретична функція розподілу має вигляд

$$F_0(x) = \begin{cases} 0, & x < 0; \\ x/12, & 0 \leq x < 12; \\ 1, & x \geq 12. \end{cases}$$

Знайдемо емпіричну функцію розподілу $\hat{F}_{14}(x)$ та обчислимо

$$D_n = \sup_x |F_0(x) - \hat{F}_{14}(x)|.$$

Зауважимо, що

$$D_n = \max\left\{|F_0(x_k) - \hat{F}_n(x_k)|, |F_0(x_k) - \hat{F}_n(x_{k+1})|, k = 1, 2, \dots, n.\right\}$$

Для зручності заповнимо таблицю:

x_k	$\hat{F}_{14}(x_k)$	$F_0(x_k)$	$ F_0(x_k) - \hat{F}_n(x_k) $	$ F_0(x_k) - \hat{F}_n(x_{k+1}) $
3,10	0	0,26	0,26	0,19
4,36	0,07	0,36	0,29	0,22

4,47	0,14	0,37	0,23	0,16
5,88	0,21	0,49	0,28	0,20
7,13	0,29	0,59	0,30	0,23
7,78	0,36	0,65	0,29	0,22
7,83	0,43	0,65	0,22	0,15
8,00	0,50	0,67	0,17	0,10
8,58	0,57	0,72	0,15	0,08
9,30	0,64	0,78	0,14	0,07
10,20	0,71	0,85	0,14	0,06
10,75	0,79	0,90	0,11	0,04
11,27	0,86	0,94	0,08	0,01
11,50	0,92	0,96	0,03	0,04
x_{k+1}	1,00			

Порівняємо знайдене відхилення $D_n=0,30$ із табличним значенням $\varepsilon_{\alpha;n} = \varepsilon_{0,01;14} = 0,4176$. Оскільки $D_n < \varepsilon_{\alpha;n}$, то при рівні значущості $\alpha = 0,01$ вибіркові дані узгоджуються з гіпотезою про рівномірний на проміжку $[0; 12]$ розподіл.

Контрольні питання для самоперевірки до теми „Перевірка статистичних гіпотез”

1. Що називається статистичною гіпотезою?
2. Які гіпотези називають параметричними, непараметричними?
3. Дати визначення нульової та альтернативної гіпотез.
4. Які статистичні гіпотези називають простим, складними?
5. Що називається статистичним критерієм?
6. Які помилки називають помилками першого та другого роду.
7. Що таке область прийняття нульової гіпотези, критична область, критична точка? Які бувають критичні області?
8. Що таке рівень значущості статистичного критерію?
9. Що таке потужність критерію?
10. Сформулюйте схему перевірки правильності нульової гіпотези.
11. Як здійснюється перевірка гіпотези про рівність середніх значень двох нормальних генеральних сукупностей у випадку відомих стандартних відхилень?
12. Як здійснюється перевірка гіпотези про рівність середніх значень двох нормальних генеральних сукупностей у випадку невідомого стандартного відхилення?
13. Як здійснюється перевірка гіпотези про рівність стандартних відхилень двох нормальних генеральних сукупностей у випадку відомих параметрів розподілів?
14. Як здійснюється перевірка гіпотези про закон розподілу за допомогою критерію χ^2 ?
15. Сформулювати алгоритм перевірки гіпотези про закон розподілу використовуючи критерій Колмогорова.

РОЗДІЛ 2

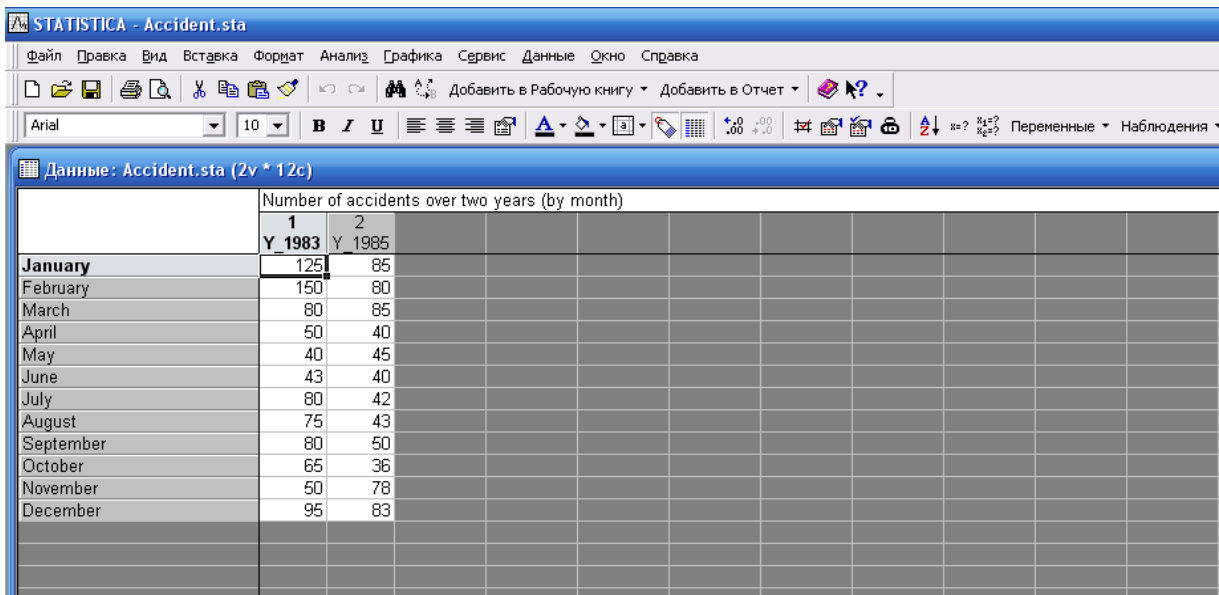
СТАТИСТИЧНИЙ АНАЛІЗ ДАНИХ З ПАКЕТОМ STATISTICA

Глава 4. Знайомство з пакетом STATISTICA

4.1. Робота з даними в пакеті STATISTICA

Сучасна статистична обробка даних практично неможлива без певних комп'ютерних програм, однією з них є STATISTICA.

Після завантаження програми перед користувачами з'являється вікно з контекстною панеллю і остання відкрита в пакеті таблиця з даними, можливо порожня.



The screenshot shows the STATISTICA software interface. The title bar reads "STATISTICA - Accident.sta". The menu bar includes "Файл", "Правка", "Вид", "Вставка", "Формат", "Анализ", "Графика", "Сервис", "Данные", "Окно", "Справка". The toolbar contains various icons for file operations and data manipulation. The main window displays a data table titled "Данные: Accident.sta (2v * 12c)". The table has the following structure:

	Number of accidents over two years (by month)	
	1 Y_1983	2 Y_1985
January	125	85
February	150	80
March	80	85
April	50	40
May	40	45
June	43	40
July	80	42
August	75	43
September	80	50
October	65	36
November	50	78
December	95	83

Рис.4.1

Таблиці в пакеті STATISTICA мають нерівноправні стовпці та рядки. Стовпці відповідають показникам, які спостерігають, а рядки – окремим спостереженням.

Створимо нову таблицю. Виберемо *Файл* → *Создать*. У діалоговому вікні, що відкрилося (див. рис. 4.2), можемо вибрати кількість змінних і спостережень.

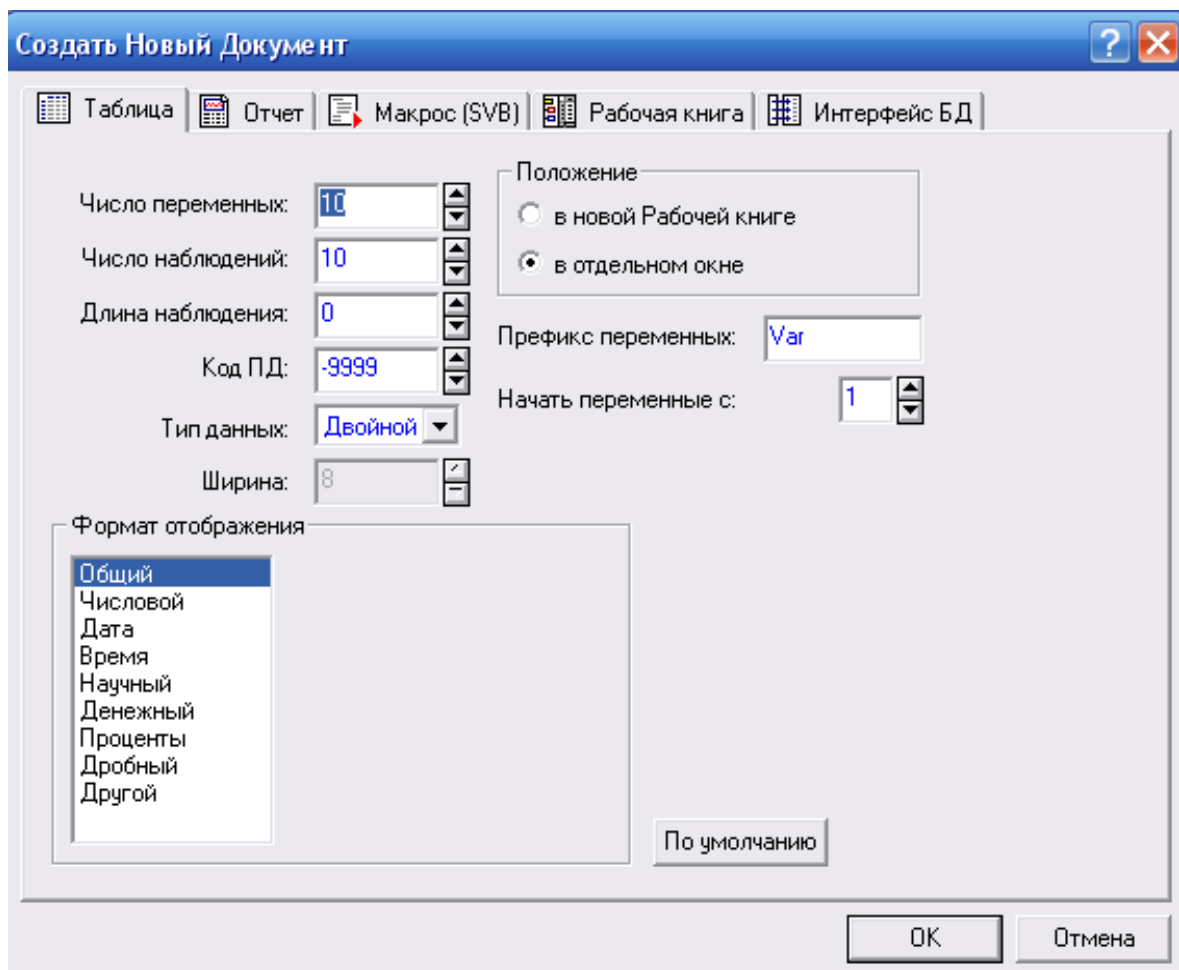


Рис. 4.2

Опція *в новой Рабочей книге* розмістить новостворену таблицю у робочій книзі, в яку також будуть записуватись всі графіки, діаграми і таблиці, отримані у процесі роботи з даними. Опція *в отдельном окне* створить таблицю в окремому вікні так, що дані можна буде зберегти окремо, виконуючи *Файл → Сохранить* при активізованій таблиці.

При виконанні певних дій в пакеті STATISTICA автоматично створюється файл звіту. На закладці *Отчет* можна вибрати розміщення даних звіту *в новой Рабочей книге* чи *в отдельном окне*.

Натиснемо *OK* внизу вікна *Создать Новый Документ*. З'явиться нова порожня таблиця. Якщо потрібно змінити кількість змінних, копіювати, перемістити чи виконати якісь інші дії над змінними, то натискаємо на верхній панелі кнопку *Переменные* і далі вибираємо потрібну опцію:

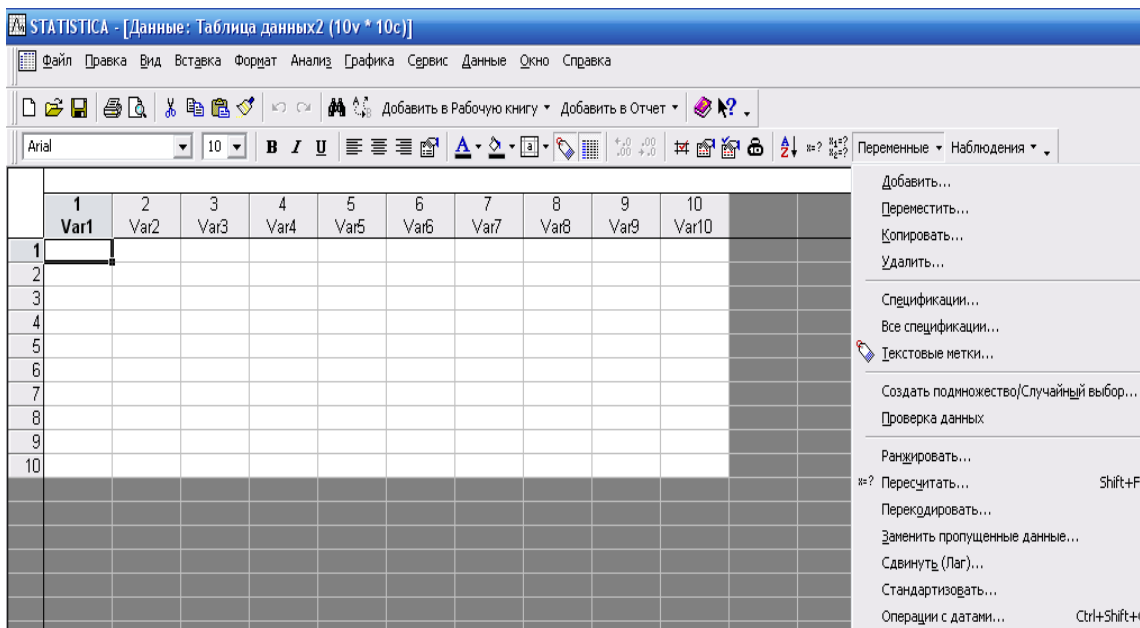


Рис. 4.3

За допомогою опції меню *Наблюдения* на верхній панелі можна зробити аналогічні операції над рядками таблиці з даними.

Виділимо якусь змінну натисканням лівої кнопки миші на імені змінної, а потім натиснемо праву кнопку миші та виберемо опцію *Спецификации переменной*. Після цього з'явиться вікно опису даної змінної (див. рис. 4.4). В останньому полі може міститись розширена інформація про дану змінну або формула, яку було використано для обчислення цієї змінної. Бачимо, що в файлі *Accident.sta* змінна, що записана в першому стовпчику, містить інформацію про щомісячну кількість аварій протягом 1983 року.

Натиснемо кнопку *Переменные* та виберемо опцію *Добавить*. У вікні, що з'явилося, вкажемо, що додаємо одну змінну після змінної *Y_1985*; назвемо її "Сума" та в останньому полі запишемо формулу для її обчислення (див. рис. 4.5).

Натиснемо кнопку *OK* та отримаємо таблицю, зображену на рис.4.6.

Можемо зберегти цей файл, вибравши *Файл* → *Сохранить как: Accident1.sta* (див. рис. 4.5).

Бачимо, що файли з даними у вигляді таблиць в пакеті STATISTICA мають розширення *.sta*. Файли з графічною інформацією мають розширення *.stg*, файли робочих книг – *.stw*, файли звіту – *.str*, файли матриць – *.stm*.

Якщо маємо дані в Excel, то їх можна скопіювати у файл з розширенням *.sta* за допомогою операцій *Копировать* в Excel та *Вставить* в STATISTICA, або *Копировать* в Excel та *Правка* → *Специальная вставка* в STATISTICA, що дозволяє встановити динамічний зв'язок між даними.

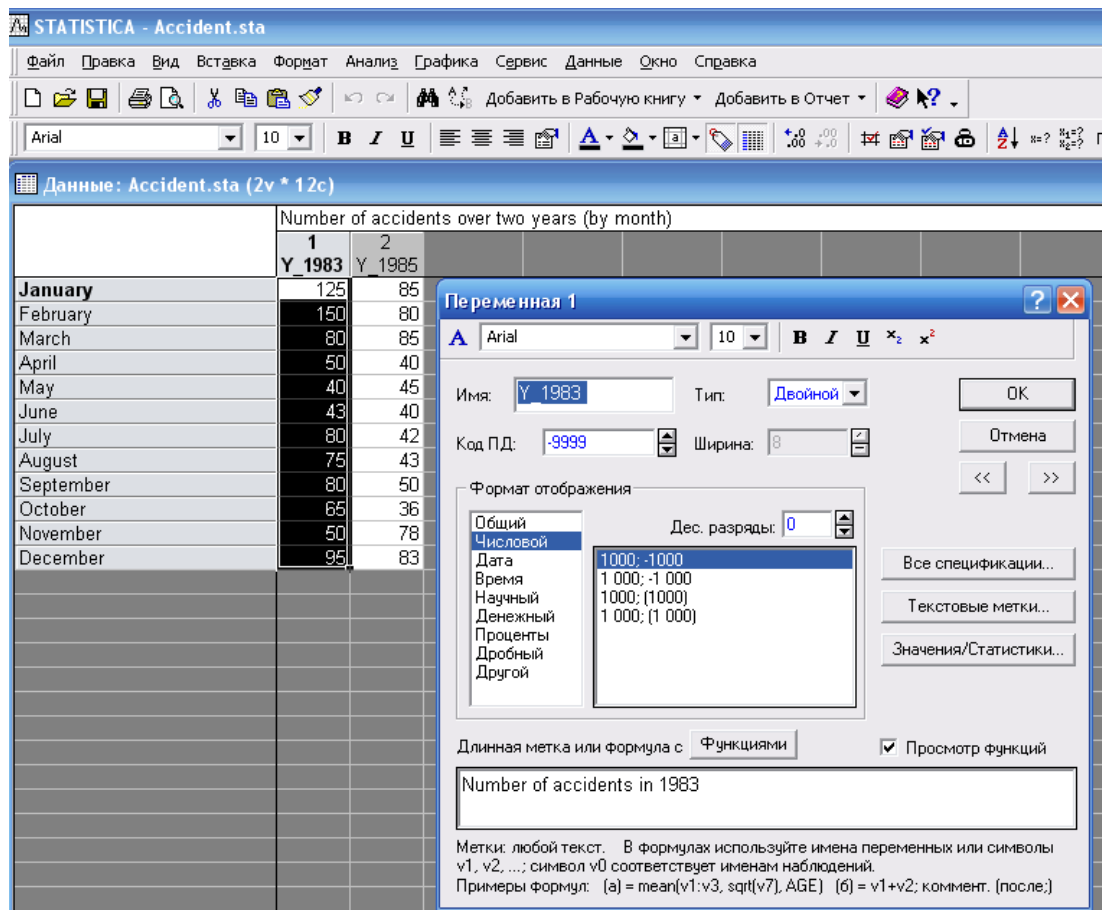


Рис. 4.4

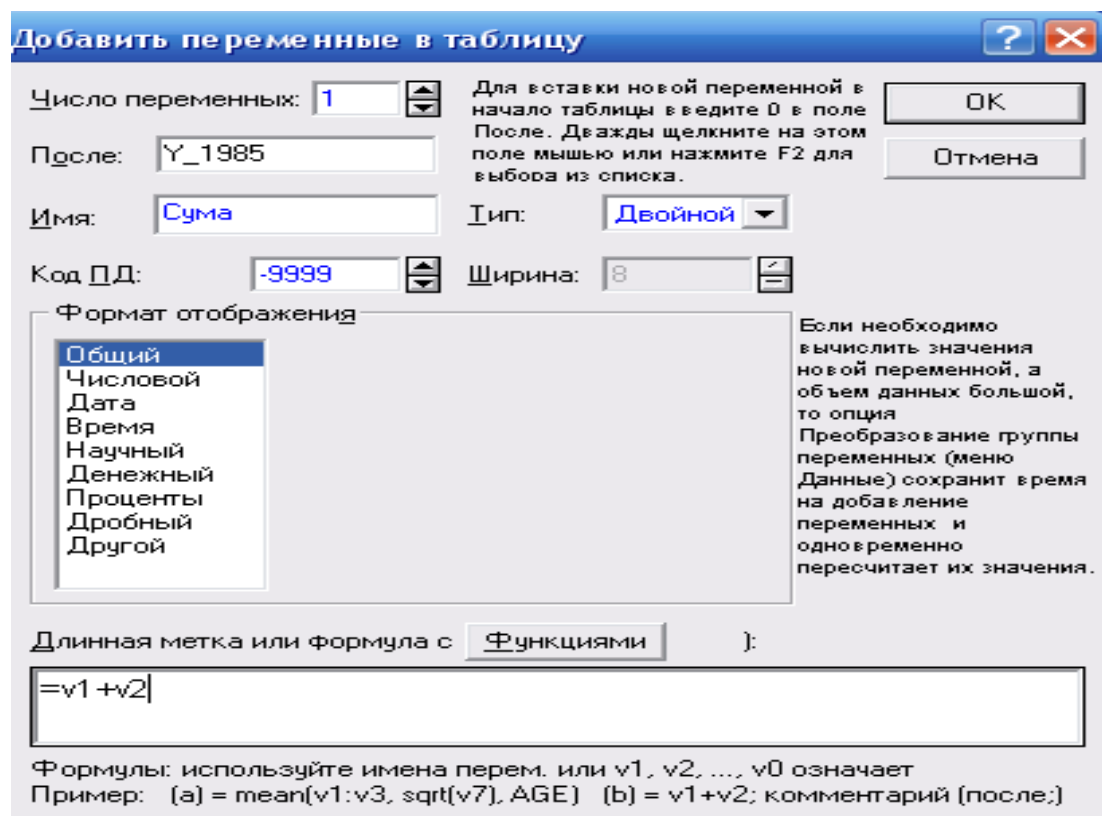


Рис. 4.5

STATISTICA - [Данные: Accident.sta* (3v * 12c)]

Файл Правка Вид Вставка Формат Анализ Графика Сервис Данные Окно Справка

Добавить в Рабочую книгу Добавить в Отчет

Number of accidents over two years (by month)

	1	2	3				
	Y_1983	Y_1985	Сума				
January	125	85	210				
February	150	80	230				
March	80	85	165				
April	50	40	90				
May	40	45	85				
June	43	40	83				
July	80	42	122				
August	75	43	118				
September	80	50	130				
October	65	36	101				
November	50	78	128				
December	95	83	178				

Рис. 4.6

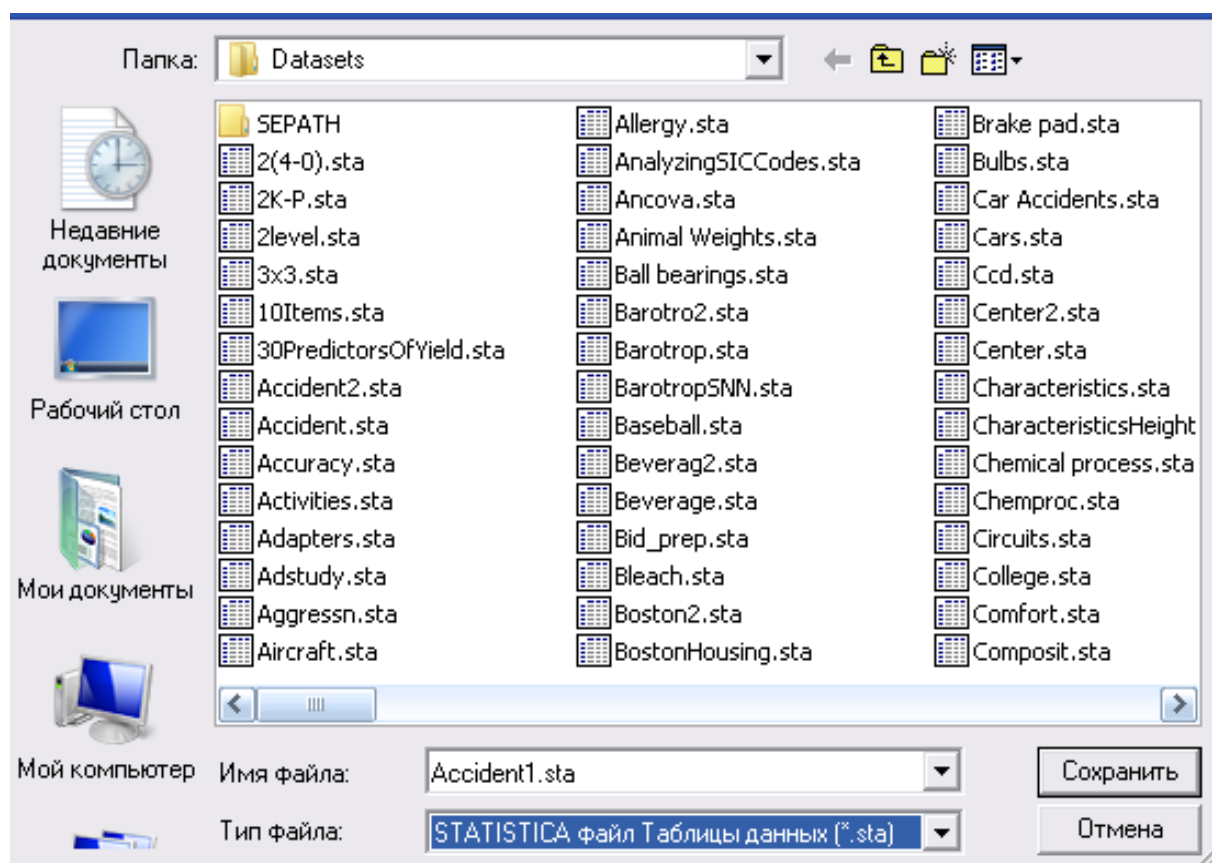


Рис. 4.7

4.2. Знаходження числових характеристик вибірки засобами пакету STATISTICA

Повернемось до збереженої таблиці Accident1.sta і виберемо *Анализ* → *Основные статистики и таблицы* → *Описательные статистики* → *ОК*:

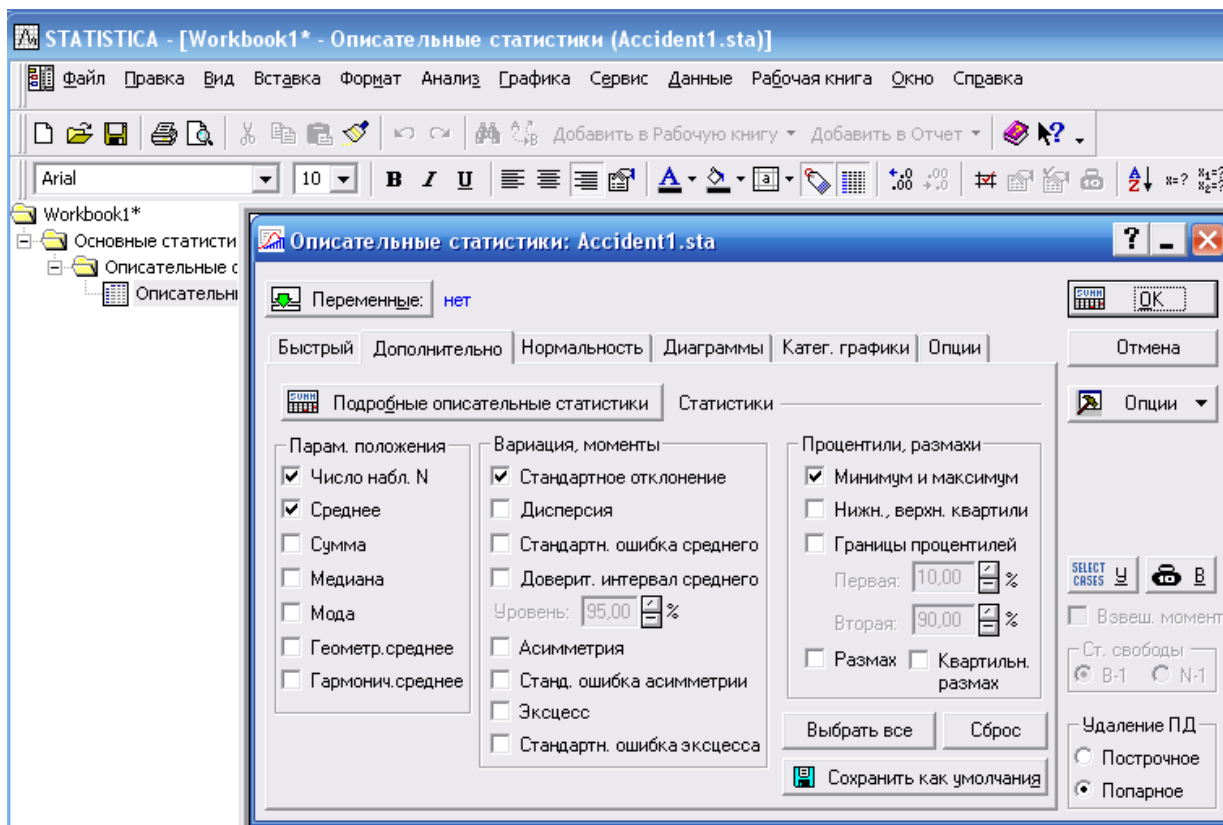


Рис. 4.8

В полі *Переменные* виберемо всі змінні та виділимо описові статистики, які потрібно обчислити. Після натискання *ОК* отримаємо:

Переменная	Среднее	Доверит. -95,000%	Доверит. +95,000%	Медиана	Сумма	Дисперс.	Стд. откл.	Асимметрия	Экссесс
Y_1983	77,7500	56,6700	98,8300	77,5000	933,000	1100,750	33,17755	1,037178	0,78159
Y_1985	58,9167	45,6402	72,1931	47,5000	707,000	436,629	20,89566	0,333954	-2,12059
Сумма	136,6667	105,6093	167,7240	125,0000	1640,000	2389,333	48,88081	0,794577	-0,43616

Рис. 4.9

Бачимо, що для кожної змінної обчислено середнє, нижня та верхня 95% надійні межі для середнього, медіана та інші числові характеристики.

Повернемося у вікно *Описательные статистики*, виберемо *Быстрый* → *Таблицы частот*:

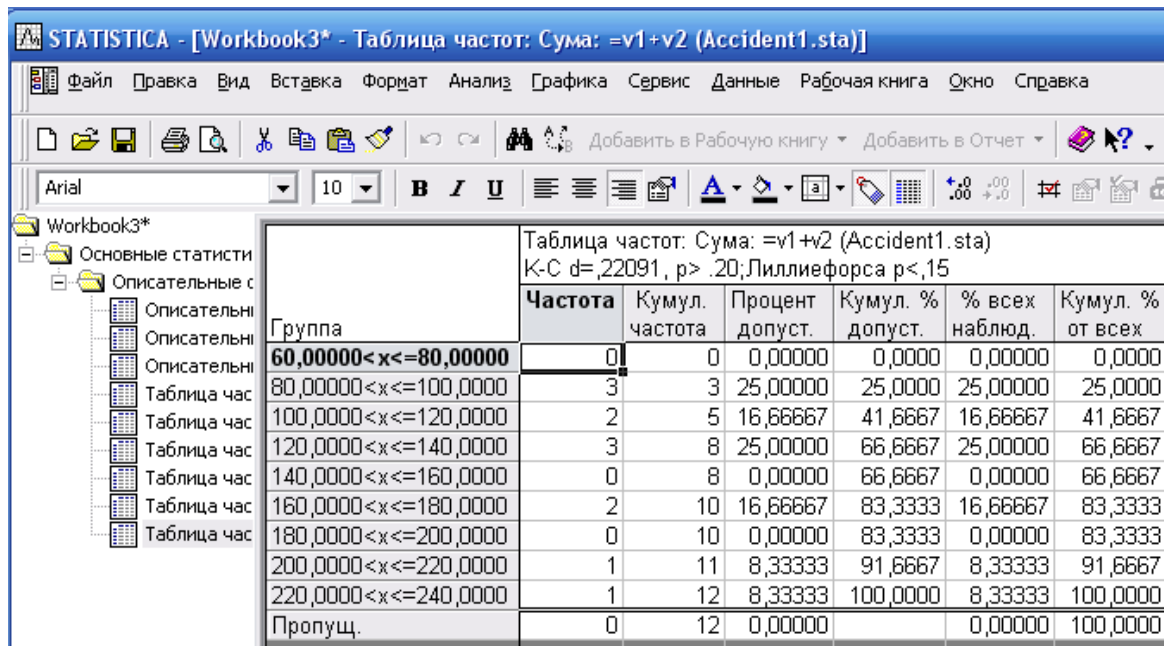


Рис. 4.10

Маємо інтервальний статистичний розподіл сумарної кількості аварій з обчисленими кумулятивними характеристиками.

Після натискання кнопки *Гистограммы* у вікні *Описательные статистики* побачимо відповідну гистограму з накладеною на неї очікуваною щільністю нормального розподілу (див. рис. 4.11).

Очевидно, що сумарна кількість аварій не узгоджується з нормальним законом розподілу. Якщо хочемо продивитись аналогічні гистограми для решти змінних, натискаємо на них внизу вікна або вибираємо в робочій книзі.

Про узгодженість даних з нормальним законом розподілу можна зробити висновок також із нормального ймовірнісного графіка. У вікні *Описательные статистики* виберемо *Диаграммы* → *Нормальные вероятностные графики*. Отримаємо графік, зображений на рис. 4.12.

Чим краще дані узгоджуються з нормальним законом розподілу, тим ближче до червоної прямої знаходяться зображені точки. Для нормальної вибірки всі точки лежать на даній прямій.

Для графічного представлення числових характеристик розміру варіації в пакеті STATISTICA можна вибрати певний тип діаграм розмаху, так званих “коробок з вусами”. У вікні *Описательные статистики* в закладці *Опции* виберемо, для прикладу, перші два типи діаграм розмаху (див. рис. 4.13).

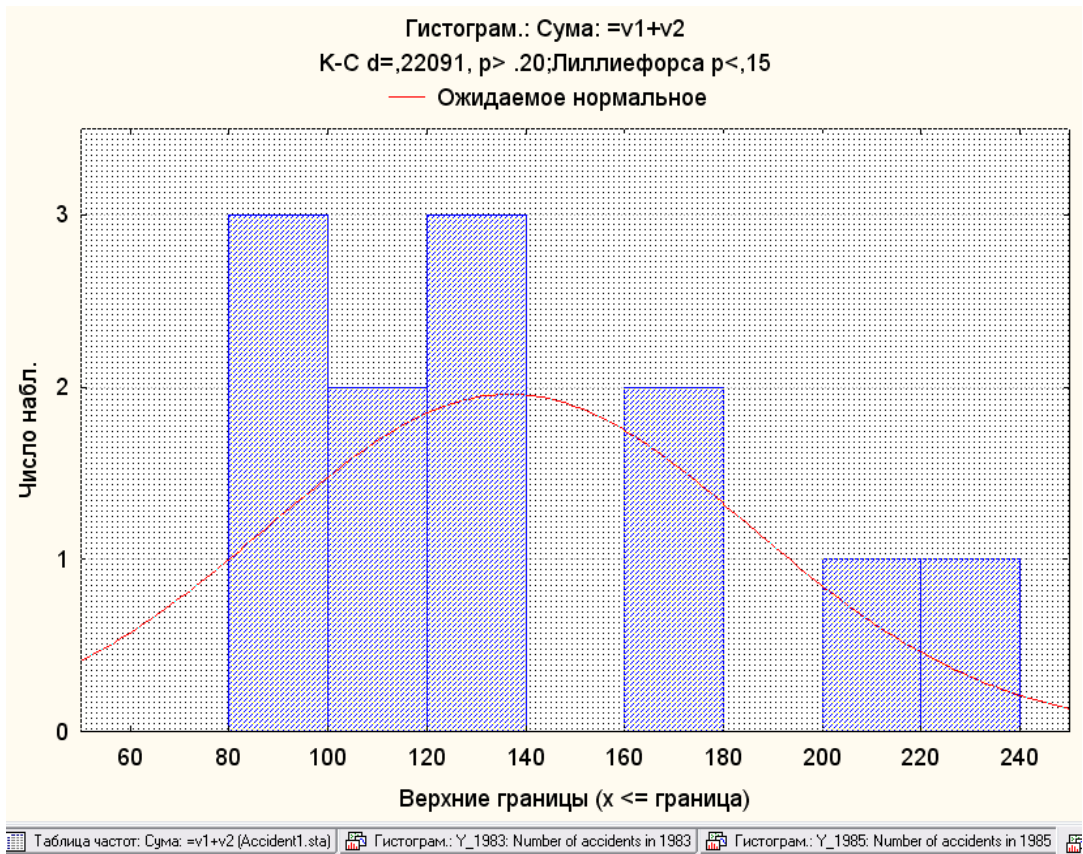


Рис. 4.11

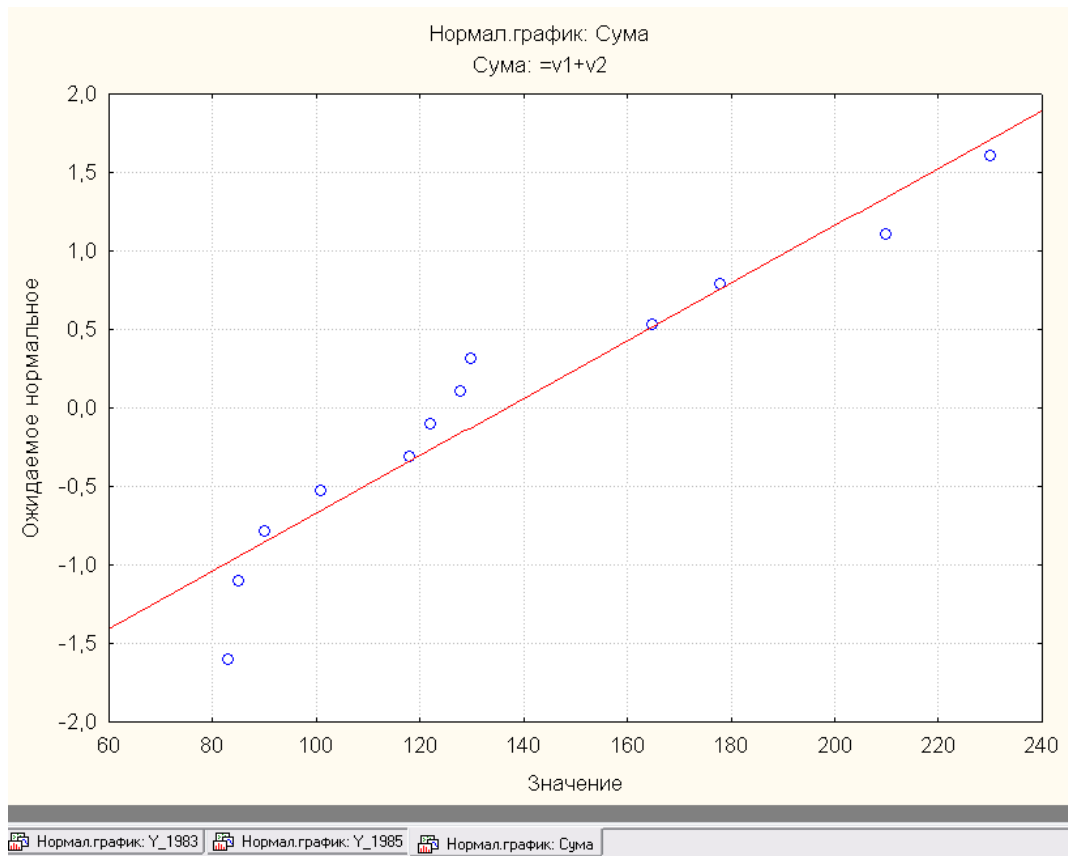


Рис. 4.12

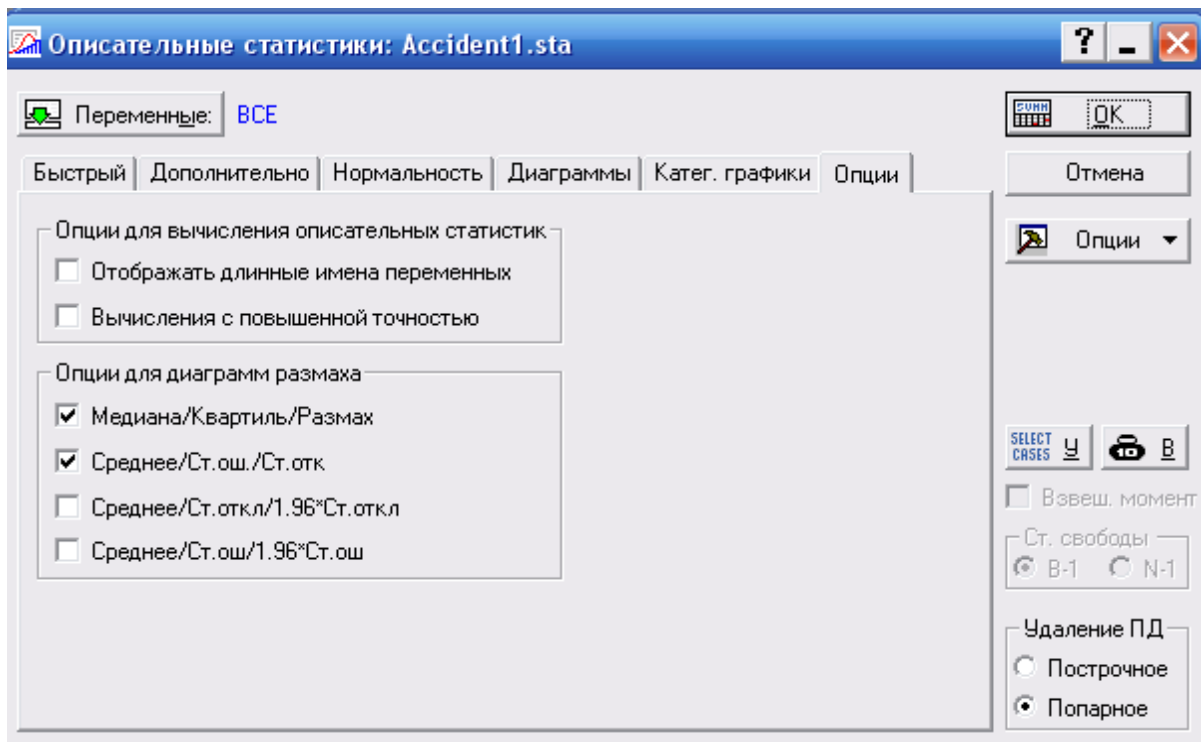


Рис. 4.13

Натиснемо *Быстрый* → *Диаграмма размаха* для всех переменных і маємо можливість продивитися “коробки з вусами” першого та другого типу:

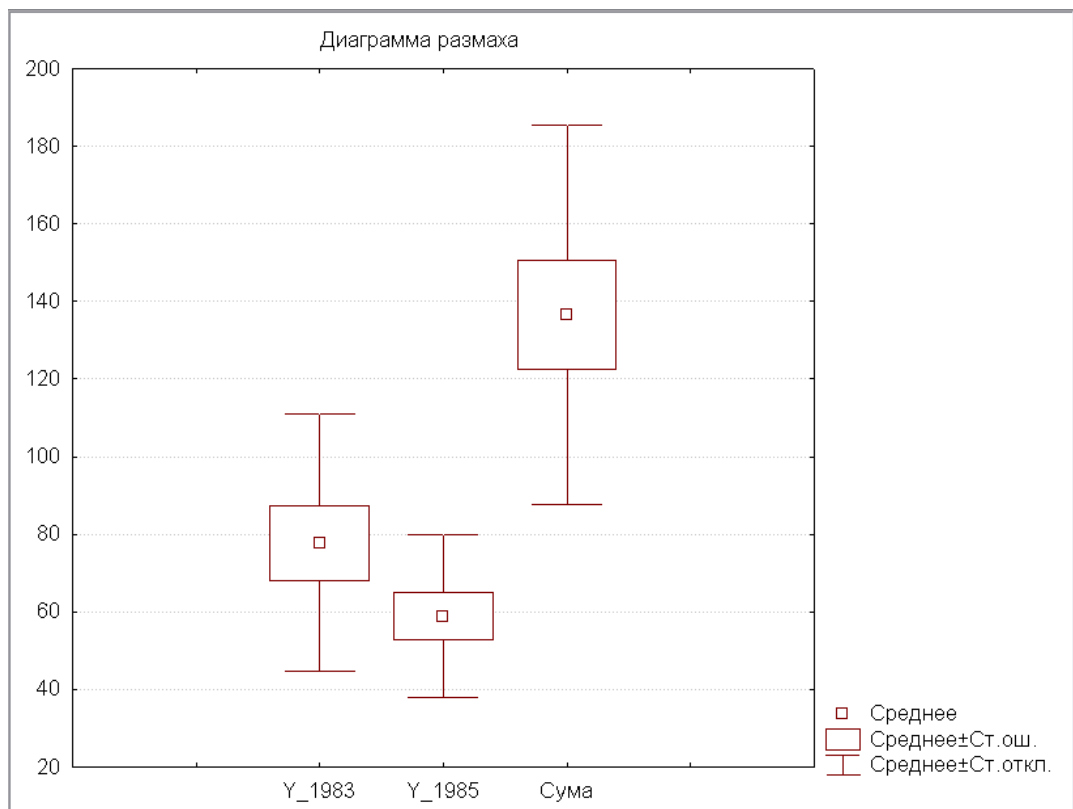


Рис. 4.14. а)

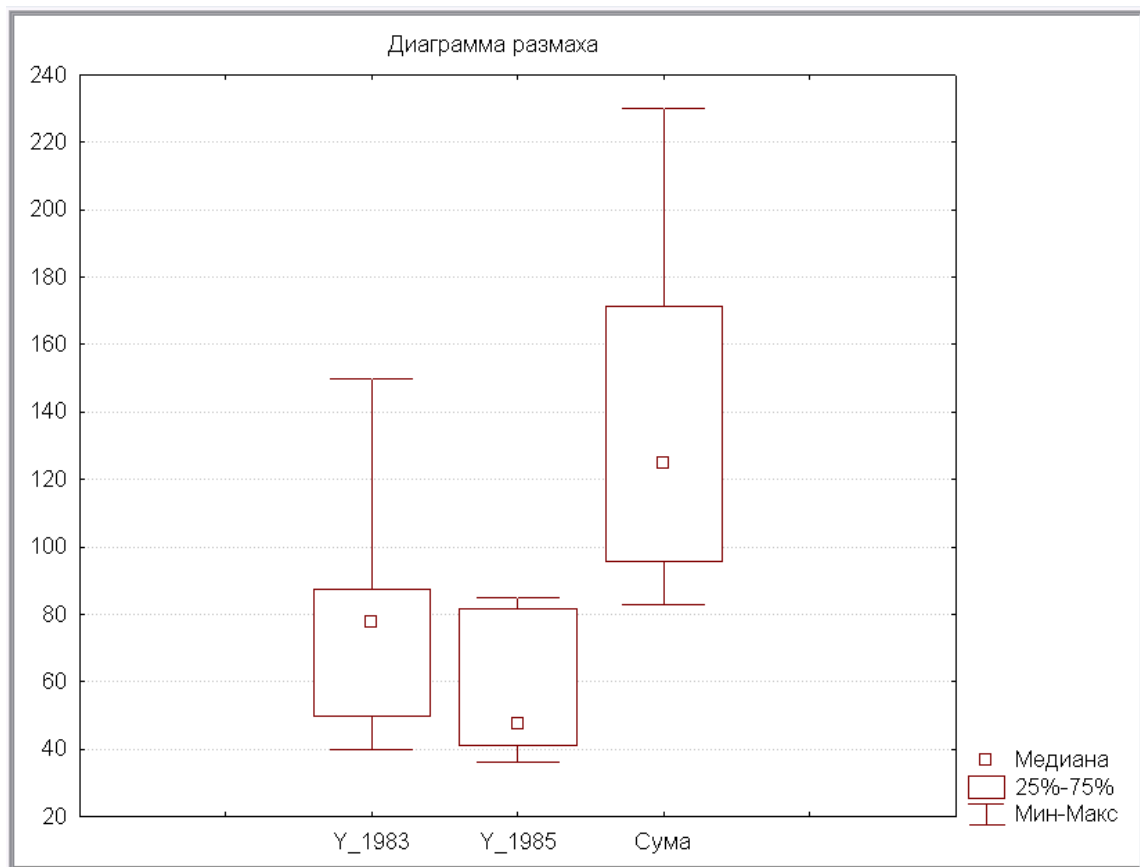


Рис. 4.14. б)

4.3. Парна лінійна регресія

Працюємо далі з таблицею Accident1.sta і виберемо *Анализ* → *Основные статистики и таблицы* → *Парные и частные корреляции* → *OK*. У вікні, що з'явилося, натиснемо *Квадратная матрица* і виберемо всі змінні для аналізу (див. рис. 4.15). Далі натискаємо *Матрица парных корреляций* та аналізуємо отриману матрицю (див. рис. 4.16).

Елементи даної матриці є вибірковими коефіцієнтами кореляції для відповідних змінних. На перетині однойменних рядків та стовпчиків (діагональні елементи кореляційної матриці) стоять одиниці, які відповідають повній кореляційній залежності між цими змінними.

Очевидно, що на діаграмі розсіювання, побудованій для однойменних змінних, всі точки будуть лежати на прямій $y = x$. Пересвідчимося у цьому. Натиснемо правою кнопкою миші на одному з елементів кореляційної матриці та в меню, що з'явилося, послідовно виберемо *Графики исходных данных* → *Диаграмма рассеяния* → *Регрессия, 95% доверит. интервал*; в списку змінних виберемо ту саму змінну **Y_1983** та отримаємо діаграму розсіювання (див. рис. 4.18).

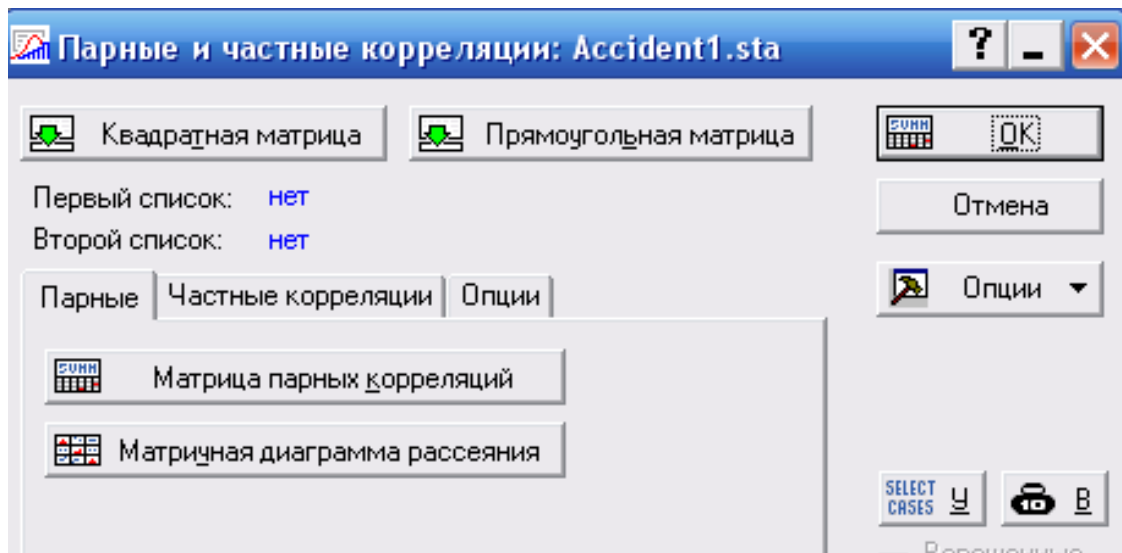


Рис. 4.15

Корреляции (Accident1.sta)			
Отмеченные корреляции значимы на уровне $p < ,05000$			
N=12 (Построчное удаление ПД)			
Переменная	Y_1983	Y_1985	Сума
Y_1983	1,00	0,61	0,94
Y_1985	0,61	1,00	0,84
Сума	0,94	0,84	1,00

Рис. 4.16

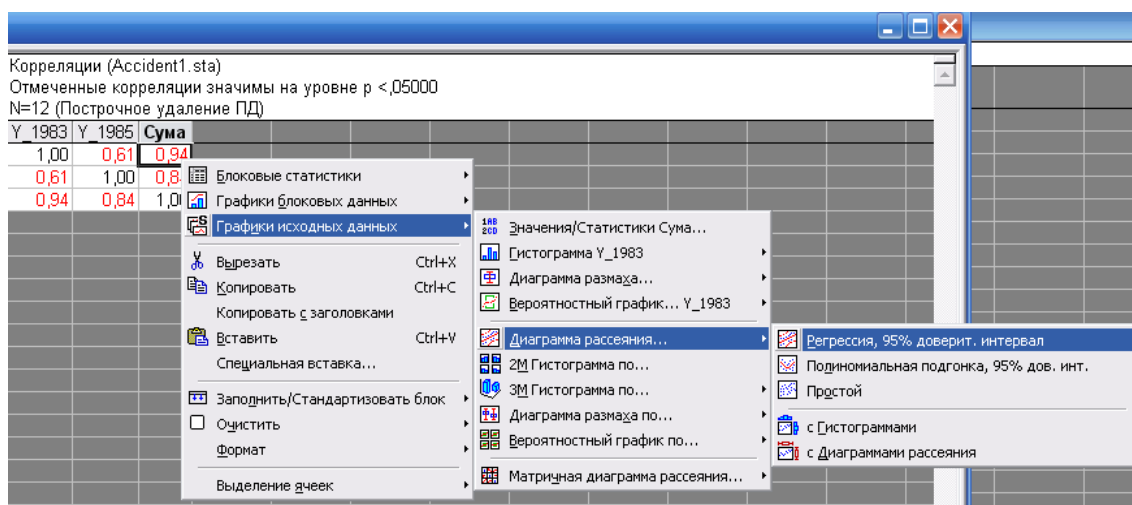


Рис. 4.17

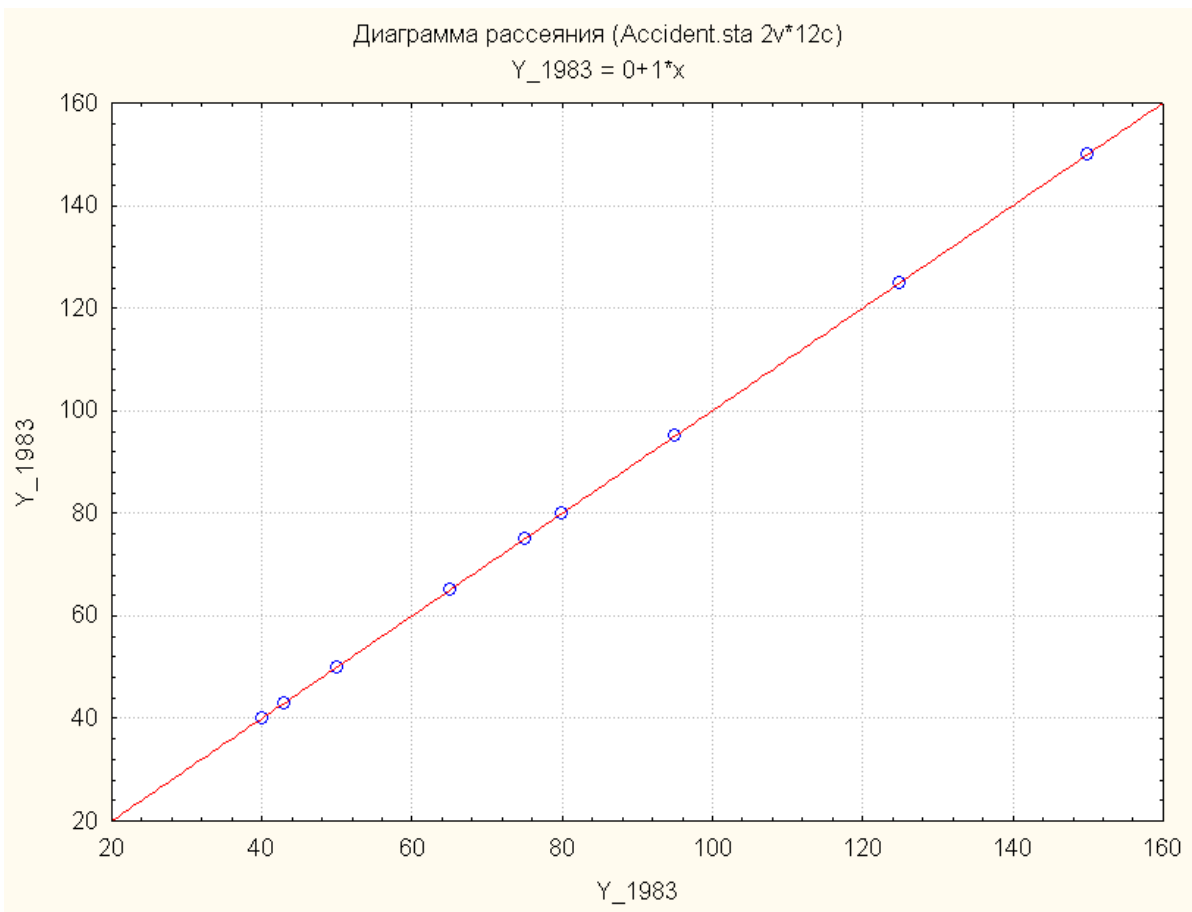


Рис. 4.18

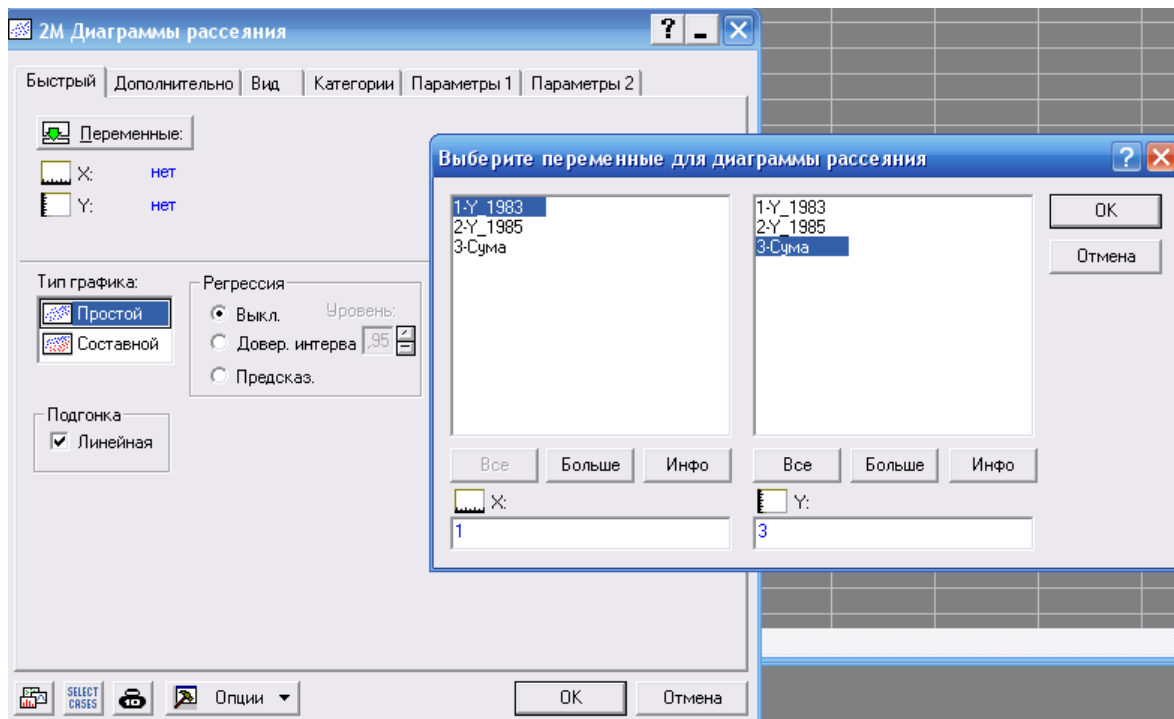


Рис. 4.19

Виберемо в основному меню *Графіка* → *Діаграмми розсіяння* і потрапимо у вікно, зображене на Рис. 4.19. У ньому натискаємо кнопку *Переменные* та вибираємо змінні для діаграми розсіяння: **Y_1983** для осі *Ox* та **Сума** для осі *Oy* (*OK*). На діаграмі розсіювання, побудованій за вибраними змінними, за умови вибору лінійної підгонки, буде зображено пряму регресії, рівняння якої виписане у заголовку даного графіка (див. рис. 4.20). Коефіцієнти в рівнянні регресії знаходяться методом найменших квадратів.

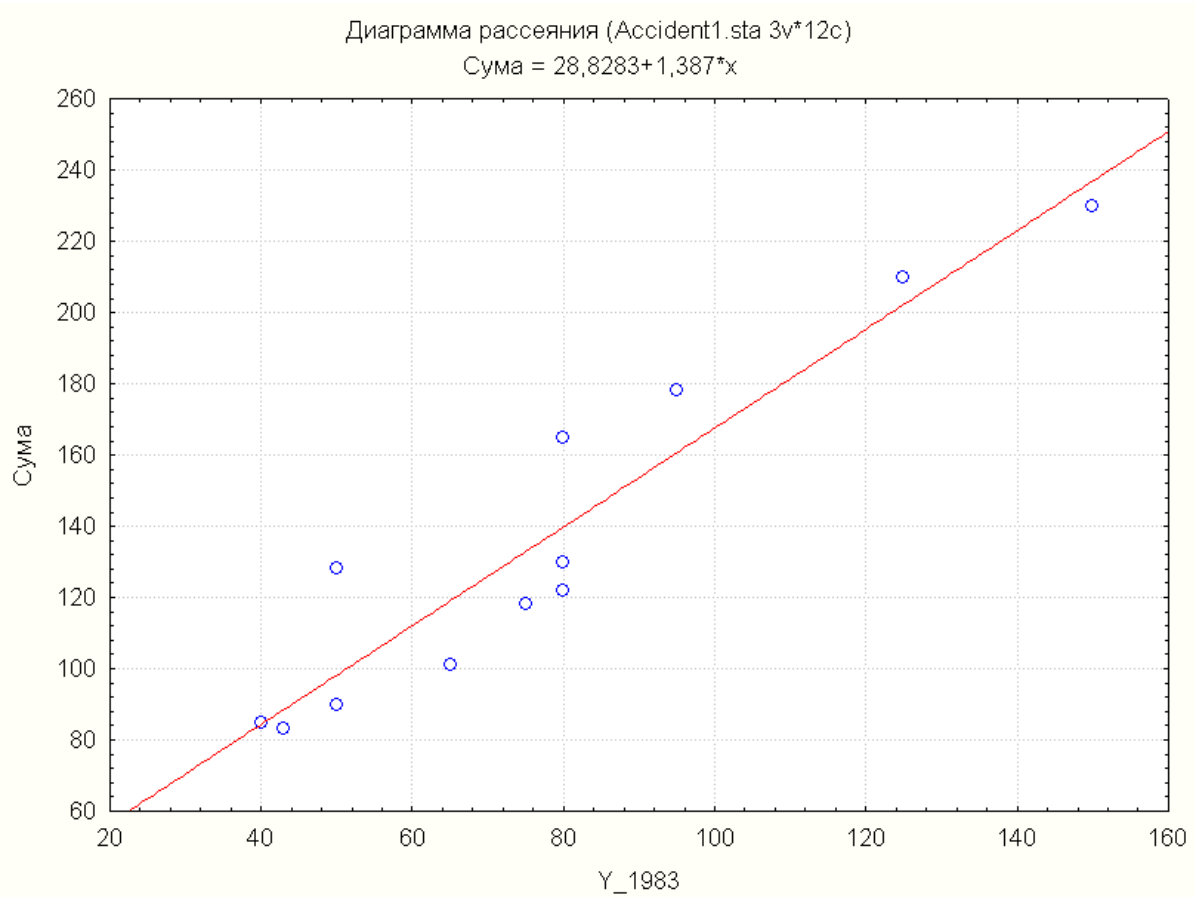


Рис. 4.20

4.4. Множинна лінійна регресія

Розглянута вище парна лінійна регресія має досить обмежену сферу застосування, оскільки на практиці часто доводиться досліджувати зв'язок між кількома незалежними змінними і однією залежною. Таку регресію називають множинною, або багатofакторною.

Модель множинної лінійної регресії можна записати у вигляді

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n,$$

де n – число незалежних змінних (факторів) множинної регресії.

Для правильного визначення коефіцієнтів b_i у множинній лінійній регресійній моделі методом найменших квадратів повинні виконуватись такі умови:

- похибки спостережень є нормально розподіленими випадковими величинами з нульовим середнім;
- дисперсія похибок є сталою;
- похибки не є автокорельованими;
- фактори множинної регресії є некорельованими (відсутність мультиколінеарності).

Якщо деякі фактори множинної регресії корелюють між собою, то варто переглянути регресійну модель для того, щоб позбутися мультиколінеарності.

За допомогою пакету STATISTICA побудуємо модель множинної лінійної регресії для даних із файлу Job_prof.sta, який знаходиться у директорії Examples пакету. Відкриємо цей файл. Перші чотири змінні (Test1-Test4) показують результати тестів на професійну придатність 25 претендентів на посаду службовців компанії (див. рис. 4.21). Після випробувального терміну робота кожного з службовців була оцінена за єдиною шкалою (змінна Job_prof).

Job proficiency data set from Neter, Wasserman, & Kutner, 1989					
	1	2	3	4	5
	TEST1	TEST2	TEST3	TEST4	JOB_PROF
1	86	110	100	87	88
2	62	97	99	100	80
3	110	107	103	103	96
4	101	117	93	95	76
5	100	101	95	88	80
6	78	85	95	84	73
7	120	77	80	74	58
8	105	122	116	102	116
9	112	119	106	105	104
10	120	89	105	97	99
11	87	81	90	88	64
12	133	120	113	108	126
13	140	121	96	89	94
14	84	113	98	78	71
15	106	102	109	109	111
16	109	129	102	108	109
17	104	83	100	102	100
18	150	118	107	110	127
19	98	125	108	95	99
20	120	94	95	90	82
21	74	121	91	85	67
22	96	114	114	103	109
23	104	73	93	80	78
24	94	121	115	104	115
25	91	129	97	83	83

Рис. 4.21

Виберемо *Анализ* → *Множественная регрессия* → *Переменные*. Як залежну змінну вибираємо Job_prof, незалежні змінні – Test1-Test4. Натиснемо *OK* у вікні вибору змінних (див. рис. 4.23).

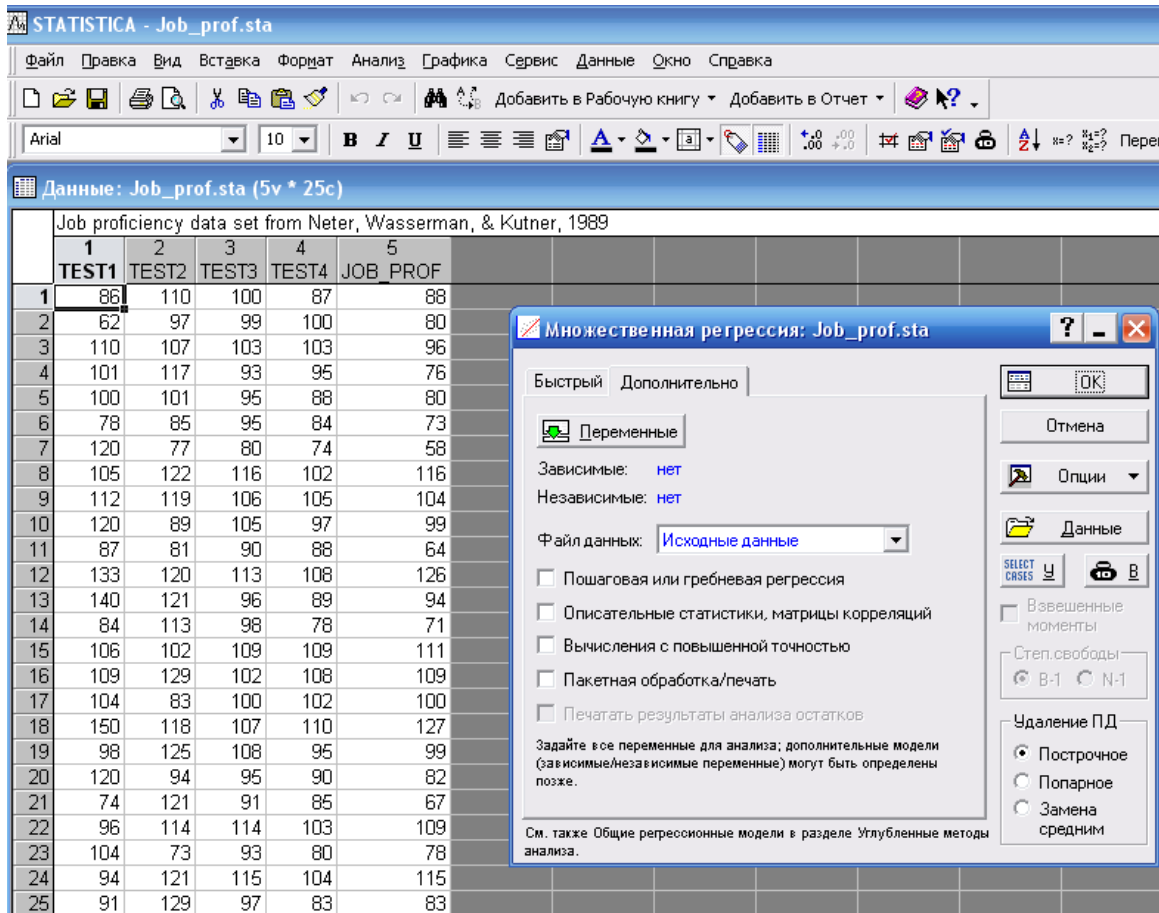


Рис. 4.22

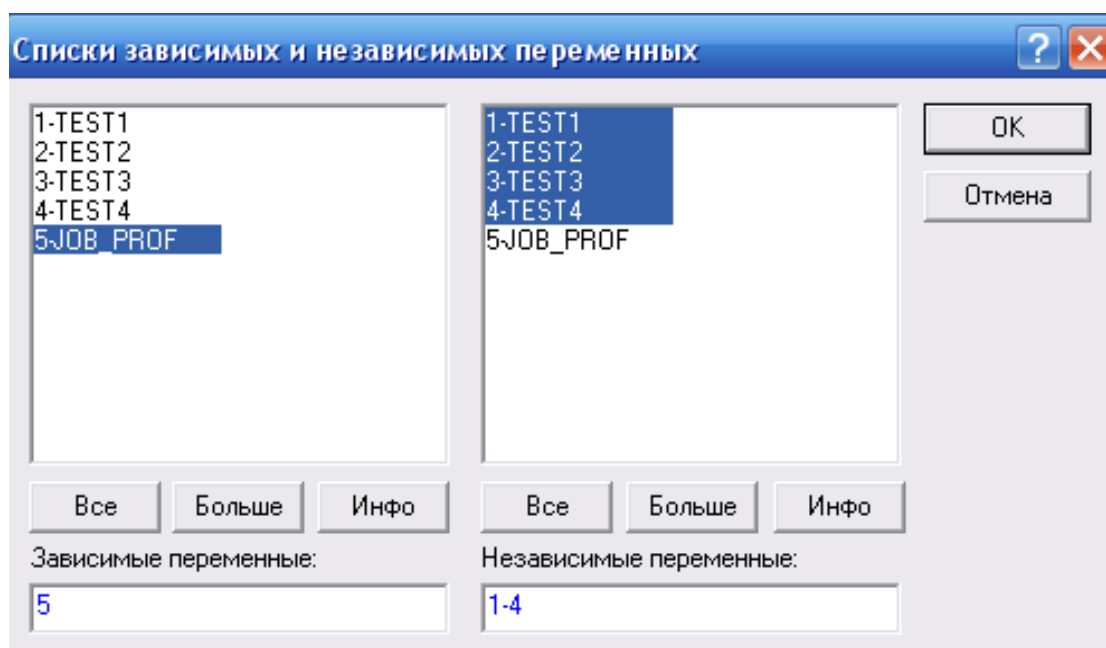


Рис. 4.23

Натиснемо ще раз *OK* у вікні *Множественная регрессия* (див. рис. 4.24). У вікні результатів роботи модуля *Множественная регрессия* (див. рис. 4.25) натиснемо *Итоговая таблица регрессии*.

Результати стандартного методу побудови регресійної моделі (див. рис. 4.26). спонукають до перегляду незалежних змінних і вилучення з розгляду змінної TEST2 (вона не виділена як значима змінна). Використаємо покрокову процедуру, кожний крок якої включатиме до регресійної моделі найбільш значиму незалежну змінну (найбільш впливовий фактор) поки не закінчатся значимі змінні.

Повернемося до вікна *Множественная регрессия* (див. рис. 4.24), де відзначимо опцію *Пошаговая или гребневая регрессия*. Після натискання *OK* потрапляємо у вікно *Определение модели* (див. рис. 4.27), де вибираємо *Процедура* → *Пошаговая с включением* → *OK*.

На початку роботи вибраної покрокової процедури до регресійної моделі не включено жодну змінну. У вікні *Результаты множ. регрессии* (*Шаг 0*) натискаємо *Далее* (див. рис. 4.28).

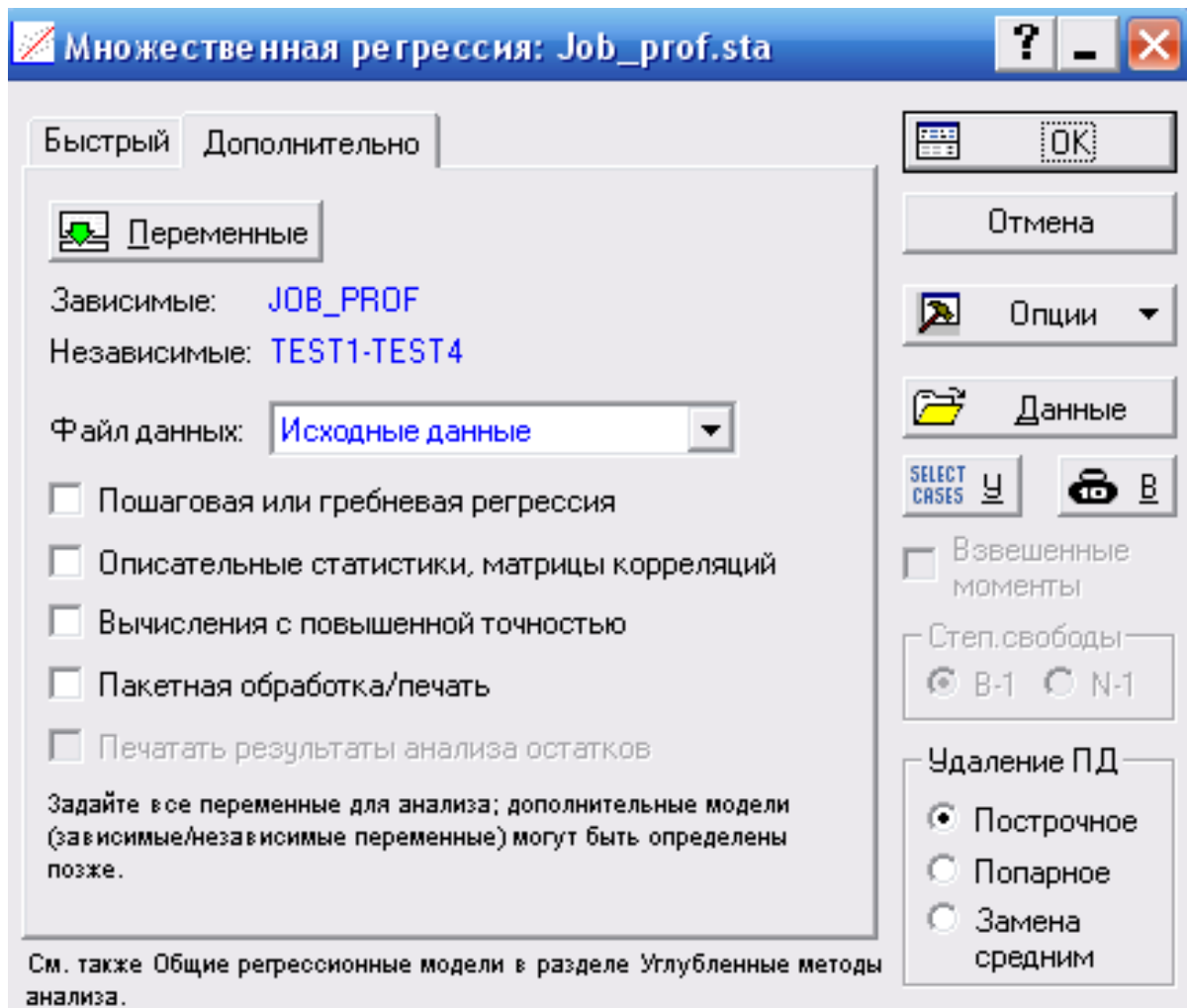


Рис. 4.24

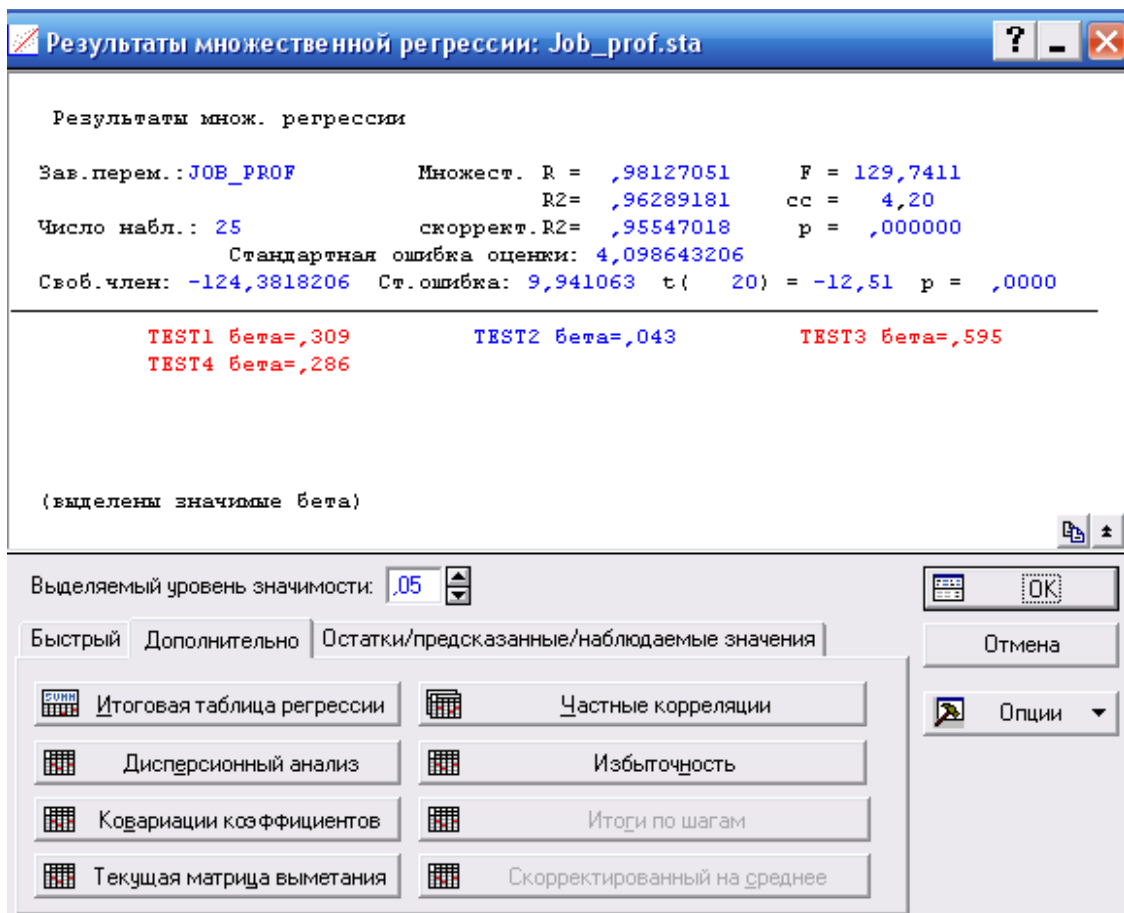


Рис. 4.25

Итоги регрессии для зависимой переменной: JOB_PROF (Job_prof.:
 R= ,98127051 R2= ,96289181 Скорректир. R2= ,95547018
 F(4,20)=129,74 p<,000000 Станд. ошибка оценки: 4,0986

N=25	БЕТА	Стд. Ош. БЕТА	B	Стд. Ош. B	t(20)	p-уров.
Св.член			-124,382	9,941063	-12,5119	0,000000
TEST1	0,309042	0,045951	0,296	0,043971	6,7254	0,000002
TEST2	0,042992	0,050408	0,048	0,056617	0,8529	0,403826
TEST3	0,595438	0,074813	1,306	0,164091	7,9591	0,000000
TEST4	0,285723	0,072524	0,520	0,131943	3,9397	0,000810

Рис. 4.26

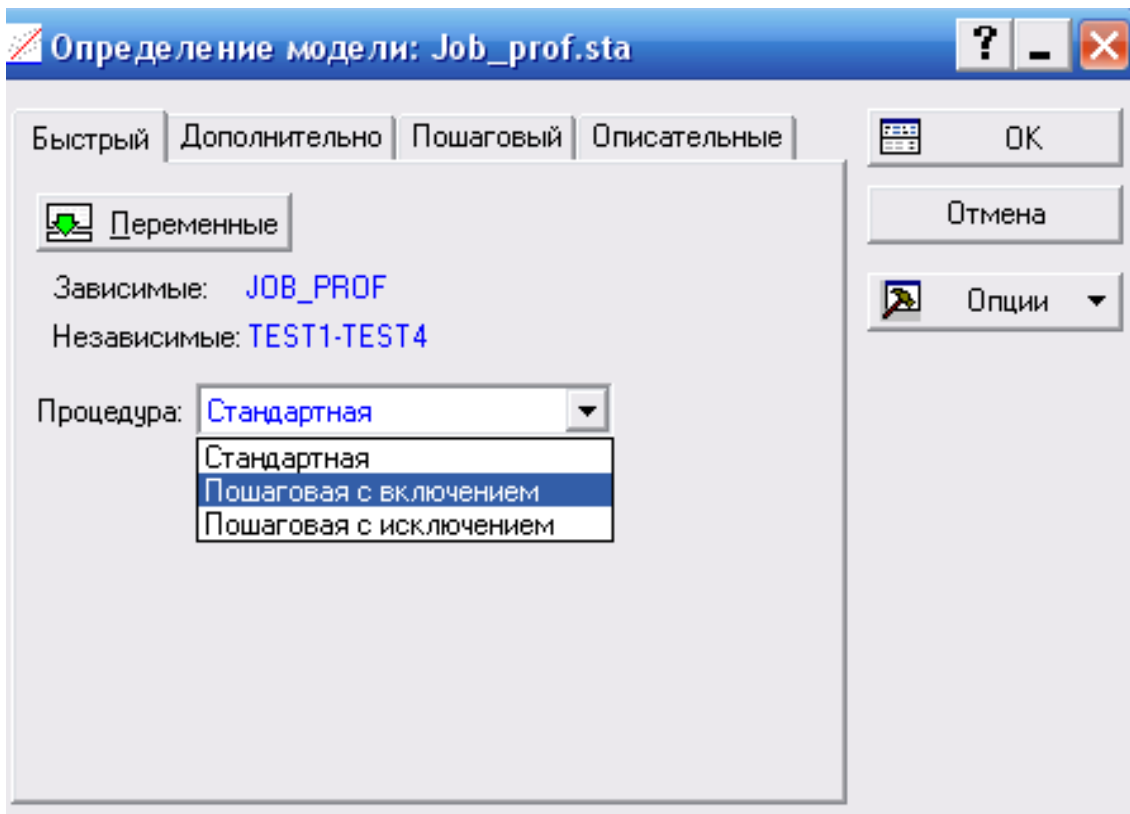


Рис. 4.27

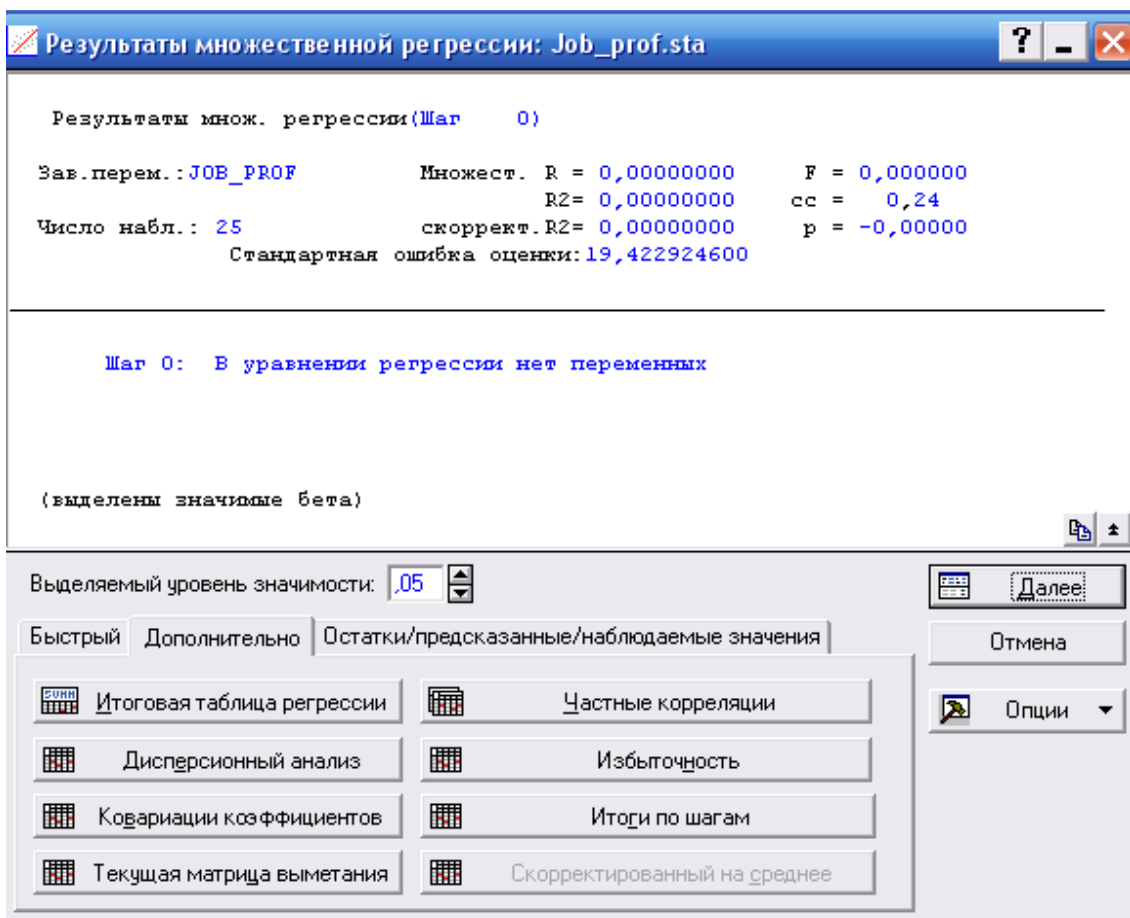


Рис. 4.28

Після першого кроку до регресійної моделі включено найбільш значиму змінну TEST3. У вікні *Результаты множ. регрессии (Шаг 1)* (див. рис. 4.29) натискаємо *Далее*.

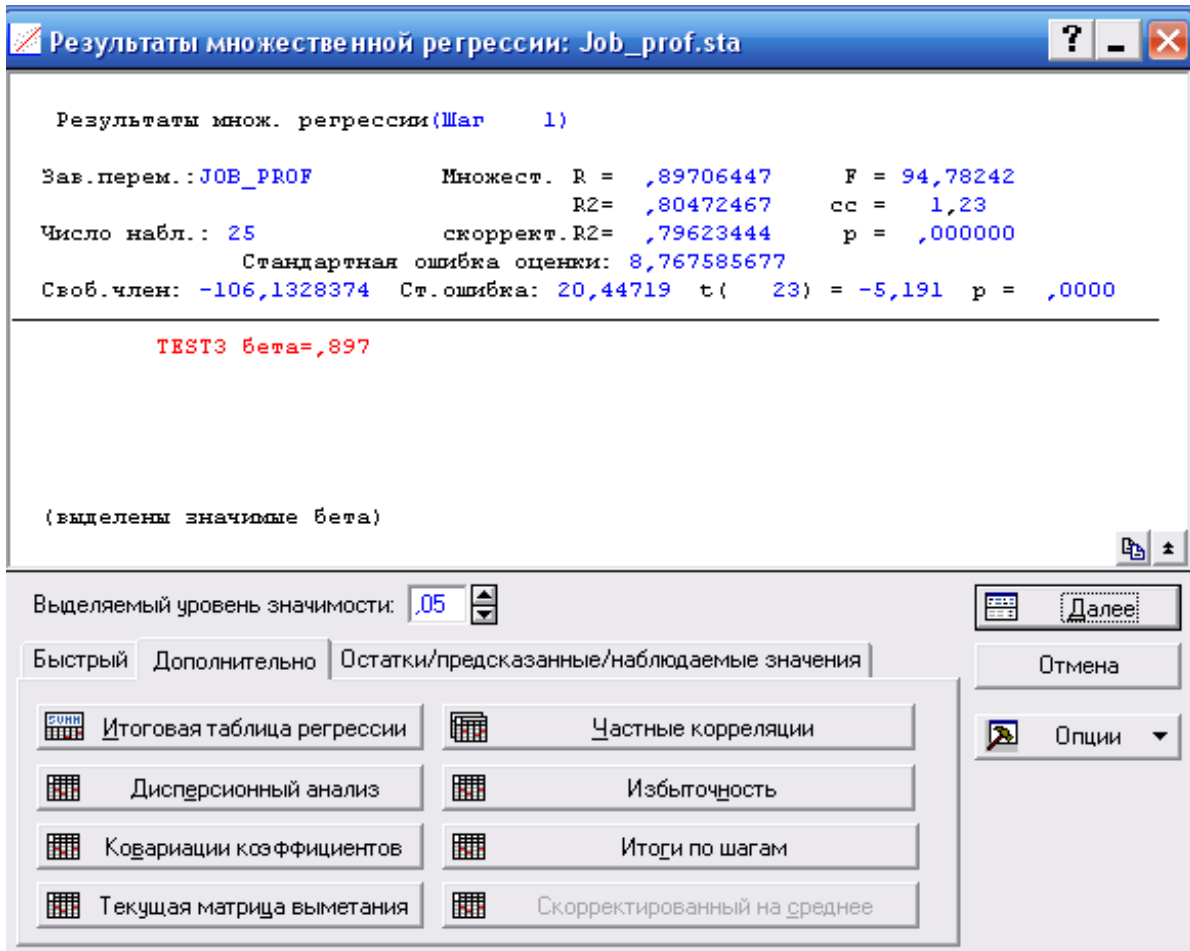


Рис. 4.29

На другому кроці до регресійної моделі включено змінну TEST1. У вікні *Результаты множ. регрессии (Шаг 2)* натискаємо *Далее*. Після цього отримуємо вікно з остаточними результатами множинної регресії, куди включено всі значимі змінні (див. рис. 4.30).

У вікні результатів роботи модуля *Множественная регрессия* натиснемо *Итоговая таблица регрессии* і отримаємо результуючу таблицю (див. рис. 4.31). У стовпчику В записані оцінки параметрів моделі, знайдені методом найменших квадратів. Залежна змінна Job_prof знаходиться за формулою

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 =$$

$$= -124,2 + 1,357 \cdot TEST3 + 0,296 \cdot TEST1 + 0,517 \cdot TEST4.$$

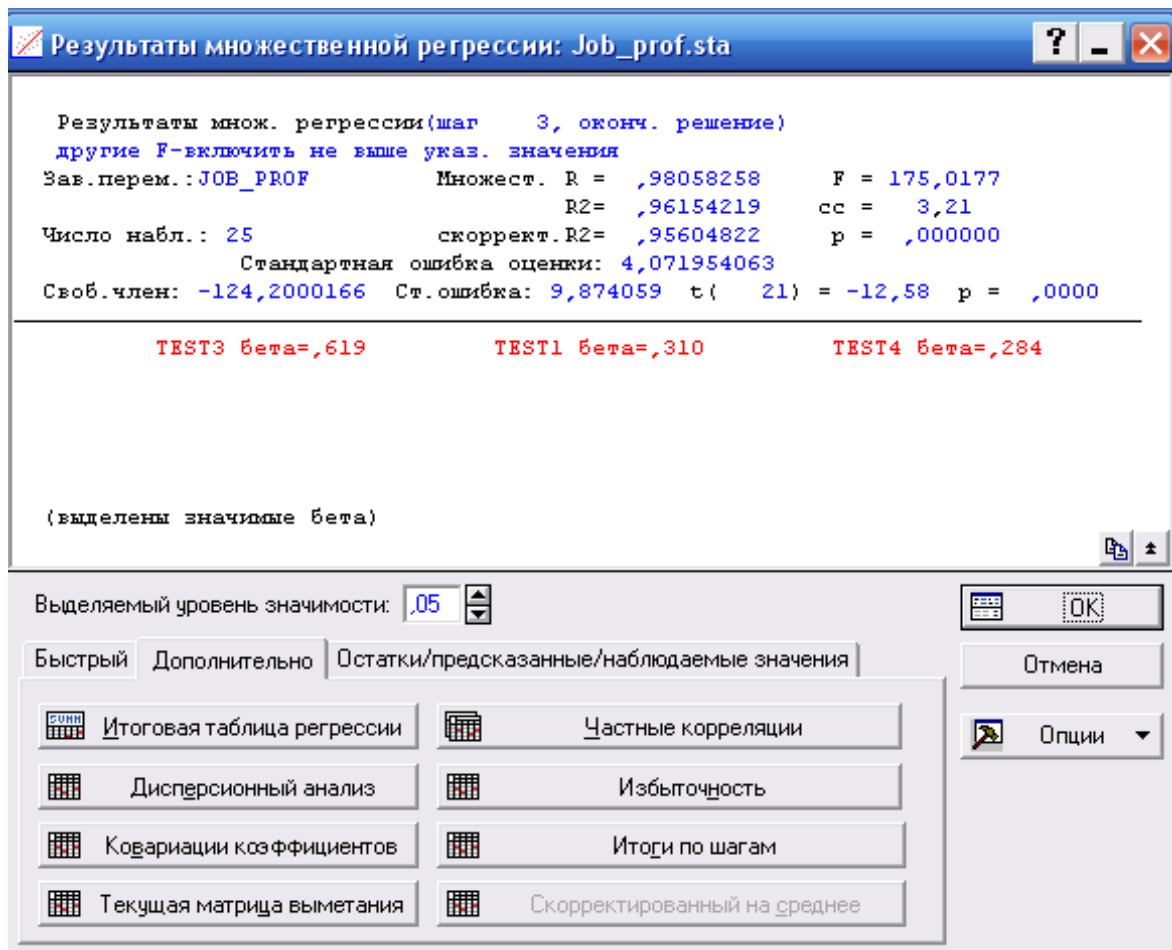


Рис. 4.30

Итоги регрессии для зависимой переменной: JOB_PROF (Job_prof.sta) R= ,98058258 R2= ,96154219 Скорректир. R2= ,95604822 F(3,21)=175,02 p<,000000 Станд. ошибка оценки: 4,0720						
N=25	БЕТА	Стд. Ош. БЕТА	B	Стд. Ош. B	t(21)	p-уров.
Св. член			-124,200	9,874059	-12,5784	0,000000
TEST3	0,618670	0,069224	1,357	0,151832	8,9373	0,000000
TEST1	0,309670	0,045646	0,296	0,043679	6,7841	0,000001
TEST4	0,284405	0,072035	0,517	0,131054	3,9482	0,000735

Рис. 4.31

Для перевірки припущення узгодженості похибок спостережень з нормальним законом можна побудувати нормальний ймовірнісний графік залишків. При задовільному узгодженні похибок спостережень з нормальним законом можна передбачити залежну змінну, надаючи при цьому певних значень незалежним змінним.

Повернемось до вікна *Результаты множественной регрессии* (див. рис. 4.30), натиснемо *Остатки/предсказанные/наблюдаемые значения* → *Анализ остатков* → *Нормальный график остатков*. Отримаємо графік, зображений на рис. 4.32.

На закладці *Остатки/предсказанные/наблюдаемые значения* виберемо *Предсказать зависимую переменную*. У вікні *Задайте значения независимых переменных* в полі *Общее значение* набираємо *100* і натискаємо *Применить* → *ОК* (див. рис. 4.33).

Передбачувані значення для залежної змінної *Job_prof* відображені на рис. 4.34.

Бачимо, що отримані 100 балів за кожний з трьох тестів на професійну придатність не гарантують, що реальна робота претендента на вакантну посаду за результатами випробувального терміну буде оцінена так само високо. Передбачувана оцінка в даному випадку близька до 93 балів. Більше того, з ймовірністю 0,95 передбачувана оцінка буде в межах між 90,4 і 95,3 бали, що є підставою замислитись роботодавцю, чи взагалі брати такого претендента на випробувальний термін.

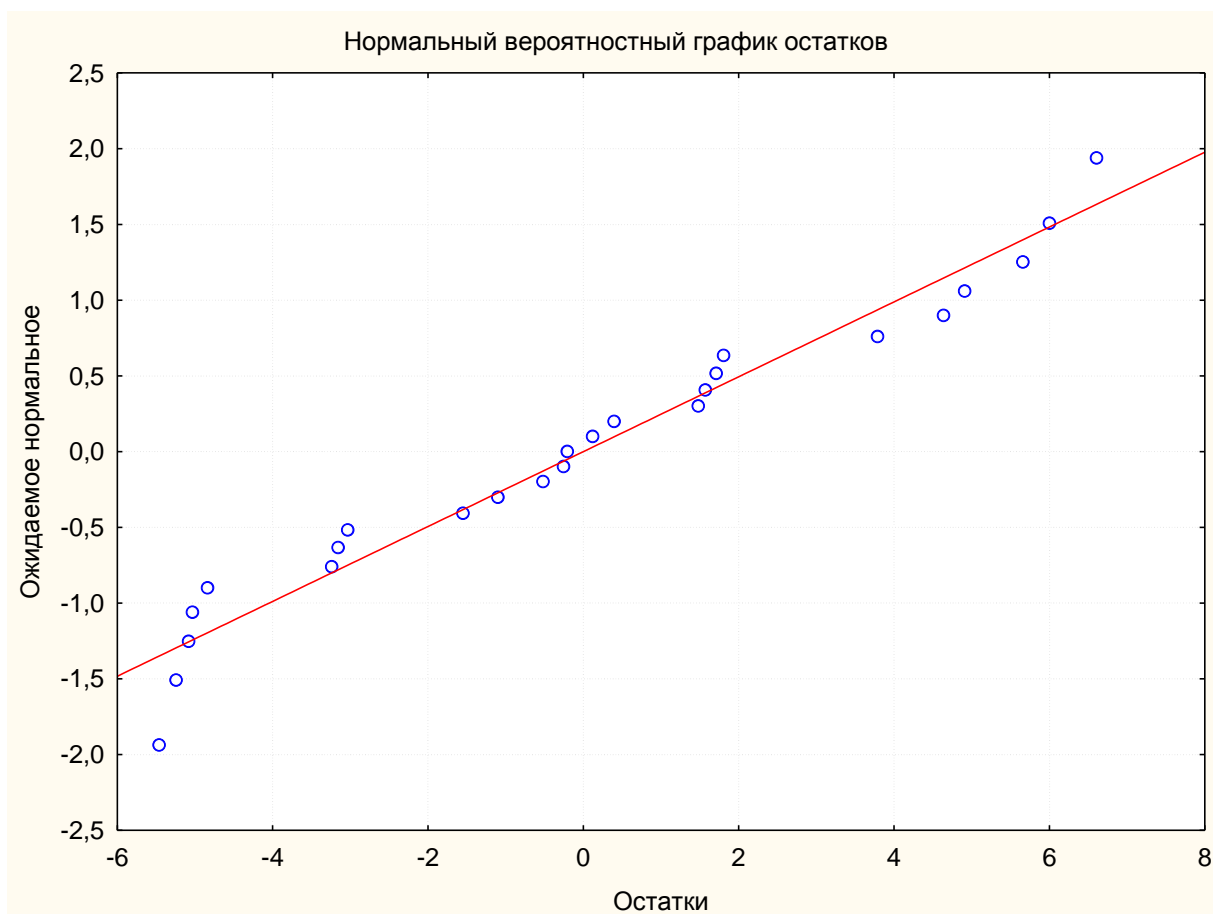


Рис. 4.32

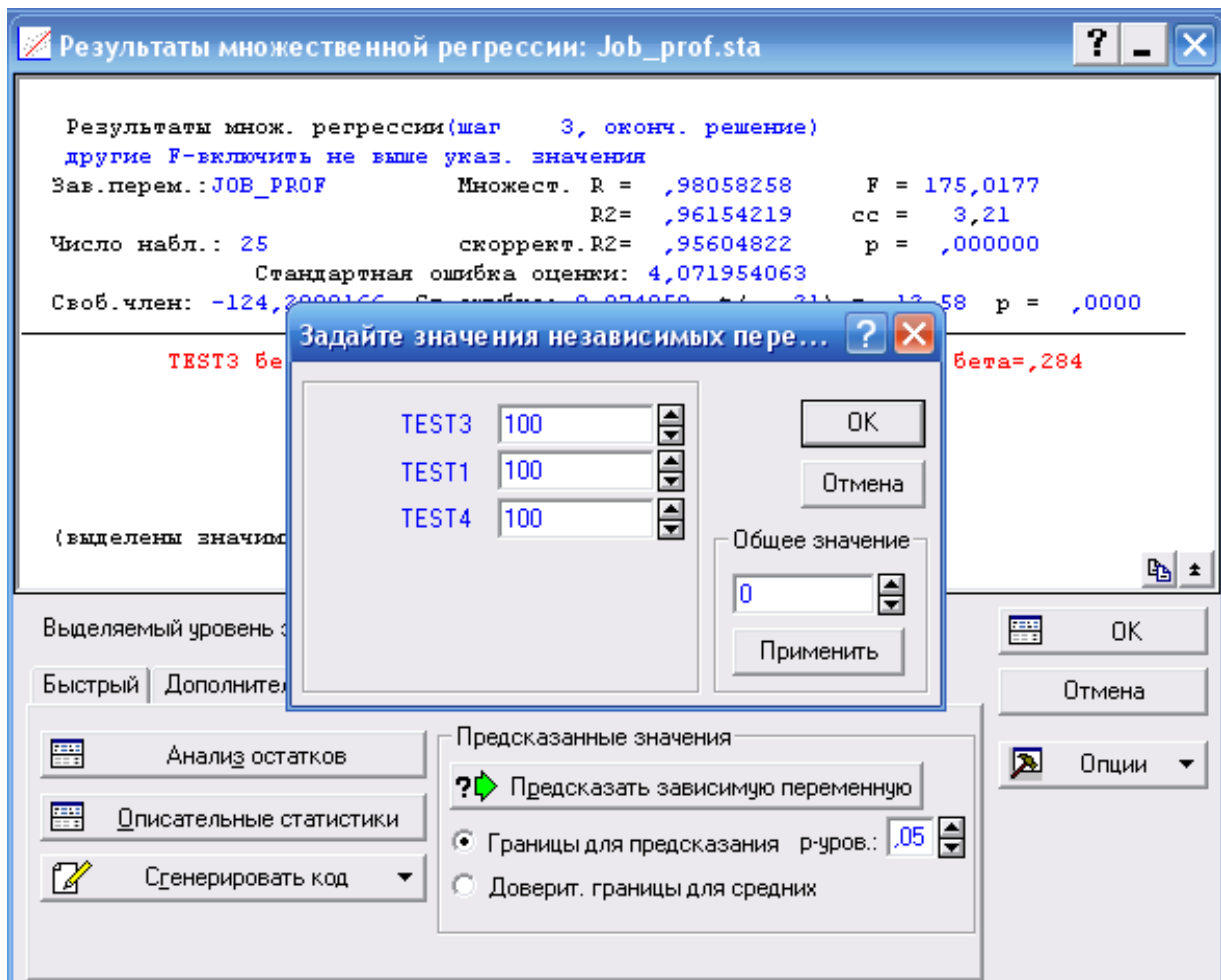


Рис. 4.33

Workbook2* - Предск.значения для (Job_prof.sta)

Workbook2*

- Множественная регрессия (Job_prof.sta)
 - Результаты множественной регрессии (Job_prof.sta)
 - Предск.значения для (Job_prof.sta)

Предск.значения для (Job_prof.sta) перемен.: JOB_PROF			
Переменная	В-Вес	Значение	В-Вес * знач.
TEST3	1,356968	100,0000	135,697
TEST1	0,296326	100,0000	29,633
TEST4	0,517421	100,0000	51,742
Св.член			-124,200
Предсказ.			92,871
-95,0%ДП			90,425
+95,0%ДП			95,318

Рис. 4.34

**Контрольні завдання для самоперевірки до теми
„Статистичний аналіз даних з пакетом STATISTICA ”**

1. Створення таблиці з даними в пакеті STATISTICA.
2. Збереження та копіювання даних.
3. Знаходження числових характеристик вибірки засобами пакету STATISTICA.
4. Перевірка узгодженості вибірових даних з нормальним законом розподілу.
5. Діаграми розмаху в пакеті STATISTICA.
6. Діаграми розсіяння в пакеті STATISTICA.
7. Побудова кореляційної матриці засобами пакету STATISTICA.
8. Знаходження рівняння регресії в пакеті STATISTICA.
9. Умови адекватності моделі множинної лінійної регресії.
10. Стандартний та покроковий методи регресійного аналізу.
11. Інтерпретація результатів роботи модуля „Множинна регресія”.
12. Прогнозування в модулі „Множинна регресія”.

Глава 5. Кластерний аналіз

5.1. Мета та застосування кластерного аналізу

Термін **кластерний аналіз** включає у себе набір різних алгоритмів класифікації. Загальне питання, що задається дослідниками в багатьох областях, полягає в тому, як класифікувати дані, що спостерігаються, при цьому наочно представити отриману класифікацію.

Наприклад, біологи ставлять мету класифікувати тварин, щоб змістовно описати розходження між ними. Відповідно до сучасної системи, прийнятої в біології, людина належить до приматів, ссавців, амніот, хребетних і тварин. Помітимо, що в цій класифікації, чим вище рівень об'єднання (агрегації), тим менше подібності між членами у відповідному класі. Людина має більше подібності з іншими приматами (тобто з мавпами), чим з “віддаленими” членами сімейства ссавців (наприклад, собаками) і т.д.

Техніка кластеризації застосовується в найрізноманітніших областях. Відомі широкі застосування кластерного аналізу в медицині, археології, географії, маркетингових дослідженнях. Відзначимо, що у випадку, коли необхідно класифікувати великий об'єм інформації і утворити придатні для подальшої обробки групи, кластерний аналіз виявляється дуже корисним і ефективним.

Існує точка зору, що на відміну від багатьох інших статистичних процедур, методи кластерного аналізу використовуються в більшості випадків тоді, коли немає яких-небудь апріорних гіпотез щодо класів об'єктів і є потреба у початковій класифікації даних.

5.2. Методи кластеризації

Ієрархічна кластерна техніка полягає у побудові деякої послідовності розбиттів множини, починаючи з розбиття на кластери, кожен з яких містить тільки один елемент даної множини, і завершуючи одним кластером, що містить усю дану множину. Якщо кластеризація відбувається у вказаному напрямі, то маємо процес об'єднання (агломерації), якщо у протилежному (на першому кроці один кластер, що містить усю множину, на останньому – кількість кластерів рівна кількості елементів даної множини), то маємо процес подрібнення.

Графічно процес ієрархічної класифікації зображують у вигляді **дендрограми** (деревоподібної діаграми), яка відображає об'єднання чи подрібнення, що відбувається на кожному кроці.

Розглянемо алгоритм об'єднання (деревоподібної кластеризації). Призначення цього алгоритму полягає в об'єднанні об'єктів у досить великі кластери, використовуючи деяку міру подібності чи відстані між об'єктами. Основні типи відстаней між об'єктами та кластерами будуть

наведені нижче, також буде аргументовано вибір тієї чи іншої відстані для оптимальної побудови послідовності кластерних розбиттів.

Проаналізуємо типовий результат кластеризації – ієрархічне дерево.

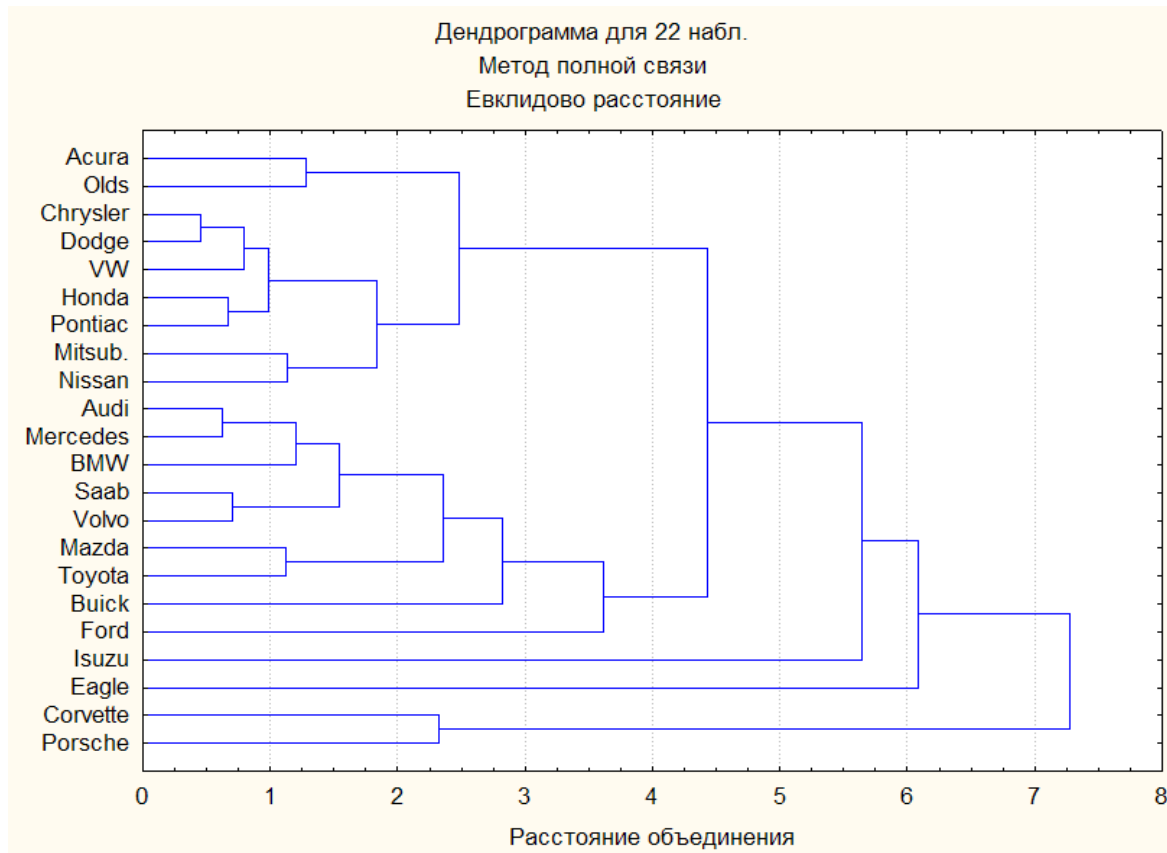


Рис.5.1

Зображена деревоподібна діаграма починається з кожного об'єкта в класі (у лівій частині діаграми). Тепер уявимо собі, що поступово (дуже малими кроками) будемо послаблювати критерій стосовно того, які об'єкти є унікальними. Іншими словами, з кожним кроком відбувається об'єднання двох чи кількох найбільш подібних об'єктів у один кластер. Далі зв'язуються разом усе більше і більше число об'єктів і утворюються кластери, що складаються з елементів, які усе сильніше відрізняються один від одного. Остаточо, на останньому кроці, всі об'єкти об'єднуються у один кластер.

На цих діаграмах горизонтальні вісі представляють відстань об'єднання (у вертикальних деревоподібних діаграмах вертикальні вісі представляють відстань об'єднання). Так, для кожного вузла у діаграмі (там, де формується новий кластер) можна бачити величину відстані, для якої відповідні елементи зв'язуються в новий єдиний кластер.

Коли дані мають чітку “структуру” у термінах кластерів об'єктів, подібних між собою, тоді ця структура, швидше за все, повинна бути відображена в ієрархічному дереві різними гілками. У результаті

успішного аналізу методом об'єднання з'являється можливість знайти кластери (гілки ієрархічного дерева) та інтерпретувати їх.

5.3. Вимірювання відстані між об'єктами

Метод деревоподібної кластеризації використовується при формуванні кластерів на основі відмінностей чи відстаней між об'єктами. Ці відстані можуть визначатися в одновимірному чи багатовимірному просторі. Наприклад, якщо потрібно кластеризувати типи продуктів, то можна взяти до уваги кількість калорій, що містяться в них, ціну, суб'єктивну оцінку смаку і т.д.

Однією з можливостей обчислення відстаней між об'єктами в багатовимірному просторі є обчислення евклідових відстаней. У дво- чи тривимірному евклідовому просторі ця відстань є реальною геометричною відстанню між об'єктами (ніби відстані між об'єктами вимірюються рулеткою). Однак алгоритм об'єднання не "підключається" про те, чи адекватно відбувається об'єднання в кластери для вибраної відстані, тому задачею дослідників є вдало підібрати правильний метод кластеризації, результат якого був би придатним для подальших використань.

Найчастіше в кластерному аналізі користуються такими відстанями.

Евклідова відстань:

$$d(x,y) = \sqrt{\sum_i (x_i - y_i)^2}.$$

Помітимо, що евклідова відстань (та її квадрат) обчислюється за вихідними, а не за стандартизованими даними. Це має певні переваги (наприклад, відстань між двома об'єктами не змінюється при введенні в аналіз нового об'єкта, що може виявитися викидом). Проте, на евклідову відстань можуть сильно впливати зміни масштабних одиниць між координатними осями. Як наслідок, після таких змін результати кластерного аналізу можуть сильно відрізнятись від попередніх.

Квадрат евклідової відстані:

$$d(x,y) = \sum_i (x_i - y_i)^2.$$

Ця відстань використовується, щоб підкреслити відмінність між більш віддаленими один від одного об'єктами.

Відстань міських кварталів (манхеттенська відстань):

$$d(x,y) = \sum_i |x_i - y_i|.$$

Вона є сумою різниць відповідних координат і геометрично означає, що ми вимірюємо відстань між об'єктами, рухаючись найкоротшим шляхом вздовж вулиць певного міста, які паралельні до координатних осей. У більшості випадків ця міра відстані приводить до близьких результатів до звичайної відстані Евкліда. Зазначимо, що для цієї відстані вплив окремих великих різниць (при наявності викидів) зменшується, тому що вони не підносяться до квадрату.

Відстань Чебишева:

$$d(x,y) = \max |x_i - y_i|.$$

Ця відстань може виявитися корисною, коли бажають визначити два об'єкти як “різні”, якщо вони суттєво відрізняються по який-небудь одній координаті (яким-небудь одним виміром).

Степенева відстань:

$$d(x,y) = \sqrt[r]{\sum_i (x_i - y_i)^p}.$$

Параметри r та p визначаються дослідником. Параметр p відповідає за поступове зважування різниць по окремих координатах, параметр r відповідає за прогресивне зважування великих відстаней між об'єктами. Кілька прикладів обчислень можуть показати, як “працює” степенева відстань. Якщо обидва параметри рівні двом, та ця відстань співпадає з відстанню Евкліда.

Відсоток незгоди:

$$d(x,y) = (\text{Кількість } x_i \neq y_i) / i.$$

Ця відстань використовується у тих випадках, коли дані не мають кількісного виразу, тобто є категоріальними.

5.4. Вимірювання відстані між кластерами

На першому кроці кластеризації, коли кожен об'єкт є окремим кластером, відстані між цими об'єктами визначаються одним із обговорених у розділі 5.3 способів. На наступних кроках, коли зв'язуються разом кілька об'єктів, виникає питання, як визначати відстані між кластерами? Іншими словами, необхідно мати правило об'єднання для двох різних кластерів.

Серед методів об'єднання кластерів можна виділити наступні.

Одиночний зв'язок (метод найближчого сусіда). У цьому методі відстань між двома кластерами визначається відстанню між двома найбільш близькими об'єктами (найближчими сусідами) у різних кластерах. Це правило буде “волокнисті” кластери, тобто такі ланцюжки, “зчеплені разом” тільки окремими елементами, які випадково опинились ближче інших один до одного.

Як альтернативу розглянутому методу для визначення відстані між двома кластерами можна використовувати об'єкти у різних кластерах, що знаходяться на найбільшій відстані один від одного серед всіх інших пар об'єктів. Відповідний метод називається **метод повного зв'язку (метод найбільш віддалених сусідів)**. Суть методу полягає у тому, що відстані між кластерами визначаються найбільшою відстанню між будь-якими двома об'єктами в різних кластерах (тобто “найбільш віддаленими сусідами”). Даний метод дуже добре працює, коли кластери нагадують розрізнені “плями”. Якщо ж кластери геометрично мають подовжену форму чи їх природний тип є “ланцюжковим”, то цей метод непридатний.

Незважене попарне середнє. Відстань між двома різними кластерами обчислюється як середня відстань між усіма парами об'єктів у них. Метод ефективний, коли об'єкти геометрично належать розрізним “плямам”, при цьому він працює однаково добре і у випадках кластерів “ланцюжкового” типу. В англійській літературі використовується аббревіатура UPGMA – метод незваженого попарного арифметичного середнього – unweighted pair-group method using arithmetic averages.

Зважене попарне середнє. Метод аналогічний методу незваженого попарного середнього, за винятком того, що при обчисленнях розмір відповідних кластерів (тобто число об'єктів, що містяться в них) використовується в якості вагового коефіцієнта. Тому пропонується методом варто користуватися, коли передбачаються нерівні розміри кластерів. Аббревіатура WPGMA – метод зваженого попарного арифметичного середнього – weighted pair-group method using arithmetic averages.

Незважений центроїдний метод. Відстань між двома кластерами визначається як відстань між їхніми центрами ваги. Аббревіатура UPGMC – метод незваженого попарного центроїдного усереднення – unweighted pair-group method using the centroid average.

Зважений центроїдний метод (медіана). Даний метод аналогічний попередньому, за винятком того, що при обчисленнях використовуються ваги для врахування різниці між розмірами кластерів. Тому даним методом доцільніше користуватися, якщо є значні відмінності у розмірах кластерів. WPGMC – метод зваженого попарного центроїдного усереднення – weighted pair-group method using the centroid average.

Метод Уорда відрізняється від всіх інших методів, оскільки він використовує методи дисперсійного аналізу для оцінки відстаней між кластерами. Він полягає в тому, що при переході від одного розбиття до іншого об'єднуються такі два кластери, що відбувається мінімальне збільшення загальної втрати інформації. За втрату інформації для однієї групи беруть, звичайно, середньоквадратичне відхилення, а для кількох груп – суму всіх групових відхилень. Тобто, даний метод мінімізує суму квадратів відхилень для будь-яких двох (гіпотетичних) кластерів, що можуть бути утворені на кожному кроці кластеризації. В цілому метод є дуже ефективним, хоча він прагне створювати кластери малого розміру.

5.5. Кластерний аналіз у програмі STATISTICA

Для прикладу відкриємо файл Cars.sta, який знаходиться у Examples/Datasets. У ньому міститься інформація про такі характеристики автомобілів різних виробників: ціна, прискорення, надійність, легкість управління, споживання пального.

	Performance, fuel economy, and approximate price for various automobiles				
	1 PRICE	2 ACCELERATION	3 BRAKING	4 HANDLING	5 MILEAGE
Acura	-0,521	0,477	-0,007	0,382	2,079
Audi	0,866	0,208	0,319	-0,091	-0,677
BMW	0,496	-0,802	0,192	-0,091	-0,154
Buick	-0,614	1,689	0,933	-0,210	-0,154
Corvette	1,235	-1,811	-0,494	0,973	-0,677
Chrysler	-0,614	0,073	0,427	-0,210	-0,154
Dodge	-0,706	-0,196	0,481	0,145	-0,154
Eagle	-0,614	1,218	-4,199	-0,210	-0,677
Ford	-0,706	-1,542	0,987	0,145	-1,724
Honda	-0,429	0,410	-0,007	0,027	0,369
Isuzu	-0,798	0,410	-0,061	-4,230	1,067
Mazda	0,126	0,679	-0,133	0,500	-1,724
Mercedes	1,051	0,006	0,120	-0,091	-0,154
Mitsub.	-0,614	-1,003	0,084	0,382	0,718
Nissan	-0,429	0,073	-0,007	0,263	0,997
Olds	-0,614	-0,734	0,409	0,382	2,114
Pontiac	-0,614	0,679	0,536	0,145	0,195
Porsche	3,454	-2,215	-0,296	0,618	-1,026
Saab	0,588	0,679	0,246	0,263	0,021
Toyota	-0,059	1,218	0,228	0,736	-0,851
VW	-0,706	-0,128	0,102	0,382	0,195
Volvo	0,219	0,612	0,138	-0,210	0,369

Рис. 5.2

У наведеній таблиці (див. рис. 5.2) у відповідних стовпчиках записані стандартизовані дані, які отримали за допомогою віднімання від кожного елемента стовпчика середнього значення (по стовпчику) та ділення на середнє квадратичне відхилення. Порівнюються саме безрозмірні стандартизовані дані для різних показників, що не можна робити, наприклад, для ціни автомобіля у певних грошових одиницях та прискорення в м/с^2 .

Виконаємо *Анализ* → *Многомерный разведочный анализ* → *Кластерный анализ* (див. рис. 5.3). Далі з'явиться вікно, у якому можна обрати різні методи кластерного аналізу (див. рис. 5.4).

Виберемо *Иерархическая классификация* і натиснемо *OK*. У вікні, яке з'явилося, на закладці *Дополнительно* (див. рис. 5.5) виберемо всі змінні (*Переменные*) і решту опцій, як показано на даному рисунку.

Якщо натиснути *OK*, то з'явиться вікно, за допомогою якого можна продивитися результати аналізу (див. рис. 5.6).

Якщо натиснути *Горизонтальная дендрограмма*, то побачимо ієрархічне дерево, зображене на рис. 5.1.

Для того, щоб побачити, як по кроках відбувалося об'єднання в кластери та при якому значенні відстані, у вікні *Результаты иерархической классификации* натиснемо кнопку *Схема объединения* і отримаємо таблицю (див. рис. 5.7) або відповідний графік (див. рис. 5.8), натиснувши *График схемы объединения*.

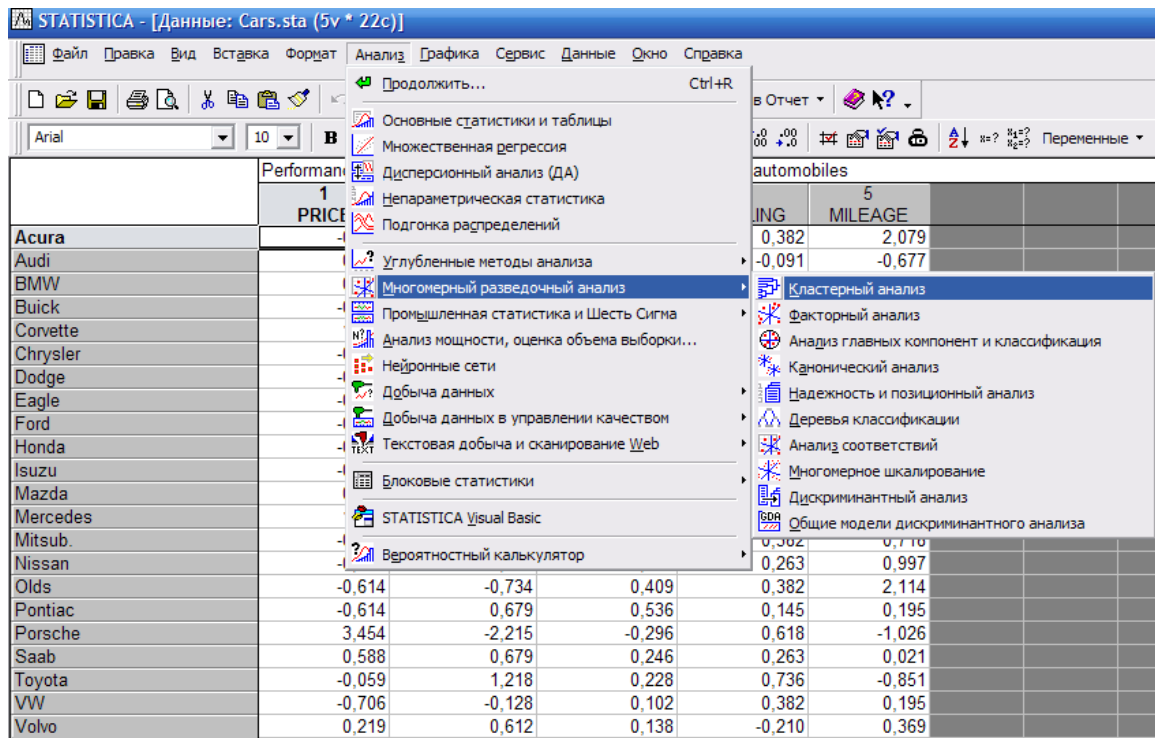


Рис. 5.3

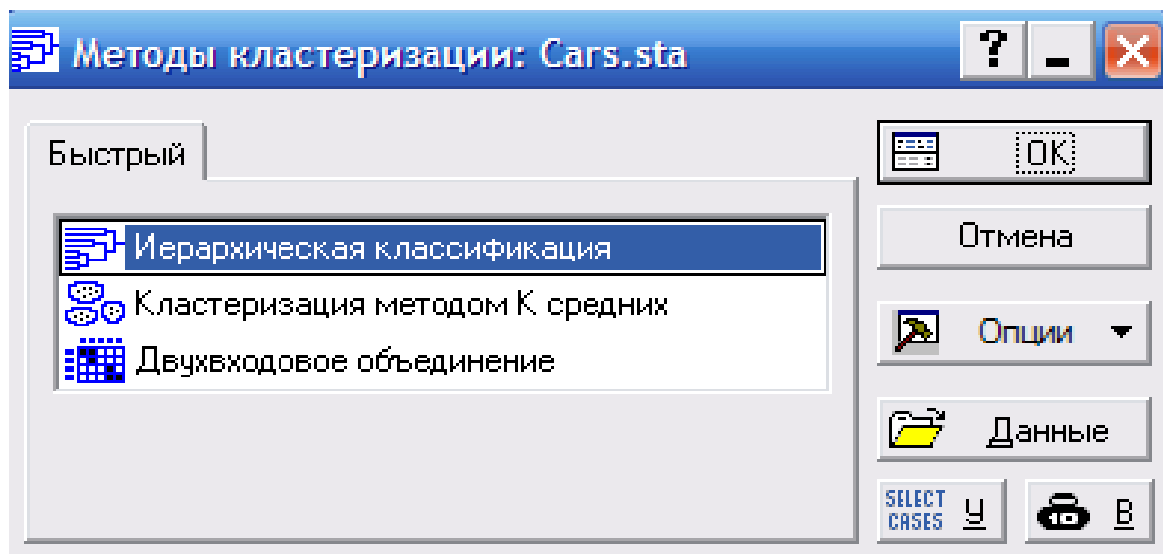


Рис. 5.4

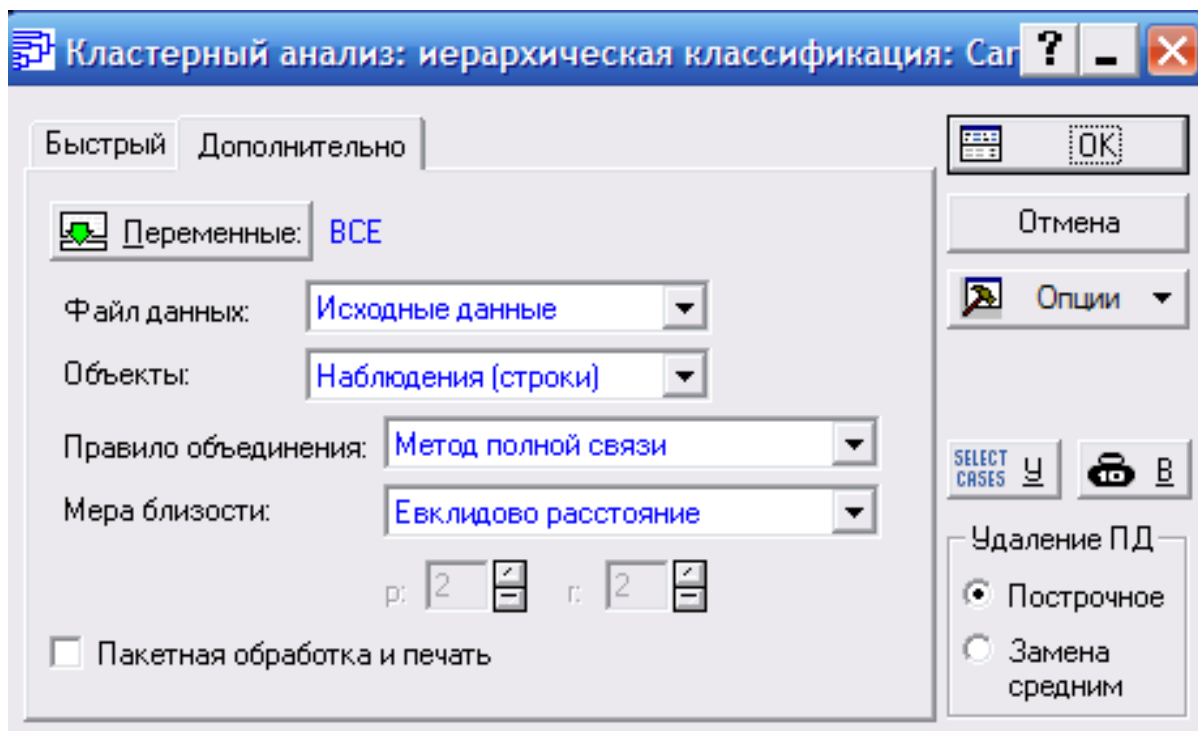


Рис. 5.5

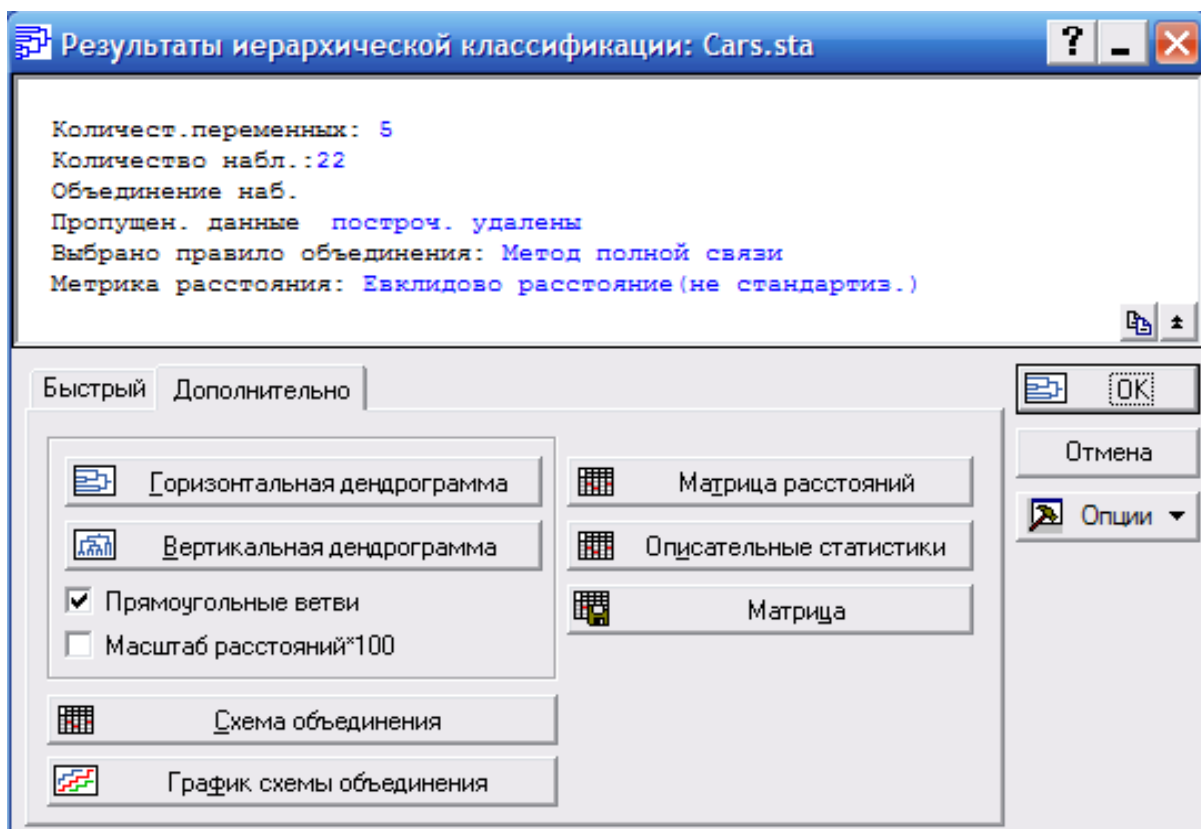


Рис. 5.6

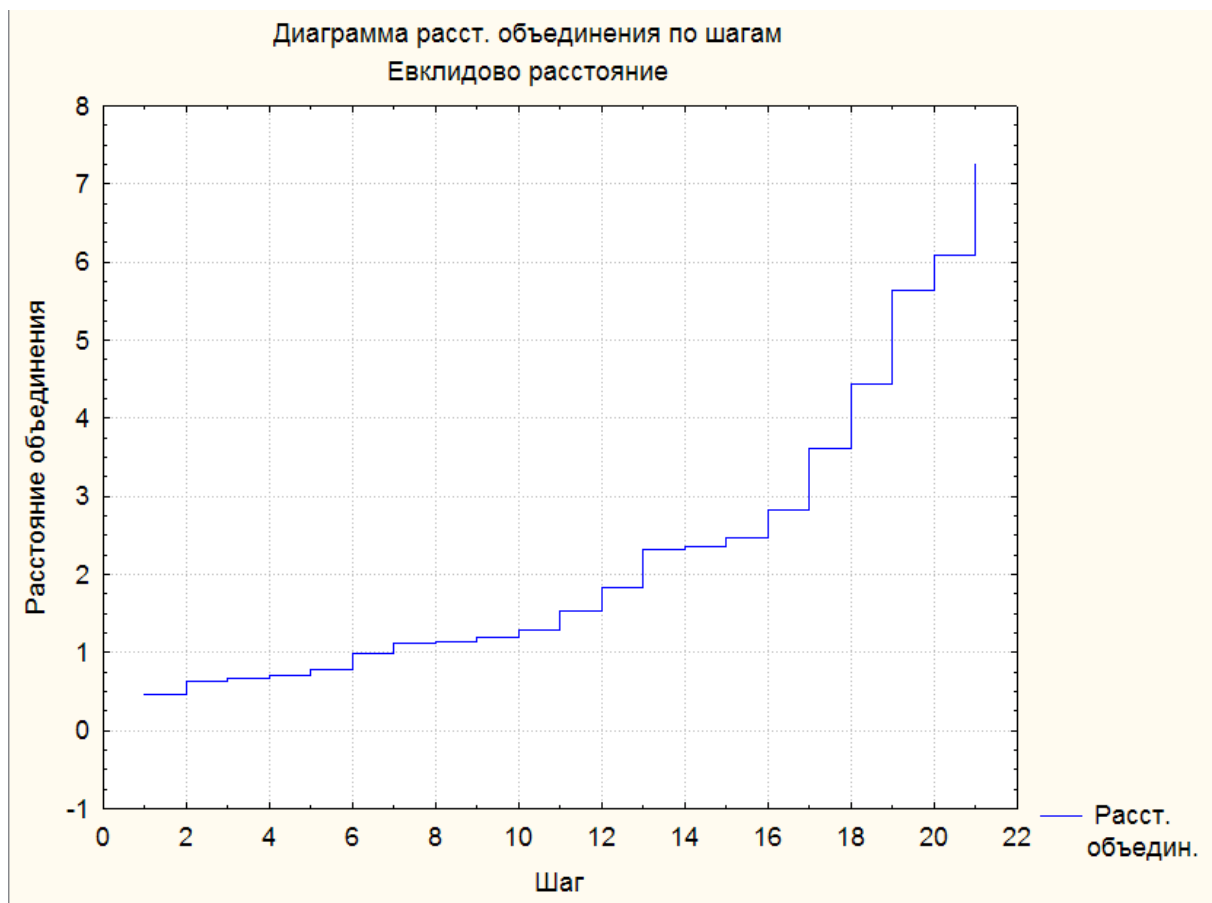


Рис. 5.8

Щоб дендрограма була вертикальною у вікні *Результаты иерархической классификации* потрібно натиснути кнопку *Вертикальная дендрограмма*, а якщо при цьому хочемо бачити відносні відстані об'єднання в кластери, то у вказаному вікні вибираємо ще й опцію *Масштаб расстояний*100* (див. рис. 5.9).

При вивченні рис. 5.9 можна прийти до висновку, що множина розглядуваних об'єктів поділилась на 3 або 4 кластери, які об'єднувались між собою при досить великих відстанях на останніх кроках кластеризації.

Для інтерпретації особливостей утворених груп повернемося до вікна *Методы кластеризации* (див. рис. 5.4), у якому виберемо *Кластеризация методом K средних* і натиснемо *OK*. У вікні, яке з'явилося, на закладці *Дополнительно* (див. рис. 5.10) виберемо всі змінні (*Переменные*) і решту опцій, як показано на даному рисунку:

Після натискання *OK* з'явиться вікно (див. рис. 5.11), у якому варто вибрати закладку *Дополнительно* → *График средних*.

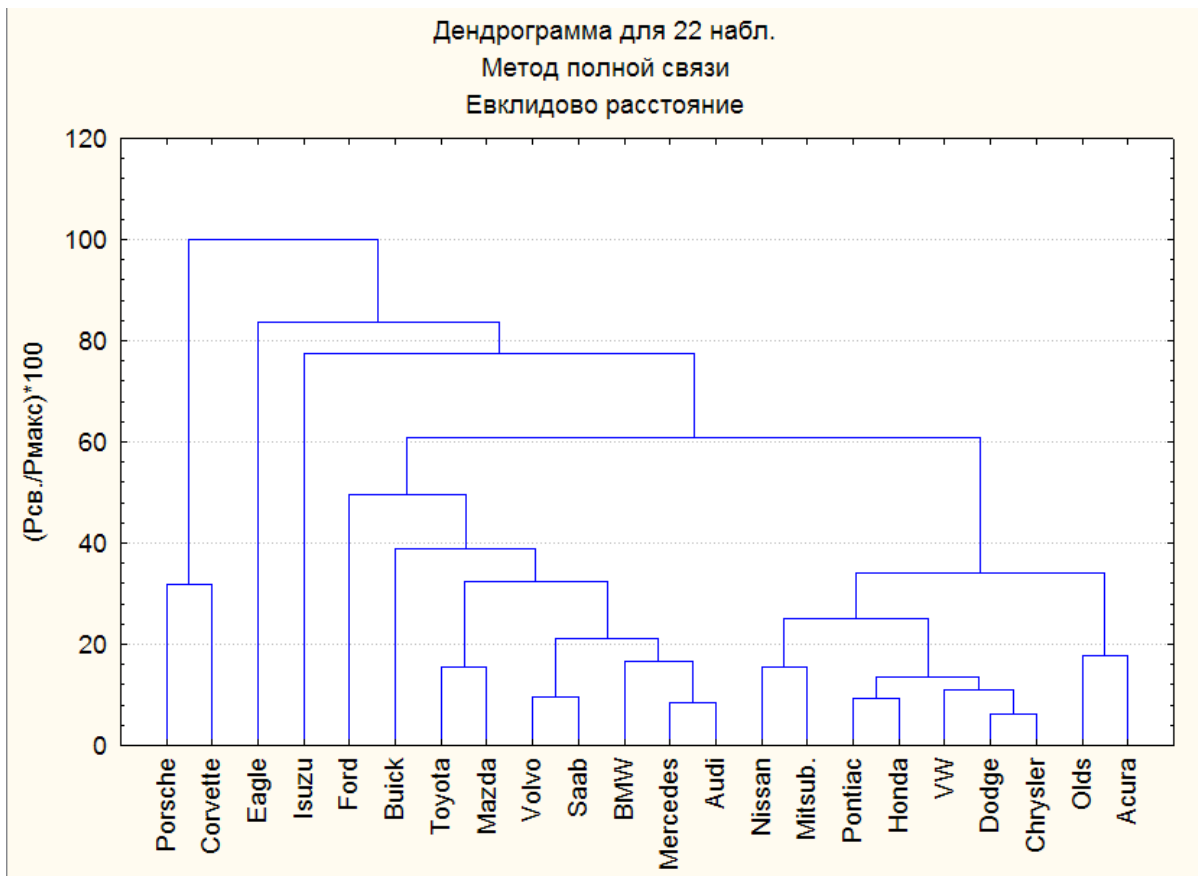


Рис. 5.9

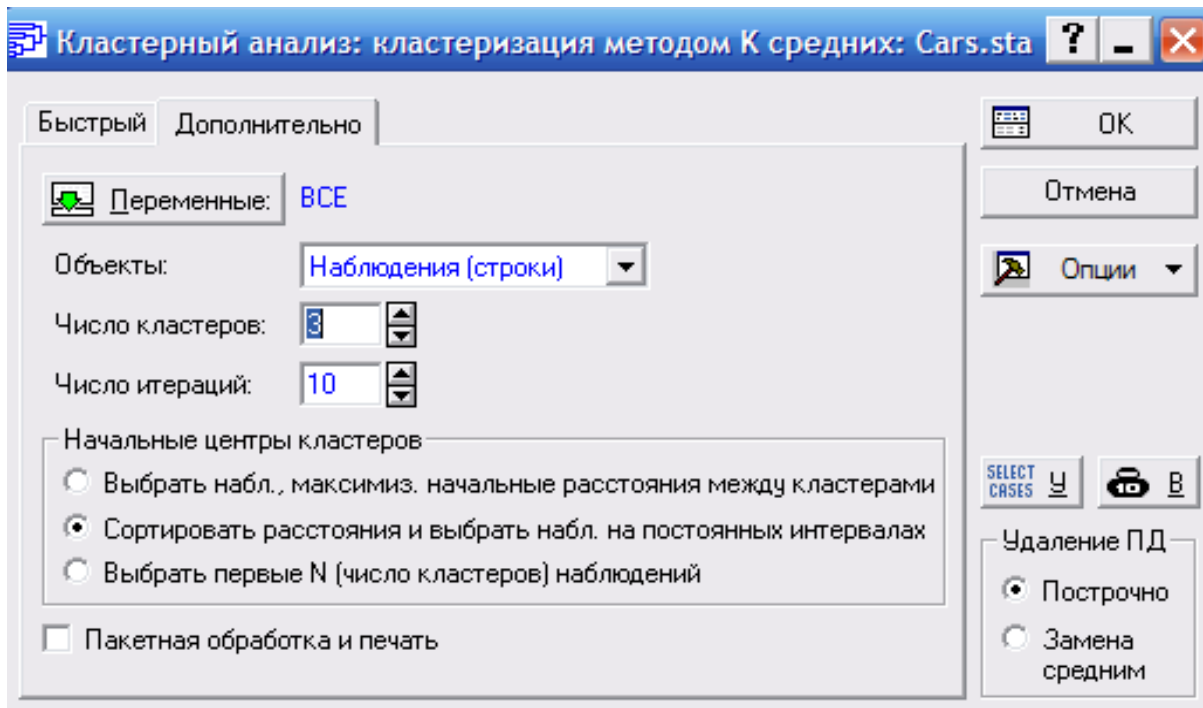


Рис. 5.10

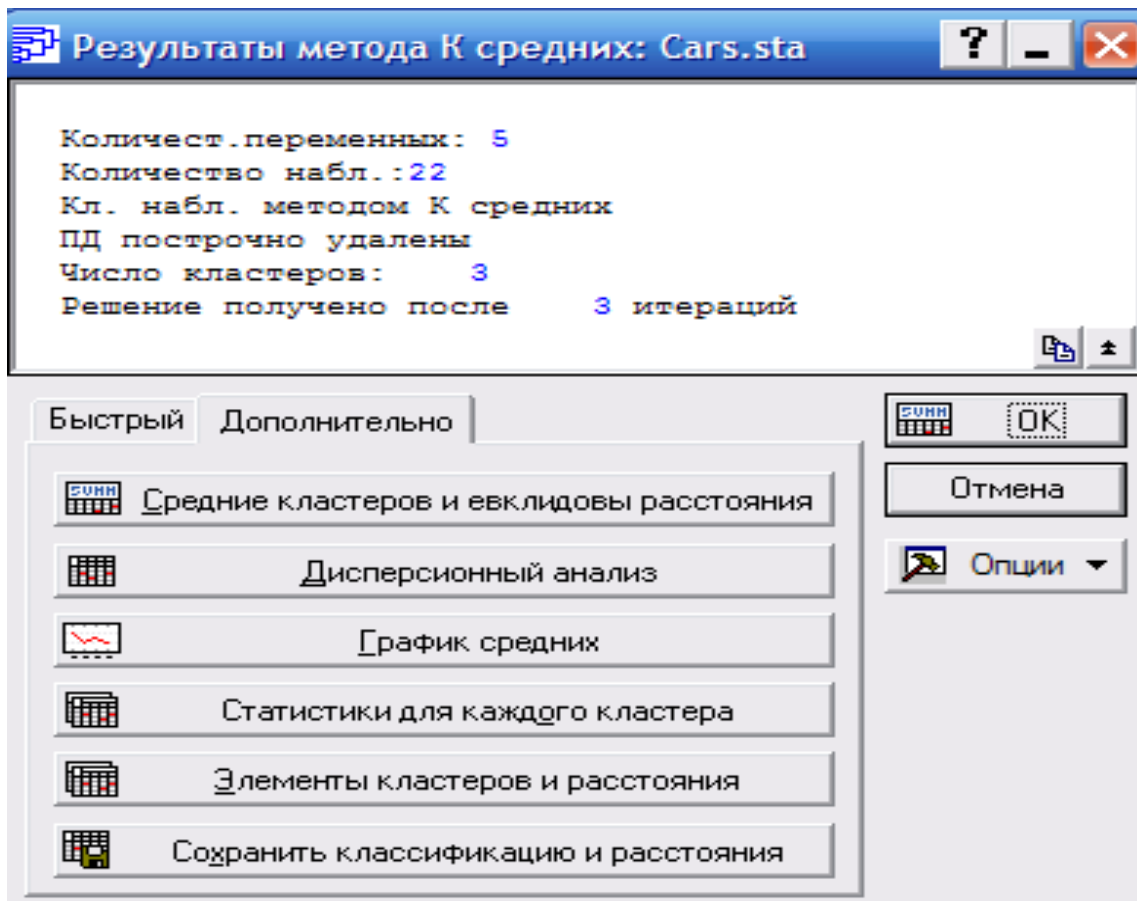


Рис. 5.11

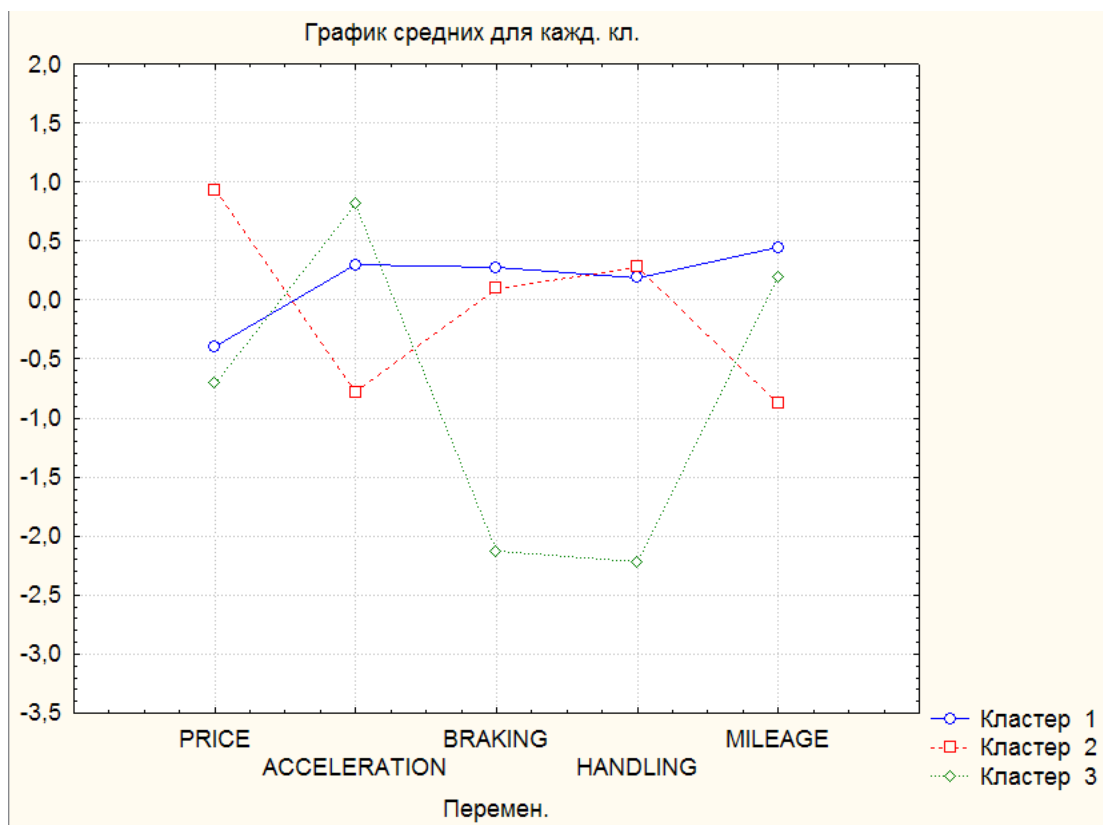


Рис. 5.12

На рис. 5.12 бачимо, що до третього кластеру потрапили виробники недорогих ненадійних автомобілів, які споживають багато пального, мають великий показник прискорення і якими важко керувати. Другий кластер характеризується високою ціною автомобілів, невеликим прискоренням, надійністю, зручністю керування та економним споживанням пального. До першого кластеру потрапили виробники надійних, економних автомобілів з рештою середніх показників.

Якщо на вікні, зображеному на рис. 5.11 натиснути *Элементы кластеров и расстояния*, то отримаємо таку інформацію по кожному кластеру:

Элементы кластера номер 1 (Cars.sta) и расстояния до центра кластера. Кластер содержит 13 набл.													
	Acura	Buick	Chrysler	Dodge	Honda	Mitsub.	Nissan	Olds	Pontiac	Saab	Toyota	VW	Volvo
Расст.	0,754166	0,766466	0,356816	0,384616	0,158199	0,614239	0,297823	0,889882	0,255611	0,508612	0,766000	0,284704	0,362700

Элементы кластера номер 2 (Cars.sta) и расстояния до центра кластера. Кластер содержит 7 набл.							
	Audi	BMW	Corvette	Ford	Mazda	Mercedes	Porsche
Расст.	0,492166	0,414537	0,636028	0,978105	0,849100	0,509201	1,319699

Элементы кластера номер 3 (Cars.sta) и расстояния до центра кластера. Кластер содержит 2 набл.				
	Eagle	Isuzu		
Расст.	1,360452	1,360452		

Рис. 5.13

Контрольні питання для самоперевірки до теми „Кластерний аналіз”

1. На якій стадії дослідження і з якою метою використовують кластерний аналіз?
2. У чому полягає суть методу деревоподібної кластеризації?
3. Чим відрізняються агломеративні та подрібнюючі методи кластеризації?
4. Як вимірюються відстані між об'єктами?
5. Як вимірюються відстані між кластерами?
6. Які особливості кластеризації у програмі STATISTICA?

Глава 6. Елементи аналізу часових рядів

6.1. Поняття часового ряду

Під **часовим рядом** будемо розуміти послідовність вимірів, зроблених у невідповідних послідовних моментах часу. На відміну від аналізу випадкових вибірок, аналіз часових рядів ґрунтується на припущенні, що послідовні значення у файлі даних спостерігаються через рівні проміжки часу.

Існують дві основні задачі аналізу часових рядів:

1) визначення природи часового ряду для побудови адекватної математичної моделі;

2) прогнозування (передбачення майбутніх значень часового ряду по сьогоденним і минулим значенням).

Як тільки модель часового ряду визначено, з її допомогою можна інтерпретувати розглянуті дані (наприклад, використовувати у економічній теорії для розуміння сезонної зміни цін на товари) або екстраполювати часовий ряд на основі знайденої моделі, тобто передбачати його майбутні значення.

Як і більшість інших видів аналізу, аналіз часових рядів припускає, що дані містять систематичну складову (яка зазвичай включає кілька компонент) і випадковий шум (помилку), що утруднює виділення регулярних компонент. Більшість методів дослідження часових рядів включає різні способи фільтрації шуму, що дозволяють побачити регулярну складову більш чітко.

Регулярні складові часових рядів належить до двох класів: вони є або детермінованою складовою, яка в економічних, географічних, біологічних, соціологічних та інших застосуваннях називається трендом, або сезонною складовою. **Тренд** є систематичною лінійною чи нелінійною детермінованою складовою, що може змінюватися в часі і визначає основну довгострокову тенденцію, яка властива розглядуваному часовому ряду. **Сезонна складова** – це періодично повторювана компонента часового ряду. Обидва ці види регулярних компонент часто присутні в часовому ряді одночасно. Наприклад, обсяги продажів певної фірми можуть зростати щорічно, але вони також містять сезонну складову (як правило, 25% річних продажів припадають на грудень і тільки 4% на серпень).

Розглянемо часовий ряд, що містить інформацію про місячні міжнародні авіап перевезення (у тисячах) протягом 12 років з 1949 року по 1960 рік певної авіакомпанії (див. рис. 6.1). Графік місячних перевезень показує майже лінійний тренд, тобто є щорічне зростання перевезень (приблизно в 4 рази більше пасажирів перевезено в 1960 році, чим у 1949). При цьому характер місячних перевезень повторюється, вони мають майже той самий характер у кожному річному періоді (наприклад, перевезень

більше у відпускні періоди, ніж в інші місяці). Цей часовий ряд є прикладом часового ряду, у якому амплітуда сезонних змін збільшується разом із трендом. Моделі таких часових рядів називаються **моделями з мультиплікативною сезонністю**.

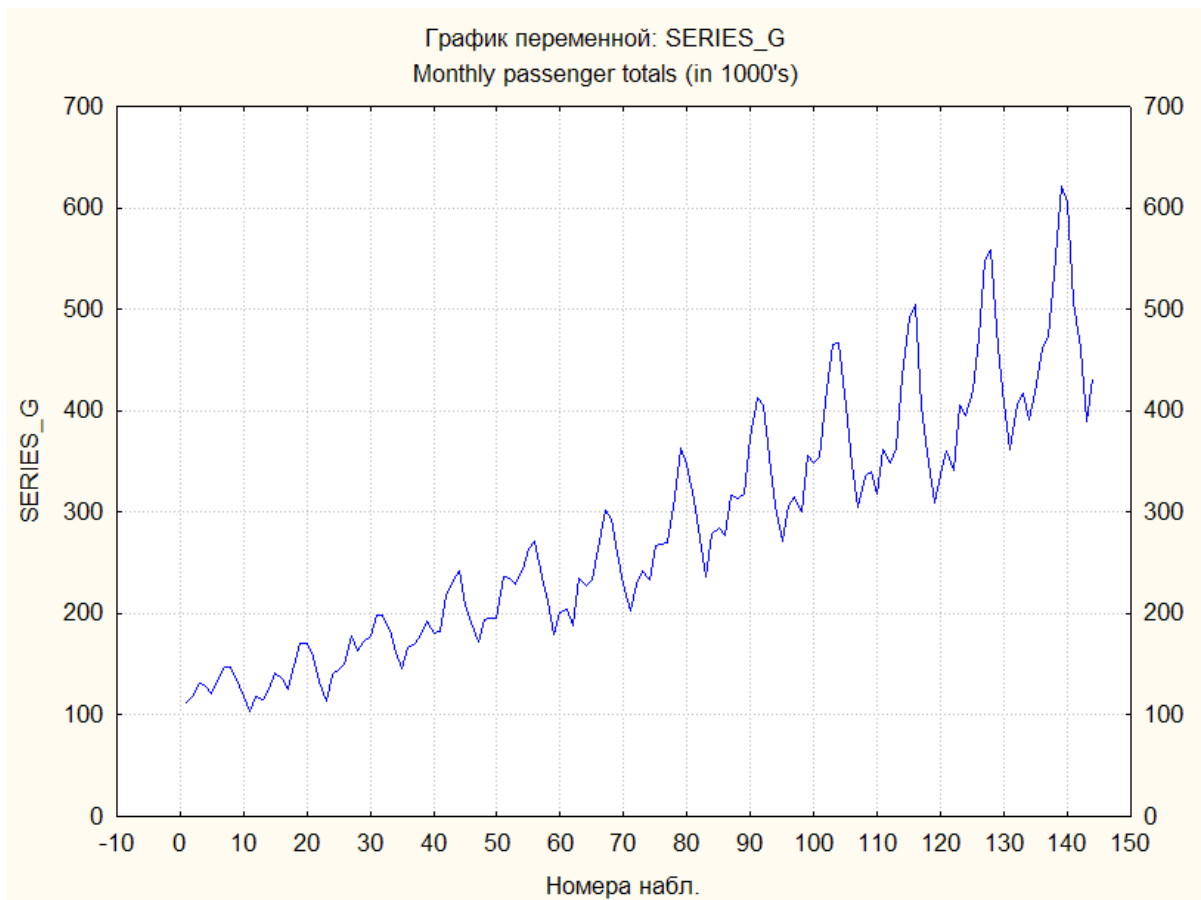


Рис. 6.1

6.2. Виділення компонент часового ряду

Якщо часові ряди містять значну помилку, то першим кроком виділення тренда є **згладжування**. Згладжування завжди включає деякий спосіб локального усереднення даних, при якому несистематичні компоненти взаємно нейтралізуються. Самий загальний метод згладжування – згладжування, що використовує **ковзне середнє**, при якому кожен член ряду замінюється простим чи зваженим середнім n сусідніх членів, де n – ширина “вікна”.

Замість середнього можна використовувати медіану значень, що потрапили у вікно. Основна перевага **медіанного згладжування** у порівнянні зі згладжуванням ковзним середнім полягає в тому, що результати стають більш стійкими до викидів (аномальних даних), якщо вони є усередині вікна. Таким чином, якщо в даних є викиди (зв'язані, наприклад, з помилками вимірів), то згладжування медіаною, звичайно,

приводить до більш гладких кривих у порівнянні з ковзним середнім з вікном такої самої ширини. Основний недолік медіанного згладжування в тому, що при відсутності явних викидів, воно приводить до більш “зубцюватих” кривих (чим згладжування ковзним середнім) і не дозволяє використовувати ваги.

Відносно рідше, коли помилки вимірів дуже великі, використовується згладжування методом найменших квадратів або методом експоненційно зваженого згладжування. Усі ці методи відфільтровують шум і перетворюють дані на такі, що графічно зображуються у вигляді відносно гладкої кривої. Ряди з відносно невеликою кількістю спостережень і систематичним розташуванням зображуваних точок можуть бути згладжені за допомогою бікубічних сплайнів.

Багато монотонних часових рядів можна добре наблизити лінійною функцією. Якщо ж у часовому ряді присутня явно монотонна нелінійна складова, то дані спочатку варто перетворити, щоб усунути нелінійність. Звичайно, для цього використовують логарифмічне, експоненційне чи поліноміальне перетворення даних.

Періодична і сезонна складові (сезонність) є іншим типом компонент часового ряду. Це поняття було проілюстроване на прикладі часового ряду, пов'язаного із авіаперевезеннями пасажирів. На графіку часового ряду можна легко бачити, що кожне спостереження близьке до сусіднього. При цьому розглядувані дані мають повторювану сезонну складову, тобто кожне спостереження подібне до спостереження, зробленого у тому самому місяці рік назад. Загалом, періодична залежність може бути формально визначена як кореляційна залежність порядку k між кожним i -м елементом ряду і $(i-k)$ -м елементом. Її можна виміряти за допомогою **автокореляції** (тобто кореляції між самими членами ряду). Число k називають **лагом**. Якщо помилка вимірів не занадто велика, то сезонність можна визначити візуально, розглядаючи поведінку членів часового ряду через кожні k часових одиниць.

Сезонні складові часового ряду можна знаходити за допомогою корелограми. **Корелограма (автокорелограма)** показує чисельно і графічно автокореляційну функцію (АКФ), іншими словами, коефіцієнти автокореляції (і їхні стандартні помилки) для послідовності лагів з визначеного діапазону (наприклад, від 1 до 30). На корелограмі, як правило, відзначається діапазон у розмірі двох стандартних помилок для кожного лагу.

При вивченні корелограм варто пам'ятати, що автокореляції послідовних лагів формально залежні між собою. Наведемо приклад. Якщо перший член ряду тісно зв'язаний із другим, а другий із третім, то перший елемент повинен також певним чином залежати від третього і т.д. Це приводить до того, що періодична залежність може істотно змінитися після видалення автокореляцій першого порядку, тобто після взяття різниці з лагом 1.

Інший метод дослідження періодичності полягає у дослідженні **частинної автокореляційної функції (ЧАКФ)**, що є узагальненням звичайної автокореляційної функції. У ЧАКФ усувається залежність між проміжними спостереженнями (спостереженнями усередині лага). Іншими словами, частинна автокореляція на даному лагу аналогічна звичайній автокореляції без врахування впливу автокореляцій з меншими лагами. На лагу 1 (коли немає проміжних елементів усередині лага), частинна автокореляція дорівнює, очевидно, звичайній автокореляції. Насправді, частинна автокореляція дає більш “чисту” картину періодичних залежностей.

Як зазначалося вище, періодична складова часового ряду для даного лага k може бути вилучена взяттям різниці відповідного порядку (застосуванням відповідного різницевого оператора). Це означає, що від кожного i -го елемента часового ряду віднімається $(i-k)$ -й елемент. На користь таких перетворень є кілька аргументів. По-перше, у такий спосіб можна визначити сховані періодичні складові часового ряду. Нагадаємо, що автокореляції на послідовних лагах залежні. Тому видалення деяких автокореляцій змінить інші автокореляції, і зробить деякі інші сезонні складові більш помітними. По-друге, видалення сезонних складових робить ряд стаціонарним (це означає, що його середнє постійне, а вибіркові дисперсія й автокореляція не змінюються з часом), що необхідно для застосування інших методів аналізу часових рядів.

6.3. Побудова моделі часового ряду

У реальних даних часто немає чітко виражених регулярних складових. Окремі спостереження містять значні помилки, що істотно утруднює виділення регулярних компонент та побудову математичної моделі часового ряду, яка давала б можливість робити адекватні прогнози досліджуваного явища.

Далі познайомимося з основними ідеями методу, розробленого Боксом і Дженкінсом (1976), що дозволяє розв’язати поставлену задачу. Даний метод надзвичайно популярний у багатьох застосуваннях і практика підтвердила його потужність і гнучкість. Потрібний певний досвід використання цього методу, щоб отримувати задовільні результати (вони залежать від кваліфікації користувача), які б узгоджувались з практикою.

Розглянемо два основних процеси.

1) **Процес авторегресії**. Більшість часових рядів містять елементи, що послідовно залежать один від іншого. Таку залежність можна виразити наступним рівнянням:

$$x_t = \xi + \phi_1 \cdot x_{t-1} + \phi_2 \cdot x_{t-2} + \phi_3 \cdot x_{t-3} + \dots + \varepsilon,$$

де ε – константа (вільний член), $\phi_1, \phi_2, \phi_3, \dots$ – параметри авторегресії, ξ – випадкова величина, що характеризує випадковий вплив на спостереження.

Кожне спостереження є сумою випадкової компоненти і лінійної комбінації попередніх спостережень. Процес авторегресії буде стаціонарним тільки тоді, коли його параметри лежать у визначеному діапазоні. Наприклад, якщо є тільки один параметр, то він повинен знаходитися в інтервалі $-1 < \phi < 1$. Інакше попередні значення будуть накопичуватися і значення наступних спостережень x_t можуть бути необмеженими, а отже, ряд не буде стаціонарним. Якщо є кілька параметрів авторегресії, то можна визначити аналогічні умови, що забезпечують стаціонарність.

2) **Процес ковзного середнього.** На відміну від процесу авторегресії, у процесі ковзного середнього кожен елемент ряду піддається сумарному впливу попередніх помилок. У загальному вигляді це можна записати в такий спосіб:

$$x_t = \mu + \xi_t + \theta_1 \cdot \xi_{t-1} + \theta_2 \cdot \xi_{t-2} + \theta_3 \cdot \xi_{t-3} + \dots,$$

де μ – константа, $\theta_1, \theta_2, \theta_3$ – параметри ковзного середнього. Іншими словами, поточне спостереження ряду є сумою випадкової компоненти (випадкового впливу) у даний момент часу і лінійної комбінації випадкових впливів у попередні моменти часу.

Не вдаючись у деталі, відзначимо, що існує “двоїстість” між процесами ковзного середнього і авторегресії. Це означає, що наведене вище рівняння ковзного середнього можна переписати (згорнути) у вигляді рівняння авторегресії (необмеженого порядку) і навпаки. Це так звана властивість оборотності. Є умови, аналогічні приведеним вище умовам стаціонарності, що забезпечують оборотність моделі.

Загальна модель авторегресії і ковзного середнього (модель АРПКС), запропонована Боксом і Дженкінсом включає як параметри авторегресії, так і параметри ковзного середнього. Отже, є три типи параметрів моделі: p – кількість параметрів авторегресії, d – порядок різниці, q – кількість параметрів ковзного середнього. У позначеннях Бокса і Дженкінса модель записується як АРПКС(p, d, q). Наприклад, модель (0,1,2) містить 0 (нуль) параметрів авторегресії (p) і 2 параметри ковзного середнього (q), що обчислюються для ряду після взяття різниці з лагом 1.

6.4. Ідентифікація порядку моделі часового ряду

Як відзначено вище, для моделі АРПКС необхідно, щоб ряд був стаціонарним. Видалення сезонних складових шляхом взяття різниці відповідного порядку робить ряд стаціонарним. Тому необхідно брати різниці ряду до тих пір, поки він не стане стаціонарним (часто також застосовують логарифмічне перетворення для стабілізації дисперсії). Число різниць, що були узяті, щоб досягти стаціонарності, визначаються параметром d .

Для того, щоб визначити необхідний порядок різниці, потрібно досліджувати графік ряду й автокорелограму. Сильні зміни рівня (сильні

стрибки вгору чи вниз), як правило, вимагають взяття несезонної різниці першого порядку (лаг = 1). Сильні зміни нахилу траєкторії часового ряду вимагають взяття різниці другого порядку. Сезонна складова вимагає взяття відповідної сезонної різниці (див. нижче). Якщо є повільне спадання вибірових коефіцієнтів автокореляції в залежності від лага, тоді також беруть різницю першого порядку.

Варто пам'ятати, що для деяких часових рядів потрібно брати різницю невеликого порядку чи зовсім не брати її. Надмірна кількість узятих різниць приводить до погіршення оцінок коефіцієнтів.

На цьому етапі (який, звичайно, називають **ідентифікацією порядку моделі**) також потрібно вирішити, як багато параметрів авторегресії (p) і ковзного середнього (q) повинно бути присутніми в ефективній і мінімальній моделі розглядуваного процесу. (Мінімальність моделі означає, що в ній є найменше число параметрів серед усіх моделей, що підганяються до даних). На практиці дуже рідко буває, що число параметрів p чи q більше 2.

Вибір виду моделі АРПКС не є простим і потрібно ґрунтовно проекспериментувати з альтернативними моделями. Проте, більшість часових рядів, що зустрічаються на практиці, можна досить точно апроксимувати однією з 5 основних моделей:

1) $p=1$: АКФ – експоненційно спадає; ЧАКФ – має значення, що різко виділяється, для лага 1 і не має кореляцій на інших лагах.

2) $p=2$: АКФ має форму синусоїди чи експоненційно спадає; ЧАКФ має великі значення на лагах 1, 2 і не має кореляцій на інших лагах.

3) $q=1$: АКФ має значення, що різко виділяється, для лага 1 і не має кореляцій на інших лагах; ЧАКФ – експоненційно спадає.

4) $q=2$: має великі значення на лагах 1, 2 і не має кореляцій на інших лагах; ЧАКФ має форму синусоїди чи експоненційно спадає.

5) $p=1$ і $q=1$: АКФ – експоненційно спадає з лага 1; ЧАКФ – експоненційно спадає з лага 1.

Зауважимо, що число параметрів кожного виду невелике, тому неважко поекспериментувати, перевіривши альтернативні моделі.

6.5. Оцінювання і прогноз

Наступний, після ідентифікації, крок аналізу часового ряду полягає в оцінюванні параметрів моделі, для чого в пакеті STATISTICA застосовуються процедури мінімізації функції втрат. Отримані оцінки параметрів використовують на останньому етапі прогнозування для того, щоб обчислити нові значення ряду і побудувати довірчий інтервал для прогнозу. Процес оцінювання проводиться по перетвореним даним (підданим застосуванню різницевого оператора). До побудови прогнозу потрібно виконати зворотну операцію (інтегрувати дані). Таким чином, прогноз методу АРПКС буде порівнюватися з відповідними вихідними

даними. На інтегрування даних вказує літера Π в загальній назві моделі (АРПКС = АвтоРегресійне Проінтегроване Ковзне Середнє).

Додатково моделі АРПКС можуть містити константу, інтерпретація якої залежить від моделі, що підганяється. Якщо у моделі немає параметрів авторегресії, то константа є середнім значенням часового ряду, якщо параметри авторегресії є, то константа буде вільним членом. Якщо бралася різниця ряду, то константа є середнім чи вільним членом перетвореного ряду. Наприклад, якщо бралася перша різниця (різниця першого порядку), а параметрів авторегресії в моделі немає, то константа є середнім значенням перетвореного ряду i , отже, кутовим коефіцієнтом лінійного тренда.

Якісна модель часового ряду повинна не тільки давати досить точний прогноз, який би узгоджувався з декількома останніми спостереженнями, але й бути мінімальною і мати незалежні залишки, що містять тільки шум без систематичних компонентів (зокрема, АКФ залишків не повинна мати якої-небудь періодичності). Тому необхідний всебічний аналіз залишків.

Якщо залишки систематично розподілені (наприклад, від'ємні в першій частині ряду і приблизно рівні нулю в другій) чи містять деяку періодичну складову, то це свідчить про неадекватність моделі. Аналіз залишків надзвичайно важливий і необхідний при аналізі часових рядів. Процедура оцінювання припускає, що залишки некорельовані і нормально розподілені.

Варто нагадати, що модель АРПКС придатна тільки для рядів, що є стаціонарними (середнє, дисперсія й автокореляція приблизно постійні в часі); для нестаціонарних рядів варто брати різниці. Рекомендується мати, як мінімум, 50 спостережень у файлі вихідних даних. Також передбачається, що параметри моделі не змінюються в часі.

6.6. Сезонні моделі

Мультиплікативна сезонна модель АРПКС є природнім узагальненням звичайної моделі АРПКС на часові ряди, у яких присутня періодична сезонна складова. Поряд з несезонними параметрами у модель вводяться сезонні параметри для визначеного лага. Аналогічно параметрам простої моделі АРПКС(p,d,q), ці параметри називаються: сезонна авторегресія (ps), сезонна різниця (ds) і сезонне ковзне середнє (qs). Таким чином, повна сезонна АРПКС може бути записана як АРПКС(p,d,q)(ps,ds,qs). Наприклад, модель АРПКС(0,1,2)(0,1,1) включає 0 регулярних параметрів авторегресії, 2 регулярних параметри ковзного середнього і 1 параметр сезонного ковзного середнього. Ці параметри обчислюються для рядів, які одержані після взяття однієї різниці з лагом 1 і однієї сезонної різниці. Сезонний лаг, що використовується для сезонних параметрів, визначається на етапі ідентифікації порядку моделі.

Наприклад, часовий ряд, що містить інформацію про місячні міжнародні авіап перевезення (див. рис. 6.1), має сезонну компоненту з лагом 12 (кількість місяців у році).

Загальні рекомендації щодо вибору звичайних параметрів (за допомогою АКФ і ЧАКФ) можуть бути застосовані і до сезонних моделей. Основна відмінність полягає в тому, що в сезонних часових рядах АКФ і ЧАКФ мають істотні значення на лагах, кратних сезонному лагу. Ця особливість накладається і доповнює характерну поведінку вказаних функцій, що описують регулярну (несезонну) компоненту моделі АРПКС.

Оскільки модель АРПКС придатна тільки для рядів, що є стаціонарними, то у випадку часового ряду з нестаціонарною амплітудою коливань потрібно робити певні перетворення часового ряду, які б стабілізували амплітуду. З цією метою можна, наприклад, застосувати логарифмічне, експоненційне, степеневе чи інше перетворення даних залежно від характеру їх нестаціонарності.

6.7. Аналіз часового ряду у програмі STATISTICA

Відкриємо файл *Series_g.sta*, який знаходиться у *Examples/Datasets*. У ньому містяться дані про кількість (у тисячах) щомісячних перевезень пасажирів певною авіакомпанією на протязі 12 років.

Виконаємо *Анализ* → *Углубленные методы анализа* → *Временные ряды и прогнозирование* (див. рис. 6.2). В результаті з'явиться вікно, у якому автоматично буде вибрано єдину змінну із нашого файла з даними і показано різні доступні методи аналізу часових рядів (див. рис. 6.3).

Почнемо з побудови графіка даного часового ряду. Для цього у вікні, зображеному на рис. 6.3, натиснемо *OK (преобразования, авто- и кросскорреляции, графики)*, після чого потрапимо у вікно, зображене на рис. 6.4. На закладці *Графики* спочатку відмітимо опцію *График после каждого преобразования* для того, щоб подальші перетворення часового ряду відображались графічно. Потім натискаємо на *График* біля *Просмотр выдел. переменной* і отримаємо графік досліджуваного часового ряду (див. рис. 6.1).

Очевидно, що наш часовий ряд не є стаціонарним, тому зробимо певні перетворення, щоб для перетворених даних можна було застосувати модель АРПКС.

Спочатку позбудемося зростання амплітуди коливань. Для цього прологарифмуємо часовий ряд. У вікні, зображеному на рис. 6.4, на закладці $x = f(x)$ виберемо опцію *Натуральный логарифм ($x = \ln(x)$)*, натиснемо *OK (Преобразовать выделенную переменную)*. Отриманий перетворений часовий ряд зображено на рис. 6.5.

Бачимо, що логарифмування збільшило амплітуду коливань для початкових спостережень (там, де вона була малою) і зменшило для кінцевих періодів (там, де вона була порівняно великою). Тобто, ми

отримали часовий ряд з майже постійною амплітудою періодичної КОМПОНЕНТИ.

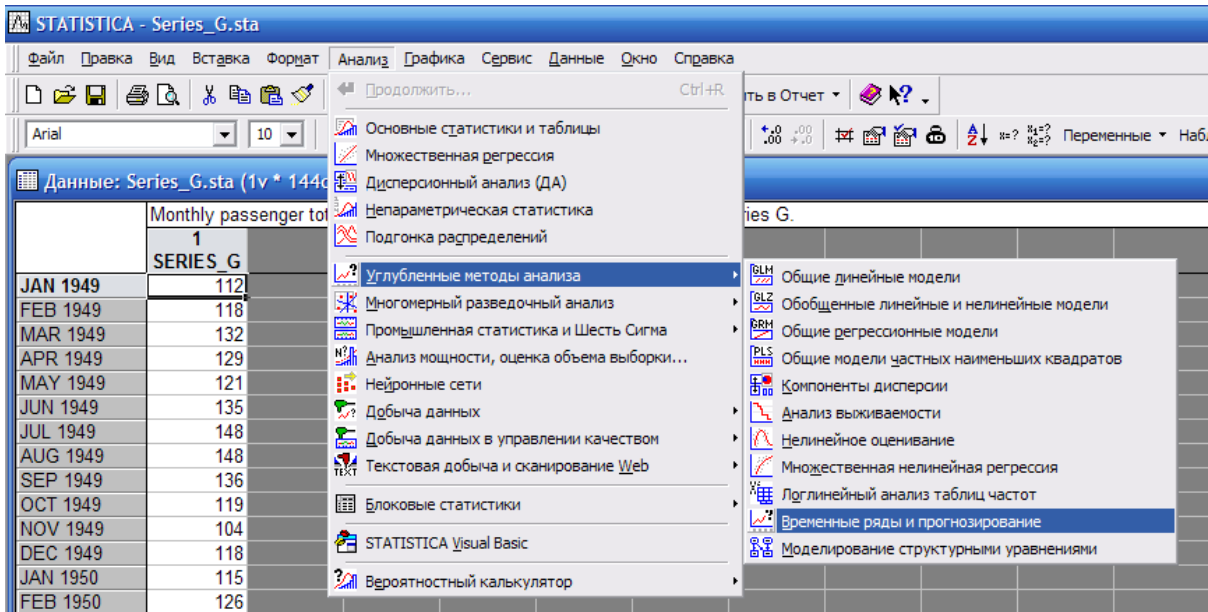


Рис. 6.2

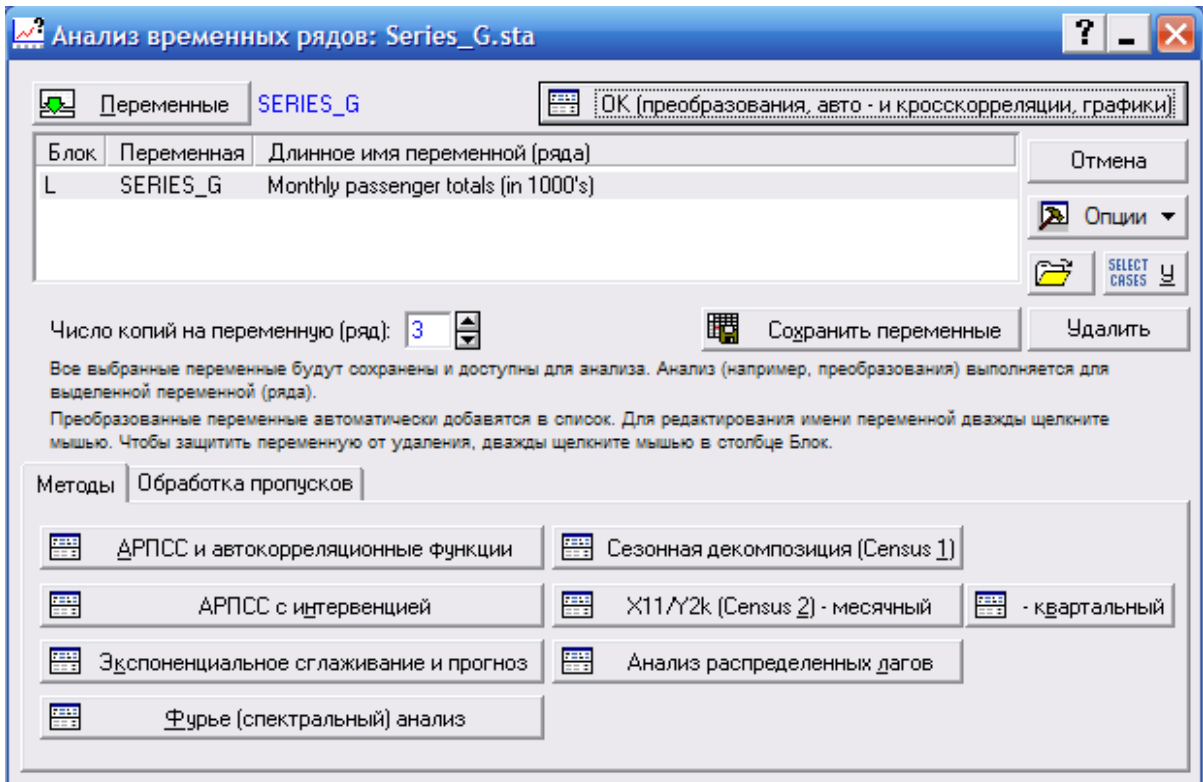


Рис. 6.3

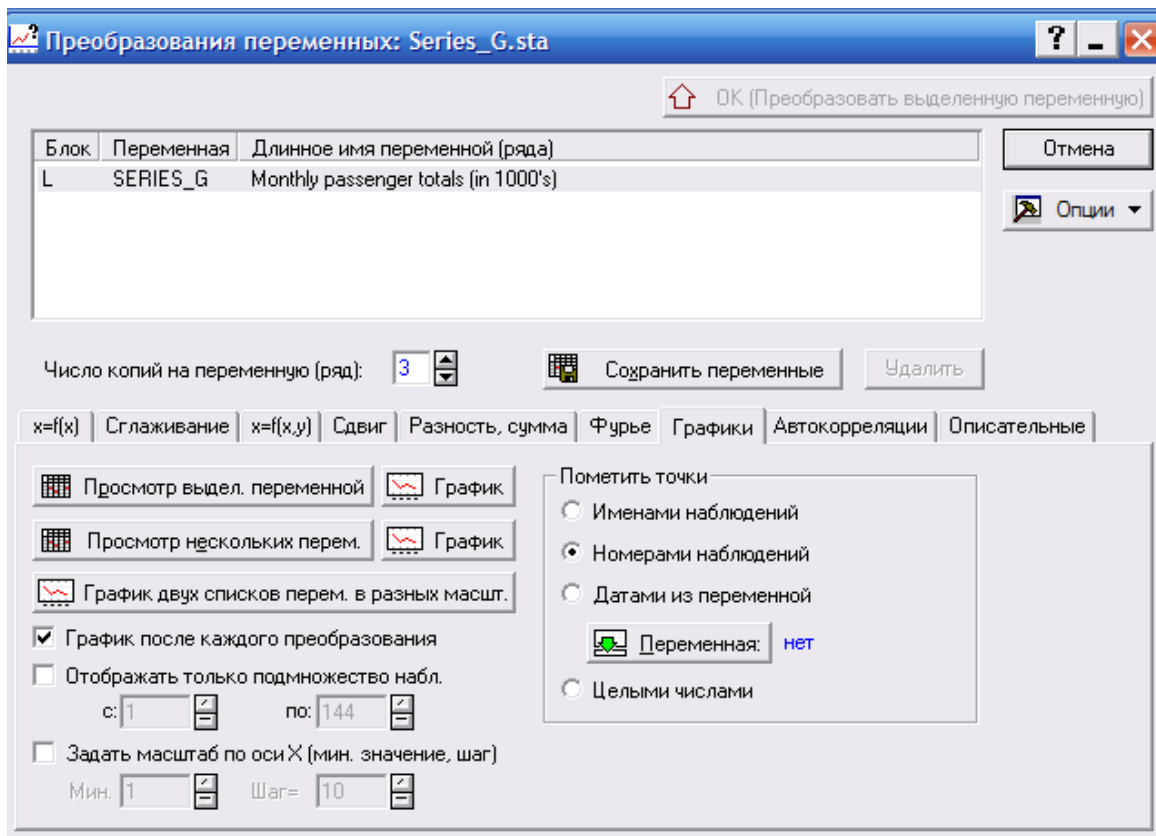


Рис. 6.4

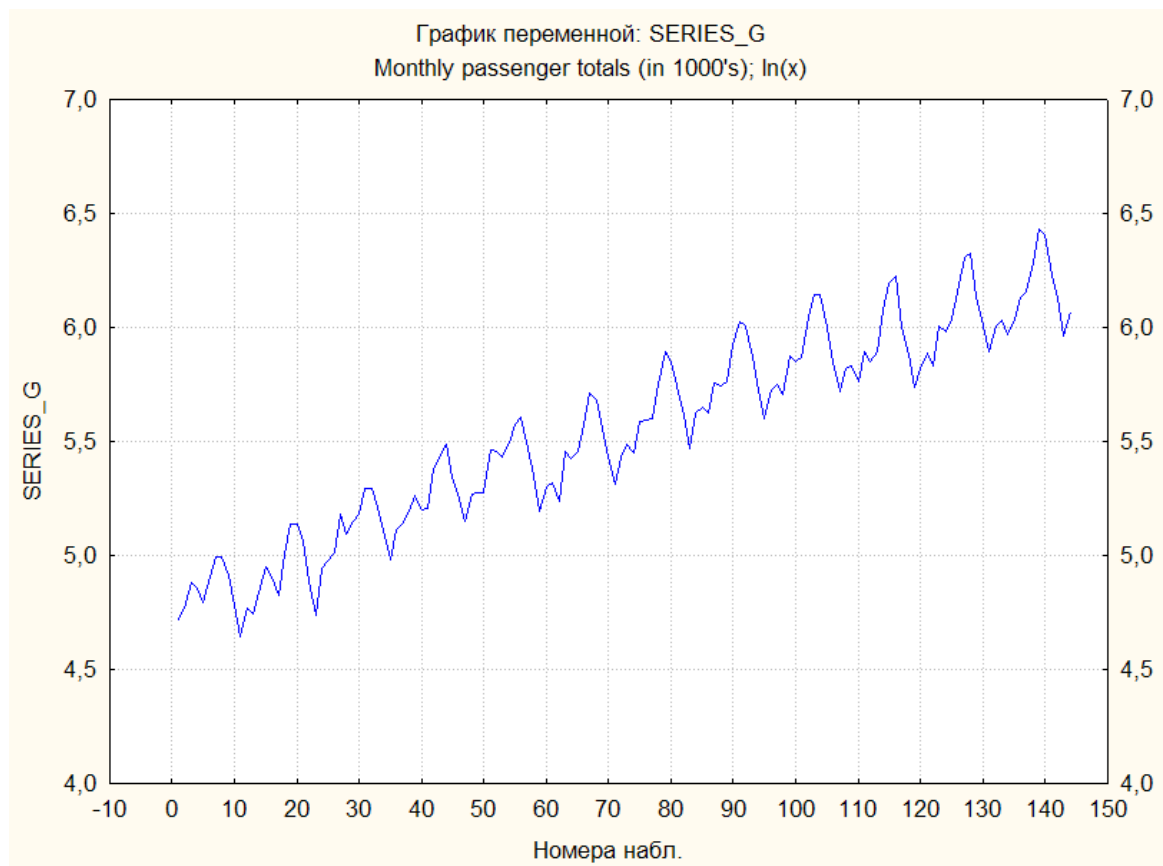


Рис. 6.5

Для усунення лінійного тренду застосуємо до часового ряду різницевий оператор з лагом 1. На закладці *Разность, сумма* (див. рис. 6.4) виберемо *Преобразование* → *Разность* $x = (x - x(\ddot{a}a))$ і вкажемо крок для різницевого оператора *лаг* = 1. Натиснемо *ОК* (*Преобразовать выделенную переменную*). Отримаємо графік перетвореного часового ряду з видаленим трендом (див. рис. 6.6).

У перетвореному часовому ряді ще залишилась періодична компонента, тому потрібно застосувати різницевий оператор з лагом 12, щоб позбутися сезонності. Ще раз на закладці *Разность, сумма* (див. рис. 6.4) виберемо *Преобразование* → *Разность* $x = (x - x(\ddot{a}a))$ і вкажемо крок для різницевого оператора *лаг* = 12. Натиснемо *ОК* (*Преобразовать выделенную переменную*). Отримаємо графік перетвореного часового ряду, в якому відсутній тренд та періодичні компоненти (див. рис. 6.7).

Переходимо до побудови моделі АРПКС для даних, що отримали після наведених перетворень. Повернемось до вікна *Анализ временных рядов* (див. рис. 6.3) і на закладці *Методы* натиснемо *АРПСС и автокорреляционные функции*. У вікні, що з'явилося (див. рис. 6.8) виберемо початкову змінну (до якої не застосовувались ніякі перетворення), а на закладці *Быстрый* вкажемо ті перетворення, які ми робили з часовим рядом для того, щоб він став стаціонарним. Також задамо параметри нашої моделі так, як показано на рис. 6.8. Після цього натиснемо *ОК* (*Начать оценивание параметров*) і з'явиться вікно, зображене на рис. 6.9. Процес оцінювання параметрів завершено.

У вікні *Результаты АРПСС* на закладці *Просмотр* можемо ще раз подивитися на графік вихідного часового ряду – *График (1)* (див. рис. 6.1), графік перетвореного часового ряду – *График (2)* (див. рис. 6.7) та графік залишків – *График (3)* (див. рис. 6.10), які треба проаналізувати, щоб з'ясувати, чи є адекватною побудована модель АРПКС(0,1,1)(0,1,1).

На закладці *Распределение остатков* натиснемо *Гистограмма* і побачимо (див. рис. 6.11), що нормальний розподіл досить добре описує залишки моделі. Натиснувши *Нормальный* на цій самій закладці, побачимо нормальний ймовірнісний графік залишків моделі АРПКС(0,1,1)(0,1,1) для досліджуваного часового ряду (див. рис. 6.12), з якого видно, що залишки гарно узгоджуються з нормальним законом розподілу.

Перейдемо на закладку *Автокорреляции* і подивимось на автокореляції та часткові автокореляції залишків, натиснувши відповідно *Автокорреляции* та *Частные автокорреляции*. Бачимо, що на обох графіках (див. рис. 6.13, 6.14) горизонтальні стовпчики не виходять за межі вертикальних ліній, тобто відсутня залежність залишків між собою, що підтверджує правильність вибору моделі часового ряду.

Після зроблених перевірок адекватності побудованої моделі АРПКС(0,1,1)(0,1,1) можемо передбачувати поведінку досліджуваного часового ряду у майбутньому.

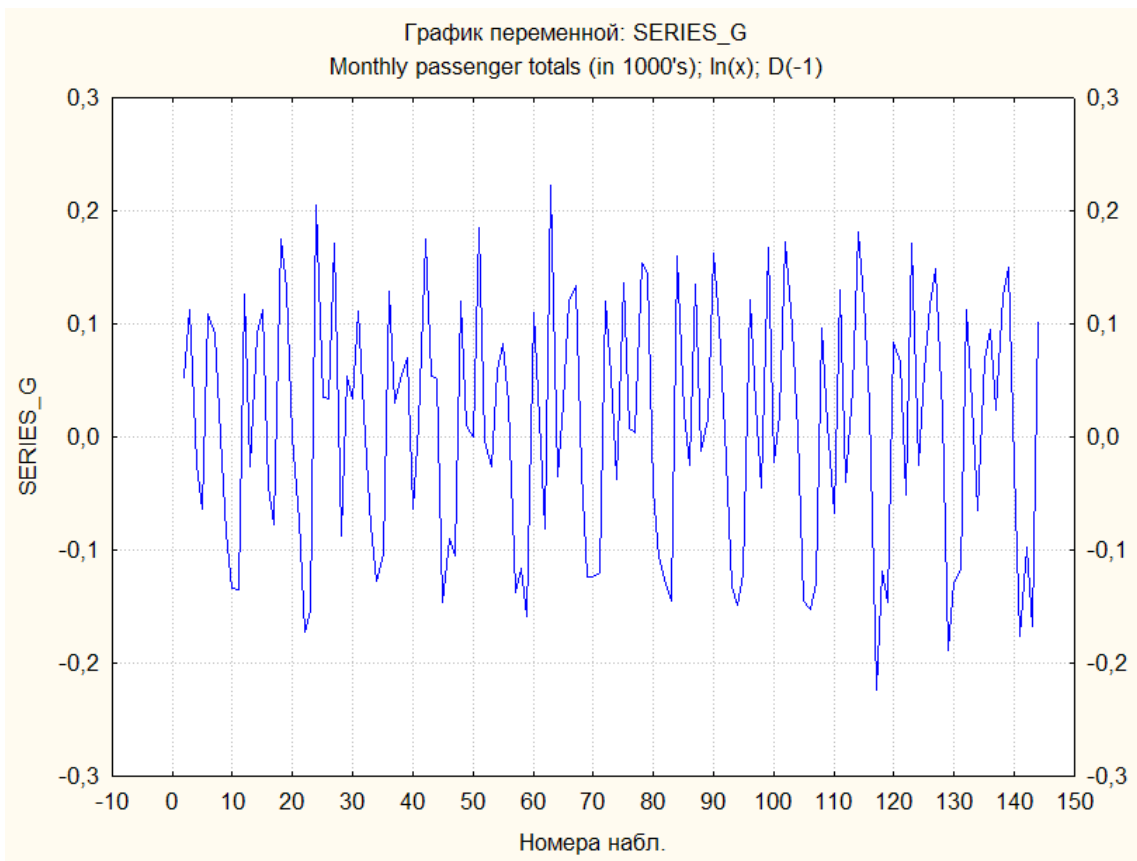


Рис. 6.6

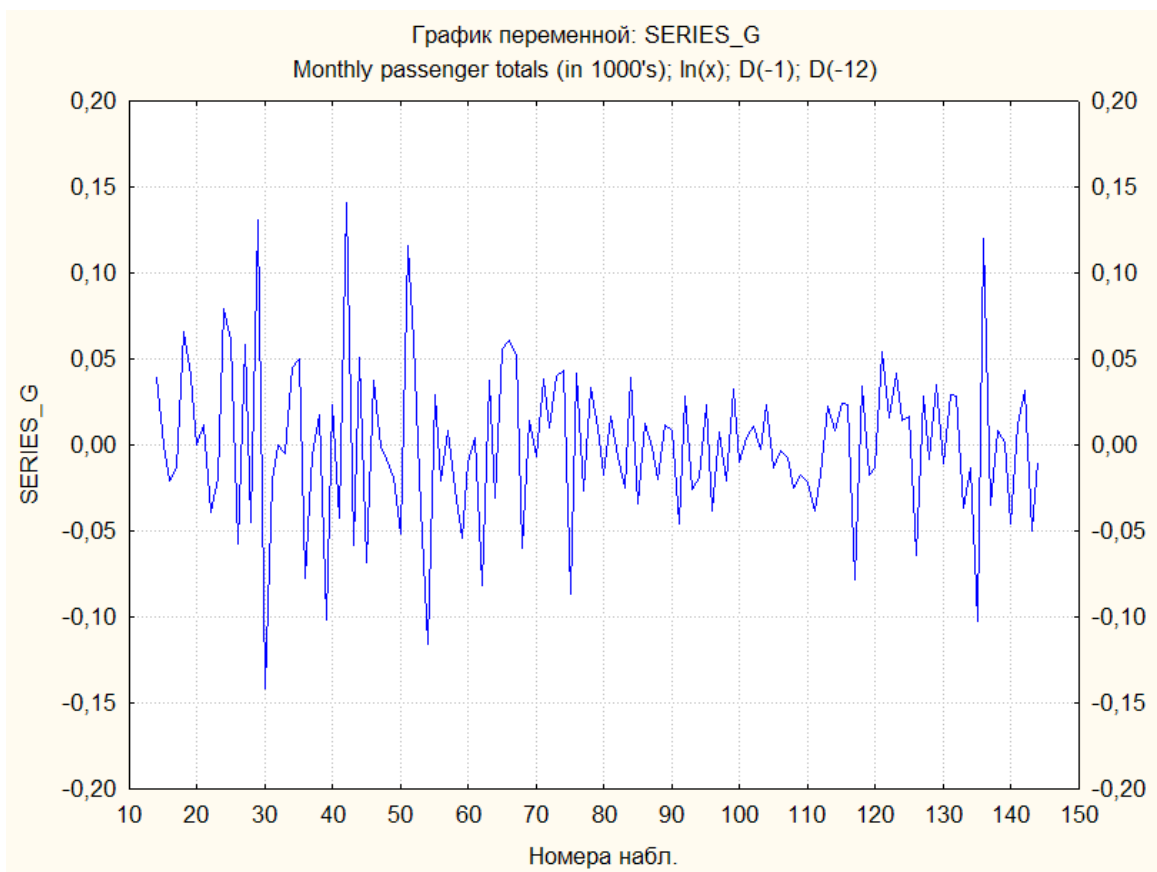


Рис. 6.7

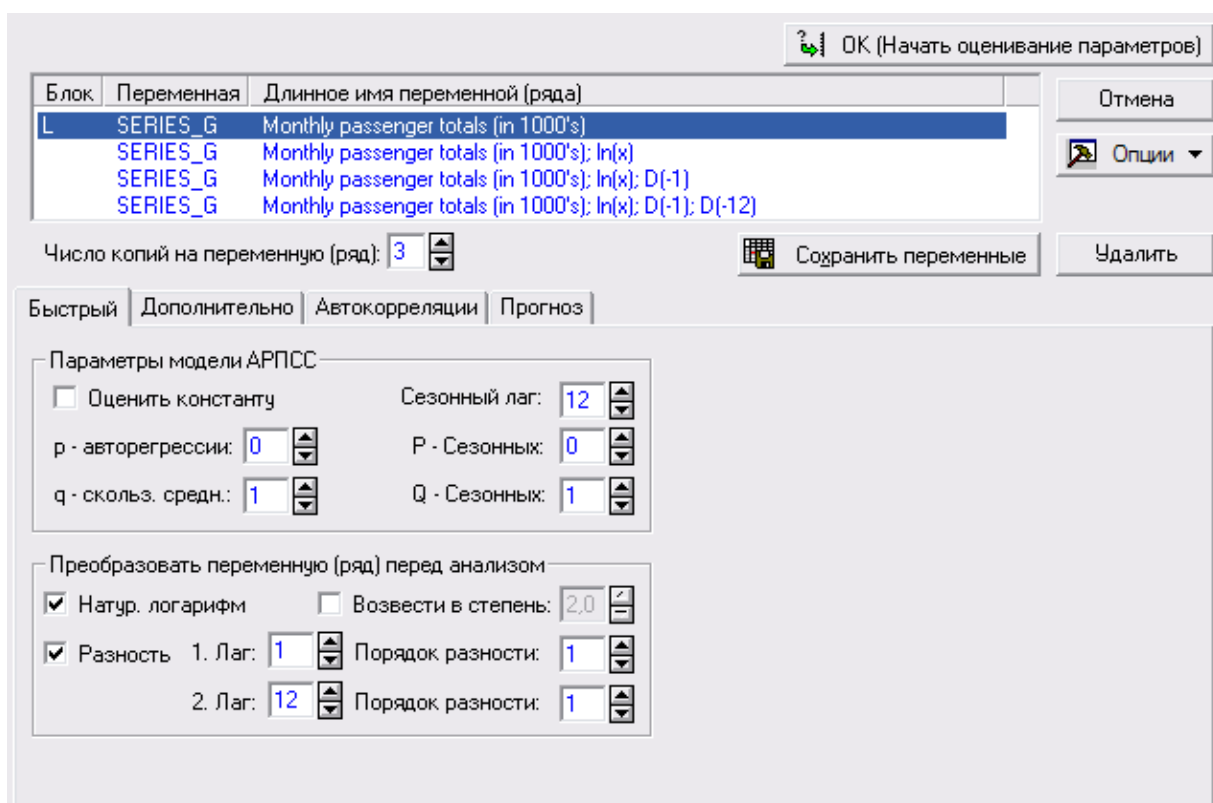


Рис. 6.8

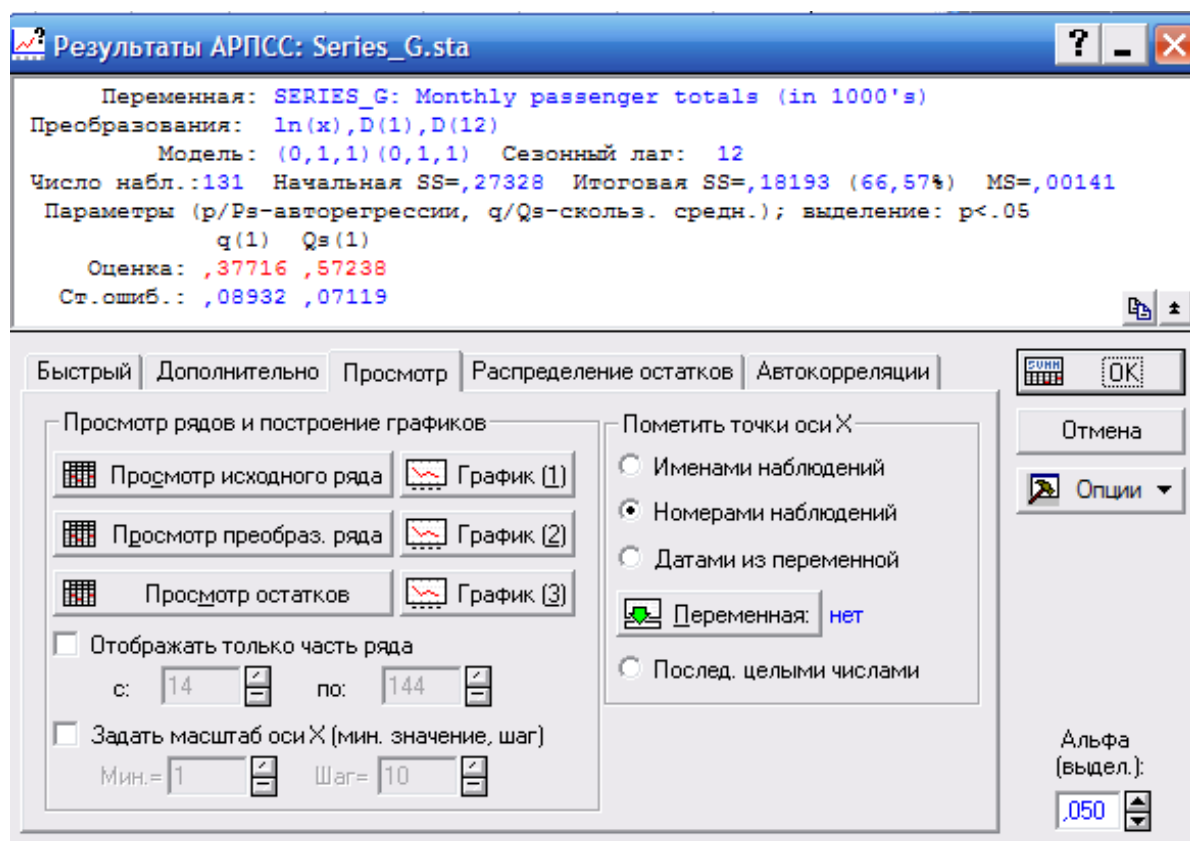


Рис. 6.9

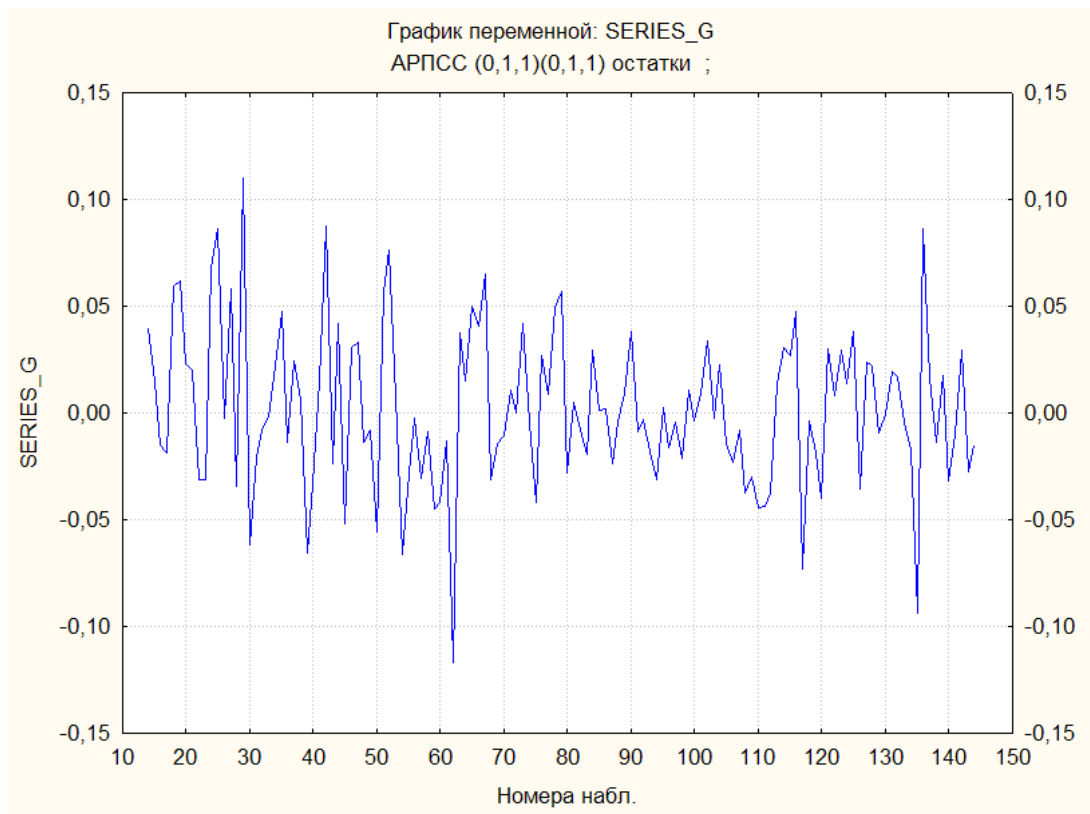


Рис. 6.10

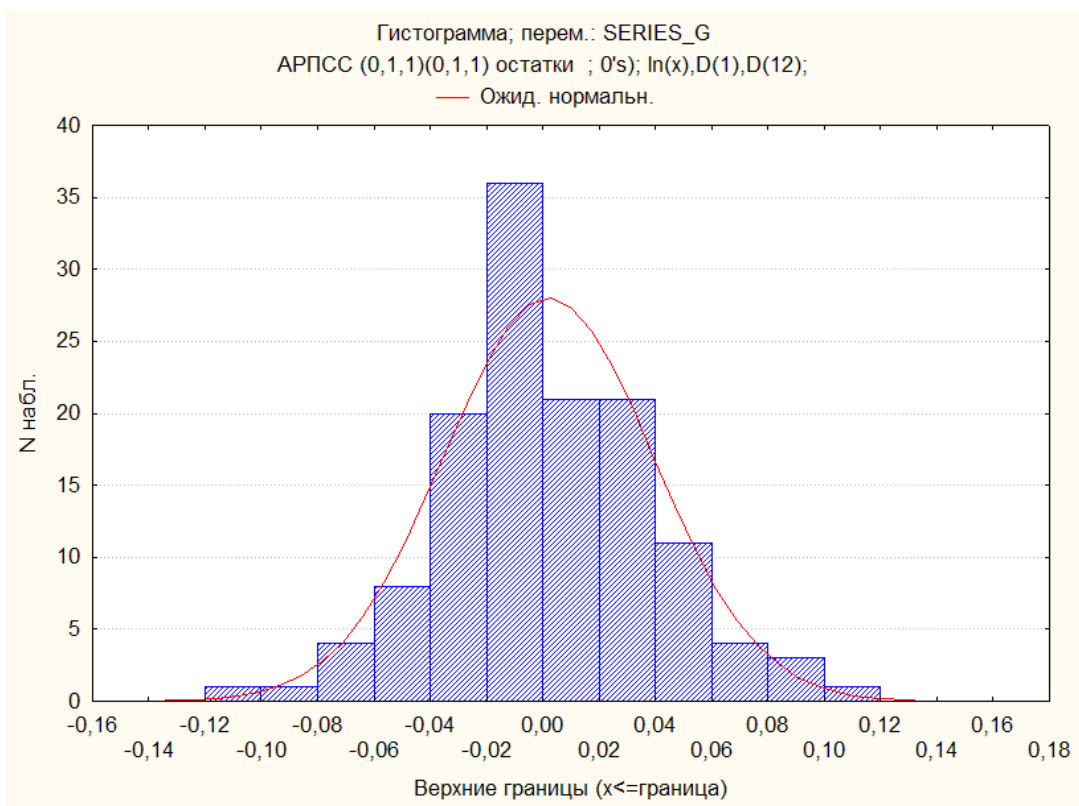


Рис. 6.11

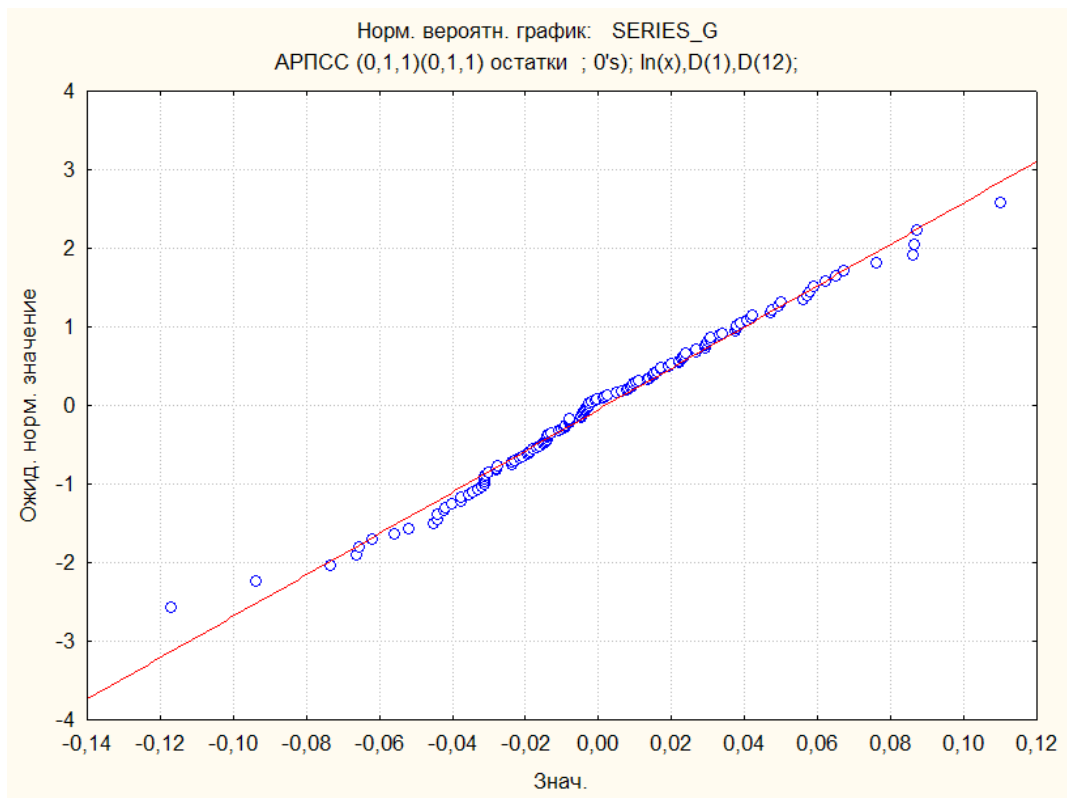


Рис. 6.12

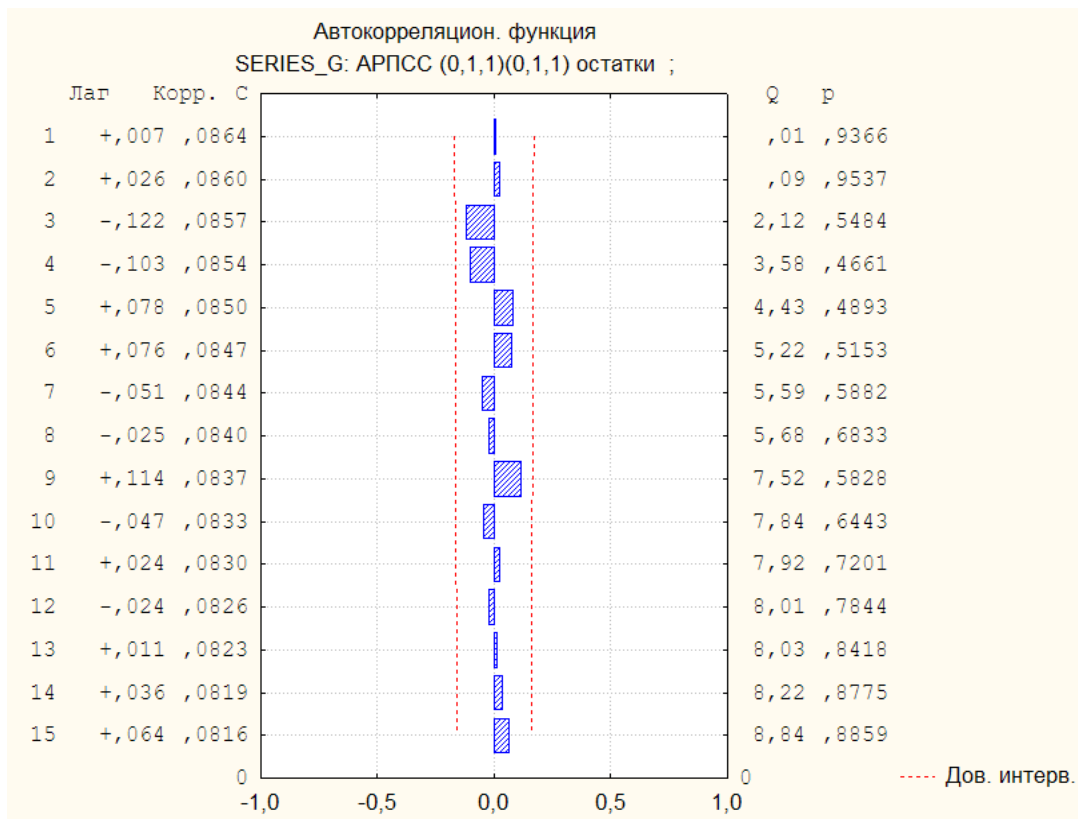


Рис. 6.13

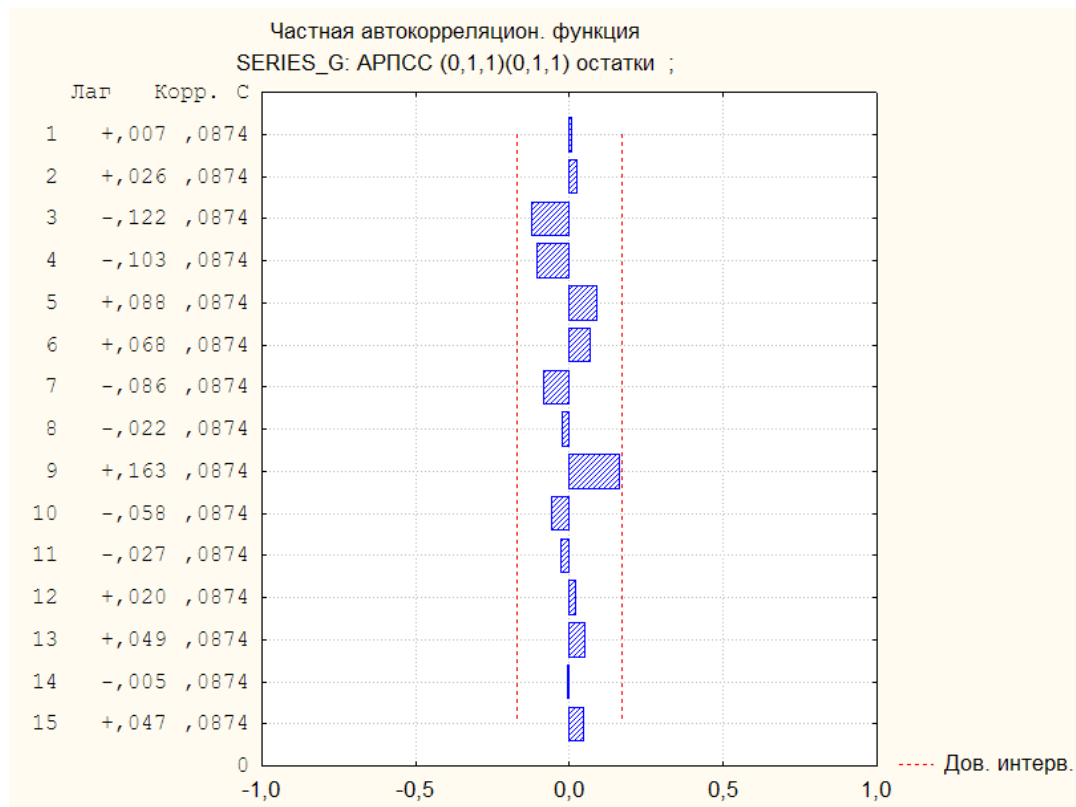


Рис. 6.14

Для прогнозування перейдемо на закладку *Прогноз* і виберемо число спостережень починаючи з наступного 145 спостереження, якого не було у початковому файлі даних, та рівень довіри майбутнього прогнозу, як показано на рис. 6.15. Натиснувши *Прогноз*, отримаємо прогнозовані кількості авіап перевезень на наступний після останнього спостереження рік та їхні нижні та верхні надійні межі (див. рис. 6.16). Після натискання *График ряда и прогнозов* побачимо графік (див. рис. 6.17) на якому червона лінія відповідає прогнозованим значенням, а зеленими лініями виділено 90-відсотковий проміжок надійності передбачення.

Якщо на закладці *Прогноз* змінимо номер спостереження, з якого хочемо зробити прогноз, на $145-12=133$ і ще раз натиснемо *График ряда и прогнозов*, то побачимо графік (див. рис. 6.18), де зображено прогноз після 132 спостереження, який дуже добре узгоджується із даними останнього річного періоду досліджуваного часового ряду. Цей графік може слугувати ще однією перевіркою адекватності побудованої моделі часового ряду.

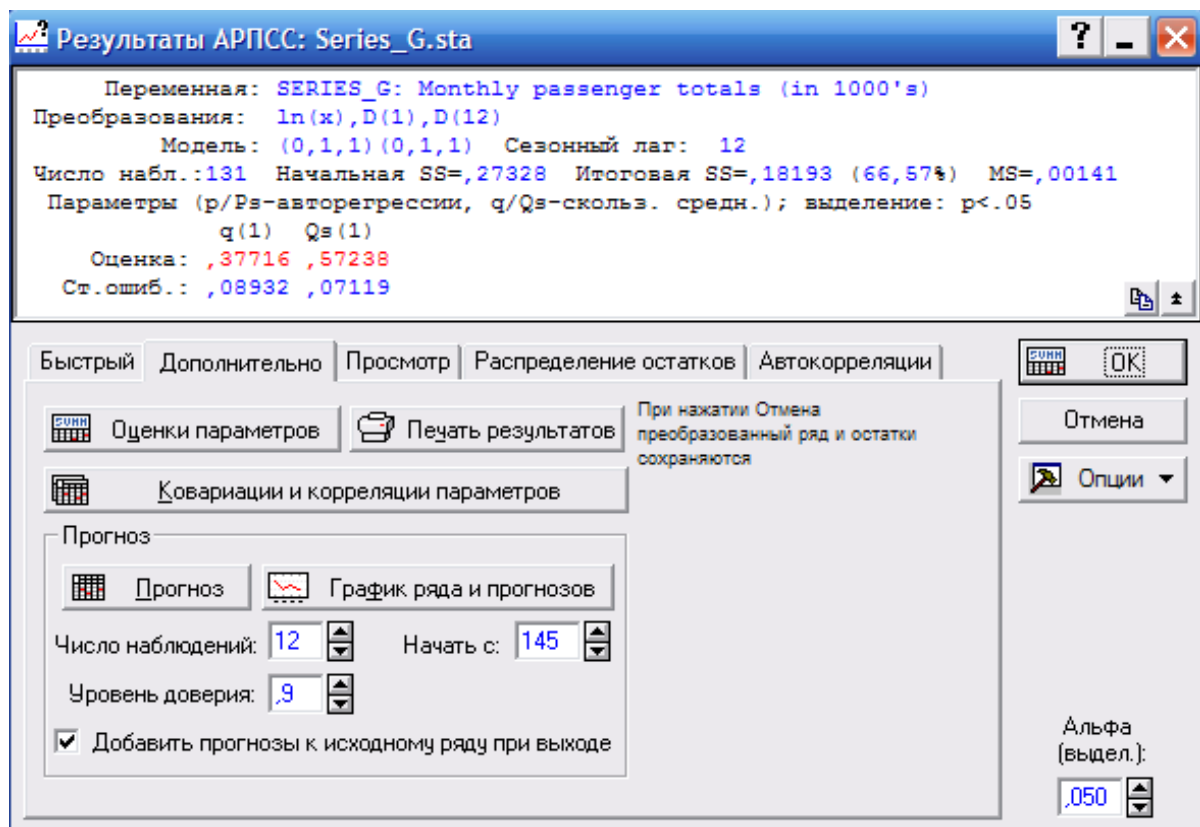


Рис. 6.15

Прогнозы; Модель:(0,1,1)(0,1,1) Сезонный лаг: 12 (Series_G.sta) Исход.:SERIES_G: Monthly passenger totals (in 1000's) Начало исходных: 1 Конец исходн.: 144			
Набл. N	Прогноз	Нижний 90,0000%	Верхний 90,0000%
145	450,1171	422,9655	479,0117
146	425,6620	395,5777	458,0341
147	479,5240	441,3696	520,9766
148	492,0412	449,0088	539,1979
149	508,5479	460,4357	561,6874
150	583,0166	524,0264	648,6473
151	669,1520	597,3584	749,5742
152	666,4152	591,1003	751,3264
153	557,9980	491,9233	632,9478
154	496,7552	435,3899	566,7696
155	429,6965	374,5207	493,0009
156	477,1535	413,6613	550,3910

Рис. 6.16

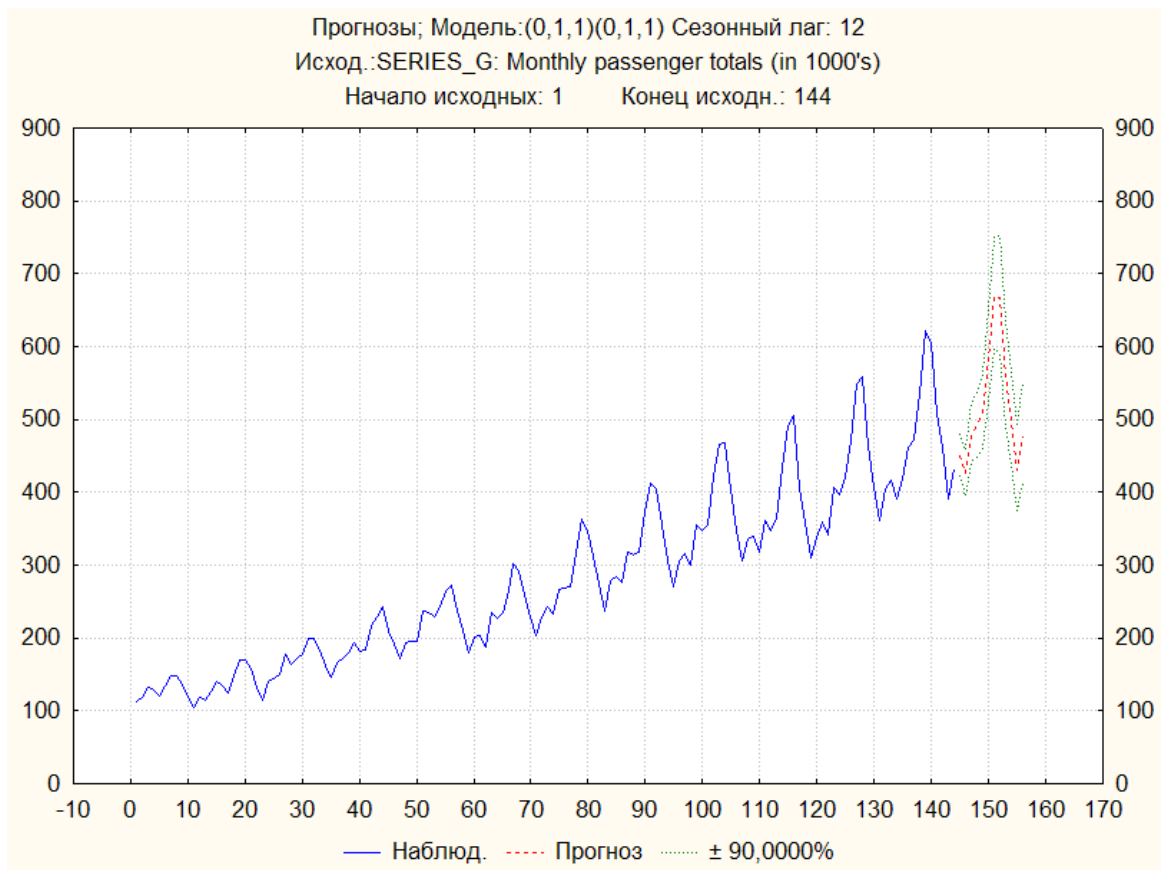


Рис. 6.17

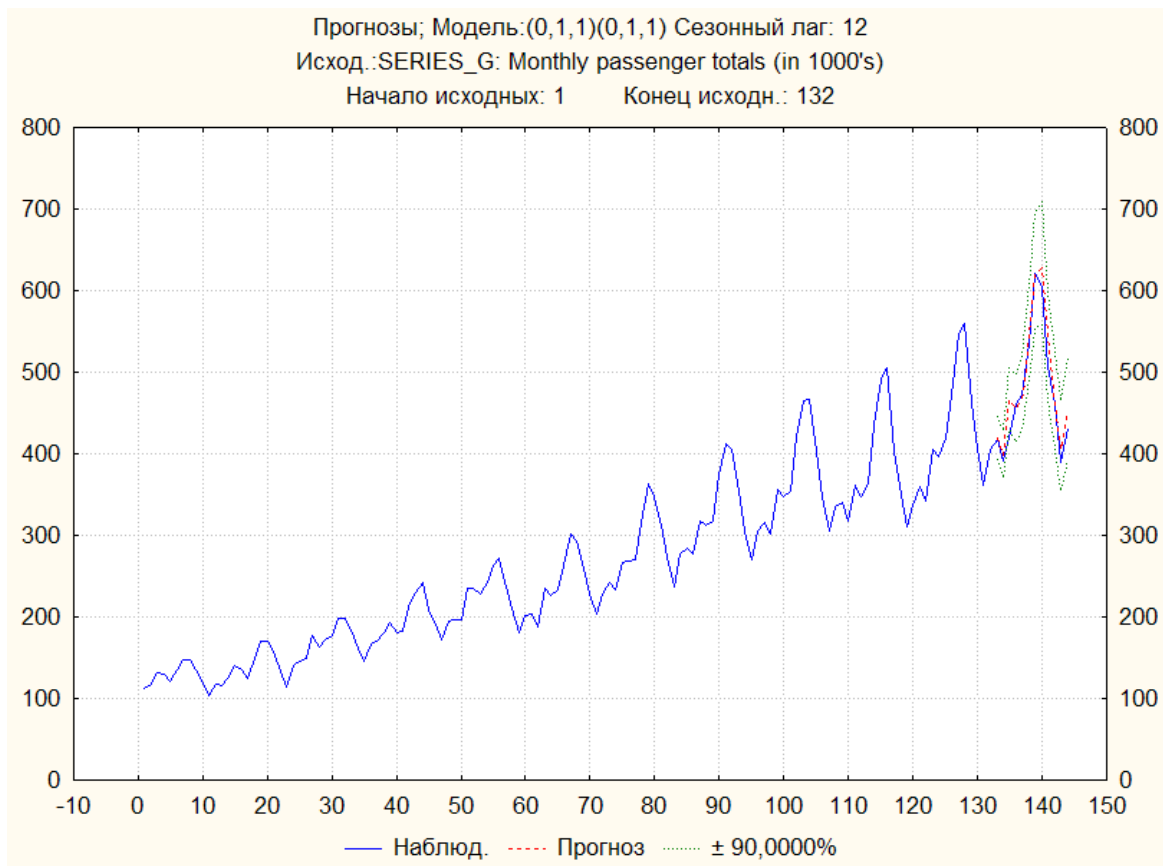


Рис. 6.18

Контрольні питання для самоперевірки до теми „Елементи аналізу часових рядів”

1. Що таке часовий ряд?
2. Які складові часового ряду є регулярними?
3. Які основні задачі аналізу часових рядів?
4. Які методи згладжування використовуються при виділенні компонент часового ряду?
5. Що таке корелограма?
6. У чому суть процесу авторегресії?
7. У чому суть процесу ковзного середнього?
8. Які основні ідеї методу аналізу часових рядів, розробленого Боксом і Дженкінсом (1976)?
9. Як визначаються параметри моделі АРПКС(p, d, q)?
10. Як визначаються параметри моделі АРПКС(p, d, q)(ps, ds, qs)?
11. Як здійснюється ідентифікація порядку моделі часового ряду?
12. Які перетворення часового ряду роблять його стаціонарним?
13. Які особливості побудови моделі часового ряду у програмі STATISTICA?
14. Які можливості для перевірки адекватності побудованої моделі часового ряду у програмі STATISTICA?
15. Як здійснюється прогнозування значень часового ряду засобами програми STATISTICA?

Глава 7. Визначення об'єму вибірки

При плануванні вибіркового обстеження постає питання, як забезпечити необхідну точність результатів і при цьому уникнути зайвих витрат. Зосередимося на проблемі визначення кількості вибірових даних, необхідних для забезпечення заданої точності при оцінюванні частки та кількості елементів з фіксованою ознакою.

7.1. Показники точності оцінювання

Наведемо найбільш розповсюджені кількісні ймовірнісно-статистичні міри точності оцінок, в основі яких лежить дисперсія або середньоквадратична похибка. Зауважимо, що остаточний вибір показника точності як основи для планування вибіркового обстеження належить не статистикам, а споживачам результатів обстеження. Суттєвим моментом повинна бути можливість змістовного тлумачення результатів. Наведемо основні показники точності оцінок:

а) **Стандартна (середньоквадратична) похибка σ** : показує порядок величини можливого відхилення оцінки від справжнього значення параметра. За умов нормального розподілу оцінок у 2/3 випадків дійсне відхилення оцінки менше за стандартну похибку.

б) **Гранична похибка вибірки e_α** – практично максимально допустиме відхилення оцінки від параметра. Рідко, з імовірністю α , можливе відхилення більше за e_α . Імовірність α вибирається дослідником малою.

в) **Величина довірчого інтервалу $l = 2e_\alpha$** – діапазон, в якому знаходиться значення оцінюваного параметра.

г) **Показники відносної точності** – коефіцієнт варіації або показники, що перераховані в пунктах б), в), поділені на математичне сподівання оцінки.

Інколи буває необхідно задавати точність результатів обстежень у відносних показниках. Зрозуміло, що треба по-різному ставитись до значення середньої квадратичної похибки $\sigma_p=0,04$, коли значення параметра $P = 1$ і коли $P = 0,1$. У першому випадку середня квадратична похибка дорівнює лише 4% від значення досліджуваного параметра P , а в другому – уже 40%.

7.2. Визначення об'єму вибірки n при оцінюванні часток P

Точність визначається середніми квадратичними похибками σ_p . Якщо вважати величини σ_p , P та N відомими, то отримаємо формулу для об'єму вибірки n :

$$n = \frac{P(1-P)N}{\sigma_p^2(N-1) + P(1-P)} \quad (1')$$

Коли вважати N досить великим, то $N \approx N-1$ і формула (1') спрощується:

$$n = \frac{P(1-P)N}{\sigma_p^2(N-1) + P(1-P)} = \frac{N}{1 + N\sigma_p^2/P(1-P)} = \frac{n_0}{1 + n_0/N}, \quad (2')$$

де $n_0 = P(1-P)/\sigma_p^2$.

Ця формула буде основою для подальших висновків. Зауважимо, що при $N \rightarrow \infty$ граничний об'єм вибірки дорівнює

$$n_0 = \lim_{N \rightarrow \infty} \frac{P(1-P)}{\sigma_p^2}. \quad (3')$$

Формулу (3') варто застосовувати, коли немає ніякої інформації стосовно N , або коли N дуже велике. Коли ж N не дуже велике, то формула (3') дає завищені результати щодо необхідної кількості спостережень. Значення n_0 також використовують як перше наближення, що потребує подальшого уточнення.

При застосуванні формул (1') – (3') виникає ще кілька проблем. У ці формули крім N входить величина P , яку як раз і треба оцінити за вибіркою. Інколи буває відома апіорна інформація відносно приблизного значення P , виходячи з теоретичних міркувань або з аналогічних досліджень. Тоді цю попередню оцінку можна використати в (1') – (3').

Якщо ж про P нічого невідомо, то варто у формулу (1') підставити значення $P = 0,5$, яке дає максимальне значення добутку $P(1-P) = 0,25$. При цьому дослідник отримує дещо завищене значення для n , яке гарантує необхідну точність.

Рекомендується проводити обстеження у два етапи. На першому етапі беруть просту випадкову вибірку обсягом m_1 , за допомогою якої знаходять оцінку частки P_1 . Цю оцінку використовують для знаходження потрібного об'єму n більшої вибірки за формулою

$$n = \frac{P_1(1-P_1)}{\sigma_p^2} + \frac{3-8P_1(1-P_1)}{P_1(1-P_1)} + \frac{1-3P_1(1-P_1)}{\sigma_p^2 m_1}.$$

Отже, потрібно додатково обстежити $n - m_1$ елементів, потім за всіма n даними знайти нову оцінку \hat{p}_1 , тобто $\hat{p}_1 = n_1/n$ і внести поправку на зсув. Остаточна оцінка для P знаходиться за формулою

$$\hat{P} = \hat{p}_1 + \frac{\sigma_p^2(1-2\hat{p}_1)}{\hat{p}_1(1-\hat{p}_1)}.$$

Така процедура дає більш надійні оцінки для P , проте подовжує терміни проведення обстежень.

7.2.1. Визначення n при заданій граничній похибці e_p

Далі вважаємо, що всі оцінки мають нормальний розподіл. Необхідний об'єм вибірки знаходять за формулою:

$$n = \frac{u_\alpha^2 P(1-P)N}{e_p^2(N-1) + u_\alpha^2 P(1-P)},$$

де u_α – квантиль рівня $1 - \alpha/2$ для стандартного нормального розподілу.

При $N \approx N-1$ маємо

$$n = \frac{u_\alpha^2 P(1-P)}{e_p^2 + u_\alpha^2 P(1-P)/N} = \frac{N}{1 + Ne_p^2 / u_\alpha^2 P(1-P)} = \frac{n_0}{1 + n_0/N} \quad (*)$$

та

$$n_0 = \lim_{N \rightarrow \infty} n = \frac{u_\alpha^2 P(1-P)}{e_p^2}.$$

7.2.2. Визначення об'єму вибірки при заданій відносній точності

При заданому коефіцієнті варіації V необхідний об'єм вибірки дорівнює

$$n = \frac{(1-P)}{V^2 P + (1-P)/N} = \frac{N}{1 + NV^2 P/(1-P)} \quad (4')$$

і при $N \rightarrow \infty$ прямує до граничного значення

$$n_0 = \frac{(1-P)}{V^2 P}. \quad (5')$$

Якщо немає апріорної інформації стосовно величини P , але з певних міркувань можна вказати нижню межу для P , то значення P_{\min} слід підставити в (4'), (5') і отримане n гарантує необхідну точність. Дозволяють також застосування двохетапної схеми. Після першого етапу, на якому за вибіркою обсягом m_1 , знайшли оцінку P_1 , треба додатково обстежити $n - m_1$ елементів, де

$$n = \frac{(1-P_1)}{V^2 p_1} + \frac{3}{P_1(1-P_1)} + \frac{1}{V^2 P_1 m_1},$$

а потім знайти

$$\hat{p} = \hat{p}_1 - V^2 \hat{p}_1(1 - \hat{p}_1).$$

7.3. Об'єм вибірки при дослідженні декількох ознак

У більшості обстежень дані збирають стосовно не однієї, а декількох ознак. Одним із методів визначення об'єму вибірки в цій ситуації є така послідовність дій: спочатку вибирають граничні значення похибок для кожної з цих ознак, визначають найважливіші для даного обстеження ознаки, знаходять необхідні значення для обсягу вибірки n_i окремо по кожній ознаці, а потім приймають компромісне рішення з урахуванням знайдених n_i для окремих ознак, вартості й термінів обстеження.

7.4. Визначення об'єму вибірки при оцінюванні середніх і сумарних значень

При прямому оцінюванні середніх і сумарних значень об'єм вибірки n – єдина величина, змінюючи яку дослідник може впливати на точність оцінки. Як і при оцінюванні частин, вимоги до точності можуть формулюватися в термінах дисперсії або середньої квадратичної похибки оцінок, граничної та відносної похибки. Розглянемо окремо ці випадки.

7.4.1. Визначення об'єму вибірки n при оцінюванні середнього при заданій середній квадратичній похибці

Нехай а рїогї задано бажане значення $s_{\bar{x}}$. Тодї, при $N \approx N - 1$ маємо формули для визначення n :

$$n = \frac{\sigma^2 N}{N\sigma_{\bar{x}}^2 + \sigma^2} = \frac{N}{1 + N(\sigma_{\bar{x}}/\sigma)^2} = \frac{\sigma^2}{\sigma_{\bar{x}}^2 + \sigma^2/N}. \quad (6')$$

Звідки

$$n_0 = \lim_{N \rightarrow \infty} n = \frac{\sigma^2}{\sigma_{\bar{x}}^2}, \quad (7')$$

де N – об'єм генеральної сукупності, σ^2 – дисперсія генеральної сукупності. Отже, коли нема точних даних про N , то у формулу (6') можна підставити число більше за N (таку верхню границю часто можна вказати) та отримати n , яке гарантує необхідну точність. Якщо ж про N нема ніякої інформації, то слід використати формулу (7'), яка дає дещо завищені значення для n , проте гарантує необхідну точність.

Певні складнощі на стадії планування вибірки можуть бути пов'язані і з відсутністю точних даних щодо σ^2 . У цьому разі можна використати:

1) дані щодо дисперсії та середньої квадратичної похибки з аналогічних обстежень, які проводилися раніше;

2) спеціально провести попереднє пробне обстеження з малим об'ємом m_0 , наприклад, з $m_0 = 30$, за результатами якого оцінити вибіркву дисперсію

$$S_0^2 = (1/(m_0 - 1)) \sum_{i=1}^{m_0} (x_i - \bar{x}_0)^2$$

і скористатися формулою

$$n \approx \frac{S_0^2}{\sigma_{\bar{x}}^2} \left(1 + \frac{2}{m_0}\right);$$

3) оцінити σ^2 на основі припущень щодо розподілу генеральної сукупності та співвідношень між середньоквадратичною похибкою та максимальним інтервалом розмаху ознаки $R = x_{\max} - x_{\min}$ для основних типів розподілів.

7.4.2. Визначення об'єму вибірки при оцінюванні середнього при заданій граничній похибці $e_{\bar{x}}$

Нехай задане бажане значення граничної похибки $e_{\bar{x}} = u_{\alpha} \sigma_{\bar{x}}$ (розподіл \bar{x} вважаємо приблизно нормальним).

У цьому разі об'єм вибірки n визначають за формулами

$$n = \frac{N u_{\alpha}^2 \sigma^2}{N e_{\bar{x}}^2 + u_{\alpha}^2 \sigma^2} = \frac{N}{1 + N(e_{\bar{x}}/u_{\alpha} \sigma)^2} = \frac{u_{\alpha}^2 \sigma^2}{\sigma_{\bar{x}}^2 + u_{\alpha}^2 \sigma^2 / N}$$

та

$$n_0 \approx (u_{\alpha} \sigma / e_{\bar{x}})^2 \text{ при } N \rightarrow \infty,$$

де u_{α} - квантиль рівня $1 - \alpha/2$ для стандартного нормального розподілу.

7.4.3. Визначення об'єму вибірки при оцінюванні сумарного значення при заданій середньоквадратичній або граничній похибці

Довірчий інтервал для сумарного значення ознаки X' можна знайти лише для скінченної генеральної сукупності з відомим обсягом N . Отже, необхідний об'єм вибірки, що гарантує задану середньоквадратичну похибку $\sigma_{X'} = N \sigma_{\bar{x}}$ визначають за формулами

$$n = \frac{N \sigma^2}{N \sigma^2 + \sigma_{X'}^2} = \frac{N}{1 + N(\sigma_{X'} / N \sigma)^2},$$

а при заданій граничній похибці $e_{X'} = u_{\alpha} \sigma_{X'}$ - за формулами

$$n = \frac{N^2 u_\alpha^2 \sigma^2}{e_{x'}^2 + N^2 u_\alpha^2 \sigma^2} = \frac{N}{1 + N(e_{x'} / Nu_\alpha \sigma)^2}.$$

7.4.4. Визначення об'єму вибірки при заданій відносній точності

Перевага відносних оцінок у тому, що вони є величинами, які не мають розмірності, і за їх допомогою можна порівнювати точність оцінок різних ознак (і в різних обстеженнях). Припустимо, що $N \approx N - 1$. Тоді при заданому коефіцієнті варіації V оцінки середнього для необхідного об'єму вибірки n маємо співвідношення

$$n = \frac{V^2}{V_{\bar{x}}^2 + V^2 / N} = \frac{N}{1 + N(V_{\bar{x}} / V)^2}, \quad (8')$$

і

$$n \approx (V/V_{\bar{x}})^2 \text{ при } N \rightarrow \infty \text{ (частина відбору } f = n/N < 0,05) \quad (9')$$

Оскільки коефіцієнти варіації оцінок середнього та сумарного значення дорівнюють один одному, то формули (8'), (9') можна використовувати і тоді, коли метою досліджень є сумарні значення ознак.

7.5. Об'єм вибірки за необхідності отримати оцінки для підрозділів сукупності

Досить часто потрібно отримати оцінки не тільки для сукупності в цілому, але й для підрозділів. Якщо ці підрозділи можна виділити заздалегідь, як, наприклад, географічні (або адміністративні) райони, то n_i знаходиться окремо по кожному підрозділу. Припустимо, що середнє треба оцінити для кожного з підрозділів із заданою середньоквадратичною

похибкою σ . Тоді для i -того підрозділу $n_i \approx s_i^2 / \sigma^2$, а загальний об'єм

вибірки $n = \sum_{i=1}^k n_i = \sum_{i=1}^k s_i^2 / \sigma^2$. Якщо можна вважати, що $s_1^2 \approx \dots \approx s_k^2 \approx S^2$, S^2 –

дисперсія всієї вибірки, то $n = kS^2 / \sigma^2$. Це означає, що коли бажано отримати оцінки середнього з заданою точністю для кожного з k підрозділів сукупності, то треба провести в k разів більше обстежень, ніж коли б ця умова ставилася лише для всієї сукупності в цілому. На цей факт слід звертати увагу при плануванні обстежень.

Приклад 7.1. Генеральна сукупність складається із 4 тис. працівників певної фірми. P – відсоток тих, хто протягом останнього року відпочивав у Карпатах. Відомо, що $35\% \leq P \leq 55\%$. Яким повинен бути

об'єм вибірки, якщо ми хочемо мати інформацію про P з точністю 2% при рівні довіри 95%?

Розв'язання. Для знаходження об'єму вибірки при заданій граничній похибці $e_p = 2\%$ скористаємось формулою (*), де $P=0,5$, яке дає максимальне значення добутку $P(1-P)=0,25$. За таблицями стандартного нормального розподілу знайдемо u_α – квантиль рівня $1 - \alpha/2$ для $\alpha = 1 - 0,95 = 0,05$.

Для знаходження u_α в пакеті STATISTICA в меню *Анализ* можна вибрати опцію *Вероятностный калькулятор* та відкрити вікно *Распределения*:

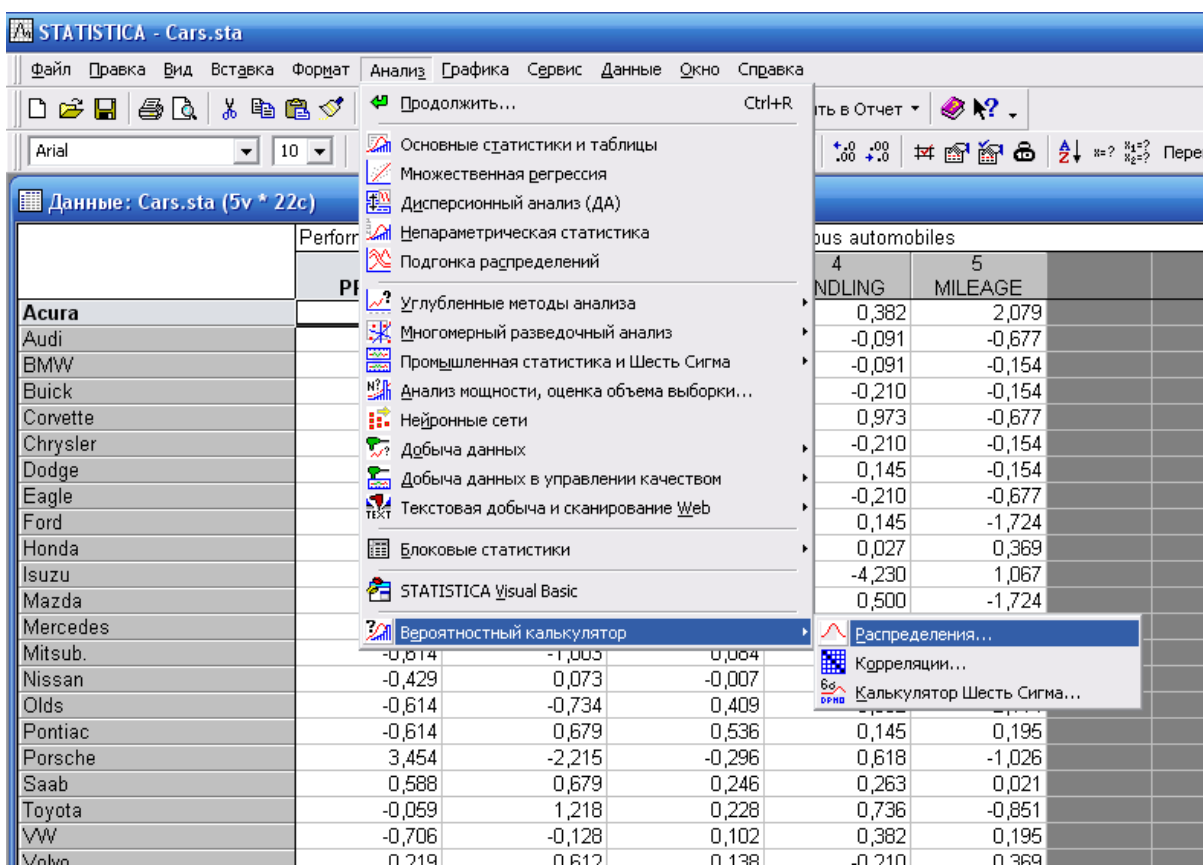


Рис. 7.1

У вікні *Распределения* вибираємо нормальний розподіл Z (Нормальное) та вказуємо параметри стандартного нормального розподілу *среднее* 0, *ст.откл.* 1 (див. рис. 7.2).

Щоб знайти u_α – квантиль рівня $1 - \alpha/2=0,975$ у поле p : вносимо значення ,975 і натискаємо кнопку *Вычислить*:

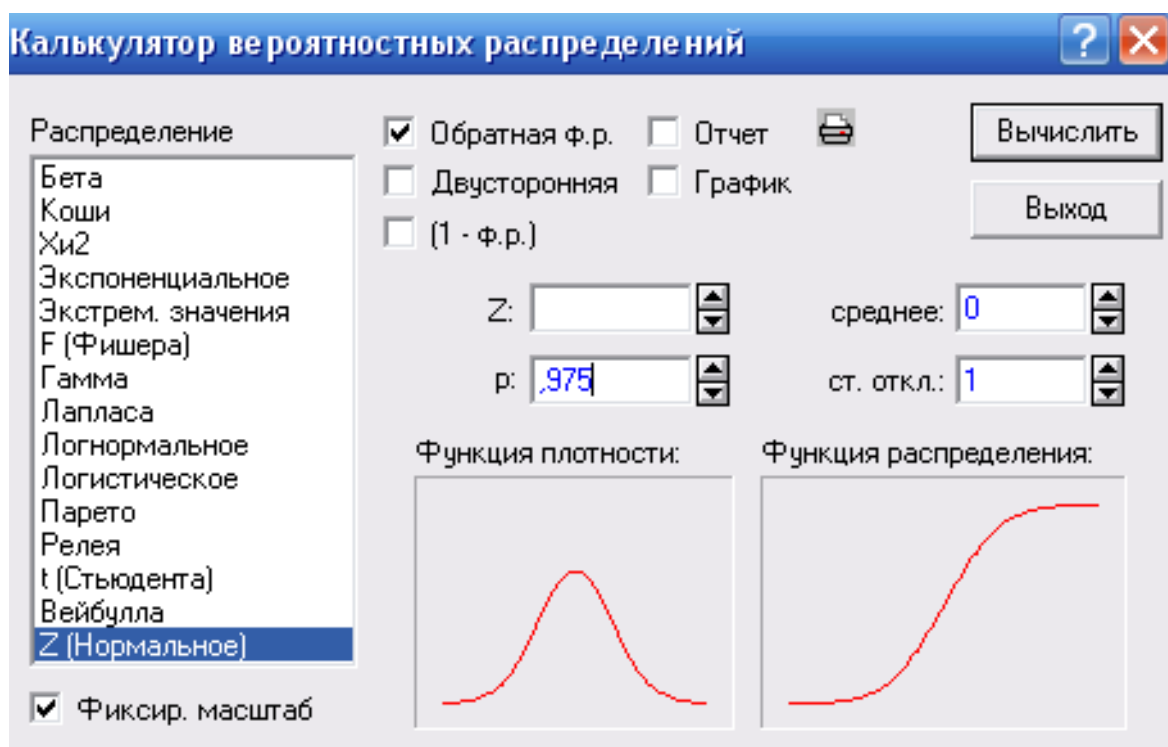


Рис. 7.2

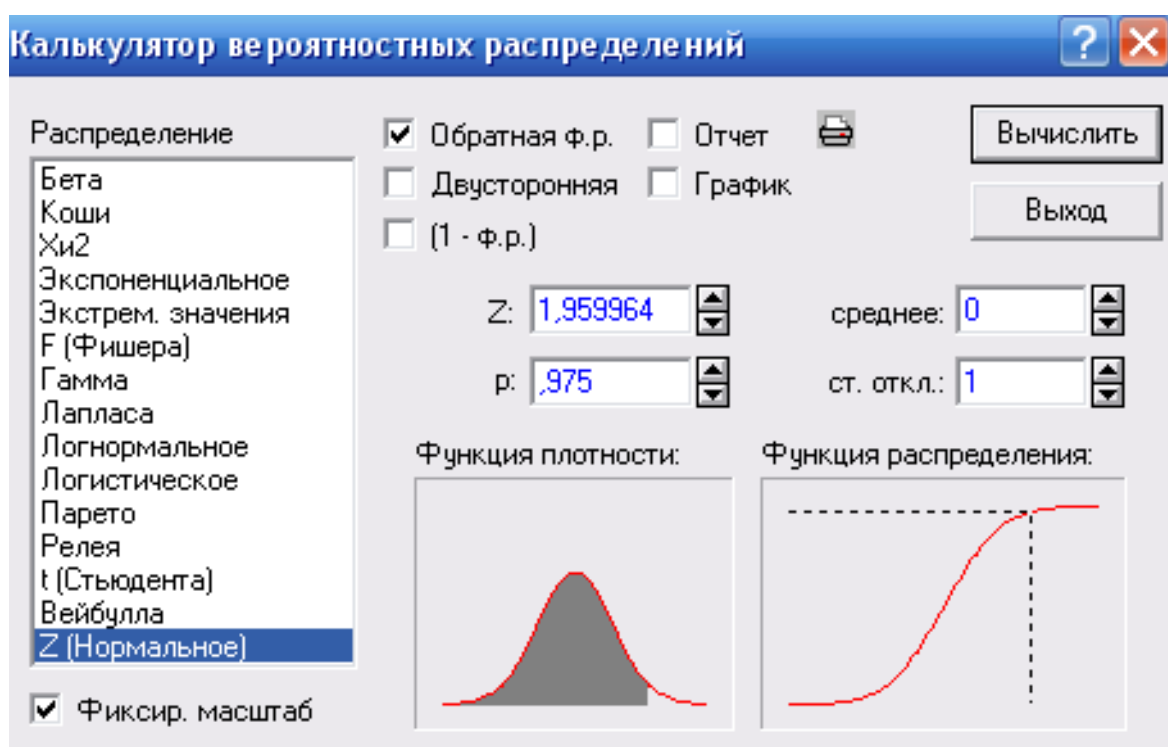


Рис. 7.3

Потрібний квантиль з'явиться у полі **Z**:

Отже, $u_\alpha \approx 1,96$. $n = \frac{N}{1 + Ne_p^2 / u_\alpha^2 P(1-P)} \approx \frac{4000}{1 + 4000 \left(\frac{0,02}{1,96}\right)^2 \cdot 4} \approx 1500$.

**Контрольні питання для самоперевірки до теми
„ Визначення об’єму вибірки ”**

1. Наведіть найбільш розповсюджені кількісні ймовірносто-статистичні міри точності оцінок.
2. Як визначається об’єм вибірки при оцінювання часток?
3. Як визначається об’єм вибірки при заданій граничній похибці?
4. Як визначається об’єм вибірки при заданій відносній точності?
5. Як визначається об’єм вибірки при дослідженні декількох ознак?
6. Як визначається об’єм вибірки при оцінюванні середнього при заданій середній квадратичній похибці?
7. Як визначається об’єм вибірки при оцінюванні середнього при заданій граничній похибці?
8. Як визначається об’єм вибірки при оцінюванні сумарного значення при заданій середньоквадратичній або граничній похибці?
9. Як визначається об’єм вибірки за необхідності отримати оцінки для підрозділів сукупності в пакеті STATISTICA?

Додаток 1. Таблиці математичної статистики

8.1. Таблиця великих чисел

P	Похибка досліду p , %									
	10	9	8	7	6	5	4	3	2	1
0,75	33	40	51	67	91	132	206	367	827	3308
0,80	41	50	64	83	114	164	256	456	1026	4105
0,85	51	63	80	105	143	207	323	575	1295	5180
0,90	67	83	105	138	187	270	422	751	1690	6763
0,91	71	88	112	146	199	287	449	798	1796	7185
0,92	76	94	119	156	212	306	478	851	1915	7662
0,93	82	101	128	167	227	328	512	911	2051	8207
0,94	88	109	138	180	245	353	552	981	2210	8843
0,95	96	118	150	195	266	384	600	1067	2400	9603
0,96	105	130	164	215	292	421	659	1171	2636	10544
0,965	111	137	173	226	308	444	694	1234	2778	11112
0,970	117	145	183	240	327	470	735	1308	2943	11773
0,975	125	155	196	256	348	502	784	1395	3139	12559
0,980	135	167	211	276	375	541	845	1503	3382	13529
0,985	147	182	231	301	410	591	924	1643	3697	14791
0,990	165	204	259	338	460	663	1036	1843	4146	16587
0,995	196	243	307	402	547	787	1288	2188	4924	19698
0,999	270	334	422	552	751	1082	1691	3009	6767	27069

8.2. Значення критерію τ в залежності від об'єму вибірки N і рівня значущості α

	α			α	
	0,05	0,01		0,05	0,01
4	0,955	0,991	17	0,359	0,460
5	0,807	0,916	18	0,349	0,449
6	0,669	0,805	19	0,341	0,439
7	0,610	0,740	20	0,334	0,430
8	0,544	0,683	21	0,327	0,421
9	0,512	0,635	22	0,320	0,414
10	0,477	0,597	23	0,314	0,407
11	0,450	0,566	24	0,309	0,400
12	0,428	0,541	25	0,304	0,394
13	0,410	0,520	26	0,299	0,389
14	0,395	0,502	27	0,295	0,383
15	0,381	0,486	28	0,291	0,378
16	0,369	0,472	29	0,287	0,374
			30	0,283	0,369

8.3. Нормальний розподіл

У таблиці 8.3 наведено значення функції $\Phi(t)$ нормального розподілу з параметрами $(0; 1)$ (квантілі стандартного нормального розподілу): для заданих t табульовані значення функції

$$N_{0;1}(t) = \Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t \exp\left\{-\frac{s^2}{2}\right\} ds.$$

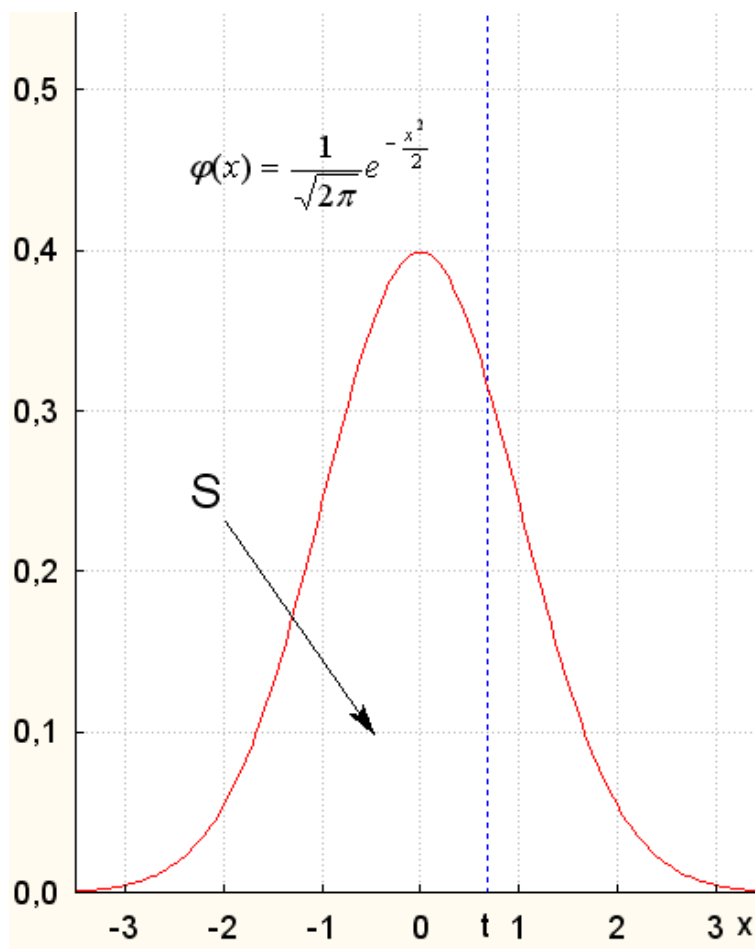


Рис. 8.1

До означення квантіля нормального розподілу;
 $\varphi(x)$ – щільність розподілу $N_{0;1}$.

Для кожного t значення $N_{0;1}(t)$ чисельно дорівнює площі S , показаної на рис. 8.1 фігури.

Значення $N_{a;\sigma^2}(x)$ – функції нормального розподілу з параметрами a і σ^2 – обчислюється за значеннями табульованої функції $N_{0;1}(t) = \Phi(t)$ нормального розподілу $N_{0;1}$:

$$N_{a;\sigma^2}(x) = N_{0;1}\left(\frac{x-a}{\sigma}\right) = \Phi\left(\frac{x-a}{\sigma}\right).$$

Таблиця 8.3 допускає лінійну інтерполяцію.

Таблиця 8.3. Значення функції $\Phi(t)$

t	0	1	2	3	4	5	6	7	8	9
-0,0	,5000	,4960	,4920	,4880	,4840	,4801	,4761	,4721	,4681	,4641
-0,1	,4602	,4562	,4522	,4483	,4443	,4404	,4364	,4325	,4286	,4247
-0,2	,4207	,4168	,4129	,4090	,4052	,4013	,3974	,3936	,3897	,3859
-0,3	,3821	,3783	,3745	,3707	,3669	,3632	,3594	,3557	,3520	,3483
-0,4	,3446	,3409	,3372	,3336	,3300	,3264	,3228	,3192	,3156	,3121
-0,5	,3085	,3050	,3015	,2981	,2946	,2912	,2877	,2843	,2810	,2776
-0,6	,2743	,2709	,2676	,2643	,2611	,2578	,2546	,2514	,2483	,2451
-0,7	,2420	,2389	,2358	,2327	,2297	,2266	,2236	,2206	,2177	,2148
-0,8	,2119	,2090	,2061	,2033	,2005	,1977	,1949	,1922	,1894	,1867
-0,9	,1841	,1814	,1788	,1762	,1736	,1711	,1685	,1660	,1635	,1611
-1,0	,1587	,1562	,1539	,1515	,1492	,1469	,1446	,1423	,1401	,1379
-1,1	,1357	,1335	,1314	,1292	,1271	,1251	,1230	,1210	,1190	,1170
-1,2	,1151	,1131	,1112	,1093	,1075	,1056	,1038	,1020	,1003	,0985
-1,3	,0968	,0951	,0934	,0918	,0901	,0885	,0869	,0853	,0838	,0823
-1,4	,0808	,0793	,0778	,0764	,0749	,0735	,0721	,0708	,0694	,0681
-1,5	,0668	,0655	,0643	,0630	,0618	,0606	,0594	,0582	,0571	,0559
-1,6	,0548	,0537	,0526	,0516	,0505	,0495	,0485	,0475	,0465	,0455
-1,7	,0446	,0436	,0427	,0418	,0409	,0401	,0392	,0384	,0375	,0367
-1,8	,0359	,0351	,0344	,0336	,0329	,0322	,0314	,0307	,0301	,0294
-1,9	,0288	,0281	,0274	,0268	,0262	,0256	,0250	,0244	,0239	,0233
-2,0	,0228	,0222	,0217	,0212	,0207	,0202	,0197	,0192	,0188	,0183
-2,1	,0179	,0174	,0170	,0166	,0162	,0158	,0154	,0150	,0146	,0143
-2,2	,0139	,0136	,0132	,0129	,0125	,0122	,0119	,0116	,0113	,0110
-2,3	,0107	,0104	,0102	,0099	,0096	,0094	,0091	,0089	,0087	,0084
-2,4	,0082	,0080	,0078	,0075	,0073	,0071	,0069	,0068	,0066	,0064
-2,5	,0062	,0060	,0059	,0057	,0055	,0054	,0052	,0051	,0049	,0048
-2,6	,0047	,0045	,0044	,0043	,0041	,0040	,0039	,0038	,0037	,0036
-2,7	,0035	,0034	,0033	,0032	,0031	,0030	,0029	,0028	,0027	,0026
-2,8	,0026	,0025	,0024	,0023	,0023	,0022	,0021	,0021	,0020	,0019
-2,9	,0019	,0018	,0018	,0017	,0016	,0016	,0015	,0015	,0014	,0014
t	-3,0	-3,1	-3,2	-3,3	-3,4	-3,5	-3,6	-3,7	-3,8	-3,9
$\Phi(t)$,0013	,0010	,0007	,0005	,0003	,0002	,0002	,0001	,0001	,0000

Таблиця 8.3 (закінчення)

t	0	1	2	3	4	5	6	7	8	9
0,0	,5000	,5040	,5080	,5120	,5160	,5199	,5239	,5279	,5319	,5359
0,1	,5398	,5438	,5478	,5517	,5557	,5596	,5636	,5675	,5714	,5753
0,2	,5793	,5832	,5871	,5910	,5948	,5987	,6026	,6064	,6103	,6141
0,3	,6179	,6217	,6255	,6293	,6331	,6368	,6406	,6443	,6480	,6517
0,4	,6554	,6591	,6628	,6664	,6700	,6736	,6772	,6808	,6844	,6879
0,5	,6915	,6950	,6985	,7019	,7054	,7088	,7123	,7157	,7190	,7224
0,6	,7257	,7291	,7324	,7357	,7389	,7422	,7454	,7486	,7517	,7549
0,7	,7580	,7611	,7642	,7673	,7703	,7734	,7764	,7794	,7823	,7852
0,8	,7881	,7910	,7939	,7967	,7995	,8023	,8051	,8078	,8106	,8133
0,9	,8159	,8186	,8212	,8238	,8264	,8289	,8315	,8340	,8365	,8389
1,0	,8413	,8438	,8461	,8485	,8508	,8531	,8554	,8577	,8599	,8621
1,1	,8643	,8665	,8686	,8708	,8729	,8749	,8770	,8790	,8810	,8830
1,2	,8849	,8869	,8888	,8907	,8925	,8944	,8962	,8980	,8997	,9015
1,3	,9032	,9049	,9066	,9082	,9099	,9115	,9131	,9147	,9162	,9177
1,4	,9192	,9207	,9222	,9236	,9251	,9265	,9279	,9292	,9306	,9319
1,5	,9332	,9345	,9357	,9370	,9382	,9394	,9406	,9418	,9429	,9441
1,6	,9452	,9463	,9474	,9484	,9495	,9505	,9515	,9525	,9535	,9545
1,7	,9554	,9564	,9573	,9582	,9591	,9599	,9608	,9616	,9625	,9633
1,8	,9641	,9649	,9656	,9664	,9671	,9678	,9686	,9693	,9699	,9706
1,9	,9713	,9719	,9726	,9732	,9738	,9744	,9750	,9756	,9761	,9767
2,0	,9772	,9778	,9783	,9788	,9793	,9798	,9803	,9808	,9812	,9817
2,1	,9821	,9826	,9830	,9834	,9838	,9842	,9846	,9850	,9854	,9857
2,2	,9861	,9864	,9868	,9871	,9875	,9878	,9881	,9884	,9887	,9890
2,3	,9893	,9896	,9898	,9900	,9904	,9906	,9909	,9911	,9913	,9916
2,4	,9918	,9920	,9922	,9925	,9927	,9929	,9931	,9932	,9934	,9936
2,5	,9938	,9940	,9941	,9943	,9945	,9946	,9948	,9949	,9951	,9952
2,6	,9953	,9955	,9956	,9957	,9959	,9960	,9961	,9962	,9963	,9964
2,7	,9965	,9966	,9967	,9968	,9969	,9970	,9971	,9972	,9973	,9974
2,8	,9974	,9975	,9976	,9977	,9977	,9978	,9979	,9979	,9980	,9981
2,9	,9981	,9982	,9982	,9983	,9984	,9984	,9985	,9985	,9986	,9986
t	3,0	3,1	3,2	3,3	3,4	3,5	3,6	3,7	3,8	3,9
$\Phi(t)$,9987	,9990	,9993	,9995	,9997	,9998	,9998	,9999	,9999	,1000

8.4. Розподіл Пірсона

У таблиці 8.4 наведено значення функції $\chi_{\alpha;n}^2$, або, що те саме, верхні α -границі розподілу Пірсона (χ^2 -розподілу) з n ступенями вільності, коротко, χ_n^2 -розподілу.

Значення $\chi_{\alpha;n}^2$ для заданих α та n визначається як розв'язок рівняння

$$\int_{\chi_{\alpha;n}^2}^{+\infty} f(x) dx = \alpha,$$

де $f(x)$ – щільність χ_n^2 -розподілу (Пірсона); $\chi_{\alpha;n}^2$ – число, що відтинає правий “хвіст” χ_n^2 -розподілу, на який припадає “маса” α (див. рис. 8.2).

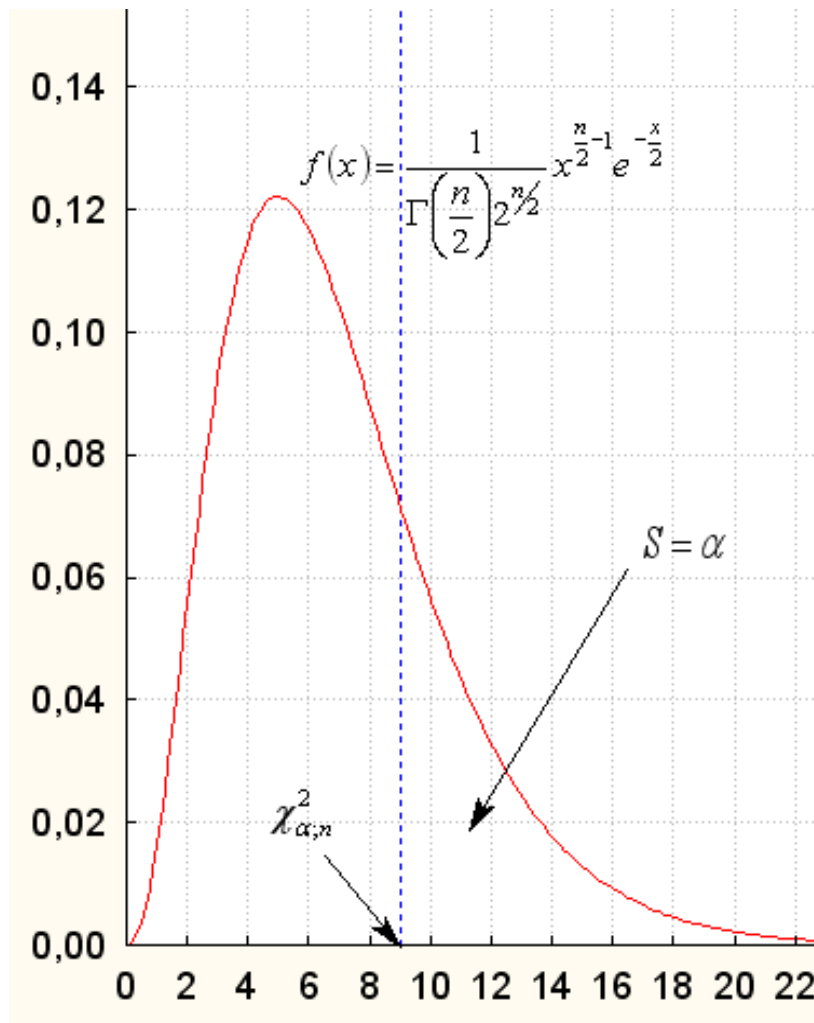


Рис. 8.2

До означення $\chi_{\alpha;n}^2$ – верхньої α –границі
розподілу Пірсона з n ступенями вільності;

$f(x)$ - щільність χ_n^2 – розподілу,

$$f(x) = \frac{1}{\Gamma\left(\frac{n}{2}\right)2^{n/2}} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, \quad x > 0, \quad \Gamma(y) = \int_0^{\infty} e^{-t} t^{y-1} dt.$$

Таблиця 8.4. Значення функції $\chi_{\alpha;n}^2$

n	Значення α							
	0,990	0,975	0,950	0,900	0,100	0,050	0,025	0,010
1	0,00	0,00	0,00	0,02	2,71	3,84	5,02	6,64
2	0,02	0,05	0,10	0,21	4,61	5,99	7,38	9,21
3	0,12	0,22	0,35	0,58	6,23	7,82	9,35	11,34
4	0,30	0,48	0,71	1,06	7,78	9,48	11,14	13,28
5	0,55	0,83	1,14	1,61	9,24	11,07	12,83	15,09
6	0,87	1,24	1,64	2,20	10,64	12,59	14,45	16,81
7	1,24	1,69	2,17	2,83	12,02	14,07	16,01	18,48
8	1,65	2,18	2,73	3,49	13,36	15,51	17,53	20,09
9	2,90	2,70	3,32	4,17	14,68	16,92	19,02	21,67
10	2,56	3,25	3,94	4,86	15,99	18,31	20,48	23,21
11	3,05	3,82	4,58	5,58	17,28	19,68	21,92	24,72
12	3,57	4,40	5,23	6,3	18,55	21,03	23,34	26,22
13	4,11	5,01	5,89	7,04	19,81	22,36	24,74	27,69
14	4,66	5,63	6,57	7,79	21,06	23,68	26,12	29,14
15	5,23	6,26	7,26	8,55	22,31	25,00	27,49	30,58
16	5,81	6,91	7,96	9,31	23,54	26,30	28,85	32,00
17	6,41	7,56	8,67	10,09	24,77	27,59	30,19	33,41
18	7,02	8,23	9,39	10,86	25,99	28,87	31,53	34,81
19	7,63	8,91	10,12	11,65	27,2	30,14	32,85	36,19
20	8,26	9,59	10,85	12,44	28,41	31,41	34,17	37,57
21	8,90	10,28	11,59	13,24	29,62	32,67	35,48	38,93
22	9,54	10,98	12,34	14,04	30,81	33,92	36,78	40,29
23	10,20	11,69	13,09	14,85	32,01	35,17	38,08	41,64
24	10,86	12,40	13,85	15,66	33,20	36,42	39,36	42,92
25	11,52	13,12	14,61	16,47	34,38	37,65	40,65	44,31
26	12,20	13,84	15,38	17,29	35,56	38,89	41,92	45,64
27	12,88	14,57	16,15	18,11	36,74	40,11	43,19	46,96
28	13,56	15,31	16,93	18,94	37,92	41,34	44,46	48,26
29	14,26	16,05	17,71	19,77	39,09	42,56	45,72	49,59
30	14,95	16,79	18,49	20,60	40,26	43,77	46,98	50,89
40	22,16	24,43	26,51	29,05	51,80	55,76	59,34	63,69
50	29,71	32,36	34,76	37,69	63,17	67,50	71,42	76,15
60	37,48	40,48	43,19	46,46	74,40	79,08	83,30	88,38
70	45,44	48,76	51,74	55,33	85,53	90,53	95,02	100,4
80	53,54	57,15	60,39	64,28	96,58	101,9	106,6	112,3
90	61,75	65,65	69,13	73,29	107,6	113,1	118,1	124,1
100	70,06	74,22	77,93	82,36	118,5	124,3	129,6	135,8

8.5. Розподіл Стюдента

У таблиці 8.5 наведено значення функції $t_{\alpha;n}$, або, що те саме, верхні α -границі розподілу Стюдента (t -розподілу) з n ступенями вільності, коротко, t_n -розподілу.

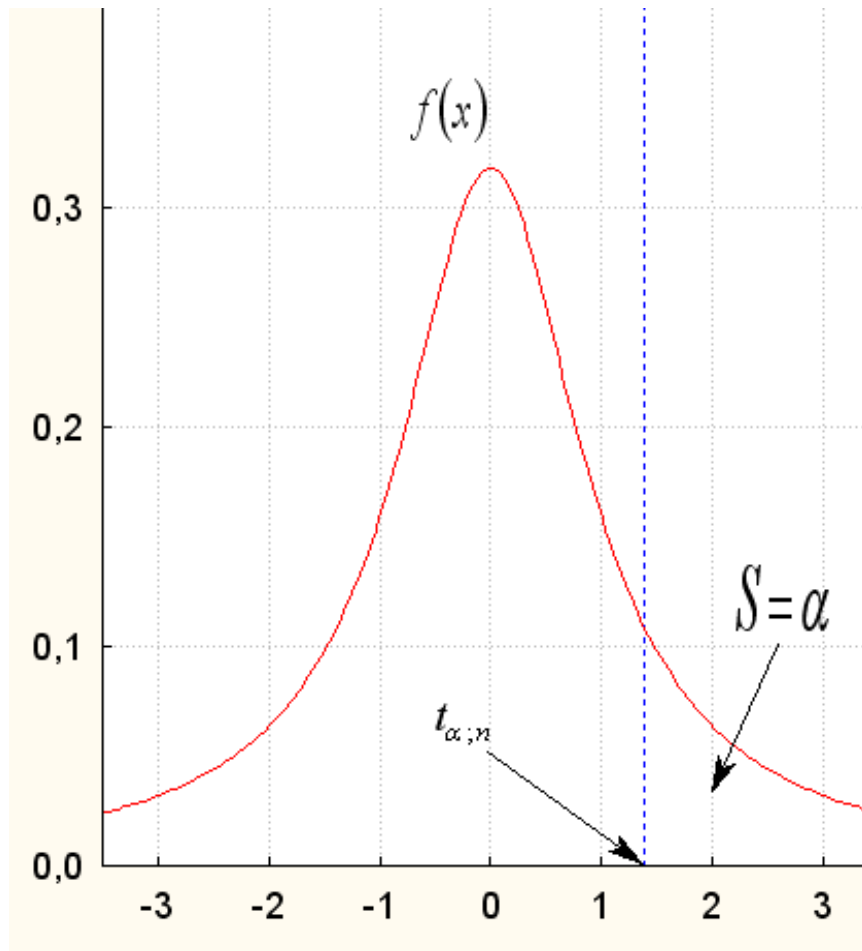


Рис. 8.3

До означення $t_{\alpha;n}$ – верхньої α -границі розподілу Стюдента з n ступенями вільності;
 $f(x)$ – щільність t_n -розподілу

Значення $t_{\alpha;n}$ для заданих α та n визначається як розв'язок рівняння

$$\int_{t_{\alpha;n}}^{+\infty} f(x) dx = \alpha,$$

де $f(x)$ – щільність t_n -розподілу (розподілу Стюдента); $t_{\alpha;n}$ – число, що відтинає правий “хвіст” t_n -розподілу, на який припадає “маса” α (див. рис. 8.3).

Щільність розподілу Стьюдента

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{\pi n}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}.$$

Таблиця 8.5. Значення функції $t_{\alpha;n}$

n	Значення α				n	Значення α			
	0,050	0,025	0,010	0,005		0,050	0,025	0,010	0,005
1	6,314	12,706	31,821	63,657	18	1,734	2,101	2,552	2,878
2	2,920	4,303	6,965	9,925	19	1,729	2,093	2,539	2,861
3	2,353	3,182	4,541	5,841	20	1,725	2,086	2,528	2,845
4	2,132	2,776	3,747	4,604	21	1,721	2,080	2,518	2,831
5	2,015	2,571	3,365	4,032	22	1,717	2,074	2,508	2,819
6	1,943	2,447	3,143	3,707	23	1,714	2,069	2,500	2,807
7	1,895	2,365	2,998	3,499	24	1,711	2,064	2,492	2,797
8	1,860	2,306	2,896	3,355	25	1,708	2,060	2,485	2,787
9	1,833	2,262	2,821	3,250	26	1,706	2,056	2,479	2,779
10	1,812	2,228	2,764	3,169	27	1,703	2,052	2,473	2,771
11	1,796	2,201	2,718	3,106	28	1,701	2,048	2,467	2,763
12	1,782	2,179	2,681	3,055	29	1,699	2,045	2,462	2,756
13	1,771	2,160	2,651	3,012	30	1,697	2,042	2,457	2,750
14	1,761	2,145	2,624	2,977	40	1,684	2,021	2,423	2,704
15	1,753	2,131	2,602	2,947	60	1,671	2,000	2,390	2,660
16	1,746	2,120	2,583	2,921	120	1,658	1,980	2,358	2,617
17	1,740	2,110	2,567	2,898	∞	1,645	1,960	2,326	2,576

8.6. Значення $t_\gamma = t(\gamma, n)$ та $q_\gamma = q(\gamma, n)$

Значення $t_\gamma = t(\gamma, n)$

n	Значення γ		
	0,95	0,99	0,999
5	2,78	4,60	8,61
6	2,57	4,03	6,86
7	2,45	3,71	5,96
8	2,37	3,50	5,41
9	2,31	3,36	5,04
10	2,26	3,25	4,78
11	2,23	3,17	4,59
12	2,20	3,11	4,44
13	2,18	3,06	4,32
14	2,16	3,01	4,22
15	2,15	2,98	4,14
16	2,13	2,95	4,07
17	2,12	2,92	4,02
18	2,11	2,90	3,97
19	2,10	2,88	3,92
20	2,093	2,861	3,883
25	2,064	2,797	3,745
30	2,045	2,756	3,659
35	2,032	2,720	3,600
40	2,023	2,708	3,558
45	2,016	2,692	3,527
50	2,008	2,679	3,502
60	2,001	2,662	3,464
70	1,996	2,649	3,439
80	1,991	2,640	3,418
90	1,987	2,633	3,403
100	1,984	2,627	3,392
120	1,980	2,617	3,374
∞	1,960	2,576	3,291

Значення $q_\gamma = q(\gamma, n)$

n	Значення γ		
	0,95	0,99	0,999
5	1,37	2,67	5,64
6	1,09	2,01	3,88
7	0,92	1,62	2,98
8	0,80	1,38	2,42
9	0,71	1,20	2,06
10	0,65	1,08	1,80
11	0,59	0,98	1,60
12	0,55	0,90	1,45
13	0,52	0,83	1,33
14	0,48	0,78	1,23
15	0,46	0,73	1,15
16	0,44	0,70	1,07
17	0,42	0,66	1,01
18	0,40	0,63	0,96
19	0,39	0,60	0,92
20	0,37	0,58	0,88
25	0,32	0,49	0,73
30	0,28	0,43	0,63
35	0,26	0,38	0,56
40	0,24	0,35	0,50
45	0,22	0,32	0,46
50	0,21	0,30	0,43
60	0,188	0,269	0,38
70	0,174	0,245	0,34
80	0,161	0,226	0,31
90	0,151	0,211	0,29
100	0,143	0,198	0,27
150	0,115	0,160	0,211
200	0,099	0,136	0,185
250	0,089	0,120	0,162

8.7. Розподіл Фішера

У таблицях 8.7.1 та 8.7.2 наведено значення функції $F_{\alpha;n;m}$, або, що те саме, верхні α -границі розподілу Фішера (F -розподілу) з n , m ступенями вільності, коротко, $F_{n;m}$ – розподілу.

Значення $F_{\alpha;n;m}$ для заданих α , n , m визначається з рівняння

$$\int_{F_{\alpha;n;m}}^{+\infty} f(x)dx = \alpha,$$

де $f(x)$ – щільність $F_{n;m}$ -розподілу (розподілу Фішера); $F_{\alpha;n;m}$ – число, що відтинає правий “хвіст” $F_{n;m}$ -розподілу, на який припадає “маса” α (див. рис. 8.4).

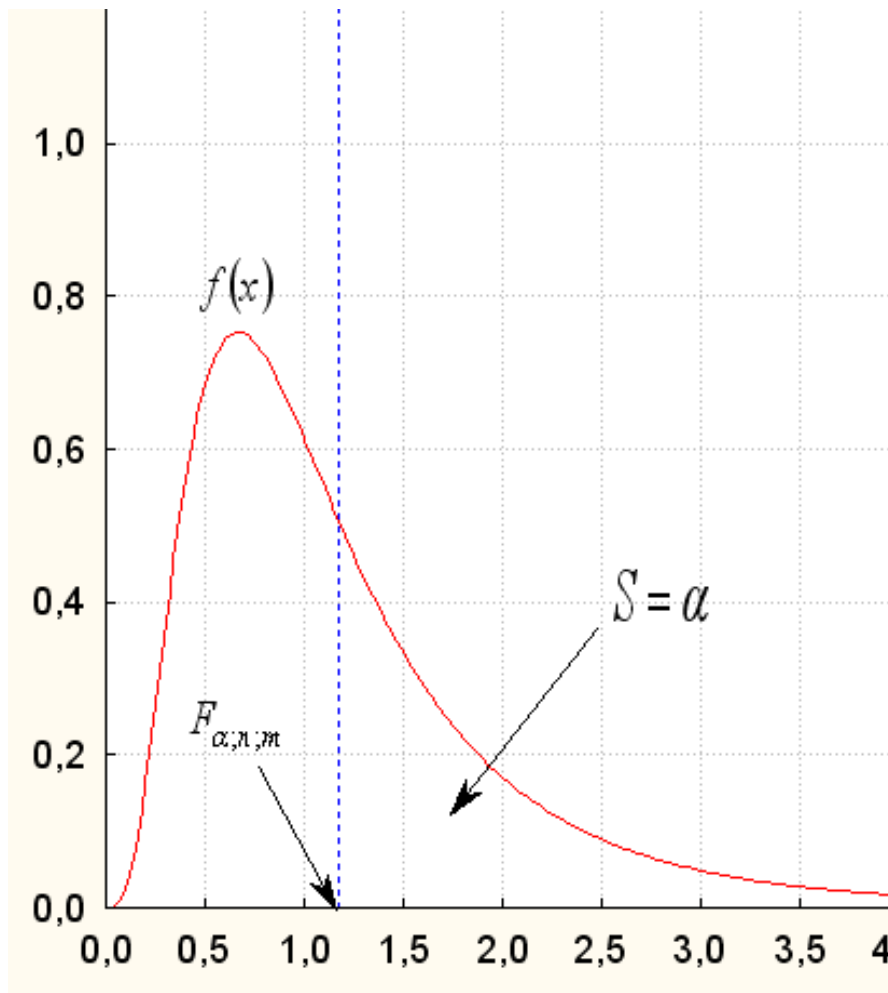


Рис.8.4

До означення $F_{\alpha;n;m}$ – верхньої α -границі розподілу Фішера з n , m ступенями вільності;
 $f(x)$ – щільність $F_{n;m}$ – розподілу

Щільність розподілу Фішера з n , m ступенями вільності

$$f(x) = \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} \left(\frac{n}{m}\right)^{\frac{n}{2}} \frac{x^{\frac{n}{2}-1}}{\left(\frac{n}{m}x+1\right)^{\frac{n+m}{2}}}, \quad x > 0.$$

Таблиця 8.7.1. Значення функції $F_{\alpha;n;m}$ (рівень значущості 0,05)

m	n (число ступенів вільності чисельника)									
	1	2	3	4	5	6	7	8	9	10
3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,57
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93

Таблиця 8.7.1 (закінчення)

<i>m</i>	<i>n</i> (число ступенів вільності чисельника)									
	12	14	16	18	20	30	40	50	60	100
3	8,74	8,71	8,69	8,67	8,66	8,62	8,59	8,58	8,57	8,55
4	5,91	5,87	5,84	5,82	5,80	5,75	5,72	5,70	5,69	5,66
5	4,68	4,64	4,60	4,58	4,56	4,50	4,46	4,44	4,43	4,41
6	4,00	3,96	3,92	3,90	3,87	3,81	3,77	3,75	3,74	3,71
7	3,57	3,53	3,49	3,47	3,44	3,38	3,34	3,32	3,30	3,27
8	3,28	3,24	3,20	3,17	3,15	3,08	3,04	3,02	3,01	2,97
9	3,07	3,03	2,99	2,96	2,94	2,86	2,83	2,80	2,79	2,76
10	2,91	2,86	2,83	2,80	2,77	2,70	2,66	2,64	2,62	2,59
11	2,79	2,74	2,70	2,67	2,65	2,57	2,53	2,51	2,49	2,46
12	2,69	2,64	2,60	2,57	2,54	2,47	2,43	2,40	2,38	2,35
13	2,60	2,55	2,51	2,48	2,46	2,38	2,34	2,31	2,30	2,26
14	2,53	2,48	2,44	2,41	2,39	2,31	2,27	2,24	2,22	2,19
15	2,48	2,42	2,38	2,35	2,33	2,25	2,20	2,18	2,16	2,12
16	2,42	2,37	2,33	2,30	2,28	2,19	2,15	2,12	2,11	2,07
17	2,38	2,33	2,29	2,26	2,23	2,15	2,10	2,08	2,06	2,02
18	2,34	2,29	2,25	2,22	2,19	2,11	2,06	2,04	2,02	1,98
19	2,31	2,26	2,21	2,18	2,16	2,07	2,03	2,00	1,98	1,94
20	2,28	2,22	2,18	2,15	2,12	2,04	1,99	1,97	1,95	1,91
22	2,23	2,17	2,13	2,10	2,07	1,98	1,94	1,91	1,89	1,85
24	2,18	2,13	2,09	2,05	2,03	1,94	1,89	1,86	1,84	1,80
26	2,15	2,09	2,05	2,02	1,99	1,90	1,85	1,82	1,80	1,76
28	2,12	2,06	2,02	1,99	1,96	1,87	1,82	1,79	1,77	1,73
30	2,09	2,04	1,99	1,96	1,93	1,84	1,79	1,76	1,74	1,70
40	2,00	1,95	1,90	1,87	1,84	1,74	1,69	1,66	1,64	1,59
50	1,95	1,89	1,85	1,81	1,78	1,69	1,63	1,60	1,58	1,52
60	1,92	1,86	1,82	1,78	1,75	1,65	1,59	1,56	1,53	1,48
100	1,85	1,79	1,75	1,71	1,68	1,57	1,52	1,48	1,45	1,39

Таблиця 8.7.2. Значення функції $F_{\alpha;n;m}$ (рівень значущості 0,01)

m	n (число ступенів вільності чисельника)									
	1	2	3	4	5	6	7	8	9	10
3	34,1	30,8	29,5	28,7	28,2	27,9	27,7	27,5	27,3	27,2
4	21,2	18,0	16,7	16,0	15,5	15,2	15,0	14,8	14,7	14,5
5	16,3	13,3	12,1	11,4	11,0	10,7	10,5	10,3	10,2	10,1
6	13,7	10,9	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87
7	12,2	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62
8	11,3	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81
9	10,6	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26
10	10,0	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80
50	7,17	5,06	4,20	3,72	3,41	3,19	3,02	2,89	2,79	2,70
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63
100	6,90	4,82	3,98	3,51	3,21	2,99	2,82	2,69	2,59	2,50

Таблиця 8.7.2 (закінчення)

<i>m</i>	<i>n</i> (число ступенів вільності чисельника)									
	12	14	16	18	20	30	40	50	60	100
3	27,1	26,9	26,8	26,8	26,7	26,5	26,4	26,4	26,3	26,2
4	14,4	14,2	14,2	14,1	14,0	13,8	13,7	13,7	13,7	13,6
5	9,89	9,77	9,68	9,61	9,55	9,38	9,29	9,24	9,20	9,13
6	7,72	7,60	7,52	7,45	7,40	7,23	7,14	7,09	7,06	6,99
7	6,47	6,36	6,27	6,21	6,16	5,99	5,91	5,86	5,82	5,75
8	5,67	5,56	5,84	5,41	5,36	5,20	5,12	5,07	5,03	4,96
9	5,11	5,00	4,92	4,86	4,81	4,65	4,57	4,52	4,48	4,42
10	4,71	4,60	4,52	4,46	4,41	4,25	4,17	4,12	4,08	4,01
11	4,40	4,29	4,21	4,15	4,10	3,94	3,86	3,81	3,78	3,71
12	4,16	4,05	4,97	3,91	3,86	3,70	3,62	3,57	3,54	3,47
13	3,96	3,86	3,78	3,72	3,66	3,51	3,43	3,38	3,34	3,27
14	3,80	3,70	3,62	3,56	3,51	3,35	3,27	3,22	3,18	3,11
15	3,67	3,56	3,49	3,42	3,37	3,21	3,13	3,08	3,05	2,98
16	3,55	3,45	3,37	3,31	3,26	3,10	3,02	2,97	2,93	2,86
17	3,46	3,35	3,27	3,21	3,16	3,00	2,92	2,87	2,83	2,76
18	3,37	3,27	3,19	3,13	3,08	2,92	2,84	2,78	2,75	2,68
19	3,30	3,19	3,12	3,05	3,00	2,84	2,76	2,71	2,67	2,60
20	3,23	3,13	3,05	2,99	2,94	2,78	2,69	2,64	2,61	2,54
22	3,12	3,02	2,94	2,88	2,83	2,67	2,58	2,53	2,50	2,42
24	3,03	2,93	2,85	2,79	2,74	2,58	2,49	2,44	2,40	2,33
26	2,96	2,86	2,78	2,72	2,66	2,50	2,42	2,36	2,33	2,25
28	2,90	2,79	2,72	2,65	2,60	2,44	2,35	2,30	2,26	2,19
30	2,84	2,74	2,66	2,60	2,55	2,39	2,30	2,25	2,21	2,13
40	2,66	2,56	2,48	2,42	2,37	2,20	2,11	2,06	2,02	1,94
50	2,56	2,46	2,38	2,32	2,27	2,10	2,01	1,95	1,91	1,82
60	2,50	2,39	2,31	2,25	2,20	2,03	1,94	1,88	1,84	1,75
100	2,37	2,26	2,19	2,12	2,07	1,89	1,80	1,73	1,69	1,60

8.8. Критерій Колмогорова

У таблиці 8.8 наведено критичні значення $\mathcal{E}_{\alpha;n}$ верхньої межі модуля різниці істинної та емпіричної функцій розподілів.

Значення $\mathcal{E}_{\alpha;n}$ для заданих α та n визначається як мінімальне \mathcal{E} , для якого

$$P\left\{\sup_x |F(x) - \hat{F}_n(x)| \geq \mathcal{E}\right\} \leq \alpha.$$

Таблиця 8.8. Критичні значення $\mathcal{E}_{\alpha;n}$ для верхньої межі модуля різниці істинної та емпіричної функцій розподілів (критерій Колмогорова)

n	Значення α			n	Значення α		
	0,05	0,02	0,01		0,05	0,02	0,01
1	0,9750	0,9900	0,9950	25	0,2640	0,2952	0,3166
2	0,8419	0,9000	0,9293	30	0,2417	0,2702	0,2899
3	0,7076	0,7846	0,8290	35	0,2243	0,2507	0,2690
4	0,6239	0,6889	0,7342	40	0,2101	0,2349	0,2520
5	0,5633	0,6272	0,6685	45	0,1984	0,2218	0,2380
6	0,5193	0,5774	0,6166	50	0,1884	0,2107	0,2260
7	0,4834	0,5384	0,5758	55	0,1798	0,2011	0,2157
8	0,4543	0,5065	0,5418	60	0,1723	0,1927	0,2067
9	0,4300	0,4796	0,5133	65	0,1657	0,1853	0,1988
10	0,4093	0,4566	0,4889	70	0,1598	0,1796	0,1917
11	0,3912	0,4367	0,4677	75	0,1544	0,1727	0,1853
12	0,3754	0,4192	0,4491	80	0,1496	0,1673	0,1795
13	0,3614	0,4036	0,4325	85	0,1452	0,1624	0,1742
14	0,3489	0,3897	0,4176	90	0,1412	0,1579	0,1694
15	0,3376	0,3771	0,4042	95	0,1375	0,1537	0,1649
20	0,2941	0,3287	0,3524	100	0,1340	0,1499	0,1608

При $n > 100$ слід користуватися асимптотичними границями

$$\mathcal{E}_{0,05;n} = \frac{1,36}{\sqrt{n}}; \quad \mathcal{E}_{0,01;n} = \frac{1,63}{\sqrt{n}},$$

для яких справжні коефіцієнти надійності навіть трохи більші від 0,95 і 0,99 відповідно.

Додаток 2. Розрахункові завдання

1. Створити файл з чотирьох чи більше стовпців реальних статистичних даних *.sta, кожен з яких міститиме не менше 12 рядків, які відповідатимуть місяцям (з січня по грудень).
2. За допомогою програми STATISTICA отримати описові статистики для кожної змінної.
3. Створити файл result.doc, куди помістити інформацію про середні значення, 0,95-надійні інтервали для середніх та коефіцієнти асиметрії та ексцесу для кожної змінної.
4. Побудувати гістограми для кожної змінної та проаналізувати їх вигляд залежно від значень коефіцієнтів асиметрії та ексцесу.
5. Визначити пару змінних, між якими найбільш виражений лінійний зв'язок, аргументувати вибір в файлі result.doc.
6. Візуалізувати цю залежність за допомогою діаграми розсіювання і зберегти графік як lin.stg.
7. У файлі *.sta додати 2 нові змінні PROGNOZ та DELTA: в стовпчику PROGNOZ перерахувати залежну змінну, використовуючи відповідне рівняння лінійної регресії з діаграми розсіювання; у стовпчику DELTA знайти різницю між залежною змінною та перерахованими значеннями.
8. Вибрати максимальне та мінімальне значення у стовпчику DELTA, показати його на діаграмі розсіювання з надійними межами.
9. Зробити висновок про узгодженість даних з лінійною моделлю.
10. Для пари змінних, між якими найбільш виражений лінійний зв'язок, у випадку узгодженості з лінійною моделлю, зробити прогноз, використовуючи рівняння регресії.
11. Результати оформити у файлі result.doc.

Список літератури

1. Боровков А.А. Теория вероятностей. – М.,1976.
2. Боровиков В.П. Популярное введение в программу STATISTICA. – М., 2001.
3. Вища математика. Спеціальні розділи . Книга 2, 2-ге вид., за ред. проф. Кулініча Г.Л. – К.: Либідь, 2003.
4. Гмурман В.Е. Теория вероятностей и математическая статистика: Уч. пособие для студентов вузов. – М.: Высш. школа, 1999.
5. Данілов В.Я., Кушніренко С.В. Теорія ймовірностей і математична статистика. Навчальний посібник. – Кам'янець-Подільський: ПП Мошак М.І., 2009. – 112 с.
6. Данілов В.Я., Кушніренко С.В. Математична статистика. Навчальний посібник. – ВГЛ"Обрії", 2012. – 152 с.
7. Жлуктенко В.І., Наконечний С.Т., Савіна С.С. Теорія ймовірностей і математична статистика. Част. 2. Математична статистика. КНЕУ. – К.,2005.
8. Кремер Н.Ш. Теория вероятностей и математическая статистика. – М., 2000.
9. Мамчич Т., Оленко А., Осипчук М., Шпортюк В. Статистичний аналіз даних з пакетом STATISTICA. – Дрогобич: Відродження, 2006.
10. Мишура Ю.С. Методические указания к изучению теории вероятностей и математической статистики. – К., изд-во КГУ, 1984.
11. Оленко А.Я. Комп'ютерна статистика. – К., ВПЦ: Київський університет, 2007.
12. Турчин В.М. Математична статистика. – К., ВЦ: Академія, 1999.
13. Чертко Н.К. Математические методы в физической географии. – Минск, изд-во: Университетское, 1987.
14. Електронний підручник: www.statsoft.ru/home/textbook

З М І С Т

Передмова	3
-----------------	---

РОЗДІЛ 1. ОСНОВИ МАТЕМАТИЧНОЇ СТАТИСТИКИ

Глава 1. Статистичні розподіли вибірок та їх числові характеристики

1.1. Поняття вибіркового обстеження	4
1.2. Дискретний статистичний розподіл вибірки та його числові характеристики	8
1.3. Інтервальний статистичний розподіл вибірки та його числові характеристики	18
1.4. Двовимірний статистичний розподіл вибірки та його числові характеристики	24
1.5. Парний статистичний розподіл вибірки та його числові характеристики	31
Контрольні питання для самоперевірки.....	34
Вправи до теми "статистичні розподіли вибірок та їх числові характеристики "	35

Глава 2. Статистичні оцінки параметрів розподілу

2.1. Постановка задачі оцінювання параметрів розподілу.....	38
2.2. Точкові статистичні оцінки параметрів розподілу	39
2.3. Методи визначення точкових статистичних оцінок параметрів генеральної сукупності.....	40
2.4. Властивості точкових оцінок для середнього та дисперсії генеральної сукупності \bar{x}_B, D_B . виправлена дисперсія, виправлене середнє квадратичне відхилення	42
2.5. Закони розподілу ймовірностей для вибіркового середнього \bar{x}_B , виправленої дисперсії S^2 , виправленого середнього квадратичного відхилення S	45
2.6. Інтервальне оцінювання	46
2.7. Надійні інтервали для параметрів нормального закону	46
2.8. Побудова довірчих інтервалів із заданою надійністю γ для дисперсії D_T та середнього квадратичного відхилення σ_T	50
2.9. Побудова довірчого інтервалу для коефіцієнта кореляції r_{xy} генеральної сукупності із заданою надійністю γ	54
2.10. Побудова довірчого інтервалу для математичного сподівання за допомогою нерівності Чебишова із заданою надійністю γ	55
Контрольні питання для самоперевірки.....	57
Вправи до теми "статистичні оцінки параметрів розподілу "	57

Глава 3. Перевірка статистичних гіпотез

3.1. Постановка задачі перевірки статистичних гіпотез	59
3.2. Схема перевірки статистичної гіпотези.....	60
3.3. Перевірка гіпотези про рівність середніх значень двох нормальних генеральних сукупностей у випадку відомих стандартних відхилень.....	61
3.4. Перевірка гіпотези про рівність середніх значень двох нормальних генеральних сукупностей у випадку невідомого стандартного відхилення	64
3.5. Перевірка гіпотези про рівність стандартних відхилень двох нормальних генеральних сукупностей у випадку невідомих параметрів розподілів	66
3.6. Перевірка гіпотези про закон розподілу за допомогою критерію χ^2	67
3.7. Критерій Колмогорова для перевірки гіпотези про закон розподілу.....	70
Контрольні питання для самоперевірки.....	72

РОЗДІЛ 2. СТАТИСТИЧНИЙ АНАЛІЗ ДАНИХ З ПАКЕТОМ STATISTICA

Глава 4. Знайомство з пакетом STATISTICA

4.1. Робота з даними в пакеті STATISTICA.....	73
4.2. Знаходження числових характеристик вибірки засобами пакету STATISTICA	78
4.3. Парна лінійна регресія.....	82
4.4. Множинна лінійна регресія	85
Контрольні завдання для самоперевірки.....	95

Глава 5. Кластерний аналіз

5.1. Мета та застосування кластерного аналізу	96
5.2. Методи кластеризації	96
5.3. Вимірювання відстані між об'єктами.....	98
5.4. Вимірювання відстані між кластерами.....	99
5.5. Кластерний аналіз у програмі STATISTICA	100
Контрольні питання для самоперевірки	108

Глава 6. Елементи аналізу часових рядів

6.1. Поняття часового ряду	109
6.2. Виділення компонент часового ряду	110
6.3. Побудова моделі часового ряду	112
6.4. Ідентифікація порядку моделі часового ряду	113
6.5. Оцінювання і прогноз	114
6.6. Сезонні моделі	115

6.7. Аналіз часового ряду у програмі STATISTICA	116
Контрольні питання для самоперевірки	127
Глава 7. Визначення об'єму вибірки	
7.1. Показники точності оцінювання	128
7.2. Визначення об'єму вибірки n при оцінюванні часток P	128
7.3. Об'єм вибірки при дослідженні декількох ознак	131
7.4. Визначення об'єму вибірки при оцінюванні середніх і сумарних значень	133
7.5. Об'єм вибірки за необхідності отримати оцінки для підрозділів сукупності	131
Контрольні питання для самоперевірки	136
Додаток 1. Таблиці математичної статистики	
8.1. Таблиця великих чисел	137
8.2. Значення критерію τ в залежності від об'єму вибірки N і рівня значущості α	137
8.3. Нормальний розподіл	138
8.4. Розподіл Пірсона	141
8.5. Розподіл Стьюдента	143
8.6. Значення $t_\gamma = t(\gamma, n)$ та $q_\gamma = q(\gamma, n)$	145
8.7. Розподіл Фішера	146
8.8. Критерій Колмогорова	151
Додаток 2. Розрахункові завдання	152
Список літератури	153