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

"НАУКОВА РОБОТА ЗА ТЕМОЮ МАГІСТЕРСЬКОЇ ДИСЕРТАЦІЇ-2. НАУКОВА РОБОТА ЗА ТЕМОЮ МАГІСТЕРСЬКОЇ ДИСЕРТАЦІЇ"

Практикум

Рекомендовано Методичною радою КПІ ім. Ігоря Сікорського
як навчальний посібник для здобувачів ступеня магістра
за освітньою програмою «Літаки і вертольоти»
спеціальності 134 АВІАЦІЙНА ТА РАКЕТНО-КОСМІЧНА ТЕХНІКА

Укладач В. В. Кабанячий, д.т.н.

Електронне мережне навчальне видання

Київ
КПІ ім. Ігоря Сікорського
2022

Наукова робота за темою магістерської дисертації - 2. Наукова робота за темою магістерської дисертації. Практикум [Електронний ресурс]: навч. посіб. для студ. спеціальності 134 «Авіаційна та ракетно-космічна техніка» / КПІ ім. Ігоря Сікорського; уклад.: В. В. Кабанячий. – Електронні текстові дані (1 файл: 1,98 Мбайт). – Київ : КПІ ім. Ігоря Сікорського, 2022. – 145 с.

Гриф надано Методичною радою КПІ ім. Ігоря Сікорського (протокол № 6 *від* 24.06.2022 *р.)*
за поданням Вченої ради навчально-наукового інституту аерокосмічних технологій (протокол № 5/22 *від* 31.05.2022 *р.)*

Електронне мережне навчальне видання

"НАУКОВА РОБОТА ЗА ТЕМОЮ МАГІСТЕРСЬКОЇ ДИСЕРТАЦІЇ-2. НАУКОВА РОБОТА ЗА ТЕМОЮ МАГІСТЕРСЬКОЇ ДИСЕРТАЦІЇ "

Практикум

Укладачі: Кабанячий Володимир Володимирович, д-р техн. наук

Відповідальний редактор: Сухов В.В. д-р техн. наук, професор

Рецензент: Пономаренко С. О., кандидат техн. наук, с.н.с., в. о. завідувача кафедри систем керування літальними апаратами

У навчальному посібнику викладено теоретичні відомості, методичні матеріали та вихідна інформація для виконання практичних занять з курсу «Наукова робота за темою магістерської дисертації - 2. Наукова робота за темою магістерської дисертації». У посібнику приділено увагу дослідженню й визначенню як закону розподілу незалежних змінних величин та їхніх числових характеристик, так і коефіцієнтів регресійної моделі процесу, діапазони змін параметрів якого відомі.

Навчальний посібник призначений для здобувачів ступеня магістра за спеціальністю 134 Авіаційна та ракетно-космічна техніка. Він може бути також корисним для здобувачів ступеня магістра інших технічних спеціальностей.

Загальні методичні вказівки

У процесі вивчення курсу «Наукова робота за темою магістерської дисертації - 2. Наукова робота за темою магістерської дисертації» студент повинен не лише засвоїти теоретичний матеріал, але й отримати теоретичні й практичні знання щодо методології та методики наукових досліджень, планування та організації наукових досліджень, методів аналізу та прийняття рішень за результатами досліджень. Зокрема, це дозволить їм самостійно ставити та творчо вирішувати різні складні питання наукових досліджень.

В аналізі й висновках підводять підсумки проведених розрахунків. Їх формулюють у вигляді окремих лаконічних і, головне, конкретних положень, які підсумовують результати проведених розрахунків. В аналіз й висновки можуть бути включені узагальнені цифрові дані. Висновки повинні містити відповідь на питання, що були сформульовані у меті практичного заняття.

Виконання практичних занять вимагає наявності у студентів навичок користування персональними комп'ютерами, які здобуваються у процесі засвоєння шкільної та загальнотеоретичної університетської програм.

Студент допускається до проведення наступного практичного заняття лише у тому випадку оформлення протоколу попередньої роботи і вивчення детально поточного практичного заняття.

Метою практичних занять є опанування методології наукових досліджень.

Переважним для практичного заняття є його виконання у спеціалізованому класі, обладнаному персональними комп’ютерами.

На кожну практичну роботу виділяється по 9 навчальних годин. Крім того, підготовча частина роботи виконується студентами вдома, у процесі підготовки до чергової навчальної пари.

ЗМІСТ

стор.

Практична робота №1

ДОСЛІДЖЕННЯ Й ВИЗНАЧЕННЯ ЗАКОНУ РОЗПОДІЛУ НЕЗАЛЕЖНИХ ЗМІННИХ ВЕЛИЧИН ТА ЇХНІХ ЧИСЛОВИХ ХАРАКТЕРИСТИК

Мета роботи – *вивчення детермінованих і випадкових процесів, основних понять теорії ймовірностей, класифікації детермінованих і випадкових процесів, дослідження незалежних змінних величин, визначення закону розподілу незалежних змінних величин і їхніх числових характеристик, здійснення аналізу проведених розрахунків.*

1.1 Детерміновані і випадкові процеси

Будь-які дані, отримані в результаті спостереження реального фізичного явища, можна віднести до детермінованого або випадкового типу. Детерміновані процеси - це процеси, які можна описати явними математичними формулами. На практиці часто зустрічаються фізичні явища, протікання яких можна з достатньою точністю описати математичними залежностями. Наприклад, зміна температури води в міру її нагрівання за своєю суттю є детермінованим процесом. Однак багато інших фізичних явищ породжують процеси, які не можна вважати детермінованими. Наприклад, висота хвиль при вітровому хвилюванні. Абсолютно неможливо передбачити точне значення такого процесу в майбутні миті часу. Цей процес випадковий по своїй суті, і для його опису потрібні ймовірнісні поняття та статистичні характеристики.

Віднесення тих чи інших фізичних процесів до детермінованого або випадкового типу часто не безперечно. Наприклад, можна стверджувати, що фізичні процеси, котрі зустрічаються на практиці, взагалі не можуть бути повною мірою детермінованими, оскільки ніколи не можна виключити можливості того, що в майбутньому відбудеться яка-небудь подія, яка вплине на явище, що породжує процес, абсолютно непередбачуваним чином. З іншого боку, можна стверджувати, що немає й дійсно випадкових процесів, оскільки може виявитися, що при досить повному вивченні основних механізмів явища породженим цим явищем процес вдасться описати точними математичними формулами. З практичної точки зору рішення про випадковість або детермінованість конкретного фізичного процесу зазвичай ґрунтується на нашій здатності відтворити процес в ході контрольованого експерименту. Якщо багаторазове повторення експерименту, в ході якого отримується процес, що цікавить нас, призводить до одних і тих же результатів (в межах помилки експерименту), то цей процес зазвичай можна вважати детермінованим. Якщо ж неможливо вказати експеримент, який давав би при його повторенні ідентичні результати, то такий процес зазвичай вважається випадковим по своїй суті.

1.2 Класифікація детермінованих процесів

Процеси, що описують детерміновані явища, поділяються на періодичні і неперіодичні. У свою чергу періодичні процеси можна підрозділити на гармонійні та полігармонічні. Неперіодичні процеси поділяються на «майже періодичні» і перехідні.

1.2.1 Синусоїдальний періодичний процес. Синусоїдальний процес - це періодичний процес, поведінка якого в часі математично виражається такою формулою:

$$x(t) = X \sin (2 \pi f_0 t + \theta), \qquad (1.1)$$

де $x(t)$ - значення в час t,

X - амплітуда,

f_0 - циклічна частота, Гц,

θ - початковий фазовий кут, рад.

Синусоїдальний процес, визначений формулою (1.1), називається гармонійним. При практичному аналізі гармонійних процесів фазовий кут θ часто ігнорується. В цьому випадку:

$$x(t) = X \sin 2 \pi f_0 t \quad (1.2)$$

Рівняння (1.2) графічно можна зобразити або у вигляді залежності поточного значення від часу, або у вигляді залежності амплітуди від частоти (частотного спектра).

Інтервал часу, протягом якого відбувається одне повне коливання або цикл гармонійного процесу, називається періодом T_p. Число циклів в одиницю часу називається частотою f_0. Частота і період пов'язані співвідношенням:

$$T_p = 1/ f_0. \quad (1.3)$$

Зауважимо, що частотний спектр складається з єдиною амплітуди, розташованої на певній частоті, і цим відрізняється від спектрів, які задають безперервну залежність амплітуди від частоти. Такі спектри називаються дискретними або лінійчатими.

Гармонійні процеси є з точки зору аналізу найпростішим видом процесів, що протікають у часі.

1.2.2 Полігармонічні процеси. До полігармонічних процесів відносяться періодичні процеси, які математично представляються функцією часу, що точно повторює свої значення через однакові інтервали часу, тобто:

$$x(t) = x(t \pm n\ T_p),\ n = 1, 2, 3, \ldots \qquad (1.4)$$

Як і в випадку гармонійних процесів, інтервал часу, протягом якого відбувається одне повне коливання, називається періодом T_p Число циклів в одиницю часу називається фундаментальною частотою f_1. Гармонійні процеси є окремим випадком полігармонічні процесів при $f_1 = f_0$.

За рідкісними на практиці винятками, полігармонічні процеси розкладаються в ряд Фур'є за формулою:

$$x(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty}\left(a_n \cos 2\pi f_1 t + b_n \sin 2\pi f_1 t\right), \qquad (1.5)$$

де

$$f_1 = 1/\ T_p,$$

$$a_n = \frac{2}{T_p}\int_0^{T_p} x(t)\cos 2\pi f_1 t dt, \quad n = 0,1,2,\ldots,$$

$$b_n = \frac{2}{T_p}\int_0^{T_p} x(t)\sin 2\pi f_1 t dt, \quad n = 1,2,3,\ldots.$$

Інше представлення полігармонічних процесів рядом Фур'є дає формула:

$$x(t) = X_0 + \sum_{n=1}^{\infty} X_n \cos(2\pi f_1 t - \vartheta_n), \qquad (1.6)$$

де

$$X_0 = a_0 / 2,$$

$$X_n = \sqrt{a_n^2 + b_n^2}, \quad n = 1,2,3,\ldots,$$

$$\vartheta_n = arctg \frac{b_n}{a_n}, \quad n = 1,2,3,\ldots.$$

інакше кажучи, формула (1.6) стверджує, що полігармонічний процес є сума постійної складової X_0 і нескінченного числа гармонійних складових, які називаються гармоніками і мають амплітуди X_n і фази θ_n. Всі частоти гармонійних складових кратні фундаментальної частоті.

При практичному аналізі періодичних процесів фазові кути θ_n часто ігноруються. У цьому випадку формулу можна охарактеризувати дискретним спектром. Іноді полігармонічні процеси містять лише кінцеве число складових. В інших випадках може бути відсутня фундаментальна складова.

Фізичні явища, що описуються полігармонічними процесами, зустрічаються набагато частіше, ніж явища, які породжують прості гармонійні процеси. Фактично багато полігармонічних процесів розглядаються як прості тільки наближено.

1.2.3 Майже періодичні процеси. Будь-який процес, утворений сумою двох і більше гармонійних процесів з порівнянними частотами, буде

періодичним. Однак якщо процес утворений сумою двох і більше гармонійних процесів з довільними частотами, то він, як правило, не буде періодичним. Точніше кажучи, сума двох і більше гармонійних процесів буде періодичним процесом тоді і лише тоді, коли відношення будь-яких двох частот є число раціональне. У цьому випадку існує фундаментальний період, що задовольняє рівняння (1.4). В іншому випадку, реалізація такого процесу носить майже періодичний характер, але співвідношення (1.4) не виконується ні при якому кінцевому значенні T_p.

На підставі цих міркувань майже періодичні процеси визначаються математично як функція часу виду:

$$x(t) = \sum_{n=1}^{\infty} X_n \sin(2\pi f_n t + \vartheta_n), \qquad (1.7)$$

Причому f_n/f_m не для всіх значень індексів є раціональними числами. На практиці майже періодичні процеси породжуються фізичними явищами, в яких одночасно діють гармонійні процеси, не пов'язані між собою. Хороший приклад дає вібрація багатомоторного гвинтового літака, в якому двигуни не синхронізовані.

Майже періодичні процеси мають таку важливу властивість. Якщо виключити з розгляду фазові кути θ_n, то формулу (1.7) можна охарактеризувати дискретним спектром, подібним спектру полігармонічного процесу. Єдина відмінність полягає в тому, що відношення частот складових не є раціональними числами (рис. 1.1).

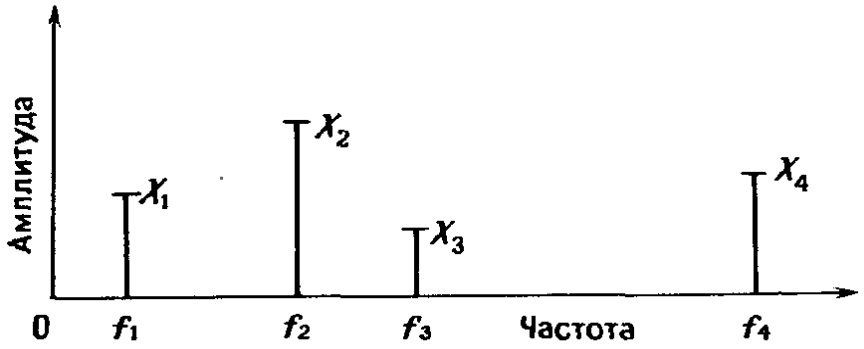

Рис. 1.1. Спектр майже періодичного процесу

1.2.4 Перехідні неперіодичні процеси. За визначенням, перехідні процеси - це всі неперіодичні процеси, за винятком майже періодичних процесів. До перехідних процесів належать численні і найрізноманітніші явища. Важлива особливість перехідних процесів, що відрізняє їх від періодичних і майже періодичних, полягає в тому, що їх не можна охарактеризувати дискретним спектром. У більшості випадків для перехідних процесів можна отримати безперервне спектральне подання, використовуючи перетворення Фур'є виду:

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft} dt . \qquad (1.8)$$

Взагалі кажучи, перетворення Фур'є є комплексною величиною, яка записується в полярній формі:

$$X(f) = |X(f)| e^{-j\vartheta(f)},$$

де $|X(f)|$ — модуль $X(f)$,

$\theta(f)$ - аргумент.

На рис. 1.3 наведені модулі безперервних спектрів трьох перехідних процесів, зображених на рис. 1.2.

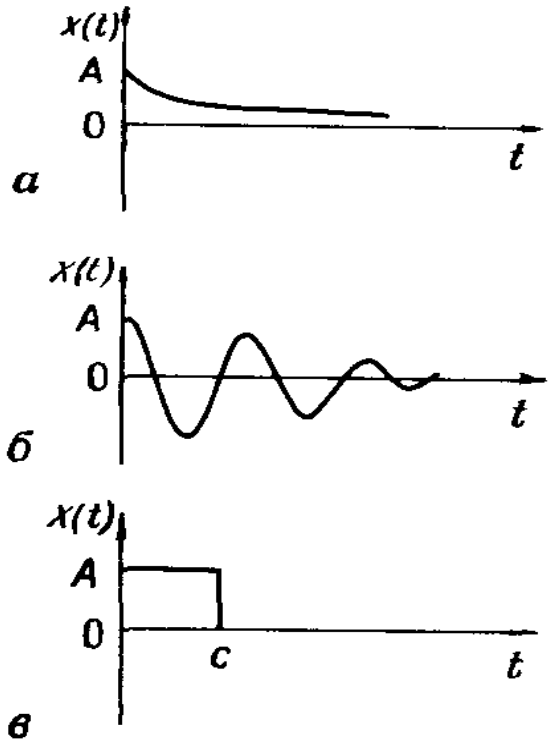

Рис. 1.2. Приклади перехідних процесів

1.3 Класифікація випадкових процесів

Як вже говорилося, процес, що описує випадкове фізичне явище, не можна задати явною математичною залежністю, оскільки кожне спостереження цього явища дає невідтворюваний результат. Іншими словами, будь-яке спостереження дає тільки один варіант з безлічі можливих.

Конкретна реалізація процесу, що описує випадкове явище, називається вибіркової функцією (або реалізацією, якщо мова йде про спостереження кінцевої тривалості). Сукупність усіх можливих вибіркових

функцій, які може дати випадкове явище, називається випадковим чи стохастичним процесом. Отже, під реалізацією випадкового фізичного явища розуміється один з можливих результатів випадкового процесу.

Випадкові процеси поділяються на стаціонарні та нестаціонарні. У свою чергу стаціонарні випадкові процеси поділяються на ергодичні і неергодичні. Подальша класифікація нестаціонарних випадкових процесів проводиться за особливостями їхньою нестаціонарності.

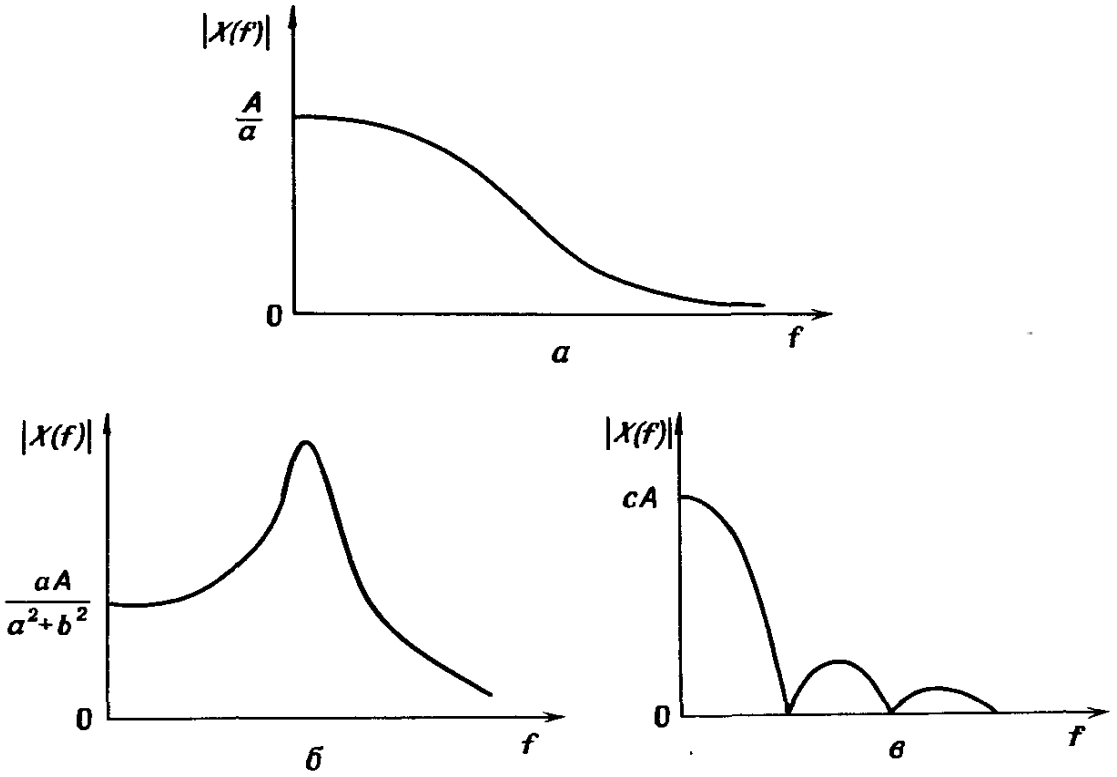

Рис. 1.3. Спектри перехідних процесів

1.3.1 Стаціонарні випадкові процеси. Якщо фізичне явище описується випадковим процесом, то властивість цього явища в принципі можна оцінити в будь-який час шляхом усереднення за сукупністю вибіркових функцій, що утворюють випадковий процес. Розглянемо,

наприклад, сукупність вибіркових функцій (звану також ансамблем), що визначають випадковий процес, зображений на рис. 1.4.

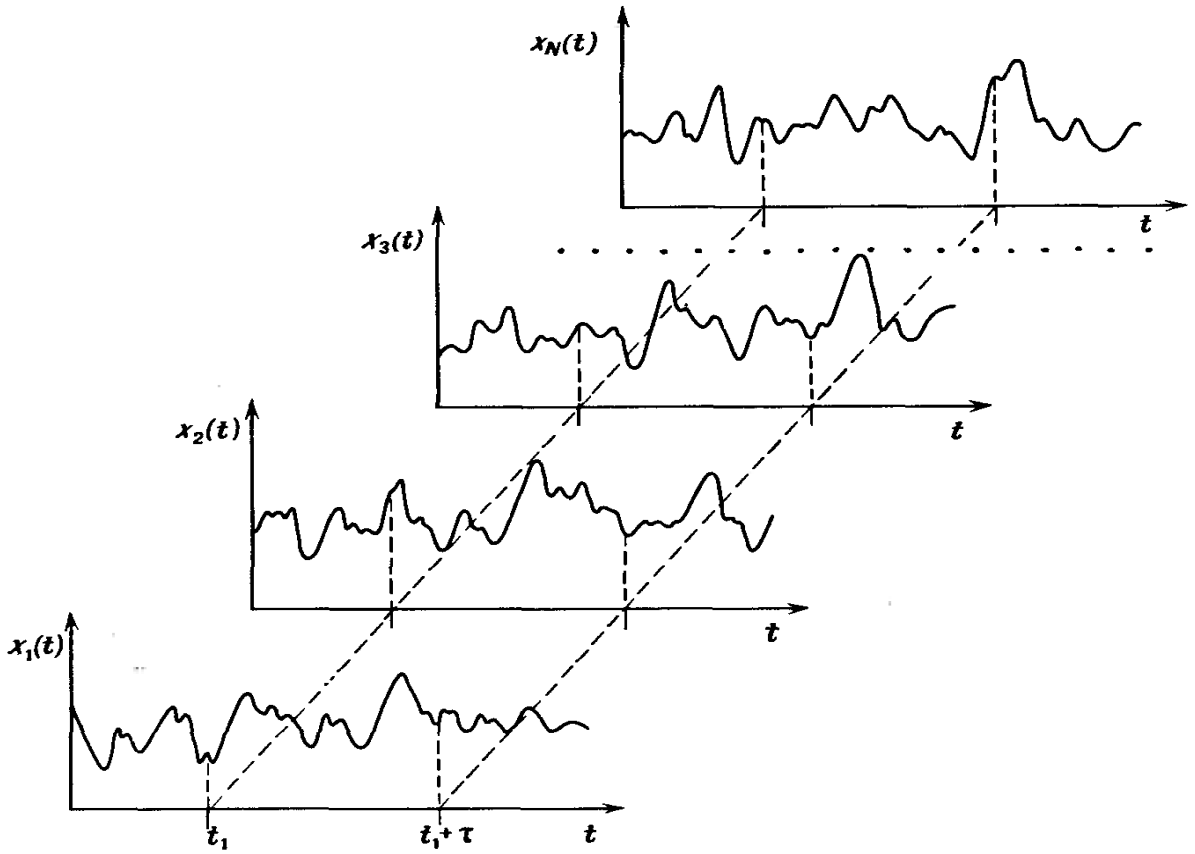

Рис. 1.4. Ансамбль реалізацій, які задають випадковий процес

.

Середнє значення (перший момент) цього випадкового процесу у час t_1 можна обчислити, взявши миттєві значення всіх вибіркових функцій ансамблю в час t_1, склавши ці значення і розділивши на число вибіркових функцій. Аналогічним чином коваріацію (змішаний момент) значень випадкового процесу у два різні часи (ця величина називається коваріаційною функцією) обчислюється шляхом усереднення по ансамблю здобутків значень в часи t_1 і $t_1 + \tau$. Отже, середнє значення $\mu_x(t_1)$ і

варіаційна функція $R_{xx}(t_1,\ t_1 + \tau)$ випадкового процесу $\{x(t)\}$, де {} позначає ансамбль вибіркових функцій, визначається формулами:

$$\mu_x(t_1) = \lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{\infty} x_k(t_1), \qquad (1.9а)$$

$$R_{xx}(t_1, t_1 + \tau) = \lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{\infty} x_k(t_1) x_k(t_1) x_k(t_1 + \tau), \quad (1.9б)$$

в яких підсумовування проводиться в припущенні рівноймовірності всіх вибіркових функцій.

У загальному випадку, коли $\mu_x(t_1)$ і $R_{xx}(t_1,\ t_1 + \tau)$ випадкового процесу $\{x(t)\}$, де {} позначає ансамбль вибіркових функцій, , що визначаються рівнянням (1.9), залежать від часу t_1, випадковий процес {x(t)} називається нестаціонарним. У тому окремому випадку, коли $\mu_x(t_1)$ і $R_{xx}(t_1,\ t_1 + \tau)$, не залежать від часу t_1, випадковий процес називається слабо стаціонарним або стаціонарним в широкому сенсі. Середнє значення слабо стаціонарного процесу постійно, а коваріаційна функція залежить тільки від зсуву часу τ, тобто

$$\mu_x(t_1) = \mu_x,$$
$$R_{xx}(t_1,\ t_1 + \tau) = R_{xx}(\tau).$$

Для визначення повного набору функцій розподілу, які задають структуру випадкового процесу $\{x(t)\}$, потрібно обчислити нескінченне число моментів і змішаних моментів вищих порядків. У тому випадку, коли всі моменти і змішані моменти інваріантні в часі, випадковий процес $\{x(t)\}$ називається строго стаціонарним або стаціонарним у вузькому сенсі.

У багатьох випадках перевірка слабкої стаціонарності дозволяє обґрунтувати сувору стаціонарність.

1.3.2 Ергодичні стаціонарні процеси. Характеристики випадкового процесу можна визначити шляхом усереднення по ансамблю в певні миті часу. Однак в більшості випадків характеристики стаціонарного випадкового процесу можна обчислити, усереднюючи за часом в межах окремих вибіркових функцій, що входять в ансамбль. Візьмемо, наприклад, *k*-у вибіркову функцію ансамблю, зображеного на рис. 1.4. Середнє значення $\mu_x(k)$ і коваріаційна функція $R_{xx}(\tau, k)$, обчислені по *k*-й реалізації дорівнюють:

$$\mu_x(k) = \lim_{T \to \infty} \frac{1}{T} \int_0^T x_k(t) dt , \qquad (1.10а)$$

$$R_{xx}(\tau, k) = \lim_{T \to \infty} \frac{1}{T} \int_0^T x_k(t) x_k(t+\tau) dt . \qquad (1.10б)$$

Якщо випадковий процес $\{x(t)\}$ стаціонарний, а $\mu_x(k)$ й $R_{xx}(\tau, k)$, обчислені за різними реалізаціями відповідно до формул (1.10), збігаються, то випадковий процес називається ергодичним. Для ергодичних процесів середні значення і коваріантні функції, отримані усередненням за часом (як і інші характеристики, обчислені усередненням за часом), дорівнюють аналогічним характеристикам, знайденим усередненням за ансамблем, тобто

$$\mu_x(k) = \mu_x,$$
$$R_{xx}(\tau, k) = R_{xx}(\tau).$$

Відзначимо, що властивість ергодичності можуть мати тільки стаціонарні процеси.

Очевидно, що ергодичні випадкові процеси утворюють дуже важливий клас випадкових процесів, оскільки всі властивості ергодичних процесів можна визначити за єдиною вибірковою функцією. На щастя, на практиці стаціонарні випадкові процеси зазвичай виявляються ергодичними. Саме з цієї причини властивості стаціонарних випадкових явищ часто можна визначити за однією спостереженою реалізацією.

1.3.3 Нестаціонарні випадкові процеси. До нестаціонарних процесів можна віднести випадкові процеси, що не відповідають умовам стаціонарності. Якщо не накладені додаткові обмеження, то властивості нестаціонарних випадкових процесів зазвичай залежать від часу і можуть бути встановлені тільки шляхом усереднення в окремі часи по ансамблю вибіркових функцій, що утворюють процес. На практиці часто не вдається отримати необхідне для точної оцінки властивостей процесу число реалізацій. Цим фактом пояснюється відставання в розвитку практичних методів вимірювання та аналізу нестаціонарних випадкових процесів.

У багатьох випадках нестаціонарні випадкові процеси, що відповідають реальним фізичним явищам, мають особливості, які спрощують їхній аналіз та вимір. Наприклад, іноді випадкові дані вдається представити у вигляді випадкового процесу $\{x(t)\}$, всі вибіркові функції якого мають вигляд:

$$x(t) = a(t)\, u(t),$$

де $u(t)$ - вибіркова функція стаціонарного випадкового процесу $\{u(t)\}$,

$a(t)$ - детермінована функція.

Іншими словами, дані представляються нестаціонарним випадковим процесом, всі вибіркові функції якого мають загальний детермінований тренд. Якщо нестаціонарний випадковий процес має такий вигляд, то для опису його властивостей не завжди потрібно усереднення по ансамблю. Іноді багато важливих властивості вдається оцінити за єдиною вибірковою функцією, як і у випадку ергодичних стаціонарних процесів.

1.3.4 Стаціонарність вибіркових функцій. Поняття «стаціонарність» відноситься до середніх по ансамблю властивостей випадкового процесу. Однак на практиці часто кажуть про стаціонарність або нестаціонарність даних, що представляють собою єдину реалізацію випадкового явища. В цьому випадку стаціонарність розуміється в дещо іншому сенсі. Якщо про єдину реалізацію говорять як про стаціонарну, то зазвичай мають на увазі, що її властивості, певні на коротких інтервалах часу, суттєво не змінюються від інтервалу до інтервалу. Слово «істотно» тут означає, що спостерігаються коливання перевершують відхилення, які можна пояснити звичайною вибірковою мінливістю статистичних оцінок.

Щоб пояснити це міркування, розглянемо одну реалізацію $x_k(t)$, що є k-ю вибірковою функцією випадкового процесу $\{x(t)\}$. Нехай середнє значення і коваріаційна функція оцінені по невеликому інтервалу довжиною T з початком в точці t_1, тобто

$$\mu_x(t_1,k)=\frac{1}{T}\int_{t_1}^{t_1+T}x_k(t)dt\,, \qquad (1.11а)$$

$$R_{xx}(t_1,t_1+\tau,k)=\frac{1}{T}\int_{t_1}^{t_1+T}x_k(t)x_k(t+\tau)dt\,. \qquad (1.11б)$$

У загальному випадку, коли вибіркові величини, що визначаються формулами (1.11), сильно змінюються зі зміною початкового часу t_1, вибіркова функція називається нестаціонарною. У тому випадку, коли вибіркові властивості величин (1.11) не змінюються суттєво зі зміною початкового часу t_1, реалізація називається стаціонарною. Зауважимо, що реалізація ергодичного випадкового процесу стаціонарна. У той же час вибіркові функції більшості практично цікавих нестаціонарних випадкових процесів не стаціонарні. Тому при вивченні властивості ергодичності (що справедливо для більшості стаціонарних фізичних явищ) перевірка стаціонарності однієї реалізації є ефективним методом перевірки в цілому припущення стаціонарності і ергодичності випадкового процесу, з якого ця реалізація отримана.

1.4 Аналіз випадкових даних

Аналіз випадкових даних заснований на інших міркуваннях, ніж аналіз детермінованих даних. Зокрема, в силу того, що реалізацію випадкового процесу можна задати явною математичною формулою, для оцінки властивостей таких даних використовуються статистичні методи. Проте, випадкові процеси задовольняють цілком певні міркування, що описує перетворення цих процесів; ці співвідношення відіграють ключову роль у багатьох випадках. У таких випадках важливо вміти виявляти і враховувати статистичні помилки, пов'язані з оцінкою і співвідношеннями між вхідними та вихідними процесами перетворень.

1.4.1 Основні характеристики випадкових процесів. Основні статистичні характеристики, що мають важливе значення для опису властивостей окремих реалізацій стаціонарних випадкових процесів, такі:

1. середні значення і середні квадрати;
2. щільності ймовірності;
3. коваріаційні функції;
4. функції спектральної щільності.

Середнє значення μ_x і дисперсія s_x^2 стаціонарної реалізації випадкового процесу характеризують центр розсіювання і величину розсіювання даних. Середній квадрат ψ_x^2, який дорівнює сумі дисперсії і квадрата середнього значення, є мірою того й іншого одночасно. Середнє значення оцінюється простим усередненням всіх значень реалізації. Аналогічним усередненням квадратів значень реалізації оцінюється середній квадрат. Якщо перед зведенням в квадрат зі значень реалізації віднімати середнє значення, то таке усереднення дасть оцінку дисперсії.

Щільність ймовірності $p(x)$ стаціонарної реалізації задає швидкість зміни ймовірності залежно від значення реалізації. Функція $p(x)$ зазвичай оцінюється шляхом обчислення ймовірності того, що миттєве значення окремої реалізації укладається у вузькому інтервалі, центр якого пробігає область значень процесу, з подальшим розподілом на ширину інтервалу. Загальна площа, обмежена графіком щільності ймовірності по всій його області визначення, дорівнює одиниці, що просто свідчить про достовірність події, яка полягає в тому, що значення реалізації містяться між $-\infty$ та $+\infty$. Частина цієї площі, що лежить лівіше даного значення x, визначає функцію розподілу, яка позначається $P(x)$. Частина площі, обмежена графіком щільності між двома довільними значеннями x_1 й x_2 і

дорівнює $P(x_2)$ - $P(x_1)$, задає ймовірність того, що значення реалізації в навмання вибраний час потраплять в цей інтервал значень процесу.

Коваріаційна функція $R_{xx}(\tau)$ стаціонарного процесу задає міру залежності його значень, зсунутих відносно один одного на однакові інтервали часу. Щоб оцінити коваріаційну функцію, слід зрушити реалізацію на час, перемножити вихідну і зсунену реалізації і усереднити отримані добутки по всій реалізації або по деякому її відрізку. Ця процедура виконується для всіх необхідних значень зсуву часу.

Спектральна щільність (інакше, спектр потужності) $G_{xx}(f)$ стаціонарної реалізації задає швидкість зміни середнього квадрата у залежності від частоти. Для оцінювання спектра обчислюється середній квадрат у вузькій смузі частот при різних центральних частотах, а потім отримане значення поділяється на ширину цієї смуги. Загальна площа, що лежить під графіком спектральної щільності по всій смузі частот, дорівнює сумарному квадрату реалізації. Частина цієї площі, яка знаходиться між частотами f_1 і f_2, дорівнює середньому квадрату, зосередженому у цій смузі частот.

На рис. 1.5 показані типові реалізації гармонійного процесу, гармонійного процесу при випадковому шумі, вузько-смугового шуму і широкосмугового шуму. На рис. 1.6-1.8 наведені обчислені теоретично відповідно щільності ймовірності, коваріаційні функції і спектральні щільності цих процесів.

Для пар реалізацій, що належать двом різним стаціонарним випадковим процесам, важливе значення мають спільні статистичні характеристики, а саме:

1. спільні щільності ймовірності;

2. взаємні коваріаційні функції;

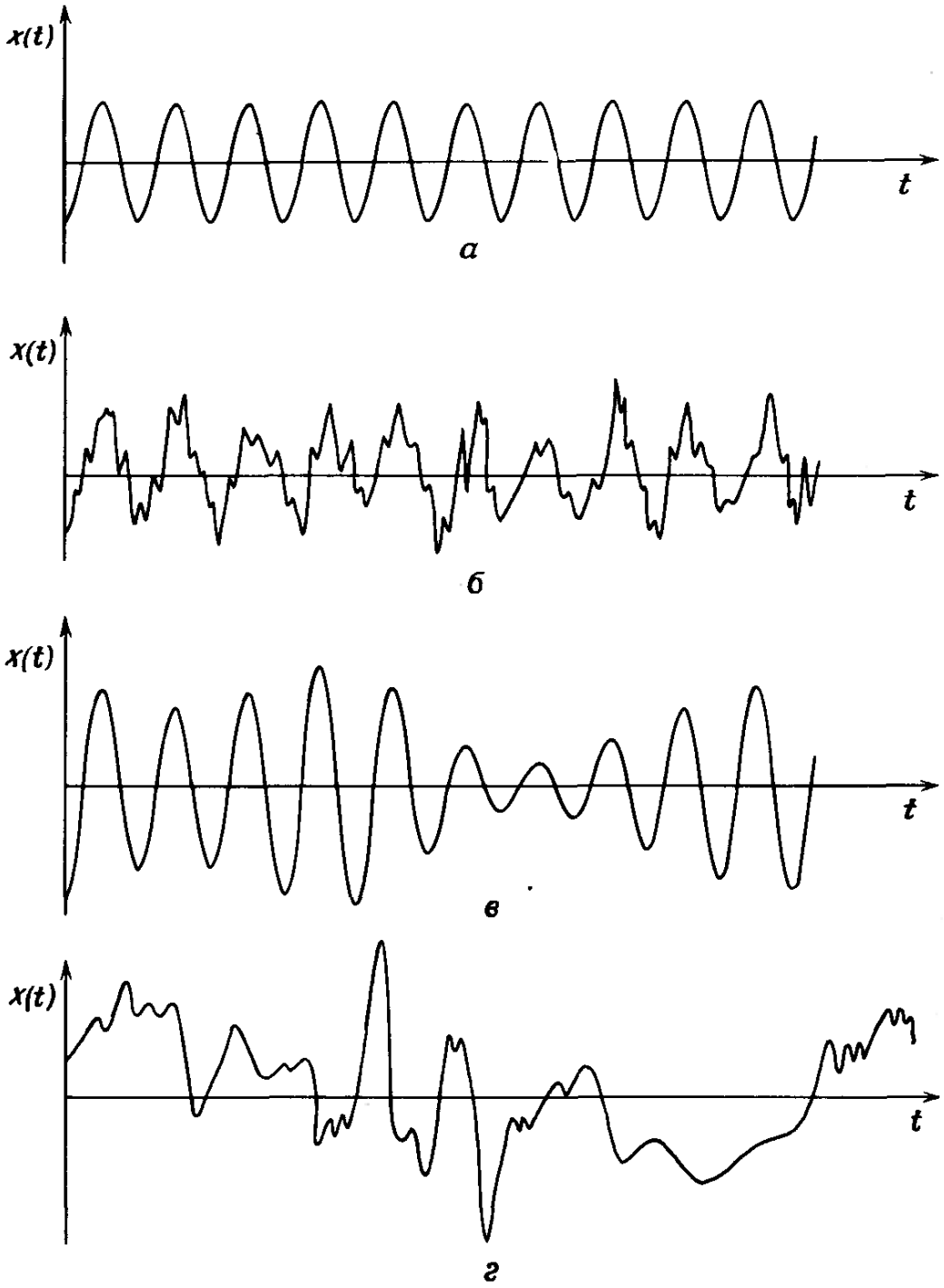

Рис. 1.5. Чотири приклади реалізацій випадкових процесів:

а - гармонійний процес;

б - гармонійний процес плюс випадковий шум;

в - вузько-смуговий випадковий шум;

г - широкосмуговий випадковий шум

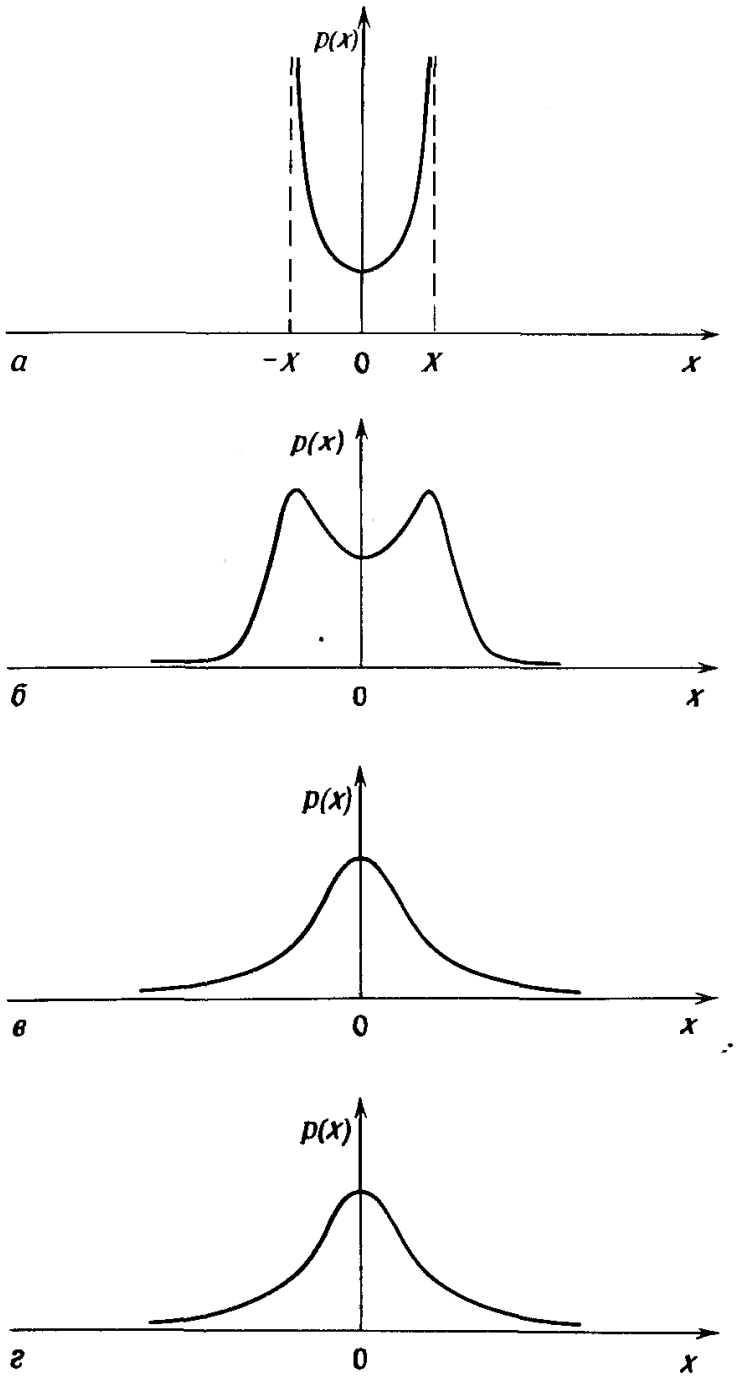

Рис. 1.6. Щільності ймовірності:

а - гармонійний процес;

б - гармонійний процес плюс випадковий шум;

в - вузько-смуговий випадковий шум;

г - широкосмуговий випадковий шум

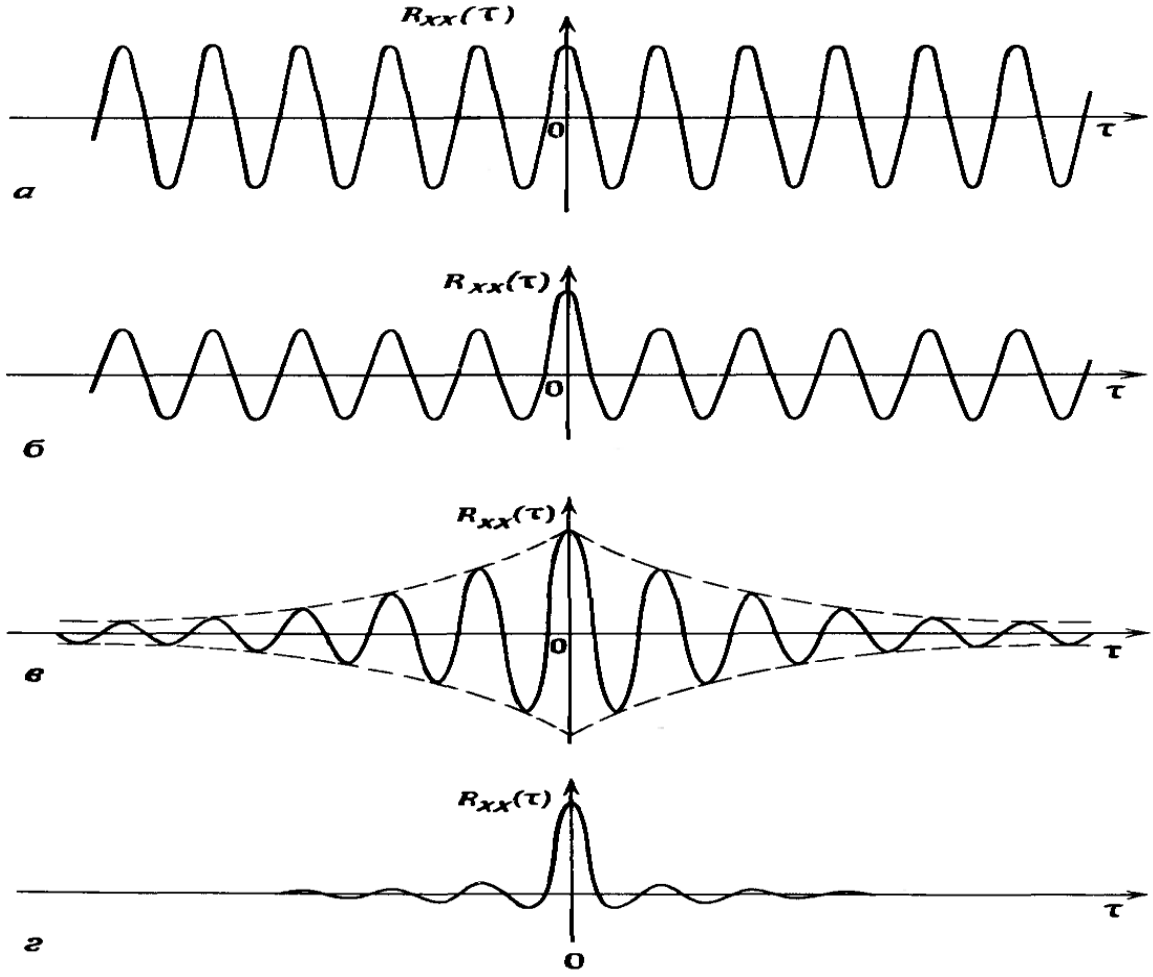

Рис. 1.7. Коваріаційні функції:

а - гармонійний процес;

б - гармонійний процес плюс випадковий шум;

в - вузько-смуговий випадковий шум;

г - широкосмуговий випадковий шум

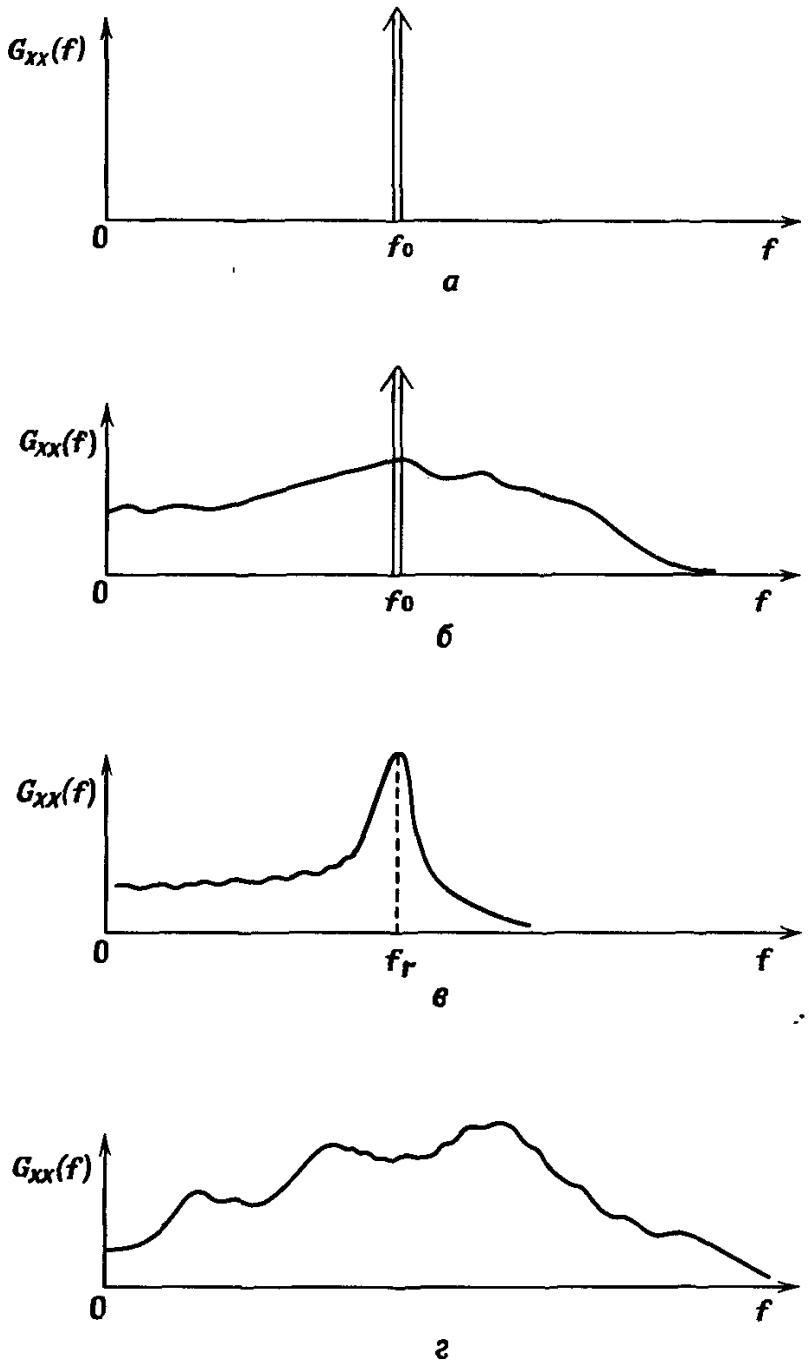

Рис. 1.8. Спектральні щільності:

а - гармонійний процес;

б - гармонійний процес плюс випадковий шум;

в - вузько-смуговий випадковий шум;

г - широкосмуговий випадковий шум

3. взаємні спектральні щільності;

4. частотні характеристики;

5. функції когерентності.

Перші три функції описують основні властивості пари реалізацій за прийнятими ними значеннями і за їхніми властивостями у часовій і частотній областях. За відомими взаємною спектральною щільністю і спектральними щільностями реалізацій можна теоретично обчислити лінійні частотні характеристики (амплітудні і фазові характеристики), що зв'язують ці дві реалізації. В цьому випадку реалізації вважаються входом і виходом деякої лінійної системи. Функція когерентності характеризує точність прийнятої лінійної моделі й теж може бути обчислена.

Щільності ймовірності та функції розподілу зазвичай застосовуються, крім опису ймовірнісної структури процесу, з метою:

1. перевірки нормальності;

2. виявлення нелінійностей;

3. аналізу екстремальних значень.

Основні застосування коваріаційних функцій охоплюють:

1. виявлення періодичностей;

2. виділення сигналів з шуму;

3. вимір запізнень;

4. локалізацію джерел перешкод;

5. ідентифікацію трактів і швидкостей поширення сигналів.

Типові застосування спектральних щільностей включають:

1. визначення властивостей систем за спостереженнями вхідних і вихідних процесів;

2. уявлення вихідних процесів за вхідними процесам і властивостями системи;

3. ідентифікація вхідних процесів за вихідними процесам і властивостями системи;

4. завдання динамічних даних для тестових програм;

5. ідентифікація джерел енергії і шуму;

6. оптимальний лінійний прогноз і фільтрацію.

1.4.3 Характеристики помилок. Оцінку величини φ позначимо через $\hat{\varphi}$. Величина $\hat{\varphi}$ - це оцінка φ, побудована зі спостереження на кінцевому інтервалі часу або за кінцевим числом вибіркових точок.

Припустимо, що оцінка $\hat{\varphi}$ може бути отримана багаторазово шляхом повторення експерименту або виконання певної програми вимірювань. Тоді у принципі можна оцінити математичне очікування $\hat{\varphi}$, що позначається $E[\hat{\varphi}]$. Наприклад, якщо експеримент повторюється багато разів і дає оцінки $\hat{\varphi}_i(i = 1,2,...,N)$, то

$$\hat{\varphi}_i(i = 1,2,...,N), \text{ то } E[\hat{\varphi}] = \frac{1}{N}\sum_{i=1}^{N}\hat{\varphi}_i . \qquad (1.12)$$

Це математичне очікування може збігтися або не збігтися з істинним значенням φ. У разі збігу оцінка $\hat{\varphi}$ називається незміщеною. В іншому випадку оцінка називається зміщеною. Зміщення оцінки, що позначається $b[\hat{\varphi}]$, дорівнює математичному очікуванню оцінки мінус справжнє значення параметра:

$$b[\hat{\varphi}] = E[\hat{\varphi}] - \varphi . \qquad (1.13)$$

Отже, зміщення оцінки - це систематична помилка, яка завжди має одну і ту ж абсолютну величину і один і той же знак, якщо виміри проводяться при незмінних умовах.

Дисперсія оцінки, що позначається $Var[\hat{\varphi}]$, визначається як математичне очікування квадрата різниці між оцінкою і її середнім значенням. Формально:

$$Var[\hat{\varphi}] = E\left[(\hat{\varphi} - E[\hat{\varphi}])^2\right]. \qquad (1.14)$$

Дисперсія характеризує випадкову помилку оцінки, тобто ту частину загальної помилки, яка не є систематичною і може мати різні знаки та різні абсолютні значення від вимірювання до вимірювання.

Сумарна помилка оцінювання характеризується середнім квадратом помилки, який визначається як математичне очікування квадрата різниці між оцінкою та її дійсним значенням. Середній квадрат помилки оцінки $\hat{\varphi}$ дорівнює:

$$E\left[(\hat{\varphi} - \varphi)^2\right]. \qquad (1.15)$$

Легко перевірити, що

$$E\left[(\hat{\varphi} - \varphi)^2\right] = Var[\hat{\varphi}] + (b[\hat{\varphi}])^2 . \qquad (1.16)$$

Інакше кажучи, середній квадрат помилки дорівнює сумі дисперсії і квадрата зміщення. Якщо зміщення дорівнює нулю або дуже мале, то середній квадрат помилки і дисперсія збігаються.

Рис. 1.9 ілюструє сенс зміщення (систематичної помилки) і дисперсію (випадкової помилки) на прикладі пристрілки двох гвинтівок. Рис. 1.9а показує, що гвинтівка А має велике зміщення і малу випадкову помилку. Рис. 1.9б показує, що гвинтівка Б має мале зміщення, але більшу випадкову помилку. Очевидно, з гвинтівки А ніколи не можна поцілити в ціль, в той час як з гвинтівки Б можна випадково вразити її. Однак більшість стрільців воліє гвинтівку А, оскільки систематичну помилку можна виключити (якщо відомо, що вона є) шляхом регулювання прицілу гвинтівки, а випадкову помилку усунути не можна. Отже, гвинтівка А потенційно має менший середній квадрат помилки.

Нарешті, важливою величиною є нормована середньоквадратична помилка оцінки, що позначається $\varepsilon[\hat{\varphi}]$. Ця безрозмірна помилка дорівнює квадратному кореню з середнього квадрата помилки, поділеній на справжнє значення параметра (зрозуміло, в припущенні, що воно не дорівнює нулю). Формально

$$\varepsilon[\hat{\varphi}] = \frac{\sqrt{E\left[(\hat{\varphi}-\varphi)^2\right]}}{\varphi}. \qquad (1.17)$$

На практиці намагаються зменшити нормовану середньоквадратичну помилку в максимально можливій мірі. У цьому випадку з'являється впевненість в тому, що довільна оцінка $\hat{\varphi}$ близька до істинного φ.

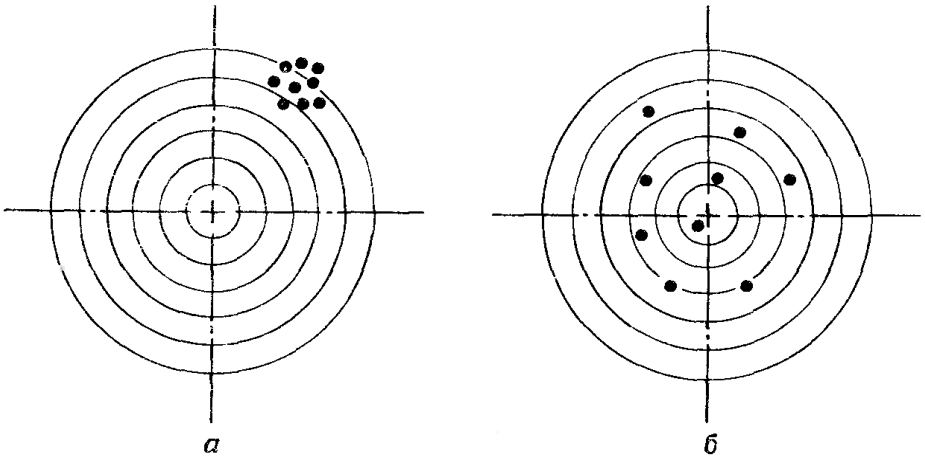

Рис. 1.9. Випадкові і систематично помилки при стрільбі з гвинтівки по мішені:

а - гвинтівка А, велика систематична помилка і мала випадкова помилка; б - гвинтівка Б, мала систематична помилка і велика випадкова помилка

1.5 Випадкові величини та закони їхнього розподілу

1.5.1 Ряд розподілу. Випадковою величиною називається величина, яка в результаті досліду може прийняти те чи інше значення, невідомо заздалегідь - яке саме. Розрізняються випадкові величини дискретного і безперервної типу. Можливі значення дискретних величин можуть бути заздалегідь визначені. Можливі значення безперервних величин не можуть бути заздалегідь перераховані і безперервно заповнюють деякий проміжок.

Розглянемо дискретну випадкову величину X з можливими значеннями x_1, x_2,…, x_n. Кожна з цих подій можлива, але не достовірна, і величина X може прийняти кожне з них з певною ймовірністю. В результаті досліду величина X прийме одне з цих значень, тобто відбудеться одна з повної групи несумісних подій:

$$\begin{pmatrix} X = x_1 \\ X = x_2 \\ \dots \\ X = x_n \end{pmatrix}. \qquad (1.18)$$

Ймовірності цих подій позначаються відповідно:

$$P(X = x_1) = p_1;$$
$$P(X = x_2) = p_2;$$
$$\dots;$$
$$P(X = x_n) = p_n.$$

(через P зазвичай позначається оператор імовірності, а через p – конкретна величина імовірності.)

Імовірність потрапляння в задану область може бути визначена таким чином:

$$p_m = \lim_{N \to \infty} \frac{N_m}{N}, \qquad (1.19)$$

де N_m – кількість спостережень випадкової величини, що потрапили у задану область;

N – загальне число спостережень (частотне визначення імовірності).

Так як несумісні події (1.18) утворюють повну групу, то $\sum_{i=1}^{n} p_i = 1$, тобто сума ймовірностей всіх можливих значень випадкової величини дорівнює одиниці. Ця сумарна ймовірність якимось чином розподілена між

окремими значеннями. Випадкова величина буде повністю описана з ймовірнісної точки часу, якщо задано цей розподіл, тобто в точності вказано яку ймовірністю має кожна з подій (1.18). Цим встановлюється закон розподілу випадкової величини.

Законом розподілу випадкової величини називається всяке співвідношення, що встановлює зв'язок між можливими значеннями випадкової величини і відповідними їм ймовірностями. Випадкова величина підпорядкована цьому закону розподілу.

Припустимо, що вивчається деяка випадкова величина X, закон розподілу якої в точності невідомий, і потрібно визначити цей закон з досліду або перевірити експериментально гіпотезу про те, що величина X підпорядкована тому чи іншому закону. З цією метою над випадковою величиною X проводиться ряд незалежних дослідів (спостережень). У кожному з цих дослідів випадкова величина X приймає певне значення. Сукупність спостережуваних значень величини і являє собою первинний статистичний матеріал, що підлягає обробці, осмисленню і науковому аналізу. Така сукупність називається «простою статистичною сукупністю» або «простим статистичним рядом». Зазвичай проста статистична сукупність оформляється у вигляді таблиці з одним входом, в першому стовпці стоїть номер досліду i, а в другому - спостережуване значення випадкової величини.

Простий статистичний ряд представляє собою первинну форму записи статистичного матеріалу і може бути оброблений різними способами. Одним із способів обробки є побудова статистичної функції розподілу випадкової величини.

Статистичної функцією розподілу випадкової величини X називається частота події $X < x$ в заданому статистичному матеріалі:

$F^*(x) = P^*(X < x)$. (1.20)

Для того щоб знайти значення статистичної функції розподілу при даному x, досить підрахувати число дослідів, в яких величина X прийняла значення, менше, ніж x, і розділити на загальне число n здійснених дослідів.

Статистична функція розподілу будь-якої випадкової величини - дискретної або безперервної - являє собою дискретну ступінчасту функцію, стрибки якої відповідають спостережуваним значенням випадкової величини і за величиною відповідають частотам цих значень. Якщо кожне окреме значення випадкової величини X спостерігалося тільки один раз, то стрибок статистичної функції розподілу, в кожному спостережуваному значенні дорівнює $1/n$, де n - число спостережень.

При збільшенні числа дослідів, відповідно до теореми Бернуллі, при будь-якому x частота події $X < x$ наближається (сходиться по ймовірності) до ймовірності цієї події. Отже, при збільшенні n статистична функція розподілу $F^*(x)$ наближається (сходиться по ймовірності) до справжньої функції розподілу $F(x)$ випадкової величини X.

Якщо X - безперервна випадкова величина, то при збільшенні числа спостережень n число стрибків функції $F^*(x)$ збільшується, самі стрибки зменшуються і графік функції $F^*(x)$ необмежено наближається до плавної кривої $F(x)$ - функції розподілу величини X.

1.5.2 Статистичний ряд. При великій кількості спостережень (порядку сотень) проста статистична сукупність перестає бути зручною формою запису статистичного матеріалу - вона стає занадто громіздкою і малонаочною. Для додання йому більшої компактності і наочності

статистичний матеріал повинен бути підданий додатковій обробці - будується так званий «статистичний ряд».

Нехай є результати спостережень над безперервною випадковою величиною X, оформлені у вигляді простої статистичної сукупності. Розділимо весь діапазон спостережень значень X на інтервали або «розряди» і підрахуємо кількість значень m_i, що припадають на кожен i-й розряд. Це число розділимо на загальне число спостережень n і знайдемо частоту, яка відповідає певному розряду:

$$p_i = \frac{m_i}{n}. \quad (1.21)$$

Сума частот усіх розрядів, очевидно, має дорівнювати одиниці.

Побудуємо таблицю (табл. 1.1), в якій наведено розряди в порядку їхнього розташування уздовж осі абсцис і відповідні частоти.

Таблиця 1.1

Статистичний ряд

l_i	$x_1; x_2$	$x_2; x_3$	…	$x_i; x_{i+1}$	…	$x_k; x_{k+1}$
p^*_i	p^*_1	p^*_2	…	p^*_i	…	p^*_k

тут l_i - позначення i-го розряду;

$x_i; x_{i+1}$ - його межі;

p^*_i - відповідна частота;

k - число розрядів.

При угрупованню спостережень значень випадкової величини за розрядами виникає питання про те, до якого розряду віднести значення, що знаходиться в точності на межі двох розрядів. У цих випадках можна

рекомендувати (чисто умовно) вважати дане значення належить в рівній мірі до обох розрядів і додавати до чисел m_i того та іншого розряду по ½.

Число розрядів, на які слід групувати статистичний матеріал, не повинно бути занадто великим (тоді ряд розподілу стає невиразним, і частоти в ньому виявляють незакономірні коливання); з іншого боку, воно не повинно бути занадто малим (при малому числі розрядів властивості розподілу описуються статистичним рядом занадто грубо). Практика показує, що в більшості випадків раціонально вибирати число розрядів порядку 10-20. Чим багатше і однорідніше статистичний матеріал, тим більше число розрядів можна вибирати при складанні статистичного ряду.

Більш наочним є графічне зображення: по осі абсцис відкладаються можливі значення випадкової величини, а по осі ординат - ймовірності цих значень. Фігура, в якій точки з'єднані відрізками прямих, називається багатокутником розподілу (рис. 1.10).

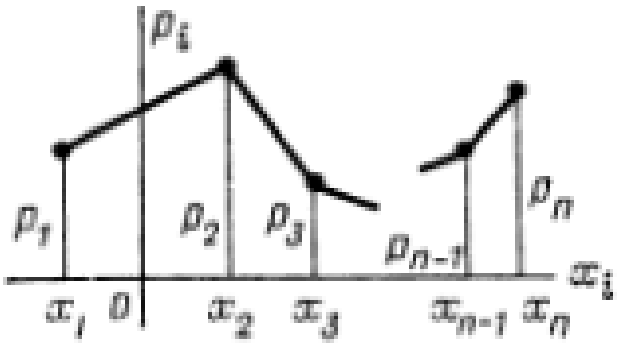

Рис. 1.10 Багатокутник розподілу

1.5.3 Функція розподілу. Для безперервної випадкової величини характеристику (закон розподілу) побудувати не можна. Дійсно, безперервна випадкова величина має безліч можливих значень, котрі суцільно заповнюють деякий проміжок (так звану «незліченну безліч»).

Скласти таблицю, в якій були б перераховані всі можливості значення такої випадкової величини, неможливо. Крім того, кожне окреме значення безперервної випадкової величини зазвичай не має ніякої відмінної від нуля ймовірності. Отже, для безперервної випадкової величини не існує ряду розподілу в тому сенсі, в якому він існує для детермінованої величини. Однак різні області можливих значень випадкової величини все ж не є однаково ймовірними, і для безперервної величини існує «розподіл ймовірностей», хоча і не в тому сенсі, як для дискретної.

Для кількісної характеристики цього розподілу ймовірностей зручно скористатися не ймовірністю події $X = x$, а ймовірністю того, що випадкова величина не перевищує деякого заданого або поточного значення x, тобто $X < x$, де x - деяка поточна змінна. Імовірність цієї події, очевидно, залежить від x, є деяка функція від x. Ця функція називається функцією розподілу випадкової величини X і позначається:

$$F(x) = P(X < x). \qquad (1.22)$$

Функцію розподілу $F(x)$ іноді називають також інтегральною функцією розподілу або інтегральним законом розподілу.

Функція розподілу - сама універсальна характеристика випадкової величини. Вона існує для всіх випадкових величин: як дискретних, так і безперервних. Функція розподілу повністю характеризує випадкову величину з ймовірнісної точки зору, тобто є однією з форм закону розподілу.

Ймовірність того, що значення випадкової величини X міститься між x_1 і x_2, дорівнює різниці значень функцій розподілу, обчислених у цих двох точках:

$$P\{ x_1 < X \leq x_2\} = F(x_1) - F(x_2). \qquad (1.23)$$

Аналогічно,

$$P\{X > x\} = 1 - F(x). \qquad (1.24)$$

Сформулюємо деякі загальні властивості функції розподілу.

1. Функція розподілу $F(x)$ є неубутна функція свого аргументу, тобто

$$F(x_2) \geq F(x_1) \mid x_2 > x_1.$$

2. На мінус нескінченності функція розподілу дорівнює нулю:

$$\lim_{x \to \infty} F(x) = F(-\infty) = 0.$$

3. На плюс нескінченності функція розподілу дорівнює одиниці:

$$\lim_{x \to \infty} F(x) = F(\infty) = 1.$$

4. Функція розподілу $F(x)$ завжди позитивна

$$F(x) \geq 0 \quad -\infty > x > \infty.$$

Для наочної геометричної інтерпретації розглянемо випадкову величину X як випадкову точку X на осі Ох (рис. 1.11), яка в результаті досліду може зайняти те чи інше положення.

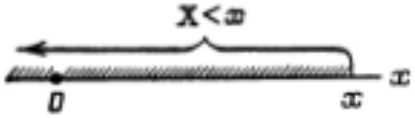

Рис. 1.11 Імовірність того, що випадкова точка X потрапить лівіше заданої точки

Тоді функція розподілу $F(x)$ є ймовірність того, що випадкова величина X в результаті досліду потрапить лівіше точки x. Очевидно, що при збільшенні x, тобто її переміщенні точки x вправо по вісі абсцис, ймовірність того, що випадкова точка X потрапить лівіше x, не може зменшуватися; отже, функція розподілу $F(x)$ зі зростанням x спадати не може. При необмеженій переміщенні точки x вліво по осі абсцис потрапляння випадкової точки X лівіше x в межі стає неможливою подією; природно вважати, що ймовірність цієї події прагнути до нуля, тобто

$$\lim_{x \to \infty} F(x) = F(-\infty) = 0 .$$

Аналогічним чином, при необмеженому переміщенні точки x вправо подія $X < x$ стає в межі достовірною:

$$\lim_{x \to \infty} F(x) = F(\infty) = 1 .$$

Графік функції розподілу $F(x)$ в загальному випадку являє собою графік неубутної функції (рис. 1.12), значення якої починаються від 0 і

доходять до 1, причому $F(-\infty) = 0$, $F(+\infty) = 1$, а в окремих точках функція може мати стрибки (розриви першого роду), на деяких ділянках вона може бути постійною, на інших — монотонно зростати.

Знаючи ряд розподілу дискретної випадкової величини, можна легко побудувати функцію розподілу цієї величини. дійсно,

$$F(x) = P(X < x) = \sum_{x_i < x} P(X = x_i),$$

де нерівність під знаком суми $x_i < x$ вказує, що підсумовування поширюється на всі ті значення x_i, які менше x.

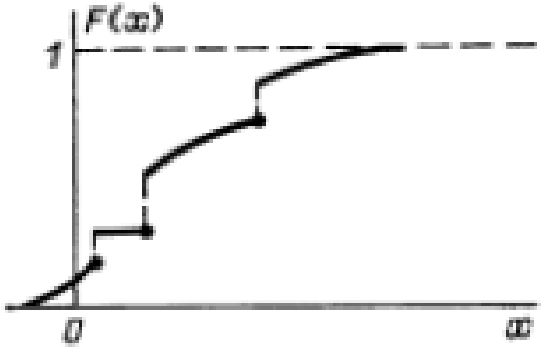

Рис. 1.12 Функція розподілу випадкової величини

Коли поточна змінна x проходить через яке-небудь з можливих значень дискретно величини X, функція розподілу змінюється стрибкоподібно, причому величина стрибка дорівнює ймовірності цього значення.

Для безперервних випадкових величин, область можливих значень яких містить усі точки з деякого інтервалу, можливий вигляд функції розподілу $F(x)$ зображено на рис. 1.13.

1.5.4 Щільність розподілу. Нехай є безперервна випадкова величина X з функцією розподілу $F(x)$, яка безперервна і диференційована. Обчислимо ймовірність потрапляння цієї випадкової величини на ділянку від x до $x + \Delta x$:

$$P\{x < X \le x + \Delta x\} = F(x + \Delta x) - F(x),$$

тобто приріст функції розподілу на цій ділянці. Розглянемо відношення цієї ймовірності до довжини ділянки, тобто середню ймовірність, що припадає на одиницю довжини на цій ділянці, і будемо наближати Δx до нуля. В межі отримаємо похідну від функції розподілу:

$$\lim_{\Delta x \to 0} \frac{P\{x < X \le x + \Delta x\}}{\Delta x} = F'(x). \quad (1.25)$$

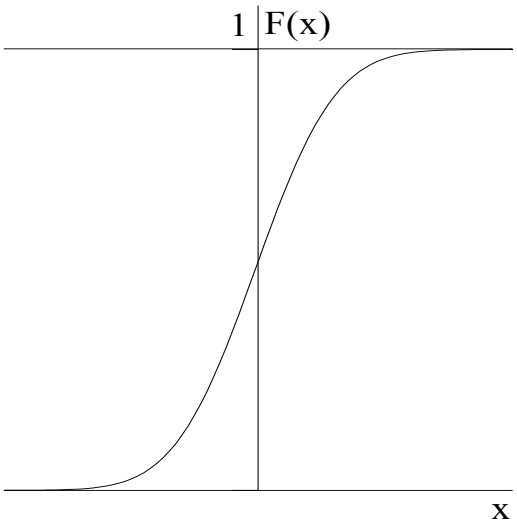

Рис. 1.13 Функцій розподілу $F(x)$ безперервних випадкових величин

Введемо позначення:

$$f(x)=\frac{dF(x)}{dx}\lim_{\Delta x\to 0}\frac{P\{x<X\le x+\Delta x\}}{\Delta x}(\Delta\Delta>0). \quad (1.26)$$

Функція $f(x)$ - похідна функції розподілу - характеризує ніби щільність, з якою розподіляються значення випадкової величини в даній точці. Ця функція називається щільністю розподілу (інакше - «щільністю ймовірності») безперервної випадкової величини X. Іноді функцію $f(x)$ називають також «диференціальної функцією розподілу» або «диференціальним законом розподілу» величини X. Крива, що зображає щільність розподілу випадкової величини, називається кривою розподілу. Щільність розподілу, так само як і функція розподілу, є однією з форм закону розподілу. На противагу функції розподілу ця форма не є універсальною: вона існує тільки для безперервних випадкових величин.

Розглянемо безперервну випадкову величину X з щільністю розподілу $f(x)$ і елементарний ділянку dx, що примикає до точки x (рис. 1.14).

Ймовірність влучення випадкової величини X на цю елементарну ділянку (з точністю до нескінченно малих вищого порядку) дорівнює:

$$f(x)\mathrm{d}x.$$

Ця величина називається елементом ймовірності. Геометрично це площа елементарного прямокутника, що спирається на відрізок dx (рис. 1.14). Висловимо ймовірність потрапляння величини X на відрізок від a_1 до a_2 (рис. 1.15) через щільність розподілу.

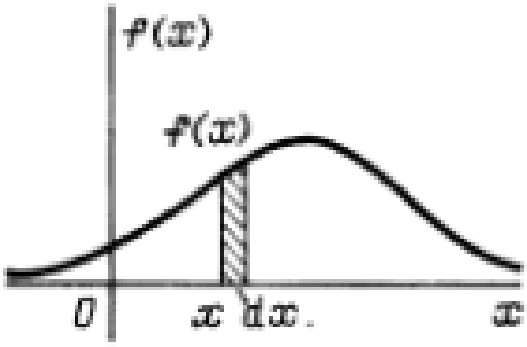

Рис. 1.14 Щільність розподілу *f*(*x*) безперервної випадкової величини *x*

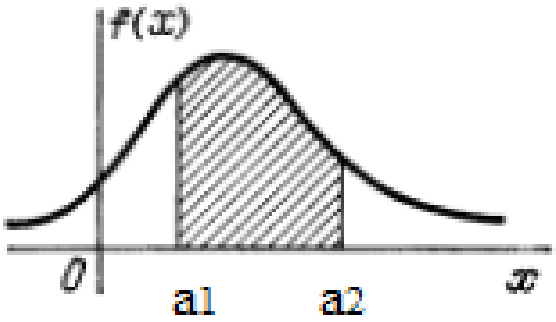

Рис. 1.15 Ймовірність потрапляння величини *x* на відрізок від a_1 до a_2

Очевидно, вона дорівнює сумі елементів ймовірності на цій ділянці, тобто інтегралу:

$$P\{a_1 < X < a_2\} = \int_{a_1}^{a_2} f(x)dx \,. \quad (1.27)$$

(Так як ймовірність будь-якого окремого значення безперервної випадкової величини дорівнює нулю, то можна розглядати тут відрізок

(a_1, a_2)

не включаючи до нього лівий кінець, тобто відкидаючи знак рівності в

$a_1 \le X < a_2$.

Геометрично ймовірність потрапляння величини X на ділянку

(a_1, a_2)

дорівнює площі кривої розподілу, що спирається на цю ділянку (рис. 1.15).

Формула (1.26) виражає щільність розподілу через функцію розподілу. Задамо зворотне завдання: виразити функцію розподілу через щільність. За визначенням:

$F(x) = P\{X < x\} = P\{-\infty < X < x\}$,

звідки за формулою (1.27) маємо:

$$F(x) = \int_{-\infty}^{x} f(x)dx \,. \quad (1.28)$$

Геометрично $F(x)$ є не що інше, як площа кривої розподілу, що лежить лівіше точки x (рис. 1.16).

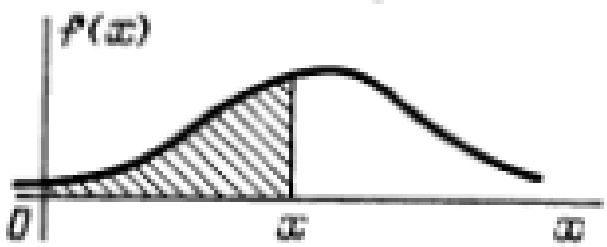

Рис. 1.16 До геометричної інтерпретації F(x)

Зазначимо основні властивості щільності розподілу.

1. Щільність розподілу є невід'ємна функція:

$$f(x) \geq 0.$$

Це властивість безпосередньо випливає з того, що функція розподілу є неубутна функція.

2. Інтеграл в нескінченних межах від щільності розподілу дорівнює одиниці:

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

Це випливає з формули (1.28) і з того, що

$$F(+\infty) = 1.$$

3) $\int_{-\infty}^{x} f(z)dz = F(x)$;

(z – змінна інтегрування).

4) $\lim f(x) = 0$
$|x| \to \infty$

За допомогою диференціальної функції розподілу обчислюється імовірність знаходження випадкової величини у будь-якій області з множини її можливих значень. Зокрема,

$$P\{X \le a_1\} = \int_{-\infty}^{a_1} f(x)dx, \quad P\{X > a_2\} = \int_{a_2}^{\infty} f(x)dx, \quad P\{a_1 X \le a_2\} = \int_{a_1}^{a_2} f(x)dx . \quad (1.29)$$

Геометрично основні властивості щільності розподілу означають, що:

- вся крива розподілу лежить не нижче осі абсцис;

- повна площа, обмежена кривою розподілу і віссю абсцис, дорівнює одиниці.

З'ясуємо розмірності основних характеристик випадкової величини - функції розподілу і щільності розподілу. Функція розподілу $F(x)$, як будь-яка ймовірність, є величина безрозмірна. Розмірність щільності розподілу f(x), як видно з формули (1.25), обернена розмірності випадкової величини.

1.5.5. Статистична функція розподілу. Якщо спостережувана випадкова величина X дискретна, то зі статистичними аналогом ряду розподілу є статистичний ряд, повністю аналогічний ряду розподілу випадкова величина X, з тією різницею, що замість ймовірностей $p_i = P\{X = x_i\}$ в ньому стоять частоти відповідних подію: $p_i^* = P^*\{X = x_i\}$. На цьому питанні ми більше не будемо зупинятися. Набагато складніше (і частіше зустрічається на практиці завдання обробки дослідів над безперервною випадковою величиною X. Займемося описом результатів серії з n

незалежних дослідів, в кожному з яких зареєстровано значення безперервної випадкова величина X, і найпростішої обробкою цих результатів.

Перше, що потрапляє в руки - це протокол, в якому зареєстровані: номер досліду k і значення x_k, яке прийняла в цьому досліді випадкова величина X. Такий протокол називають первинною статистичною сукупністю. Це - зовсім ще не оброблений статистичний матеріал.

Розгляд та осмислення таблиці такого типу (особливо при великому числі дослідів n) важко, і по ній практично не можна уявити собі, характер розподілу випадкова величина X. Перший крок до осмислення матеріалу - це його впорядкування, розташування в порядку зростання значень випадкової величини. Протокол результатів досліду, в якому вони пронумеровані і розташовані в порядку зростання, будемо називати впорядкованою статистичною сукупністю. Номер значення позначається і (на відміну від номера досліду k). Якщо в таблиці одне і те ж значення зустрічається кілька разів, пишуть його стільки разів, скільки воно зустрілося.

За упорядкованої статистичної сукупності можна вже побудувати статистичну функцію розподілу:

$$F^*(x) = P^*\{X < x\}. \quad (1.30)$$

Функція $F^*(x)$ - розривна ступінчаста функція, безперервна зліва, дорівнює нулю лівіше найменшого спостереженого значення випадкова величина X і одиниці — направо від найбільшого. Теоретично вона повинна мати n стрибків, де n - число дослідів, а величина кожного

стрибка повинна дорівнювати $1/n$ - частоті спостереженого значення випадкової величини.

Практично, якщо одне і те ж значення спостерігалося кілька разів, відповідні стрибки зливаються в один, так що загальне число стрибків дорівнює числу різних спостережених значень випадкової величини. Кожен стрибок в точці X_i дорівнює «кратності» l_i значення x_i в статистичній сукупності, поділеній на число дослідів n.

Обчислюючи таким чином функцію $F^*(x)$, отримаємо таблицю її значень на інтервалах між стрибками.

За матеріалами цієї таблиці будуємо графік функції $F^*(x)$ (рис. 1.17). Жирними точками, як завжди, позначені значення функції в точках розриву.

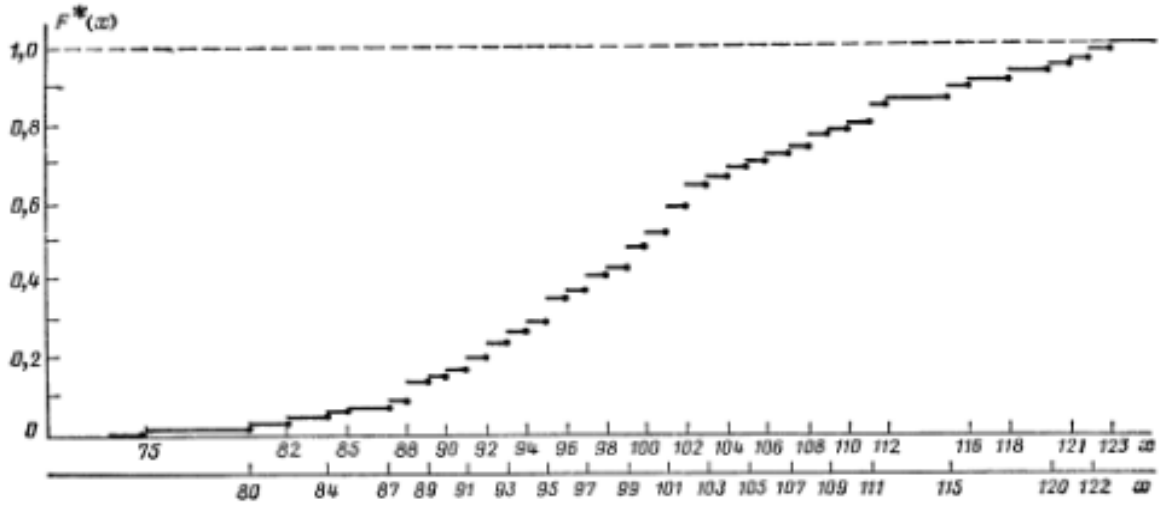

Рис. 1.17 Графік функції $F^*(x)$

Рис. 1.17, на відміну від таблиці 1.1, вже дає певне уявлення про характер розподілу випадкової величини X; зрозуміло, саме загальне уявлення, так як ясно, що деякі особливості кривої $F^*(x)$ випадкові і пов'язані з вибором цих, а не інших дослідів. Інші досліди дали б дещо

інший графік функції $F^*(x)$, але загальна тенденція збереглася б. При необмеженому збільшенні n стрибки кривої $F^*(x)$ стануть дрібнішими; крива $F^*(x)$ стане плавнішою, буде наближатися (сходитися по ймовірності) до функції розподілу $F(x)$ випадкової величини X. Проте, такий громіздкий і трудомісткий спосіб отримання функції розподілу $F(x)$ навряд чи може бути рекомендований. На практиці застосовуються інші, простіші способи побудови законів розподілу випадкових величин за дослідними даними.

1.5.6. Груповий статистичний ряд. Гістограма. Для того, щоб скласти собі загальне уявлення про закон розподілу випадкова величина X, нема чого фіксувати кожне спостережене значення і будувати статистичну функцію розподілу $F^*(x)$. Цим цілям краще слугують груповий статистичний ряд і гістограма.

Для побудови групованого статистичного ряду всю ділянку осі абсцис, на якій розташовані значення випадкової величини X, що спостерігалися в досліді, поділяється на ділянки або «розряди». Довжини розрядів необов'язково брати однаковими: бувають випадки, коли на тих ділянках осі абсцис, де спостережені значення X розташовуються густіше, зручніше брати розряди дрібнішими, а там, де рідше - більшими (або об'єднувати два або більше однакових по довжині розрядів в один). Межі розрядів зручно брати «круглими» числами.

Груповаим статистичним рядом називається таблиця 1.2, де у верхньому рядку вказані розряди: від - до +, в нижній - відповідні їм частоти.

Таблиця 1.2

Груповий статистичний ряд

X:	$x_1 \div x_2$	$x_2 \div x_3$	…	$x_i \div x_{i+1}$	…	$x_{k-1} \div x_k$
	p^*_1	p^*_2	…	p^*_i	…	p^*_k

Причому

$$\sum_{i=1}^{n} p_i = 1$$

Частота p_i^* події $\{X \in (x_i, x_i+l)\}$ обчислюється як відношення числа l_i дослідів, в яких значення випадкова величина X потрапило в i-й розряд $\{X \in (x_i, x_i+l)\}$, до загальної кількості n здійснених дослідів.

Відразу ж виникає питання: а як бути, якщо значення випадкова величина X потрапило в точності на межу між розрядами? До якого розряду його віднести? Це не має значення: можна віднести значення або до лівого розряду, або до правого (адже ймовірність того, що безперервна випадкова величина X прийме заздалегідь задане значення, дорівнює нулю); зупинимося на більш симетричному «справедливому» правилі: якщо значення випадкова величина потрапило в точності на кордон розрядів, розділити його порівну між сусідніми розрядами і додати по 1/2 до числам l_i для обох розрядів.

Побудуємо груповий статистичний ряд для випадкових величин X. Виберемо межі розрядів «круглими».

Підраховуючи кількість значень випадкової величини, що потрапили в кожен розряд (вважаючи половинки від тих, що потрапили в межу між

розрядами) і ділячи на число дослідів n, отримаємо групований статистичний ряд.

Розділяючи кожну частоту p_i* на довжину відповідного розряду $\Delta_i = x_i + + 1 - x_i$, отримаємо таблицю щільності частоти f_i*.

Відкладаючи по осі абсцис розряди і будуючи на кожному розряді як на підставі прямокутник площі P_i, отримаємо гістограму - статистичний аналог кривою розподілу (рис. 1.18).

Маючи в своєму розпорядженні групований статистичний ряд, ми можемо наближено побудувати статистичну функцію розподілу $F^*(x)$. Як тих значень x, для яких обчислюється $F^*(x)$, природно вжити заходів розрядів.

Графік статистичної функції розподілу показаний на рис. 1.19 (точки з'єднані відрізками прямих).

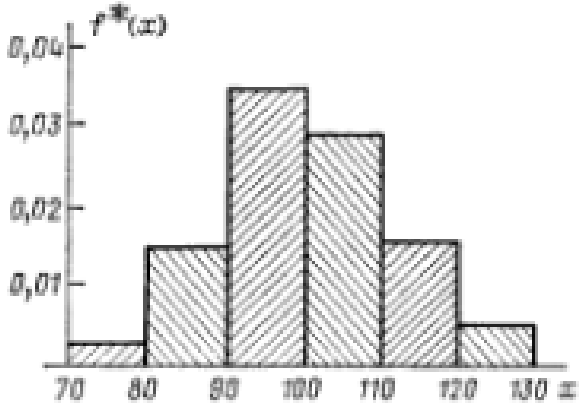

Рис. 1.18 Статистичний аналог кривою розподілу

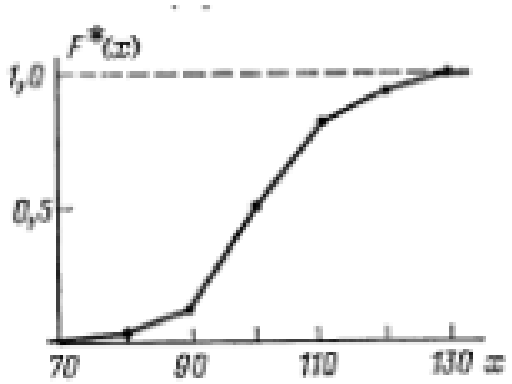

Рис. 1.19 Графік статистичної функції розподілу

1.5.7. Вирівнювання статистичних розподілів. У всякому статистичному розподілі неминуче присутні елементи випадковості, пов'язані з тим, що число дослідів обмежене, що зроблені саме ті, а не інші досліди, що дали саме ті, а не інші результати. Тільки при дуже великому числі дослідів ці випадковості згладжуються, і явище виявляє в повній мірі властиві йому закономірності. На практиці ми майже ніколи не маємо в своєму розпорядженні таке велике число дослідів (спостережень) і змушені зважати на те, що будь-якому статистичному розподілу притаманні в тій чи іншій мірі риси випадковості. Чи випадковий ступінчастий вид статистичної функції розподілу безперервної випадкової величини; випадкова форма гістограми, обмеженою теж ступінчастою лінією? Незручно користуватися такими негладкими функціями при подальшому їхньому перетворенні. Тому на практиці часто доводиться вирішувати питання про те, як підібрати для даного статистичного розподілу аналітичну формулу, яка має лише істотні риси статистичного матеріалу, через випадковості, пов'язані з недостатнім обсягом дослідних даних. Таке завдання називається завданням вирівнювання статистичних розподілів. Зазвичай вирівнюванню піддаються гістограми. Завдання зводиться до того, щоб замінити гістограму плавною кривою, що має досить простий

аналітичний вираз, і в подальшому користуватися нею як щільністю розподілу $f(x)$ (рис. 1.20).

Як підібрати найкращим чином плавну криву, котра вирівнює гістограму? Це завдання значною мірою невизначена, як і будь-яке завдання про аналітичне подання емпіричних функцій.

Наприклад, якщо кілька отриманих в досліді точок на площині xOy розташовані приблизно по прямій (рис. 1.21), природно виникає ідея замінити цю залежність лінійною функцією. Якщо залежність явно нелінійна (рис. 1.22), в якості апроксимувальної кривої вибирають параболу тощо.

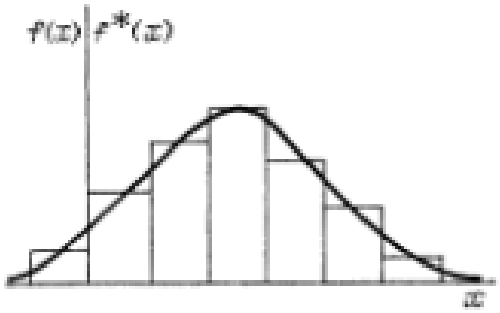

Рис. 1.20 Щільність розподілу $f(x)$

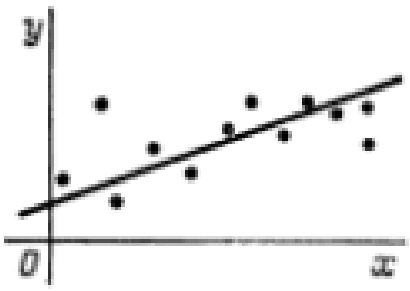

Рис. 1.21 Отримані в досліді точки на площині xOy

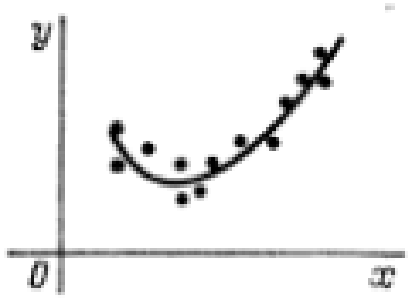

Рис. 1.22 Апроксимація параболою

При згладжуванні емпіричних залежностей дуже часто виходять з «принципу найменших квадратів», вважаючи, що найкращим наближенням в даному класі функцій є та, для якої сума квадратів відхилень звертається в мінімум. Питання про те, в якому саме класі функцій слід шукати найкраще наближення, вирішується вже не математично, а виходячи з міркувань, пов'язаних з фізикою розв'язуваної задачі, з урахуванням характеру емпіричної кривої і ступеня точності спостережень. Іноді принциповий вид функції, що виражає досліджувану залежність, відомий заздалегідь з теоретичних міркувань; з досліду же потрібно отримати лише деякі чисельні параметри, що входять у вираз функції.

Аналогічна справа і з завданням вирівнювання статистичних розподілів. Принциповий вигляд вирівнює плавною кривою $f(x)$ вибирається заздалегідь, виходячи з умов виникнення випадкової величини X, а іноді просто з міркувань, пов'язаних із зовнішнім виглядом гістограми. Наприклад, гістограма, зображена на рис. 1.20, явно наводить на думку про нормальний розподіл, а на рис. 1.24 - показовий.

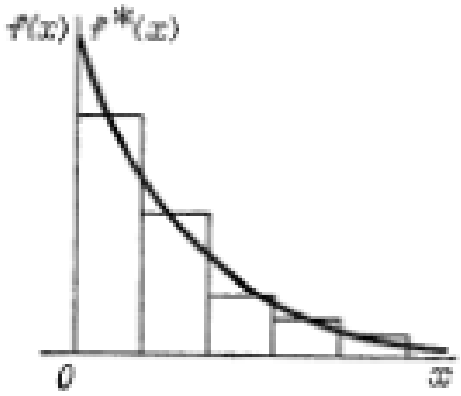

Рис. 1.24 Показовий розподіл

При цьому необхідно мати на увазі, що будь-яка аналітична функція $f(x)$, за допомогою якої вирівнюється гістограма, повинна мати основні властивості щільності:

$$f(x) \geq 0; \int_{-\infty}^{\infty} f(x)dx = 1.$$

Що стосується параметрів, що входять у вираз функції f(x), то їх підбирають так, щоб найкращим чином узгодити компенсаційний аналітичний розподіл зі статистичним.

Гістограма і нормальна крива розподілу f(x), що вирівнює її, показані на рис. 1.25.

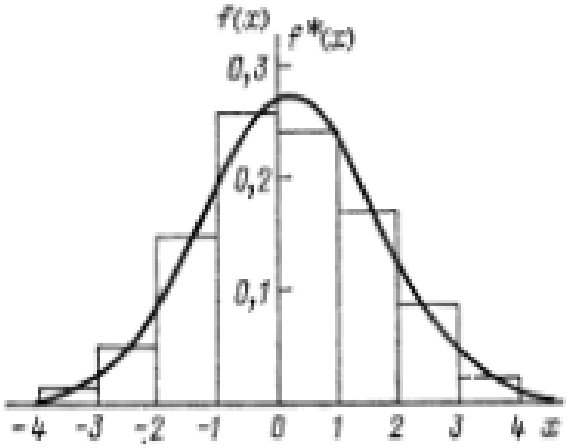

Рис. 1.25 Гістограма і нормальна крива розподілу f(x)

1.5.7 Числові характеристики випадкових величин. Повними, вичерпними характеристиками випадкових величин - так званих законів розподілу - є:

Для дискретної випадкової величини

а) функція розподілу;

б) ряд розподілу (графічно - багатокутник розподілу);

Для безперервної величини

а) функція розподілу;

б) щільність розподілу (графічно - крива розподілу).

Кожен закон розподілу являє собою деяку функцію, і вказівка цієї функції повністю описує випадкову величину з деякою ймовірнісної точки зору.

Однак у багатьох питаннях практики немає необхідності характеризувати випадкову величину повністю, вичерпно. Найчастіше буває досить вказати тільки окремі числові параметри, які до деякої міри характеризують істотні риси розподілу випадкової величини: наприклад, якесь середнє значення, поблизу якого групуються можливі значення випадкової величини; якесь число, що характеризує ступінь розкиданості

цих значень відносно середнього, тощо. Користуючись такими характеристиками, всі істотні наявні відомості щодо випадкової величини можна виразити найбільш компактно за допомогою мінімальної кількості числових параметрів. Такі характеристики, призначення яких - висловити у стислій формі найбільш істотні особливості розподілу, називаються числовими характеристиками випадкової величини.

У теорії ймовірностей числові характеристики і операції з ними грають величезну роль. За допомогою числових характеристик істотно полегшується вирішення багатьох ймовірнісних задач. Дуже часто вдається вирішити задачу до кінця, залишаючи осторонь закони розподілу і оперуючи одними числовими характеристиками. При цьому дуже важливу роль грає та обставина, що коли в завданні фігурує велика кількість випадкових величин, кожна з яких має певний вплив на чисельний результат досліду, то закон розподілу цього результату значною мірою можна вважати незалежними від законів розподілу окремих випадкових величин (виникає так званий нормальний закон розподілу). У цих випадках по суті завдання для вичерпного судження про результуючому законі розподілу не потрібно знати закони розподілу окремих випадкових величин.

1.5.8 Характеристики положення. Серед числових характеристик випадкових величин потрібно, перш за все, відзначити ті, які характеризують положення випадкової величини на числовій осі, тобто вказують деяке середнє орієнтовне значення, навколо якого групуються всі можливі значення випадкової величини.

З характеристик положення найважливішу роль відіграє математичне очікування випадкової величини, яке іноді називають просто середнім значенням випадкової величини.

Розглянемо дискретну випадкову величину X, що має можливі значення $x_1, x_2, \dots, x_n$ з вірогідністю $p_1, p_2, \dots, p_n$.

Потрібно охарактеризувати якимось числом положень значення випадкової величини на осі абсцис з урахуванням того, що ці значення мають різні ймовірності. Для цієї мети природно скористатися так званим «середнім зважуванням» з значень x_1, причому кожне значення при усередненні має враховуватися з «вагою», пропорційним ймовірності цього значення. Таким чином, обчислюється середнє значення випадкової величини X, яке позначається m_x:

$$m_x = \frac{x_1 p_1 + x_2 p_2 + \dots + x_n p_n}{p_1 + p_2 + \dots + p_n} = \frac{\sum_{i=1}^{n} x_i p_i}{\sum_{i=1}^{n} p_i}$$

або, враховуючи, що

$$\sum_{i=1}^{n} p_i = 1,$$

$$m_x = \sum_{i=1}^{n} x_i p_i . \qquad (1.31)$$

Це середнє зважене значення і називається математичним очікуванням дискретної випадкової величини: математичним очікуванням випадкової величини називається сума добутків всіх можливих значень випадкової величини на ймовірність цих значень.

Математичне очікування випадкової величини X пов'язано своєрідною залежністю з середнім арифметичним спостережень значень випадкової величини при великому числі дослідів. Ця залежність того ж

типу, як і залежність між частотою і ймовірністю, а саме: при великому числі дослідів середнє арифметичне спостережень значень випадкової величини наближається (сходиться по ймовірності) до її математичного очікування. З наявності зв'язку між частотою і ймовірністю можна вивести як наслідок наявність подібної ж зв'язку між середнім арифметичним і математичним очікуванням.

Для безперервної величини X математичне очікування, природно, виражається вже не сумою, а інтегралом: Нехай проводиться N незалежних дослідів, в кожному з яких величина X приймає певне значення. Припустимо, що значення x_1 з'явилося m_1 раз, значення x_2 з'явилося m_2 раз, взагалі значення x_i з'явилося m_i раз. очевидно,

$$m_x = \int_{-\infty}^{\infty} xf(x)dx. \quad (1.32)$$

Формула (1.32) виходить з формули (1.31), якщо в ній замінити окремі значення x_i, що є безперервно змінюваним параметром x, відповідні ймовірності p_i - елементом ймовірності $f(x)dx$, кінцеву суму — інтегралом.

Степінь розсіювання випадкової величини X відносно m_x може бути охарактеризована за допомогою *генеральної дисперсії* σ_x^2 :

$$\sigma_x^2 = \int_{-\infty}^{\infty} (x - m_x)^2 f(x)dx. \quad (1.33)$$

Якщо $f(x)$ усе у більшої степені концентрується поблизу m_x, то значення σ_x^2 зменшуються. Якщо ж є дуже віддалені від m_x значення випадкової величини X і для них $f(x)$ не занадто мала, то дисперсія σ_x^2

збільшується. Квадратний корінь із дисперсії σ_x^2 називається середнім квадратичним відхиленням σ_x.

1.6 Визначення законів розподілу випадкових величин на засадах дослідних даних

1.6.1 Основні задачі математичної статистики. Математичні закони теорії ймовірностей не є безпредметними абстракціями, позбавленими фізичного змісту; вони являють собою математичний вираз реальних закономірностей, фактично існуючих в масових випадкових явищах природи.

Закони розподілу випадкових величин визначаються на основі досліду. Кожне дослідження випадкових явищ, що виконується методами теорії ймовірностей, прямо або побічно спирається на експериментальні дані. Оперуючи такими поняттями, як події і їхні ймовірності, випадкові величини, їхні закони розподілу та числові характеристики, теорія ймовірностей дає можливість теоретичним шляхом визначати ймовірності одних подій через ймовірності інших, закони розподілу та числові характеристики одних випадкових величин через закони розподілу і числові характеристики інших. Такі непрямі методи дозволяють значно економити час і кошти, що витрачаються на експеримент, але аж ніяк не виключають самого експерименту. Кожне дослідження в області випадкових явищ, як би абстрактно воно не було, корінням своїми завжди йде в експеримент, в дослідні дані, в систему спостережень.

Розробка методів реєстрації, опису та аналізу статистичних експериментальних даних, одержуваних в результаті спостереження масових випадкових явищ, становить предмет математичної статистики.

Охарактеризуємо коротко деякі типові завдання математичної статистики, котрі часто зустрічаються на практиці.

1) Завдання визначення закону розподілу випадкової величини (або системи випадкових величин) за статистичними даними. Закономірності, які спостерігаються в масових випадкових явищах, проявляються тим точніше і виразніше, чим більший об'єм статистичного матеріалу. При обробці великих за своїм обсягом статистичних даних часто виникає питання про визначення законів розподілу тих чи інших випадкових величин. Теоретично при достатній кількості дослідів властиві цим випадковим величинам закономірності будуть здійснюватися як завгодно точно. На практиці нам завжди доводиться мати справу з обмеженою кількістю експериментальних даних; в зв'язку з цим результати спостережень і їхньої обробки завжди містять більший чи менший елемент випадковості. Виникає питання про те, які риси спостережуваного явища відносяться до постійних, стійким і дійсно притаманні йому, а які є випадковими і проявляються в даній серії спостережень тільки за рахунок обмеженого обсягу експериментальних даних. Природно, до методики обробки експериментальних даних слід пред'явити такі вимоги, щоб вона, по можливості, зберігала типові, характерні риси спостережуваного явища і відкидала всі несуттєве, другорядне, пов'язане з недостатнім обсягом дослідного матеріалу. У зв'язку з цим виникає характерна для математичної статистики задача згладжування або вирівнювання статистичних даних, подання їх в найбільш компактному вигляді за допомогою простих залежностей.

2) Завдання перевірки правдоподібності гіпотез. Це завдання тісно пов'язано з попередньою; при вирішенні такого роду завдань зазвичай відсутній настільки великий статистичний матеріал, щоб виявлені в ньому

статистичні закономірності були в достатній мірі вільні від елементів випадковості. Статистичний матеріал може з більшою чи меншою правдоподібністю підтверджувати або не підтверджувати справедливість тієї чи іншої гіпотези. Наприклад, може виникнути таке питання: чи узгоджуються результати експерименту з гіпотезою про те, що дана випадкова величина підпорядкована закону розподілу? Інше подібне питання: вказує чи спостерігається в досліді тенденція до залежності між двома випадковими величинами на наявність дійсної об'єктивної залежності між ними або ж вона пояснюється випадковими причинами, пов'язаними з недостатнім обсягом спостережень? Для вирішення подібних питань математична статистика виробила ряд спеціальних прийомів.

3) Завдання знаходження невідомих параметрів розподілу. Часто при обробці статистичного матеріалу зовсім не виникає питання про визначення законів розподілу досліджуваних випадкових величин. Звичайно це буває пов'язано з вкрай недостатнім обсягом експериментального матеріалу. Іноді ж характер закону розподілу якісно відомий з досліду, з теоретичних міркувань; наприклад, часто можна стверджувати заздалегідь, що випадкова величина підпорядкована нормальному закону. Тоді виникає вужче завдання обробки спостережень - визначення тільки деяких параметрів (числових характеристик) випадкової величини або системи випадкових величин. При невеликому числі дослідів завдання більш-менш точного визначення цих параметрів не може бути вирішено; у цих випадках експериментальний матеріал містить в собі неминуче значний елемент випадковості; тому випадковими виявляються і всі параметри, обчислені на основі цих даних. У цих умовах може бути поставлена тільки завдання про визначення так званих «оцінок» або

«відповідних значень для шуканих параметрів, тобто таких наближених значень, які при масовому застосуванні приводили б в середньому до менших помилок, ніж всякі інші. Із завданням відшукання «потрібних значень» числових характеристик тісно пов'язана задача оцінки їх точності і надійності.

1.6.2. Критерій згоди χ^2. Припустимо, ми хочемо встановити, чи суперечать дослідні дані гіпотезі про те, що випадкова величина X розподілена за таким-то закону? Для відповіді на таке питання користуються так званими критеріями згоди, з яких ми зупинимося тільки на одному, найчастіше застосовуваному критерії χ^2 Пірсона.

Викладемо ідею цього критерію спочатку для випадку дискретної випадкової величини X з можливими значеннями x_1, x_2 ..., x_n. Припустимо, що здійснено n незалежних дослідів, в кожному з яких випадкова величина X прийняла певне значення. На основі цих дослідів складено статистичний ряд розподілу випадкової величини X (табл. 1.3).

Таблиця 1.3

Статистичний ряд розподілу випадкової величини

X:	x_1	x_2	…	x_i	…	x_k
	p^*_1	p^*_2	…	p^*_i	…	p^*_k

де $p_i^* = n_i/n$ (1.34) — частота події $\{X = x_i\}$,

n_i - число дослідів, в яких з'явилася ця подія (i = l, 2, ..., k).

Ми висуваємо гіпотезу H, яка полягає в тому, що випадкова величина X має ряд розподілу (табл. 1.4).

Таблиця 1.4

Гіпотеза Н ряду розподілу випадкової величини

X:	x_1	x_2	…	x_i	…	x_k
	P_1	P_2	…	P_i	…	P_k

а відхилення частот p_i^* від ймовірностей p_i пояснюються випадковими причинами. Щоб перевірити правдоподібність цієї гіпотези, треба вибрати якусь міру розбіжності статистичного розподілу з гіпотетичним.

В якості запобіжного розбіжності R між гіпотетичним розподілом (табл. 1.4) і статистичними (табл. 1.3) при користуванні критерієм χ^2 береться сума квадратів відхилень $p_i^* - p$ статистичних ймовірностей p_i^* від гіпотетичних p_i, взятих з деякими «вагами» c_i:

$$R = \sum_{i=1}^{k} c_i \left(p_i - p_i\right)^2.$$

Коефіцієнти c_i вводяться тому, що відхилення, які відносяться до різних значень p_i, не можна вважати рівноправними за значимістю: одне і те ж по абсолютній величині відхилення $p_i^* - p_i$ може бути незначним, якщо сама ймовірність p_i велика, і дуже помітним, якщо вона мала. Тому природно ваги c_i взяти обернено пропорційними можливостям p_i. Але з яким коефіцієнтом пропорційності?

Пірсон довів, що якщо покласти

$c_i = n/p_i$, (1.35);

то при великому числі дослідів n закон розподілу величини R має досить прості властивості: він практично не залежить від закону розподілу випадкової величини X і мало залежить від числа дослідів n, а залежить тільки від числа значень випадкової величини k і при збільшенні n наближається до розподілу χ^2. При такому виборі коефіцієнтів c_i міра розбіжності R звичайно позначається χ^2:

$$\chi^2 = n\sum_{i=1}^{k}(p_i - p_i)^2 / p_i$$

або, вводячи величину n під знак суми і з огляду на (1.34);

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

отримаємо

$$R = \chi^2 = n\sum_{i=1}^{k}(n_i - np_i)^2 /(np_i). \quad (1.36)$$

Розподіл χ^2, як ми знаємо, залежить від параметра r, званого «числом ступенів свободи». При користуванні критерієм (1.36) число ступенів свободи покладається як числу розрядів k мінус число незалежних умов ("зв'язків"), накладених на частоти p_i. Прикладами таких умов можуть бути:

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

якщо ми вимагаємо тільки того, щоб сума частот дорівнювала одиниці (це вимога накладається у всіх випадках); або ж

$$\sum_{i=1}^{k} x_i p_i = m_x,$$

якщо ми вимагаємо, щоб збігалося статистичне середнє з гіпотетичним, або ж

$$\sum_{i=1}^{k} (x_i - m_x)^2 p_i = D_x,$$

якщо ми вимагаємо, крім того, ще й збіги дисперсій тощо.

Для розподілу χ^2 складені таблиці (див. таблицю 1.10). Користуючись ними, можна для кожного значення χ^2 і числа ступенів вільності r знайти імовірність p того, що величина, розподілена за законом χ^2, перевершить це значення. У таблиці 3 входами є: значення ймовірності p і число ступенів свободи r; числа, що стоять в таблиці, являють собою відповідні значення χ^2.

Розподіл χ^2 дає можливість оцінити розбіжність між гіпотетичним розподілом (табл. 1.4) і статистичними (табл. 1.3). Якщо ймовірність p дуже мала (не перевищує обраного нами значення «рівня значущості» α, такого, що подія з ймовірністю α вважається вже практично неможливою), це означає, що дослідні дані суперечать гіпотезі H, яка полягає у тому, що випадкова величина X має розподіл (табл. 1.4): цю гіпотезу треба відкинути. Якщо ж ймовірність p не мала, можна визнати розбіжності між

теоретичним і гіпотетичним розподілами несуттєвими і віднести їх за рахунок випадкових причин. Гіпотезу H можна вважати правдоподібною, або, що найменше, не суперечить дослідним даним.

Підкреслимо, що велике значення ймовірності p (наприклад, близьке до одиниці) аж ніяк не свідчить про велику правдоподібність гіпотези H. Це може говорити, наприклад, про те, що дослідні дані свідомо «підганялися» під бажаний нам розподіл (або просто про те, що число дослідів n недостатньо велике, щоб розподіл величини R став близький до χ^2).

Критерій згоди χ^2 можна застосовувати і для безперервних випадкових величин, якщо, групуючи статистичний ряд, приблизно замінити безперервну випадкову величину X дискретною з можливими значеннями $x_1, x_2, ..., x_i, ..., x_k$, де x_i - середина i-го розряду (табл. 1.5).

Таблиця 1.5

Згрупований статистичний ряд безперервних випадкових величин

X:	x_1	x_2	…	x_i	…	x_k
	P^*_1	P^*_2	…	P^*_i	…	P^*_k

а p_i^* - частота потрапляння випадкової величини X в i-й розряд:

$p_i^* = n_i/n$;

n_i - число значень випадкової величини, що потрапили в i-й розряд ($i = 1, 2, ..., k$);

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

Нехай потрібно вирівняти (згладити) статистичний розподіл (табл. 1.5) за допомогою гіпотетичної щільності $f(x)$. Будемо чинити так само, як для дискретної випадкової величини X, замінюючи частоти p_i^* їхніми гіпотетичними значеннями: $p_i = P\{X \in (x_i, x_{i+1})\}$ - тобто ймовірність потрапляння випадкової величини X в i-й розряд, що обчислюється за формулою:

$$p_i = \int_{x_i}^{x_{i+1}} f(x)dx;$$

замість числа значенні випадкової величини береться число розрядів k. У всьому іншому чинимо і розмірковуємо так само, як для дискретної випадкової величини.

1.7 Порядок виконання роботи

1) Відповідно до варіанту (за номером у списку) взяти вибірку одномірної випадкової величини (табл. 1.6).

2) Побудувати варіаційний ряд.

Варіаційний ряд $z_1, z_2, \ldots, z_n$ отримується з вихідних даних шляхом розташування x_m ($m = 1, 2, \ldots, N$) у порядку зростання від x_{min} до x_{max} так, щоб $x_{min} = z_1 \leq z_2 \leq \ldots \leq z_n = x_{max}$.

3) За допомогою варіаційного ряду побудувати діаграму накопичених частот.

Діаграма будується відповідно до формули:

$$\hat{F}_N(x) = \sum_{j=1}^{\mu_N(x)} \frac{1}{N},$$

де $\mu_N(x)$ – число елементів у вибірці, для яких значення $x_j < x$.

Практично це здійснюється так. На осі абсцис указується значення спостережень x_m (або z_l). Значення за віссю ординат дорівнює нулю лівіше точки x_{min}; у точці x_{min} і далі в усіх інших точках x_m діаграма має стрибок, що дорівнює $1/N$. Якщо існує декілька значень x_m, що збігаються, то у цьому місці на діаграмі відбувається стрибок, що дорівнює λ/N, де λ – кількість точок, що збігаються. Ясно, що для величин $x > x_{max}$ значення діаграми накопичених частот дорівнює 1. Відзначимо, що якщо $N \to \infty$, то $\hat{F}_N(x) \to F(x)$. Будуємо відповідну діаграму (рис. 1.26).

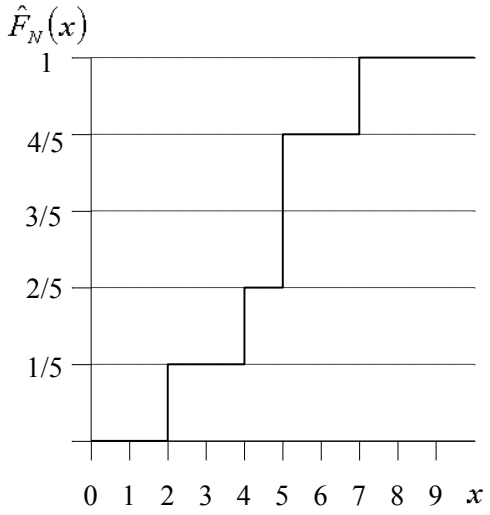

Рис. 1.26 Діаграма

4) Побудувати гістограму вибірки.

Гістограма $\hat{f}_N(x)$ будується таким способом:

1. Знаходиться попередня кількість квантів (інтервалів), на які повинна бути розбита вісь Ох. Ця кількість K визначається за допомогою оцінної формули:

$$K = 1 + 3{,}2 \lg N,$$

де знайдене значення округлюється до найближчого цілого числа.

2. Визначається довжина інтервалу:

$$\Delta x = (x_{max} - x_{min})/K$$

Величину Δx можна дещо округлити для зручності обчислень.

3. Середина області зміни вибірки (центр розподілу) $(x_{max} + x_{min})/2$ приймається за центр деякого інтервалу, після чого легко знаходяться межі й остаточна кількість зазначених інтервалів так, щоб у сукупності вони перекривали всю область від x_{min} до x_{max}.

4. Підраховується кількість спостережень N_m, що потрапили у кожний квант: N_m дорівнює числу членів варіаційного ряду, для котрих справедлива нерівність:

$$x_m \le z_l < x_m + \Delta x,$$

де x_m і $x_m + \Delta x$ – межі m-го інтервалу.

Відзначимо, що значення z_l, що потрапили на межу між $(m - 1)$-м і m-м інтервалами, зараховують до m-го інтервалу.

5. Підраховується відносна кількість (відносна частота) спостережень N_m/N, що потрапили у даний квант.

6. Будується гістограма, що представляє собою східчасту криву, значення якої на m-м інтервалі (x_m, $x_m + \Delta x$) ($m = \overline{1,K}$) постійно і N_m/N дорівнює , або з урахуванням умови $\int_{-\infty}^{\infty} \hat{f}(z)dz = 1$ дорівнює $(N_m/N)\Delta x$.

Відповідна гістограма зображена на рис. 1.27.

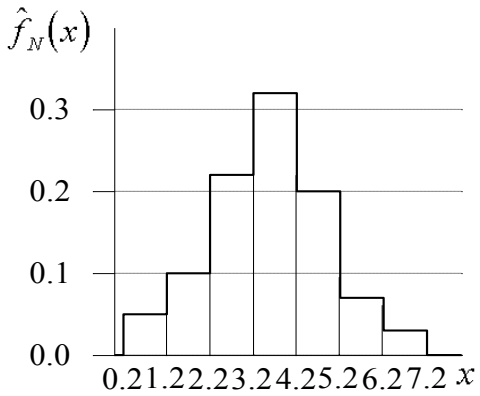

Рис. 1.27 Гістограма

4) Розрахувати оцінки математичного очікування, дисперсії й середнього квадратичного відхилення випадкової величини за формулами

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i ,$$

$$s_x^2 = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \bar{x})^2 ,$$

$$s_x = +\sqrt{s_x^2} .$$

6) Перевірка гіпотези про нормальний закон розподілу випадкової величини.

1.8 Контрольні запитання

1. Що таке детерміновані й що таке випадкові процеси?

2. Яка класифікація детермінованих процесів?

3. Яка класифікація випадкових процесів?

4. У чому полягає аналіз випадкових даних?

5. Які основні статистичні характеристики мають важливе значення для опису властивостей окремих реалізацій стаціонарних випадкових процесів?

6. Які характеристики помилок?

7. Що таке випадкова величина і що таке ряд розподілу?

8. Що таке статистичний ряд і що таке багатокутник розподілу?

9. Що таке функція розподілу і які її загальні властивості?

10. Що таке щільність розподілу випадкової величини і які її загальні властивості?

11. Що таке статистична функція розподілу?

12. Що таке групований статистичний ряд і гістограма?

13. Що таке вирівнювання статистичних розподілів?

14. Які є вичерпні характеристики випадкових величин?

15. Що таке характеристики положення?

16. Що таке основні задачі математичної статистики?

17. Що таке критерій згоди χ^2 Пірсона?

1.9 Приклад виконання роботи

1. З таблиці 1.10 вибираємо данні відповідно варіанту (табл. 1.6).

Таблиця 1.6

Вихідні дані для розрахунку

1	2	3	4	5	6	7	8	9	10	11	12
23,8	29	31,4	43,2	48,6	52	53,4	51,6	47,2	40,4	33	28,2
13	14	15	16	17	18	19	20	21	22	23	24
25	22,2	21,4	19,6	22	22,6	24,4	24,4	25,6	27,6	24,4	24

2. Побудова гістограми вибірки:

Знаходимо попередню кількість квантів (інтервалів), на які повинно бути розбита вісь Ох. Ця кількість K визначається за допомогою оціночної формули:

$$K = 1 + 3{,}2 \log(N) = 5.417$$

Визначаємо довжину інтервалу:

$$\Delta x = (x_{\max} - x_{\min})/K = (53.4 - 19.6)/5 = 6.76$$

Розраховуємо середину області зміни вибірки (центр розподілу):

$$m = (x_{\max} - x_{\min})/2 = (53.4 - 19.6)/2 = 36.5$$

Підраховуємо кількість спостережень, що потрапили у кожний квант, а також відносну кількість (відносну частоту) спостережень, що потрапили у даний інтервал (табл. 1.7).

Таблиця 1.7

Відносна частота і частота потрапляння спостережень у кожний квант

інтервали	0-3,8	3,8-7,7	7,7-11,5	11,5-15,3	15,3-19,2	19,2-23
Відн. частота	2,7	3,1083	2,85	1,325	1,9	2,0833
Част. потрапл.	64,8	74,6	68,4	31,8	45,6	50

5. Будуємо гістограму (рис. 1.28).

Рис. 1.28 Гістограма

3. Визначаємо оцінки математичного очікування $X_{сер}$, дисперсії S^2 і середнього квадратичного очікування S:

$X_{сер} = 31,87$

$S^2 = 47,91667$

$S = 6,922$

4. Перевірка гіпотези про узгодження вибіркового розподілу з нормальним законом на рівні значимості $q = 0.05$ та $q = 0.01$ (табл. 1.12).

Усі розрахунки наведені у табл. 1.8.

$D = (n*h)/S$

$n = 24$ – кількість спостережень;

$h = 3.8$ – крок;

$S = 6.922187$ — середнє квадратичне;

$D = (24*3.8)/6{,}922187 = 13{,}29$

P_m – визначаємо за допомогою таблиць інтегрального закону розподілу (табл. 1.11).

Таблиця 1.8

Номер (А)	C=A-m/S	Pm=F(C)	B'=Pm*D	Дані (В)	$(B-B')^2$	$(B-B')^2/B'$
0	-4,60476	0.097	1,289188	23,8	506,7367	393,0665245
1	-4,4603	0,1238	1,645376	29	748,2755	454,7747301
2	4,31583	0,1518	2,017513	31,4	863,3306	427,9182527
3	-4,1713	0,1849	2,457432	43,2	1659,957	675,4844681
4	-4,02691	0,2203	2,927919	48,6	2085,939	712,4306241
5	-3,88244	0,2541	3,377141	52	2364,182	700,0544008
6	3,73798	0,2897	3,850286	53,4	2455,174	637,6601483
7	-3,59352	0,3209	4,264953	51,6	2240,607	525,3531989
8	-3,44905	0,3503	4,655697	47,2	1810,018	388,7748568
9	-3,30459	0,3739	4,969355	40,4	1255,331	252,6144168
10	-3,16013	0,3894	5,175359	33	774,2107	149,5955471
11	-3,01567	0,398	5,289658	28,2	524,8838	99,22829785
12	-2,8712	0,398	5,289658	25	388,4976	73,44474316

13	-2,72674	0,3894	5,175359	22,2	289,8384	56,00353663
14	-2,58228	0,3739	4,969355	21,4	269,9661	54,32619018
15	-2,43781	0,3503	4,655697	19,6	223,3322	47,9696655
16	-2,29335	0,3209	4,264953	22	314,5319	73,74803354
17	-2,14889	0,2897	3,850286	22,6	351,5518	91,30535689
18	-2,00442	0,2541	3,377141	24,4	441,9606	130,8682702
19	-1,85996	0,2203	2,927919	24,4	461,0503	157,4668926
20	-1,7155	0,1849	2,457432	25,6	535,5785	27,9423717
21	-1,57104	0,1518	2,017513	27,6	654,4637	324,3913248
22	-1,42657	0,1238	1,645376	24,4	517,7729	314,6836308
23	-1,28211	0,097	1,289188	24	515,781	400,0820448
					g(кр)=	7359,187527

Проводимо розрахунки (див. табл. 1.9) і будуємо графік відхилення від нормального закону розподілу (рис. 1.29).

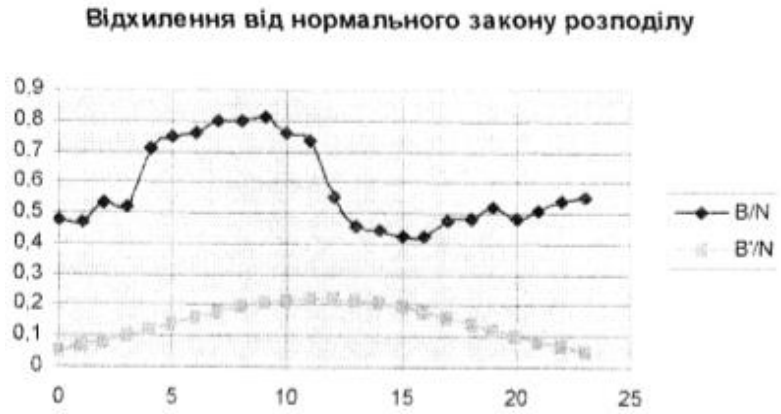

Рис. 1.29 Графік відхилення від нормального закону розподілу

Визначаємо ступінь вільності: v = 24 - 2 -1 =21.

За таблицею χ^2-розподілу для ступеня вільності 21 при

$q = 0,05$ $g_{кр} = 38,9$

$q = 0,01$ $g_{кр} = 32,7$

Таблиця 1.9

B/N	B’/N
0,991667	0,053716
1,208333	0,068557
1,308333	0,084063
1,8	0,102393
2,025	0,121997
2,166667	0,140714
2,225	0,160429
2,15	0,177706
1,966667	0,193987
1,683333	0,207056
1,375	0,21564
1,175	0,220402
1,041667	0,220402
0,925	0,21564
0,891667	0,207056
0,816667	0,193987
0,916667	0,177706

0,941667	0,160429
1,016667	0,140714
1,066667	0,102393
1,15	0,084063
1,016667	0,068557
1	0,053716

Отримане значення g = 997,03

Оскільки g = 997,03 >> $g_{кр}$ = 38,9, то гіпотеза про нормальність випадкової величини суперечить спостереженням і повинна бути відхилена.

1.10 Звіт з практичної роботи

Звіт повинен містити такі матеріали:

- назву, мету, порядок проведення роботи;
- проведені розрахунки і графіки розрахованих залежностей;
- аналіз отриманих результатів, зроблені висновки.

Таблиця 1.10

Вибірка однорідної випадкової величини

Номер	Варіант						
	1	2	3	4	5	6	7
1	11,0	16,8	30,8	41,6	52,0	75,6	93,4
2	4,6	10,8	19,4	23,8	29,0	42,6	54,4
3	10,4	11,8	18,6	25,2	31,4	44,4	56,6
4	4,8	10,6	18,0	24,8	27,6	44,6	55,0
5	10,0	15,2	38,0	41,0	48,6	69,4	87,6
6	5,2	8,8	15,0	19,0	23,8	36,8	48,0
7	11,2	17,0	31,4	43,4	53,4	77,6	96,4
8	4,6	8,6	15,6	19,8	24,4	36,6	45,2
9	10,0	9,0	30,8	40,4	47,2	75,0	97,4
10	7,8	15,8	25,8	32,4	40,4	68,0	91,0
11	3,4	9,0	13,8	19,0	22,0	38,6	51,4
12	5,6	9,2	17,0	22,6	28,2	48,4	67,2
13	4,2	8,2	14,8	19,4	25,0	44,0	64,0
14	3,8	6,8	12,4	17,4	22,2	41,0	56,4
15	4,0	7,0	11,8	17,2	21,4	39,4	53,0
16	7,4	6,8	13,0	17,2	19,6	36,4	50,6
17	4,2	12,4	20,6	27,8	33,0	57,8	79,4
18	4,6	9,2	13,6	18,8	22,6	39,0	50,0
19	6,0	10,4	15,8	21,8	24,4	41,4	50,2
20	5,0	9,6	15,2	19,8	24,4	39,6	52,0
21	3,8	8,8	16,4	22,4	25,6	42,6	54,6
22	8,8	14,6	25,0	34,4	43,2	61,2	74,6
23	4,2	18,4	31,2	40,8	51,6	80,6	101,6
24	4,8	7,6	15,0	20,6	24,0	33,2	43,2

Таблиця 1.10 (продовження)

Вибірка однорідної випадкової величини

Номер	Варіант						
	8	9	10	11	12	13	14
1	4,6	8,8	15,0	19,0	23,8	36,8	48,0
2	4,6	10,8	19,4	23,8	29,0	42,6	54,4
3	5,2	11,8	18,6	25,2	31,4	44,4	56,6
4	8,8	14,6	25,0	34,4	43,2	61,2	74,6
5	10,0	15,2	38,0	41,0	48,6	69,4	87,6
6	10,4	16,8	30,8	41,6	52,0	75,6	93,4
7	11,2	17,0	31,4	43,4	53,4	77,6	96,4
8	11,0	18,4	31,2	40,8	51,6	80,6	101,6
9	10,0	9,0	30,8	40,4	47,2	75,0	97,4
10	7,8	15,8	25,8	32,4	40,4	68,0	91,0
11	7,4	12,4	20,6	27,8	33,0	57,8	79,4
12	5,6	9,2	17,0	22,6	28,2	48,4	67,2
13	4,2	8,2	14,8	19,4	25,0	44,0	64,0
14	3,8	6,8	12,4	17,4	22,2	41,0	56,4
15	4,0	7,0	11,8	17,2	21,4	39,4	53,0
16	3,4	6,8	13,0	17,2	19,6	36,4	50,6
17	4,2	9,0	13,8	19,0	22,0	38,6	51,4
18	4,6	9,2	13,6	18,8	22,6	39,0	50,0
19	6,0	10,4	15,8	21,8	24,4	41,4	50,2
20	5,0	9,6	15,2	19,8	24,4	39,6	52,0
21	4,8	8,8	16,4	22,4	25,6	42,6	54,6
22	4,8	10,6	18,0	24,8	27,6	44,6	55,0
23	4,2	8,6	15,6	19,8	24,4	36,6	45,2
24	3,8	7,6	15,0	20,6	24,0	33,2	43,2

Таблиця 1.10 (продовження)

Вибірка однорідної випадкової величини

Номер	Варіант						
	15	16	17	18	19	20	21
1	2,2	4,0	7,8	11,4	12,8	17,4	21,2
2	2,6	5,6	9,6	11,2	13,8	18,8	24,2
3	2,6	5,8	9,6	12,8	15,0	20,8	25,4
4	2,4	5,6	8,8	12,4	15,4	20,2	25,0
5	3,2	7,0	11,8	17,0	20,2	28,8	32,4
6	3,8	6,4	14,0	18,0	20,6	29,2	34,2
7	3,4	6,8	13,8	18,2	23,0	32,6	38,4
8	3,8	6,4	13,6	19,2	22,2	32,6	39,0
9	4,0	8,0	14,4	19,2	24,2	35,6	41,4
10	4,4	9,8	15,8	19,4	22,4	33,8	41,4
11	4,2	9,4	15,2	18,2	21,6	32,0	39,6
12	4,6	9,2	14,2	17,6	19,2	29,8	37,8
13	3,6	5,8	9,4	13,2	15,4	24,4	32,6
14	2,2	5,2	9,2	11,0	14,0	23,2	31,6
15	1,8	3,6	8,0	10,6	12,8	23,2	30,4
16	2,4	4,2	6,6	10,2	12,2	22,6	28,8
17	1,6	3,6	7,2	10,2	11,2	21,2	27,0
18	2,2	5,6	8,6	11,4	12,4	20,8	26,4
19	2,2	5,2	8,2	11,6	13,8	21,4	25,8
20	3,0	5,4	8,6	12,4	13,6	22,2	26,6
21	2,8	4,6	8,0	11,6	14,2	22,0	28,9
22	2,6	4,4	8,6	12,2	14,2	22,8	26,8
23	2,0	5,0	9,2	13,0	14,0	22,6	26,8
24	2,0	6,4	10,4	13,2	15,2	21,2	23,6

Таблиця 1.10 (продовження)

Вибірка однорідної випадкової величини

Номер	Варіант						
	22	23	24	25	26	27	28
1	0,3	0,3	0,4	0,8	1,3	2,6	3,2
2	0,2	0,3	0,4	1,7	2,7	1,9	2,5
3	0,3	0,6	1,0	2,2	3,5	2,6	3,9
4	0,4	0,9	1,9	2,5	3,9	5,2	7,1
5	1,1	1,5	2,6	4,4	6,0	5,2	6,5
6	1,8	2,8	3,6	6,1	7,9	4,5	7,1
7	1,7	2,6	3,8	7,0	8,6	4,5	7,1
8	1,7	2,9	4,2	6,9	9,1	5,8	7,7
9	1,7	2,5	3,6	6,6	9,3	5,2	7,1
10	1,2	2,2	3,0	5,8	8,6	3,2	5,1
11	0,4	0,8	1,6	4,5	7,2	1,9	3,2
12	0,2	0,4	0,9	3,1	5,5	1,3	1,9
13	0,1	0	0,7	2,2	3,7	0	0,6
14	0,1	0,2	0,4	1,3	3,4	0	0,6
15	0	0	0,4	1,3	2,2	0	0,6
16	0,1	0,2	0,5	1,4	1,9	0	0
17	0,1	0,2	0,3	1,2	1,9	0	0
18	0,1	0	0,4	1,1	1,9	0,6	1,2
19	0,1	0,3	0,5	1,1	1,4	1,3	1,3
20	0,3	0,4	0,6	1,5	1,9	1,3	1,9
21	0,3	0,3	0,4	1,1	1,5	1,3	3,2
22	0,3	0,3	0,5	1,6	2,2	1,9	3,8
23	0,4	0,5	0,6	1,7	2,1	2,6	3,9
24	0,3	0,3	0,3	1,1	1,8	2,6	3,9

Таблиця 1.10 (продовження)

Вибірка однорідної випадкової величини

Номер	Варіант						
	29	30	31	32	33	34	35
1	3,2	5,1	7,0	8,9	10,8	3,4	8,1
2	5,7	7,0	8,9	14,1	16,0	6,0	8,0
3	6,4	8,3	10,2	15,4	17,3	4,7	8,0
4	12,3	13,6	16,2	22,0	24,6	7,3	12,0
5	10,4	13,0	16,2	22,0	28,5	11,2	17,2
6	11,0	14,9	18,1	22,0	28,5	12,0	17,3
7	11,6	14,2	18,7	23,9	28,4	13,2	15,9
8	11,6	13,5	17,4	24,5	31,1	13,4	16,7
9	12,3	14,2	14,8	23,8	30,3	12,7	15,4
10	7,0	11,5	13,4	19,2	28,9	10,0	15,3
11	5,8	10,3	10,9	18,6	24,4	10,0	13,3
12	9,0	9,6	10,2	16,7	22,5	7,4	10,7
13	3,8	5,7	7,0	17,3	19,9	5,4	9,4
14	1,2	3,1	3,7	8,2	14,0	4,7	6,0
15	1,2	2,5	4,4	7,6	11,5	2,7	6,7
16	0,6	1,9	2,5	7,0	11,5	2,7	5,4
17	1,3	1,9	3,2	5,8	11,0	5,3	6,6
18	1,8	3,1	3,1	7,0	15,4	5,3	6,6
19	1,9	4,5	6,4	8,3	11,5	3,4	5,4
20	3,2	3,2	3,8	6,4	10,9	4,3	6,3
21	5,1	6,4	7,0	10,9	14,8	10,9	14,8
22	5,1	6,0	6,6	9,8	14,3	6,7	8,7
23	3,9	5,2	6,5	9,7	12,9	4,7	6,7
24	3,9	5,2	6,5	7,8	11,0	3,4	5,4

Таблиця 1.11

Значення інтегральної функції нормованого нормального розподілу

$$F(u)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u}e^{-u^2/2}du=\int_{-\infty}^{u}f(u)du$$

u	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7703	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,90147
1,3	0,90320	0,90490	0,90658	0,90824	0,90988	0,91149	0,91309	0,91466	0,91621	0,91774
1,4	0,91924	0,92073	0,92220	0,92364	0,92507	0,92647	0,92786	0,92922	0,93056	0,93189
1,5	0,93319	0,93448	0,93574	0,93699	0,93822	0,93943	0,94062	0,94179	0,94295	0,94408
1,6	0,94520	0,94630	0,94738	0,94845	0,94950	0,95053	0,95154	0,95254	0,95352	0,95449
1,7	0,95543	0,95637	0,95728	0,95818	0,95907	0,95994	0,96080	0,96164	0,96246	0,96327
1,8	0,96407	0,96485	0,96562	0,96638	0,96712	0,96784	0,96856	0,96926	0,96995	0,97062
1,9	0,97128	0,97193	0,97257	0,97320	0,97381	0,97441	0,97500	0,97558	0,97615	0,97670
2,0	0,97725	0,97778	0,97831	0,97882	0,97932	0,97982	0,98030	0,98077	0,98124	0,98169
2,1	0,98214	0,98257	0,98300	0,98341	0,98382	0,98422	0,98461	0,98500	0,98537	0,98574
2,2	0,98610	0,98645	0,98679	0,98713	0,98745	0,98778	0,98809	0,98840	0,98870	0,98899
2,3	0,98928	0,98956	0,98983	0,990097	0,990358	0,990613	0,990863	0,9901106	0,991344	0,991576
2,4	0,991802	0,992024	0,992240	0,992451	0,992656	0,992857	0,993053	0,993244	0,993431	0,993613
2,5	0,993790	0,993963	0,994132	0,994297	0,994457	0,994614	0,994766	0,994915	0,995060	0,995201

Таблиця 1.12

q%-ні межі для величини $\chi^2_{кр}$ в залежності від числа v ступенів вільності й рівня значимості q для розподілу Пірсона

$v \backslash q$	99,5 %	97,5 %	95 %	5 %	2,5 %	0,5 %
1	$0{,}39*10^{-4}$	$0{,}98*10^{-3}$	$0{,}39*10^{-2}$	3,841	5,024	7,879
2	0,010	0,050	0,103	5,991	7,378	10,597
3	0,071	0,216	0,352	7,815	9,348	12,838
4	0,207	0,484	0,711	9,488	11,143	14,830
5	0,412	0,831	1,145	11,070	12,832	16,750
6	0,676	1,237	1,635	12,592	14,449	18,475
7	0,989	1,690	2,167	14,067	16,013	20,278
8	1,314	2,180	2,733	15,507	17,535	21,955
9	1,735	2,700	3,3325	16,919	19,023	23,589
10	2,156	3,247	3,940	18,307	20,483	25,188
11	2,603	3,816	4,575	19,675	21,920	26,757
12	3,074	4,404	5,226	21,026	23,336	28,300
13	3,565	5,009	5,892	22,362	24,736	29,819
14	4,075	5,629	6,571	23,685	26,119	31,319
15	4,601	6,262	7,261	24,996	27,488	32,804
16	5,142	6,908	7,962	26,296	28,845	34,267
17	5,697	7,564	8,672	27,587	30,191	35,713
18	6,265	8,231	9,390	28,869	31,526	37,156
19	6,844	8,907	10,117	30,144	32,852	38,582
20	7,434	9,591	10,851	31,410	34,170	39,897
21	8,034	10,283	11,591	32,671	35,479	41,401
22	8,643	10,982	12,338	33,924	36,781	42,796
23	9,260	11,688	13,091	35,172	38,076	44,181
24	9,886	12,401	13,848	36,415	39,364	45,558
25	10,520	13,120	14,671	37,652	40,646	46,928
26	11,760	13,844	15,379	38,885	41,923	48,290
27	11,808	14,573	16,151	40,113	42,194	49,645

28	12,461	15,308	16,928	41,337	44,461	50,993
29	13,121	16,047	17,708	42,557	45,722	52,336
30	13,787	16,791	18,498	43,773	46,979	53,672
31	14,458	17,539	19,281	44,985	48,232	53,003
32	15,134	18,291	20,072	46,194	49,483	56,328
33	15,815	19,047	20,867	47,400	50,725	57,648
34	16,501	19,803	21,664	48,602	51,966	58,964
35	17,192	20,569	22,465	49,802	53,203	60,275
36	17,887	21,336	23,269	50,998	54,437	61,581
37	18,586	22,106	24,075	52,192	55,668	62,882
38	19,289	22,878	24,884	53,384	56,895	64,181
39	19,996	23,654	25,695	54,572	58,120	65,476
40	20,707	24,433	26,509	55,758	59,342	66,766

Практична робота №2

Дослідження і визначення коефіцієнтів регресійної моделі процесу, діапазони змін параметрів якого відомі

Мета роботи - *вивчення методів регресійного аналізу для дослідження і визначення коефіцієнтів регресійної моделі процесу.*

2.1 Загальні положення

Самою ранньою формою регресії був метод найменших квадратів, який був опублікований Лежандром в 1805 році і Гауссом в 1809 році. І Лежандр, і Гаусс застосували цей метод до проблеми визначення з астрономічних спостережень орбіт тіл навколо Сонця (в основному комет, але також і недавно виявлених малих планет). Гаусс опублікував подальший розвиток теорії найменших квадратів в 1821 році, включаючи версію теореми Гаусса - Маркова.

Термін «регресія» був придуманий Френсісом Гальтоном в дев'ятнадцятому столітті для опису біологічного феномена (рис. 2.1). Феномен полягав у тому, що висоти нащадків високих предків мають тенденцію регресувати вниз до нормального середнього (явище, також відоме як регресія до середнього). Для Гальтона регресія мала тільки це біологічне значення, але пізніше його робота була розширена Удні Юлом і Карлом Пірсоном для більш загального статистичного контексту. В роботі Юла і Пірсона спільний розподіл відповідних і пояснювальних змінних передбачається гауссовским. Це припущення було ослаблено Р.А. Фішером. В своїх роботах 1922 і 1925 рр. Фішер припустив, що умовний

розподіл змінної відгуку є гауссовським, але спільний розподіл необов'язково. В цьому відношенні припущення Фішера ближче до формулювання Гаусса 1821 року.

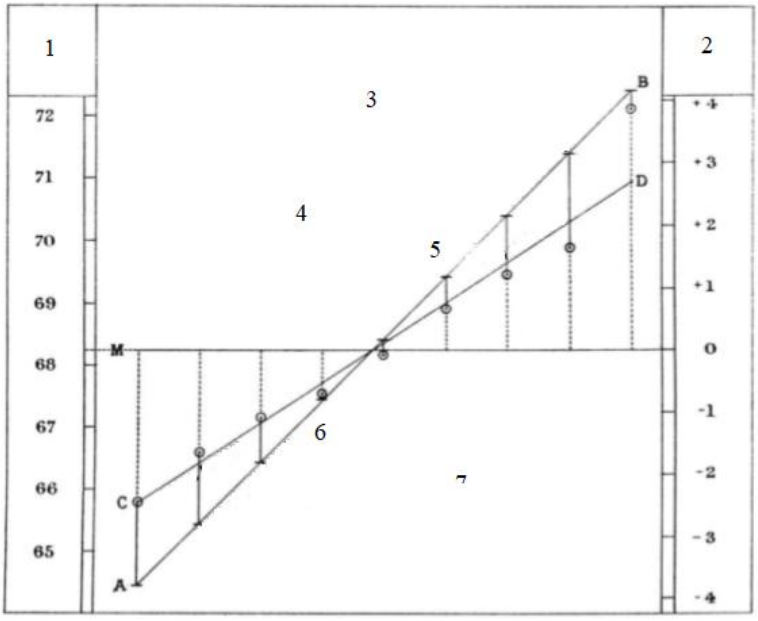

Рис. 2.1 Біологічний феномен

1 — висота у дюймах; 2 — приріст у дюймах; 3 — прирости дітей до їхніх батьків: 2/3; 4 — коли батьки вище середнього зросту, їхні діти мають тенденцію бути нижчими за них; 5 — батьки; 6 — діти; 7 — коли батьки нижче середнього зросту, їхні діти мають тенденцію бути вищими за них

Регресійний аналіз - метод моделювання вимірюваних даних і дослідження їхніх властивостей. Дані складаються з пар значень залежної змінної (змінної відгуку) і незалежної змінної (що пояснює змінну). Регресійна модель є функцією незалежної змінної і параметрів з доданою випадковою змінною. Параметри моделі налаштовуються таким чином, що модель найкращим чином наближала дані. Критерієм якості наближення (цільовою функцією) зазвичай є середньоквадратична помилка: сума

квадратів різниці значень моделі і залежної змінної для всіх значень незалежної змінної як аргументу. Регресійний аналіз - розділ математичної статистики. Передбачається, що залежна змінна є сума значень деякої моделі і деякої випадкової величини. Припущення про характер розподілу цієї величини називається гіпотезою породження даних. Для підтвердження або спростування цієї гіпотези виконуються статистичні тести. При цьому передбачається, що незалежна змінна не містить помилок. Регресійний аналіз використовується для прогнозу, аналізу часових рядів, перевірок гіпотез і виявлення прихованих закономірностей в даних.

Методи регресії продовжують залишатися областю активних досліджень. В останні десятиліття були розроблені нові методи для стійкої регресії, що включає корельовані відгуки, такі як часові ряди і криві зростання, регресії, в якій предиктор (незалежна змінна) або змінні відгуку представляють собою криві, зображення, графіки або інші складні об'єкти даних, методи регресії, що враховують різні типи пропущених даних, непараметрична регресія, байєсовські методи регресії, регресія, в якій змінні предиктора вимірюються з помилкою, регресія з більшою кількістю змінних предиктора, ніж спостереження, і причинний висновок з регресією.

На практиці дослідники спочатку вибирають модель, яку вони хотіли б оцінити, а потім використовують обраний ними метод (наприклад, звичайні найменші квадрати) для оцінки параметрів цієї моделі. Моделі регресії включають в себе такі компоненти:

• невідомі параметри, часто позначаються як скаляр або вектор.

• незалежні змінні, які спостерігаються в даних і часто позначаються як вектор (що позначає рядок даних).

• залежна змінна, яка спостерігається в даних і часто позначається за допомогою скаляра.

• члени помилок, які безпосередньо не спостерігаються в даних і часто позначаються за допомогою скаляра.

У різних областях застосування замість залежних і незалежних змінних використовуються різні терміни.

Мета дослідників - оцінити функцію, яка найточніше відповідає даним. Для проведення регресійного аналізу необхідно вказати форму функції. Іноді форма цієї функції заснована на знаннях про відносини між ними і не залежить від даних. Якщо таких знань немає, вибирається гнучка або зручна форма. Наприклад, можна запропонувати просту одновимірну регресію, припускаючи, що дослідник вважає розумним наближенням статистичний процес, що генерує дані.

Як тільки дослідники визначають свою кращу статистичну модель, різні форми регресійного аналізу надають інструменти для оцінки параметрів. Наприклад, метод найменших квадратів знаходить значення, яке мінімізує суму квадратів помилок. Даний метод регресії в кінцевому підсумку дає оцінку, яка зазвичай позначається для того, щоб відрізнити оцінку від істинного (невідомого) значення параметра, що згенерував дані. Використовуючи цю оцінку, дослідник може потім використовувати підібране значення для прогнозування або для оцінки точності моделі при поясненні даних. Чи буде дослідник зацікавлений в оцінці або прогнозованої цінності, залежить від контексту і цілей дослідження. Широко використовується метод найменших квадратів, оскільки оцінна функція наближається до умовного очікуванню. Однак альтернативні варіанти (наприклад, найменші абсолютні відхилення або квантільна регресія) корисні, коли дослідники хочуть змоделювати інші функції.

Важливо відзначити, що повинно бути достатньо даних для оцінки регресійній моделі. Наприклад, припустимо, що дослідник має доступ до рядків даних з трьома незалежними змінними. Припустимо далі, що дослідник хоче оцінити двовимірну лінійну модель за методом найменших квадратів. Якщо є доступ тільки до точок даних, то можна знайти нескінченно багато комбінацій, які однаково добре пояснюють дані: можна вибрати будь-яку задовільну комбінацію, і всі вони є дійсними рішеннями, які мінімізують суму квадратів нев'язок. Щоб зрозуміти, чому нескінченно багато варіантів, слід звернути увагу на систему рівнянь і необхідність знайти 3 невідомі, що робить систему недовизначеною. Крім того, можна візуалізувати нескінченну кількість тривимірних площин, що проходять через фіксовані точки.

У загальнішому плані, для оцінки моделі найменших квадратів з різними параметрами необхідно мати різні точки даних. Однак, як правило, не існує набору параметрів, який ідеально відповідає даним. Кількість параметрів часто з'являється в регресійному аналізі і називається ступенями свободи в моделі. Крім того, для оцінки моделі найменших квадратів незалежні змінні повинні бути лінійно незалежними: не можна реконструювати будь-яку з незалежних змінних шляхом додавання і множення незалежних змінних, що залишилися. Ця умова гарантує, що це оборотна матриця і, отже, рішення існує.

Регресія - залежність математичного очікування (наприклад, середнього значення) випадкової величини від однієї або декількох інших випадкових величин (вільних змінних), тобто $E\,(y \mid x) = f\,(x)$. Регресійний аналізом називається пошук такої функції f, яка описує цю залежність. Регресія може бути представлена у вигляді суми невипадковою і випадкової складових.

$$y = f(x) + v,$$

де f - функція регресійної залежності,

v - адитивна випадкова величина з нульовим маточікуванням.

Припущення про характер розподілу цієї величини називається гіпотезою породження даних. Зазвичай передбачається, що величина v має гаусовий розподіл з нульовим середнім і дисперсією σ^2_v.

Завдання знаходження регресійній моделі декількох вільних змінних ставиться таким чином. Задана вибірка - множина $\{x_1, \ldots, x_N \mid x \in R^N\}$ значень вільних змінних і множина $\{y_1, ..., y_N \mid y \in R\}$ відповідних їм значень залежної змінної. Ці множини позначаються як D, множина вихідних даних $\{(x, y)_i\}$. Задана регресійна модель - параметричне сімейство функцій $f(w, x)$ залежить від параметрів $w \in R$ і вільних змінних x. Потрібно знайти найбільш ймовірні параметри w^-:

$$w^- = \underset{w \in R^W}{\operatorname{argmax}}\, p(y \mid x, w, f) = p(D \mid w, f)$$

Функція ймовірності p залежить від гіпотези породження даних і задається баєсовим виведенням або методом найбільшої правдоподібності.

Розрізняють регресію за участю однієї вільної змінної і з декількома вільними змінними - одновимірну і багатовимірну регресію. Передбачається, що використовується кілька вільних змінних, тобто, вільна змінна - вектор $x \in R^N$. В окремих випадках, коли вільна змінна є скаляром, вона позначається x. Розрізняють лінійну і нелінійну регресію. Якщо регресійна модель не є лінійною комбінацією функцій від

параметрів, то говорять про нелінійної регресії. При цьому модель може бути довільною суперпозицією функцій g з деякого набору. Нелінійними моделями є, експоненціальні, тригонометричні та інші (наприклад, радіальні базисні функції), які вважають залежність між параметрами і залежною змінною нелінійною.

Розрізняють параметричну і непараметричну регресію. Сувору межу між цими двома типами регресій провести складно. Зараз не існує загальноприйнятого критерію відмінності одного типу моделей від іншого. Наприклад, вважається, що лінійні моделі є параметричними, а моделі, що включають усереднення залежної змінної по простору вільної змінної - непараметричними. Приклад параметричної регресійної моделі: лінійний предиктор, багатошаровий персептрон. Приклад змішаної регресійної моделі: функції радіального базису. Непараметрична модель - ковзке усереднення у вікні деякої ширини. В цілому, непараметрична регресія відрізняється від параметричної тим, що залежна змінна залежить не від одного значення вільної змінної, а від деякої заданої околиці цього значення.

Інтерполяція: функція f задана значеннями вузлових точок (рис. 2.2).

Вибірка може бути не функцією, а відношенням. У такій вибірці одному значенню змінної x відповідає кілька значень змінної y (рис. 2.3).

Лінійна регресія передбачає, що функція f залежить від параметрів w лінійно. При цьому лінійна залежність від вільної змінної x необов'язкова,

$$y = f(w, x) + v = \Sigma w_j g_j(x) + v.$$

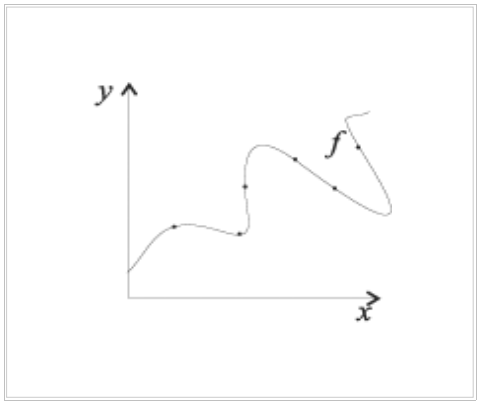

Рис. 2.2 Інтерполяція

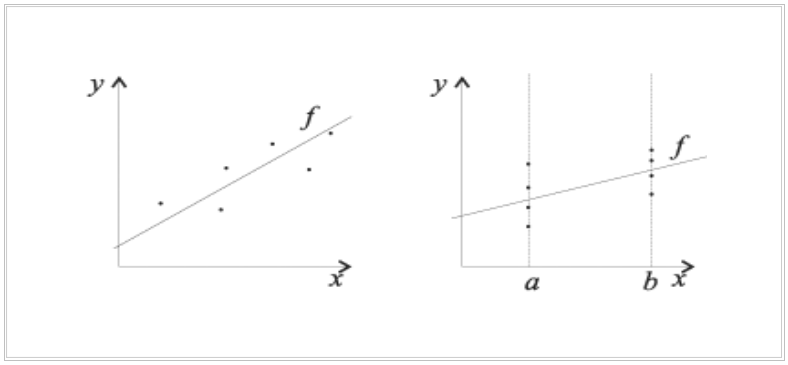

Рис. 2.3 Вибірка

У разі, коли функція $g \equiv \mathrm{id}$ лінійна регресія має вигляд

$$y = \Sigma w_j x_j + v = \langle w, x \rangle + v,$$

тут x_j - компоненти вектора x.

Значення параметрів в разі лінійної регресії знаходять за допомогою методу найменших квадратів. Використання цього методу обґрунтовано припущенням про гаусовський розподілі випадкової змінної.

Різниці y_i - $f(x_i)$ між фактичними значеннями залежної змінної і відновленими називаються регресійний залишками (використовуються

також синоніми: нев'язки і помилки). Однією з важливих оцінок критерію якості отриманої залежності є сума квадратів залишків:

$$SSE = |f(x_i) - y_i|_2 = \Sigma (y_i - f(w, x_i))^2$$

Тут SSE — сума квадратів помилок (Sum of Squared Errors)

Дисперсія залишків обчислюється за формулою

$$\sigma^{-2}_v = SSE/(N\text{-}2) = MSE.$$

Тут MSE - середньоквадратична помилка.

На графіках (рис. 2.4) представлені вибірки, позначені синіми крапками, і регресивні залежності, позначені суцільними лініями. По осі абсцис відкладена вільна змінна, а по осі ординат - залежна. Всі три залежності лінійні щодо параметрів.

Нелінійні регресійні моделі - моделі виду

$$y = f(w, x) + v,$$

які не можуть бути представлені у вигляді скалярного добутку

$$f(w, x) = (w, g(x)) = \Sigma w_j g_j(x),$$

де $w = [w_1, \ldots, w_n]$ - параметри регресійної моделі,

x - вільна змінна з простору R^n,

y - залежна змінна,

v - випадкова величина,

$g = [g_1, \ldots, g_n]$ - функція з деякої заданої множини.

Значення параметрів в разі нелінійної регресії знаходять за допомогою одного з методів градієнтного спуску.

Апроксимація функцій (рис. 2.5): безперервна функція f наближає безперервну або дискретну функцію u.

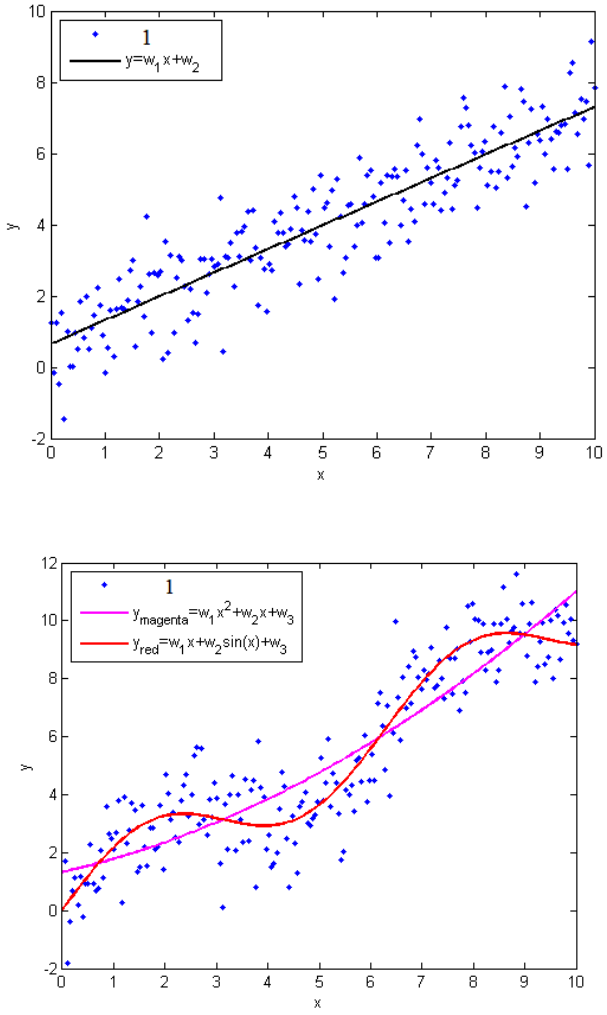

Рис. 2.4 Вибірка і регресивні залежності

1 — експериментальні данні

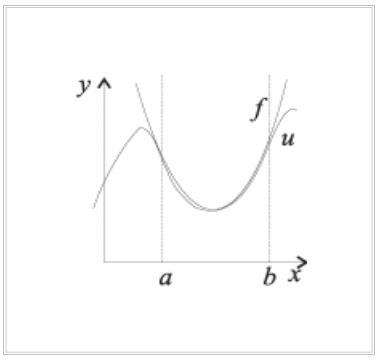

Рис. 2.5 Апроксимація функцій

Є відмінність між термінами: "наближення функцій", "апроксимація", "інтерполяція", і "регресія". Воно полягає в такому.

Наближення функцій. Дана функція *u* дискретного або безперервного аргументу. Потрібно знайти функцію *f* з деякого параметричного сімейства, наприклад, серед алгебраїчних поліномів заданого ступеня. Параметри функції *f* повинні доставляти мінімум деякого функціоналу, наприклад,

$$\rho(u,f)=\left(\frac{1}{b-a}\int_a^b |f(x)-g(x)|^2 dx\right)^{\frac{1}{2}}.$$

Термін апроксимація - синонім терміну "наближення функцій". Найчастіше використовується тоді, коли мова йде про задану функцію, як про функцію дискретного аргументу. Тут також потрібно відшукати таку функцію *f*, яка проходить найближче до всіх точок заданої функції. При цьому вводиться поняття нев'язки - відстані між точками безперервної функції *f* і відповідними точками функції дискретного аргументу.

Інтерполяція функцій - окремий випадок завдання наближення, коли потрібно, щоб в певних точках, званих вузлами інтерполяції співпадали

значення функції *u* і функція *f*, що наближає її. У більш загальному випадку накладаються обмеження на значення деяких похідних *f*, тобто, дана функція *u* дискретного аргументу. Потрібно знайти таку функцію *f*, яка проходить через всі точки *u*. При цьому метрика зазвичай не використовується, проте часто вводиться поняття "гладкості" шуканої функції.

2.2 Елементи регресійного аналізу

Статистичні методи планування активного експерименту є одним з емпіричних методів отримання математичного опису статистики складних об'єктів дослідження, тобто рівняння зв'язку відгуку об'єкта y і незалежних керованих нормованих вхідних змінних (чинників) $\vec{z}^T$ = (z_1, z_2,.., z_n). При цьому математичний опис представляється у вигляді деякого полінома - відрізка ряду Тейлора, в який розкладається невідома залежність в околиці основний точки z_0:

$$M\{y\} = (z_1, z_2, ..., z_n) = \beta_0 + \sum_{i=1}^{n} \beta_i z_i + \sum_{i:l=1}^{n} \beta_{il} z_i z_l + \sum_{i=1}^{n} \beta_{ii} z_i^2 + \ldots, \quad (2.1):$$

де $\beta_i = \left.\frac{\partial \varphi}{\partial z_i}\right|_{\vec{z} = \vec{z}_0}$; $\beta_{il} = \left.\frac{\partial^2 \varphi}{\partial z_i \partial z_l}\right|_{\vec{z} = \vec{z}_0}$; $\beta_{ii} = \frac{1}{2}\left.\frac{\partial^2 \varphi}{\partial z_i^2}\right|_{\vec{z} = \vec{z}_0}$ – теоретичні коефіцієнти.

Внаслідок наявності некерованих і навіть неконтрольованих чинників зміна величина *y* носить випадковий характер, тому функціональна залежність $\varphi(z)$ не дає точної зв'язку між керованими чинниками і відгуком y_g об'єкта в кожному g-м випробуванні, а лише між

керованими чинниками і математичним очікуванням випадкової величини y:

$$M\{y_g\} = \varphi(z_g). \qquad (2.2)$$

Тут $\vec{z}_g^T = (z_{1g}, z_{2g}, ..., z_{ng})$ - g-я точка простору незалежних керованих чинників (факторного простору). У такому випадку за результатами експерименту можна відшукати рівняння регресії у формі деякого полінома:

$$\hat{y} = b_0 + \sum_{i=1}^{n} b_i z_i + \sum_{\substack{il=1 \\ i<l}}^{n} b_{il} z_i z_l + \sum_{i=1}^{n} b_{ii} z_i^2 + \ldots, \quad (2.3)$$

де вибіркові коефіцієнти регресії b_0, b_i, b_{il}, b_{ii},... є лише оцінками для теоретичних коефіцієнтів, відповіднoβ_0, β_i, β_{il}, β_{ii}, ..., а $\hat{y}$ - оцінкою для $M\{y\}$. Нехай $\vec{z}_g$ (g = 1, 2,…, N) – точки факторного простору, в яких проводиться експеримент. Тоді задача відшукання оцінок коефіцієнтів рівняння регресії (2.3) за результатами випробувань в N точках факторного простору є типовою завданням множинного регресійного аналізу в тому випадку, якщо виконуються такі передумови:

1. Результати спостережень y_1, y_2,…, y_N відгуку в N точках факторного простору являють собою незалежні нормально розподілені випадкові величини, тобто на них впливають нормально розподілені випадкові перешкоди ε з нульовим математичним очікуванням $M\{\varepsilon\} = 0$.

2. Дисперсії $\sigma^2\{y_g\}$ (g = 1, 2,…, N) однакові. Це означає, що отримуються при проведенні багаторазових повторних спостережень над величиною y_g в точках z_g вибіркові оцінки $s_g^2\{y\}$ однорідні, дисперсія ж

$\sigma^2\{y_g\}$ не залежить від математичного очікування $M\{y_g\}$, тобто не відрізняється від дисперсії $\sigma^2\{y_g\}$, отриманої при повторних спостереженнях в будь-якій точці z_q факторного простору (відтворюваність з однаковою точністю).

3. Незалежні керовані чинники z_1, z_2,.., z_n вимірюються з нехтувано малими помилками в порівнянні з помилкою у визначенні y (мається на увазі вплив їхніх помилок на величину у порівнянні з впливом некерованих і неконтрольованих чинників ε).

Розглянемо задачу знаходження коефіцієнтів рівняння регресії (3) на прикладі рівняння другого порядку з чотирма незалежними чинниками, при цьому, природно, вся процедура і зроблені висновки можуть бути поширені на рівняння будь-якого ступеня з n незалежними чинниками. Перш за все, спростимо систему позначень: введемо фіктивну змінну $z_0 = 1$ і позначимо $z_0 = f_0$, $z_1 = f_1$, … , $z_4 = f_4$, $z_1^2 = f_5$, …, $z_4^2 = f_8$, $z_1 z_4 = f_9$, …, $z_3 z_4 = f_{14}$ ... У новій системі позначень поліном другого ступеня записується як лінійне однорідне рівняння:

$$\hat{y} = \sum_{i=1}^{14} b_i f_i \text{ . (2.4)}$$

Нехай проводяться випробування в N точках 4-факторного простору X, причому y_g - значення відгуку при випробуванні в точці z_g (g = 1, 2, ..., N). Коефіцієнти рівняння (2.4) знаходяться на підставі методу найменших квадратів, тобто з умови мінімуму суми квадратів відхилень значень відгуку $\hat{y}_g$, передбачених рівнянням (2.4) для умов випробувань в точках z_g, від спостережуваних значень y_g, які виходять при дослідах в цих точках:

$$\sum_{g=1}^{N}\left(y_g - y_g\right)^2 = \sum_{g=1}^{N}\left(y_g - \sum_{j=1}^{14} b_j f_{gj}\right)^2 . \quad (2.5)$$

Оскільки завдання полягає в знаходженні значень коефіцієнтів b_j, що мінімізують вираз (2.5), вона вирішується за допомогою системи так званих нормальних рівнянь, отриманих прирівнянням нулю часткових похідних від квадратичної форми (2.5) по змінним параметрами b_j (j = 0, 1, 2,. .., 14). Система нормальних рівнянь має вигляд:

$$(c_{00}b_0 + c_{01}b_1 + ... + c_{0.14}b_{14} = \alpha_0)(c_{10}b_0 + c_{11}b_1 + ... + c_{1.14}b_{14} = \alpha_1)(.....) \quad (2.6)$$

де

$$C = \{c_{jl}\} = (c_{00} \quad c_{01} \quad ...c_{0.14})(c_{10} \quad c_{11} \quad ...c_{1.14})(.....) \quad (2.7)$$

- матриця коефіцієнтів системи (2.6), елементи якої знаходяться в такий спосіб:

$$c_{jl} = \sum_{g=1}^{N} f_{gj} f_{gl} . \quad (2.8)$$

Матриця $F\{f_{gj}\}$ називається матрицею незалежних змінних, тоді $F^T\{f_{gj}\}$ - матриця, отримана транспонуванням матриці F, причому:

$$F = \{f_{gj}\} = (f_{10} \quad f_{11} \quad ...f_{1.14})(f_{20} \quad f_{21} \quad ...f_{2.14})(.....). \quad (2.9)$$

Можна показати, що $C = F^T F$.

Вільні члени α_j системи нормальних рівнянь визначаються за допомогою рівності:

$$\alpha_j = \sum_{g=1}^{N} f_{gj} y_g \text{ . (2.10)}$$

Для того щоб система (2.6) мала єдине рішення, необхідно і достатньо, щоб матриця С була невироджена, тобто її визначник повинен бути відмінний від нуля: $|C| \neq 0$. Неважко показати, що ця умова зводиться до умови лінійної незалежності вектор-стовпців матриці F, тобто для того щоб система (2.6) мала єдине рішення, необхідно і достатньо, щоб вектор-стовпці матриці F були лінійно незалежні. Аналіз рішення системи (2.6), отримуваного, наприклад, за формулою Крамера:

$$b_l = |C_l|/|C| \quad \left(l = \overline{0,14}\right), \text{ (2.11)}$$

де $|C_l|$ - визначник, що виходить з $|C|$ при заміні елементів c_{il} l-то стовпці відповідними вільними членами α_j, показує, що значення коефіцієнтів b_i зменшення числа членів рівняння впливає на значення коефіцієнтів всіх включених в рівняння членів. Така невизначеність в оцінюванні коефіцієнтів регресії вкрай ускладнює їхню фізичну інтерпретацію. У тому випадку, коли матриця С діагональна, тобто виконується умова:

$$c_{jl} = \sum_{g=1}^{N} f_{gj} f_{gl} = 0 \quad \text{при} \quad j \neq l \text{ , (2.12)}$$

система (2.6) розпадається на незалежні нормальні рівняння:

$$(c_{00}b_0 = \alpha_{0,})(c_{11}b_1 = \alpha_{1,})(.....), \quad (2.13)$$

вирішення яких знаходиться з співвідношень:

$$b_j = \alpha_j / c_{jj} \quad (j = \overline{0,14}). \quad (2.14)$$

При цьому вдається позбутися від невизначеності, пов'язаної з неоднозначним оцінюванням коефіцієнтів регресії. Відзначимо, що співвідношення (2.12) є умова ортогональності вектор-стовпців матриці F незалежних змінних. Таким чином, для отримання незалежних один від одного оцінок коефіцієнтів регресії потрібно спланувати експеримент так, щоб виконувалися умови лінійної незалежності і ортогональності вектор-стовпців матриці F незалежних змінних, або, як будемо її називати, матриці планування.

2.3 Експеримент - основні поняття і терміни

Експеримент - це спеціальним чином спланована і організована процедура вивчення деякого об'єкту дослідження, при якій на цей об'єкт надають заплановані дії і реєструють його реакції на ці дії. Впливу на об'єкт називають чинниками і позначають величинами x_1, x_2, …, $x_к$. Реакції об'єкта називають відгуками і позначають символом y. Експеримент складається з ряду випробувань або спостережень, при яких чинники x_1, x_2, …, $x_к$ мають різне значення. Номер випробування відображають індексом при чинниках і відгуках, тобто для g-ого, наприклад, спостереження матимемо x_{1g}, x_{2g}, …, $x_{кg}$ і y_g.

При організації та плануванні експерименту параметри поведінки об'єкта дослідження, що цікавлять дослідника - тобто майбутні відгуки y_g, грають роль функції невідомої залежності виду $y= \phi(x_1, x_2, \ldots, x_к)$. Аргументи - експериментальні чинники впливу на об'єкт - визначають шляхом професійної експертизи при побудові логічної моделі об'єкта дослідження. Зрозуміло, це в певній мірі обумовлює суб'єктивний характер майбутньої моделі об'єкта дослідження. Але головна особливість ситуації не в цьому, а в тому, що поведінка реальних об'єктів зазвичай визначається такою безліччю чинників, що все їх включити в модель неможливо. І справа не тільки в тому, що список чинників невичерпний, але і ще і в тому, що багато хто з них можуть бути невідомими навіть професійним експертам. Крім того, збільшення кількості чинників, включених в математичну модель об'єкта, ускладнюють експеримент як за термінами проведення, так і за витратами, аж до того, що може зробити здійснення експерименту неможливим. Зважаючи на викладене, прийнята модель об'єкта за такими чинниками завжди (або майже завжди) є неповною. А тим часом реальну поведінку об'єкта складається під впливом всіх чинників - і включених в експеримент, і не включених у нього, тобто це поведінка відповідає не залежності $y= \phi(x_1, x_2, \ldots, x_к)$, а залежності $y= \phi(x_1, x_2, \ldots, x_к, w_1, w_2, \ldots, w_к)$, де w_n - невраховані чинники. Вплив неврахованих чинників робить відгук об'єкта y_g непередбачуваних за значенням величиною, тобто величиною випадковою. Значення випадкової величини, таким чином, складається з рівняння

$$y = \varphi(\bar{x}) + \delta(\bar{w}),$$

де $\varphi(\bar{x})$ - функція істинного відгуку, що відображає вплив включених в модель чинників;

$\delta(\bar{w})$ - функція неврахованих чинників або функція шуму.

2.4 Особливості зв'язку між випадковими величинами

В математиці поняття залежності між величинами виражається поняттям функції y= ϕ(x), коли одному значенню аргументу x відповідає одне, і тільки одне, значення функції y. Якщо зі зміною величини x величина y не змінює свого значення, ці величини є незалежними.

Але бувають і інші ситуації. Графічний вид такої залежності наведено на рис. 2.6.

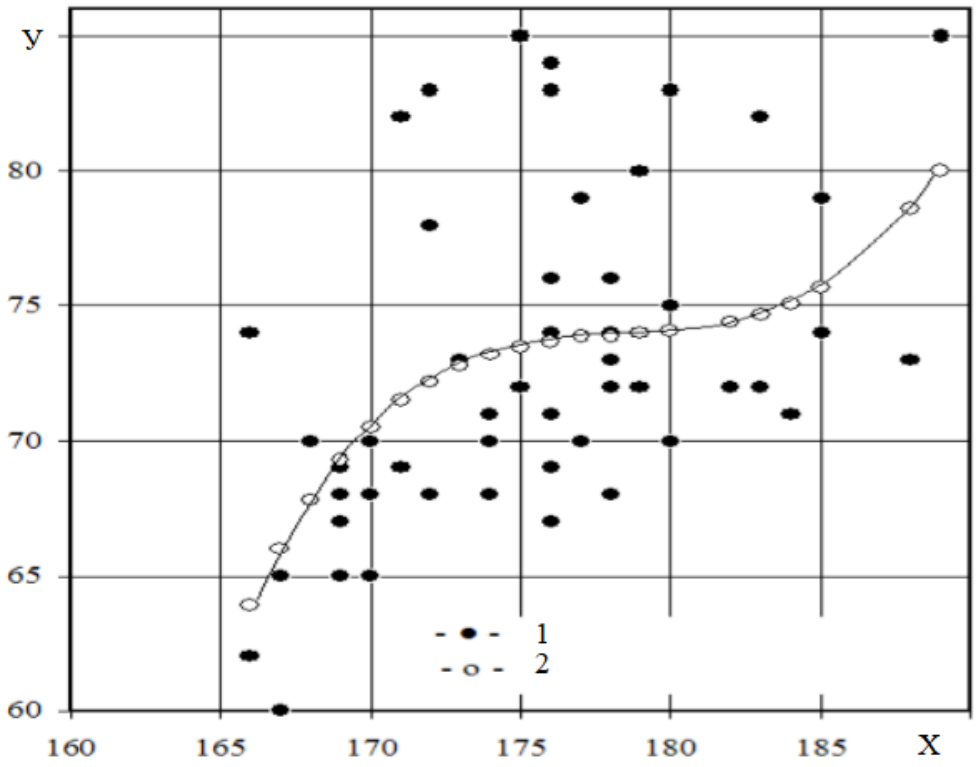

Рис. 2.6. Залежність y = f(x) при слабкому зв'язку

1 - експериментальні точки; 2 - розрахункові точки

Незважаючи на очевидну залежність між величинами x і y, є рівняння, що описує цю залежність. Особливості таких залежностей полягають перш за все в тому, що графік має вигляд слабкоорієнтованої хмари точок і в тому, що одному значенню аргументу може відповідати кілька значень функції. Виходить, що даному значенню аргументу може відповідати кілька значень функції. Виходить, що для даного значення аргументу може випасти або одне, або інше значення функції - тобто з'являється ймовірність того чи іншого значення. Тому такий вид зв'язку між величинами називається ймовірнісним або стохастичним зв'язком.

Такий вид зв'язку може бути обумовлений тим, що в математичну модель об'єкта і в експеримент був включений тільки один аргумент-чинник, хоча існують і інші чинники, що впливають на функцію. У загальному випадку стохастичний зв'язок між випадковими величинами має місце тоді, коли вони мають як загальні, так і різні аргументи, наприклад, $y = f(u, \varepsilon)$ и $x = \phi(u, \gamma)$. Якщо вплив загального аргументу буде нульовим, x і y будуть незалежними. Якщо вплив різних аргументів буде нульовим, зв'язок x і y буде функціональним. Це два крайніх положення, а між ними знаходиться множина різних по силі станів стохастичною зв'язку. При цьому зміна величин x і y складатиметься з двох складових:

- власне стохастичною під дією загального аргументу u;

- випадкової складової під дією різних аргументів ε і γ.

Співвідношення між цими складовими може бути різними, відповідно до цього стохастична зв'язок може бути сильним або слабким, що проілюстровано на графіку. Сильний зв'язок на графіку дає щільну доріжку точок, тобто хмару їх вузька і має виражену спрямованість. У межі ця ситуація зводиться до лінії, тобто функції. Слабкий зв'язок ілюструється рис. 2.6 - хмара розмита, орієнтованість напрямку

проявляється слабо. У межі ситуація зводиться до повної хаотичності в розташуванні точок - тоді залежність між випадковими величинами відсутня. Приклад сильної стохастичного зв'язку ілюструється рисунком 2.7.

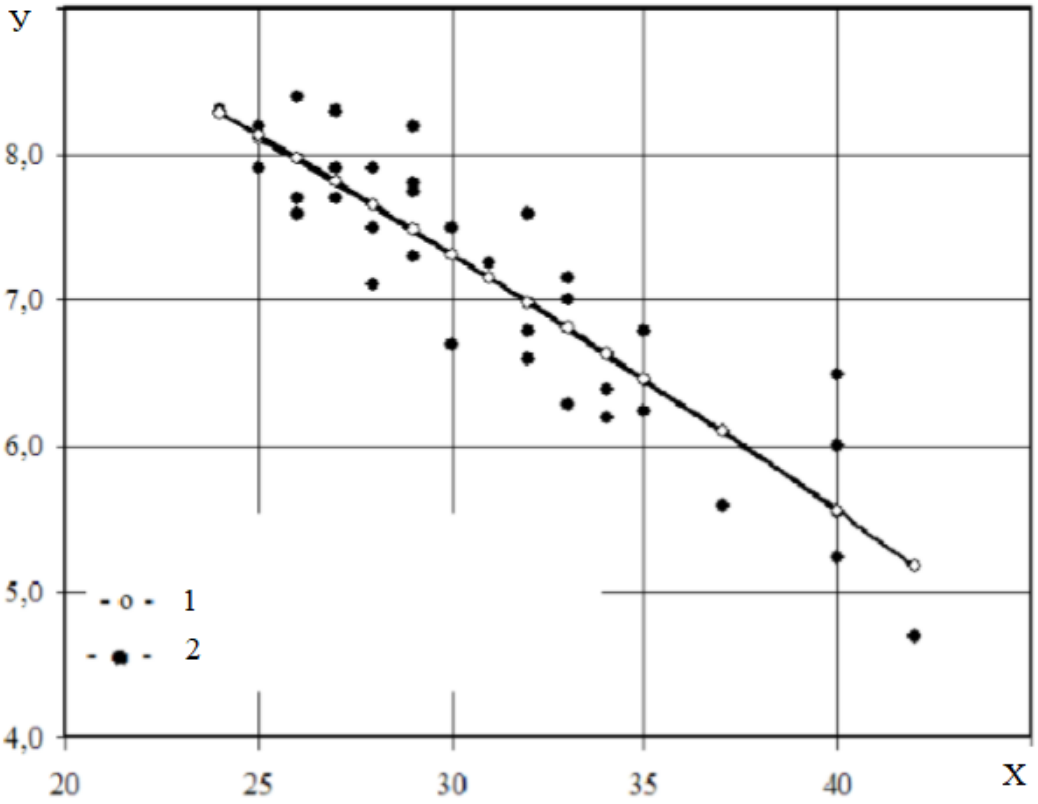

Рис. 2.7. Залежність y = f(x) при сильному зв'язку

1 - експериментальні точки; 2 - розрахункові точки

Оскільки значення випадкової величини при даних аргументах непостійна і його повна характеристика вимагає розсіювання щодо генерального середнього - математичного очікування, остільки стохастичний зв'язок визначають як такий зв'язок, при якому зміна однієї величини викликає зміну закону розподілу іншої.

Наведені приклади показують необхідність кількісної оцінки зв'язку. Таку оцінку можна вивести з положення математичної статистики, що дисперсія суми незалежних величин дорівнює сумі їх дисперсій, тобто $D\{x + y\} = Dx + Dy$. Оскільки $Dz = M\{(z - Mz)^2\}$, можна записати

$$D\{x + y\} = M\{[(x + y) - M\{(x + y)\}]^2\}.$$

Символ математичного очікування суми розноситься по складовим цієї суми, тому

$$\begin{aligned} D\{x + y\} &= M\{(x + y + Mx + My)^2\} = M\{[(x + Mx) + (y + My)]^2\} = \\ &= M\{(x + Mx)^2 + 2(x + Mx)(y + My) + (y + My)^2\} = \\ &= M\{(x + Mx)^2\} + 2M\{(x + Mx)(y + My)\} + M\{(y + My)^2\} = \\ &= Dx + 2M\{(x + Mx)(y + My)\} + Dy. \end{aligned}$$

Як бачимо, у порівнянні з вихідним рівнянням $D\{x + y\} = Dx + Dy$ з'являється інший результат, який містить елемент $2M\{(x + Mx)(y + My)\}$. Очевидно, що при незалежних змінних x і y ця величина дорівнюватиме нулю. При наявності ж стохастичною зв'язку між x і y вона прийме чисельне значення, яке буде тим більше, чим сильніше зв'язок.

Величина $2M\{(x + Mx)(y + My)\}$ називається другим змішаним центральним моментом і позначається як

$$\mu_{11}\{x, y\} = 2M\{(x + Mx)(y + My)\}.$$

Вона і є показником сили стохастичною зв'язку, але тільки не в початковому вигляді, а у вигляді безрозмірною функції - коефіцієнта кореляції

$$\rho\{x, y\} = \mu_{11}\{x, y\}/\sigma_x\sigma_y,$$

де σ - середньоквадратичне відхилення.

При функціональної залежності $y = f(x)$ коефіцієнт кореляції по модулю дорівнює одиниці; при відсутності залежності - нулю. Між цими крайніми значеннями лежить перехідна область стохастичного зв'язку різної сили. Але цей показник справедливий тільки в області лінійного зв'язку.

2.5 Таблиця експериментальних даних

Кожен чинник x для реального натурного об'єкта досліджень має допустимий діапазон значень - від x_{min} до x_{max}. Записавши значення чинників x в таблицю по колонках x_1, x_2,..., x_k і включивши в неї колонку для відгуків y, отримаємо таблицю плану експерименту.

Чисельні значення чинників x і відгуків y і є експериментальними даними. Заповнивши таблицю планування експерименту експериментальними значеннями y_1, y_2,..., y_n, отримаємо таблицю експериментальних даних. Вона і є предметом процесу обробки експериментальних даних.

Будь-яка залежність між змінними x і y може бути представлена різними способами, наприклад у вигляді графіка або в аналітичному вигляді - у вигляді математичної моделі - рівняння, системи рівнянь або

алгоритму. При проведенні експерименту його результатом є уявлення об'єктивно існуючої залежності $y = \varphi (x_1, x_2, ..., x_k, w_1, w_2, ..., w_k)$ у вигляді таблиці експериментальних даних.

Метою обробки експериментальних даних є представлення цієї табличної, аналітично невідомої залежності між змінними x і відгуками y у вигляді математичної моделі, тобто рівняння, що досить точно узгоджує розрахункові і табличні значення відгуку об'єкта y.

2.6 Дисперсія відтворюваності

Звідси випливає, що при багаторазовому повторенні випробування по режиму одного і того ж рядка таблиці експериментальних даних знімаються різні значення відгуку об'єкта при однакових значеннях чинників x. Таким чином, за одиничним випадковим значенням відгуку об'єкта дослідження знаходиться масив випадкових величин (рис. 2.8).

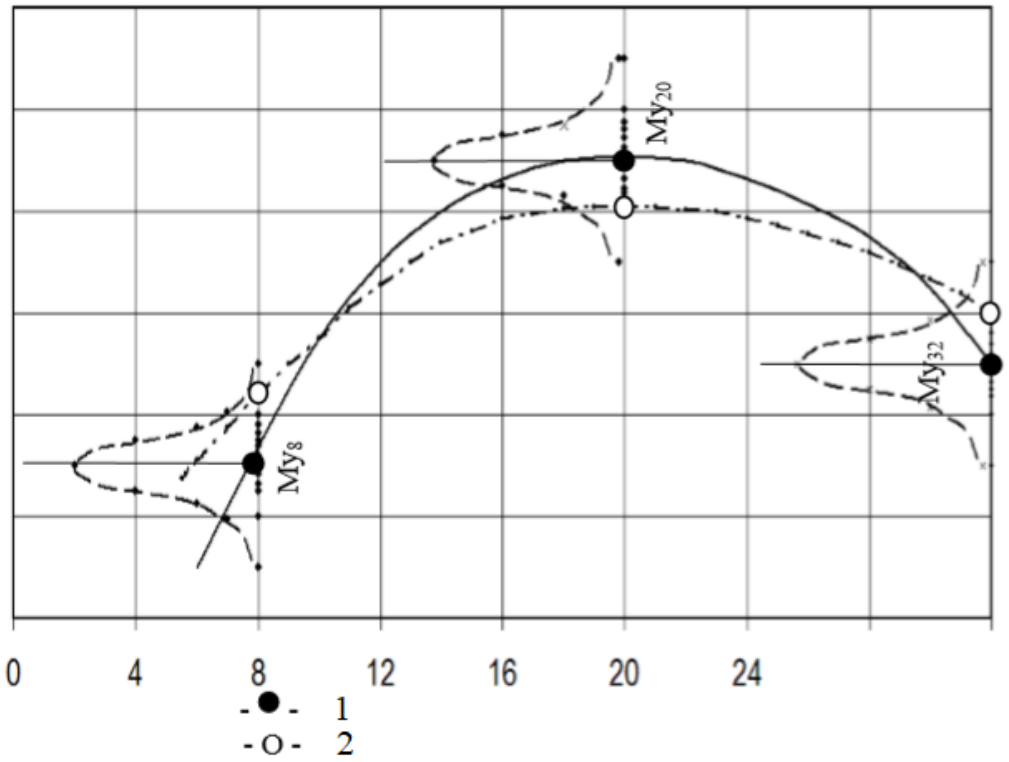

Рис. 2.8 Співвідношення істинної математичної моделі (1) і експериментального рівняння регресії (2)

На рис. 2.8 на горизонтальній осі відкладені номери рядків таблиці, на вертикальній - умовний масив можливих значень відгуків y_g по 10-му, 20-му і 30-му рядку (тобто масиви значень величин y_{10}, y_{20} і y_{30}), що виникає при повторенні одного і того ж спостереження. Кожна з випадкових y_{10}, y_{20} і y_{30} має своє математичне очікування $M\{y_g\}$ і дисперсію σ^2_{yg}. Відповідно до цього побудовані на масивах значень величини y_g графіки законів розподілу цих величин (вертикаль центрів розподілу розташована горизонтально). Позначимо експериментальні значення відгуку світлими точками y_g і з'єднаємо їх лінією, яка буде імітувати експериментально знайдену залежність. Лінія, що проходить через координати математичного очікування $M\{y_g\}$, буде відповідати цій функції істинного відгуку $\varphi(x)$, яку необхідно апроксимувати поліномом

регресії. Звідси випливає, що якби в таблиці експериментальних даних замість випадкової величини y_g стояли постійні величини $M\{y_g\}$, таблична залежність $\varphi(x_1, x_2, ..., x_k)$ втратила б свій стохастичний характер. В цьому випадку система мала єдине рішення в вигляді ідеальної математичної моделі функції істинного відгуку $\varphi(x)$, а саме у вигляді полінома $\eta(x, \beta)$, де β - справжні коефіцієнти «ідеальної» регресії. Модель $\eta(x, \beta)$ адекватна функції $\varphi(x)$ і, таким чином, $\eta(x, \beta) = \varphi(x)$. Але в силу випадкового характеру відгуку об'єкта дослідження, поліном регресії $\eta(x, b)$, знайдений за експериментальними даними, є тільки статистичною оцінкою ідеальної моделі $\eta(x, \beta)$. Звідси випливає, що розраховане за рівнянням регресії значення y_g (позначимо його як y_{gr}) є оцінкою математичного очікування $M\{y_g\}$. Лінія, що проходить через світлі точки, і буде графічною інтерпретацією експериментально знайденого полінома $\eta(x, b)$.

Дисперсія випадкової величини y_g на цьому рядку таблиці σ^2_{yg} є характеристикою поведінки об'єкта дослідження і визначається тільки його природою. Тому значення величини σ^2_{yg} однаково на всіх рядках таблиці даних

$$\sigma^2_1 = \sigma^2_2 = \ldots = \sigma^2_g = \ldots = \sigma^2_k,$$

а сама дисперсія називається дисперсією відтворюваності σ^2_{vos}. Таким чином, графіки розподілу величини y_g на рис. 2.8 відрізняється тільки математичними очікуваннями $M\{y_g\}$, а дисперсії їхні однакові.

Табличне значення величини y_g є експериментальною оцінкою $M\{yg\}$. Надійність оцінок залежить від двох чинників: обсягу вибірки і дисперсії оцінюваної випадкової величини. На рис. 2.9 представлені графіки законів розподілу трьох випадкових величин при одному значенні

математичного очікування і різних значеннях дисперсії. Наочно ілюструється то становище, що чим більше дисперсія, тим більше згладжена крива розподілу і тим більша ймовірність того, що експериментальне значення відгуку y_g буде далі від «ідеального» значення $M\{y_g\}$. Тому різниця (y_g - $M\{y_g\}$), яка обумовлена впливом шуму $\delta\{w\}$, можна розглядати як «помилку» експериментального визначення значення відгуку y_g, а дисперсію σ^2_{vos} як міру цієї помилки. Це обумовлює особливе значення дисперсії відтворюваності для обробки експериментальних даних.

Дисперсія відтворюваності є мірою початкової помилки всієї процедури обробки експериментальних даних, початком «координат помилки». Тому, порівнюючи її по ходу виконання процедури з такими показниками міри помилки, можна оцінити ступінь точності досягнутих результатів.

2.7 Перша частина процедури регресійного аналізу. Знаходження рівняння регресії

Умови (передумови) застосування методу регресійного аналізу. Найбільш поширеним способом обробки експериментальних даних є метод регресійного аналізу, зокрема його варіант, що включає:

- використання методу найменших квадратів;

- відображення невідомої функції істинного відгуку $\varphi(x)$, «захованої» в таблиці експериментальних даних, алгебраїчним степеневим поліномом $\eta(x, b)$.

Метод регресійного аналізу застосовний при дотриманні таких умов:

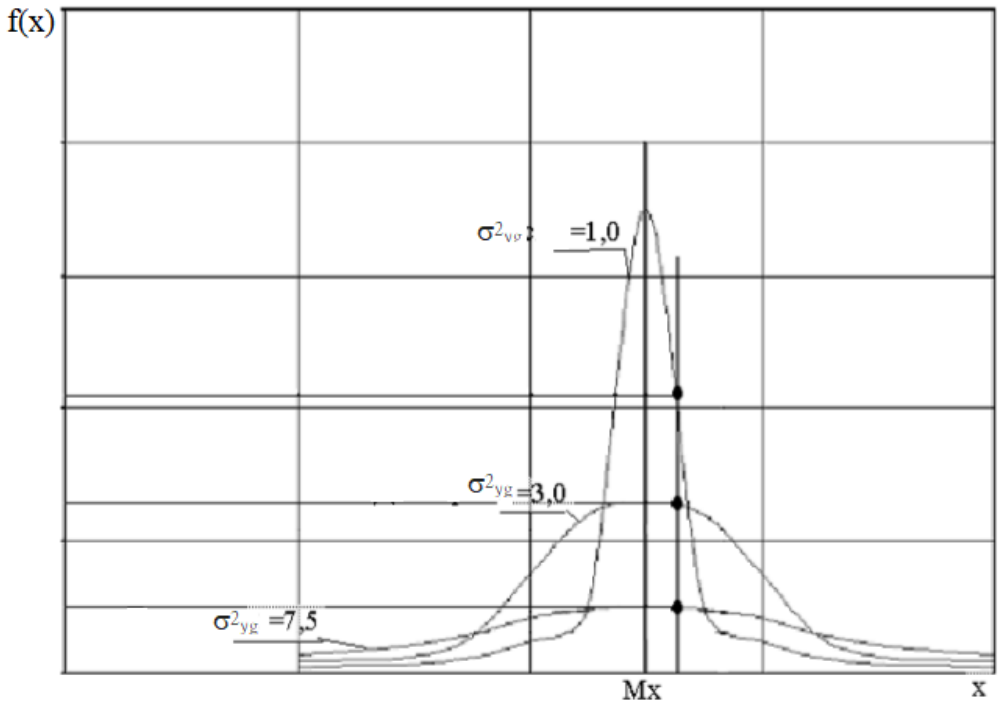

Рис. 2.9 Вірогідність випадання даного значення X в залежності від значення дисперсії

а) масив значень відгуків об'єкта дослідження на даній g-рядку має нормальний розподіл з математичним очікуванням $M\{y_g\} = \varphi(x)$ і дисперсією σ^2_{vos};

б) дисперсії σ^2_{vos} для g = 1, 2, 3, ..., n однакові. Оскільки дисперсія спостереження характеризує точність, з якою отримані спостереження, оскільки випробування при g = 1, 2, 3, ..., n однаковоточні, тобто експеримент проводиться при різних спостереженнях з однаковою точністю;

в) результати спостереження відгуку y_g і їх помилки δ_g в різних дослідах незалежні, тобто $\mu_{11}\{y_j y_q\}$ и $\mu_{11}\{\delta_j \delta_q\}$ дорівнюють нулю;

г) незалежні від відгуку чинника впливу на об'єкт x і похідні від них базисні функції f (x) визначаються в експерименті без помилок в силу двох чинників:

- у разі наявності таких помилок вони «стікають» на відгук об'єкта, збільшуючи розсіювання хмари експериментальних точок;

- вплив цих помилок на розсіювання хмари точок дуже малий в порівнянні з вплив шуму;

д) вектори чинників впливу на об'єкт x і вектори похідних від них базисних функцій f(x) лінійно незалежні, тобто жоден вектор не можна отримати як лінійну комбінацію інших. В іншому випадку визначники похідних від них матриць дорівнюватимуть нулю і матричні розрахунки стануть неможливі;

е) математична модель відгуку об'єкта дослідження $\eta(x, \beta)$ адекватна функції $\varphi(x)$ і, таким чином, $\eta(x, \beta) = \varphi(x)$.

Сформоване таким чином завдання носить назву завдання регресії, експеримент називається регресійним, рівняння (поліноми) - рівняннями (поліномами) регресії, а сам метод вирішення називається регресійним аналізом. Цей термін відображає той факт, що зі збільшенням ступеня полінома, тобто зі збільшенням кількості його членів, в загальному випадку помилка рівняння зменшується - «регресує».

Поліном регресії і система умовних рівнянь. Метод регресійного аналізу використовує опис об'єкта дослідження у вигляді деякого полінома - відрізка ряду Тейлора, в який розкладається невідоме рівняння зв'язку відгуку об'єкта y і вхідних чинників x. При цьому рекомендується така форма полінома, яка містить всі можливі поєднання чинників в першій ступені (одиничні, парні, потрійні тощо), а при ступені більше одиниці - тільки їхні індивідуальні комбінації. Тоді поліном має вигляд

$$M\{y\} = \varphi(x_1, x_2, ..., x_k) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1; j>i}^{k} \beta_{ij} x_i x_j + \sum_{i=1; j>i; q>j}^{k} \beta_{ijq} x_i x_j x_q + ... + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i=1}^{k} \beta_{iii} x_i^3 + ... \quad , (2.15)$$

де β - коефіцієнти, які є похідними вигляду $\partial\varphi/\partial x_i$.

Оскільки за кількістю чинників математична модель об'єкта не може бути вичерпною і зазвичай є неповною, вплив неврахованих чинників робить відгук об'єкта y_g випадковою величиною. Тому залежність φ(x) не дає точної зв'язку між y_g і чинниками, включеними в математичну модель, і за результатами експерименту знаходиться не рівняння (2.15), а рівняння

$$y_g = b_0 + \sum_{i=1}^{k1} \beta_i x_i + \sum_{i=1; j>i}^{k2} \beta_{ij} x_i x_j + ... + \sum_{i=1}^{k3} \beta_{iii} x_i^3 + ... \quad (2.16)$$

де b - вибіркові емпіричні коефіцієнти регресії.

Останні є лише оцінками для теоретичних коефіцієнтів β, а відгук об'єкта y_g - оцінкою для математичного очікування $M\{y_g\}$.

Практика обробки експериментальних даних показала, що результати експерименту у вигляді табличній функції в більшості випадків з достатнім наближенням відображаються повним кубічним поліномом за формою рівняння (2.16). Часто третя ступінь полінома не тільки достатня, але і надлишкова, тобто кількість членів полінома можна і зменшити без істотної втрати точності. Тому при побудові і виборі апроксимовного рівняння будують систему альтернативних рівнянь з повного кубічного полінома і його окремих частин. Порівнюючи характеристики цих рівнянь, вибирають найприйнятніше.

Конкретний вид полінома регресії для даної таблиці даних зазвичай невідомий, як і об'єктивна функція, яка «закодована» цією таблицею. Тому процедура регресійного аналізу починається з висунення гіпотези про конкретний вид рівняння, яке зможе відобразити експериментальну табличну залежність. Вид рівняння регресії задається на основі якихось математичних, фізичних або професійних міркувань, або, при відсутності останніх, - в порядку альтернативи - знаходження для даної таблиці декількох варіантів рівнянь і порівняння їх за точністю відтворення табличного значення відгуку y_g.

Таблиця експериментальних даних і прийнята у вигляді гіпотези форма рівняння регресії є основними відправними умовами завдання і визначають подальший хід її рішення.

Процедура обробки експериментальних даних починається з поєднання прийнятої форми рівняння з таблицею, для чого в рівняння підставляють значення чинників x_{gk} відповідно до рядками таблиці даних, де g - номер рядка таблиці, а k - номер вектора x. Це дає систему рівнянь відповідно до кількості рядків в таблиці експериментальних даних.

Уявімо багаторазово повторюване спостереження, задаючи значення чинників x_{1g}, x_{2g}, ..., x_{kg} для одного і того ж g-го рядка таблиці експериментальних даних. Значення відгуків при цьому в силу наявності шуму в цілому буде різними, тобто значення випадкової помилки спостереження при повторних дослідах буде змінюватися. Розподіл таких помилок має важливу особливість - помилки, протилежні за знаком і близькі по абсолютній величині, в середньому зустрічаються однаково часто, тобто розподіл випадкових помилок симетрично відносно нуля.

Звідси випливає, що якщо всі допустимі значення по y_g по цьому рядку є генеральна сукупність, то істинний результат спостереження є

математичне очікування випадкової величини y_g по цьому рядку. Третя передумова регресійного аналізу свідчить, що спостережуване значення відгуку y_g є нормально розподілена випадкова величина з центром

$$M\{y_g\} = \varphi(x_g).$$

Таким чином, рівняння регресії, отримане в результаті обробки експериментальних даних, є залежність оцінки математичного очікування відповіді від чинників x. Сумарної характеристикою відхилення розрахункового значення відгуку від експериментального його значення є залишкова сума

$$SUM_{ost} = \sum_{g=1}^{n} \left(y_g - y_{gr}\right)^2 = \sum e^2 g \,. \quad (2.17)$$

Ця величина дозволяє сформулювати поняття найкращого рішення системи рівнянь, яка не має однозначної відповіді. Найкращим буде рішення, яке мінімізує залишкову суму. Таке рішення можна отримати методом найменших квадратів. У точці мінімуму функції (2.17) її похідні $\partial SUM_{ost}/\partial b_j$ дорівнюють нулю. Диференціюючи рівняння (2.17) за всіма коефіцієнтами регресії і прирівнюючи нулю похідні, одержимо систему нормальних рівнянь (2.15), яка сумісна, має єдине рішення і мінімізує залишкову суму. Але для багатофакторних поліномів високих ступенів спосіб створення системи нормальних рівнянь через частинні похідні складний і трудомісткий. Існує простіший спосіб побудови системи нормальних рівнянь шляхом покрокового перетворення системи умовних рівнянь.

Перетворення системи умовних рівнянь за методом найменших квадратів. Система нормальних рівнянь. Покрокова процедура перетворення системи умовних рівнянь в систему нормальних рівнянь була розроблена Гаусом. На першому кроці процедури кожне рівняння системи (9) множиться на свій множник при першому коефіцієнті регресії b_0, після чого всі перетворені таким чином умовні рівняння складаються зверху вниз; сумарне рівняння і буде першим нормальним рівнянням.

На другому кроці кожне вихідне умовне рівняння множиться на свій множник при другому коефіцієнті b з наступним складанням отриманих рівнянь і утворенням другого нормального рівняння - тощо, до вичерпання всіх множників при коефіцієнтах b. В результаті формується система нормальних рівнянь, число яких дорівнює числу коефіцієнтів регресії.

Система нормальних рівнянь сумісна, має єдине рішення і мінімізує залишкову суму (2.17), тобто забезпечує найкраще рішення системи рівнянь з усіх можливих рішень.

Основне рівняння процедури регресійного аналізу. Ліва частина системи рівнянь являє собою добуток матриці на вектор коефіцієнтів b. Ця матриця називається матрицею базисних функцій, що позначається F. Кількість рядків в ній дорівнює кількості рядків в таблиці, а кількість стовпців - числу коефіцієнтів b в рівнянні регресії неважко бачити, що її зміст визначається формою полінома, а точніше - вектором базисних функцій.

Ліва частина системи нормальних рівнянь являє собою добуток матриці на вектор коефіцієнтів b. Виділяючи матрицю з системи рівнянь, отримаємо квадратну симетричну матрицю, розмірність якої дорівнює числу коефіцієнтів b в рівнянні регресії. Ця матриця називається матрицею моментів M.

Таким чином, ліву частину системи рівнянь можна представити у вигляді добутку $\bar{b}M$. Можна показати, що матриця моментів

$$M = F^{\mathrm{T}}F,$$

Права частина системи рівнянь являє собою суми парних добутків. Звідси видно, що права частина системи нормальних рівнянь є $\bar{y}_g F^T$ добутком матриці на вектор відгуків yg. Виділяючи матрицю отримаємо транспоновану матрицю F^{T}. Таким чином, права частина рівнянь є добутком і вся система нормальних рівнянь може бути представлена матричним рівнянням

$$\bar{b}M = F^T \bar{y}_g ,$$

звідси випливає

$$\bar{b} = M^{-1}(F^T \bar{y}_g) = (F^T F)^{-1}(F^T \bar{y}_g).$$

Це рівняння називається основним рівнянням процедури регресійного аналізу. З рівняння випливає, що рішення задачі регресії визначається видом матриці F і вектором y_{g}.

Знаходження вектора коефіцієнтів b, тобто отримання рівняння регресії, і становить першу частину процедури регресійного аналізу. Після знаходження полінома регресії слід оцінити адекватність його функції істинного відгуку, тобто точність, з якою рівняння регресії відображає таблицю експериментальних даних. Вирішення цього завдання і становить другу частину процедури регресійного аналізу.

Коефіцієнт регресії $\bar{b}$ *як статистичні оцінки та їх властивості*. Вектор відгуків об'єкта дослідження y_g- є випадкова величина в зв'язку з дією неврахованих в експерименті чинників. Вектор коефіцієнтів регресії b- пов'язаний з вектором y_g- лінійно, і в силу того має той же випадковий характер з тим же законом розподілу. Випадкової є і розрахункові значення y_{gr} за рівнянням регресії.

Рішення системи нормальних рівнянь за формулою Крамера дозволяє зробити висновок, що значення коефіцієнтів $\bar{b}$ залежить від кількості членів рівняння регресії, тобто всі коефіцієнти є взаємозалежними випадковими величинами. У рівнянні можуть бути коефіцієнти, значення яких близькі нулю. Проте, просто виключати їх з рівняння не можна; потрібно робити повністю новий розрахунок для іншої форми полінома регресії, тобто без членів, близьких нулю. При цьому значення всіх збережених коефіцієнтів змінюються. Іншими словами, можлива група різних поліномів з приблизно однаковими характеристиками точності для однієї таблиці даних, тобто саме значення j-го коефіцієнта b невизначено і не має фізичного сенсу, що відображає сутність об'єкта дослідження. Звідси випливає, що рівняння регресії слід трактувати лише як певну інтерполяційну формулу, що дозволяє прогнозувати значення відгуку об'єкта в факторному просторі без додаткового випробування.

Проте, завжди потрібно мати на увазі, що поліном регресії може збігтися з змістовною фізико-математичною моделлю об'єкта дослідження. Це зазвичай відразу різко підвищує інформаційну цінність регресійній моделі об'єкту дослідження.

Відзначимо, що точність поліномів зростає зі збільшенням ступеня, тобто кількості коефіцієнтів b в рівнянні. Тому з ростом числа n значення b прагне до β.

2.8 Друга частина процедури регресійного аналізу - статистичний аналіз якості рівнянь регресії

Залишкова дисперсія полінома регресії. Узгодженість між експериментальними (y_g) і обчисленими по знайденому рівнянню регресії значення відгуку y_{gr} в загальному випадку оцінюють не за значенням залишкової суми SUM_{ost}, а по так званій залишкової дисперсії рівняння регресії, яка позначається як S^2_{ost}:

$$S^2_{ost} = \frac{SUM_{ost}}{n-(k+1)} = \frac{\sum_{g=1}^{n}\left(y_g - y_{gr}\right)^2}{n-(k+1)}, \quad (2.18)$$

де (k + 1) - кількість коефіцієнтів b в рівнянні регресії,

n - число рядків в таблиці експериментальних даних, тобто знаменник рівняння є числом ступенів свободи системи.

Оскільки величина y_{gr} є оцінка $M\{y_g\}$, оскільки змінна S^2_{ost} за своїм змістом є сумарною характеристикою відхилення поточних значень випадкової величини від середнього, тобто дисперсією. Таким чином, залишкова дисперсія характеризує розсіювання спостережень щодо оцінки математичної моделі

$$\eta\left(\bar{x}, \bar{b}\right) = \hat{\eta}\left(\bar{x}, \bar{\beta}\right).$$

Залишкова дисперсія є випадковою величиною, так як вона є функція випадкових величин y_g і y_{gr}, тобто вона має своє математичне очікування і свою дисперсію. Можна показати, що

$$M\{S^2_{ost}\} = \sigma^2_{vos},$$

тобто що S^2_{ost} є незміщена оцінка дисперсії відтворюваності.

Залишкова дисперсія S^2_{ost} так само, як і дисперсія відтворюваності σ^2_{vos}, є мірою помилки всієї попередньої процедури обробки даних, але тепер, на відміну від σ^2_{vos}, ця помилка має два джерела. По-перше, як і S^2_{ost} σ^2_{vos}, вона містить помилку експериментального визначення значення y_g. По-друге, вона містить помилку розрахункового визначення значення y_{gr}, тобто помилку рівняння регресії. Таким чином, співвідношення значень σ^2_{vos} і S^2_{ost} може мати два результати. Якщо поліном регресії має помилку, залишкова дисперсія більше дисперсії відтворюваності, причому чим більше помилка рівняння, тим більше різниця між σ^2_{vos} і S^2_{ost}. Якщо ж поліном регресії $\eta(x_-, b_-)$ адекватний функції істинного відгуку $\varphi(x)$, помилка рівняння відсутня, $S^2_{ost} = \sigma^2_{vos}$. Таким чином, зіставлення цих дисперсій дозволяє оцінити точність отриманого рівняння. Оскільки обидві ці змінні є випадковими величинами, порівнювати їх потрібно не за фактичними одиничним значенням, а з урахуванням розсіювання і з використанням інтервальних оцінок, що дозволяє встановити - значимо чи статистично відмінність між порівнюваними величинами. Ця значимість перевіряється за критерієм Фішера F-розподілу, тобто помилка рівняння визнається значущою якщо

$$\frac{S^2_{ost}}{\sigma^2_{vos}} > F_{1-p}, \quad (2.19)$$

де F_{1-p} - значення табличного квантиля розподілу Фішера при прийнятої ймовірності p і ступенях свободи $m_1 = n - (k + 1)$, $m_2 = \infty$,
$(k + 1)$ – кількість коефіцієнтів регресії в поліномі.

Для навчальних розрахунків при p = 0,95 і n = 50 критичною межею довірчого інтервалу орієнтовно можна вважати $F_{1-p} = 1,5$. Якщо відношення (2.19) дорівнює або менше 1,5 - дисперсії статистично невиразні, тобто їх можна вважати однаковими і поліном буде адекватний функції істинного відгуку φ(x). Факт статистичної незначущості відмінності між S^2_{ost} і σ^2_{vos} є абсолютним показником адекватності рівняння регресії функції істинного відгуку, тобто того факту, що знайдене рівняння слід прийняти «в експлуатацію». Якщо умова (2.19) дотримується, рівняння має помилку і необхідно зважити - чи прийнятний рівень цієї помилки або потрібно шукати інше рівняння.

Оцінку точності рівняння регресії за умовою (2.19) можна здійснити тільки при відомому значенні дисперсії відтворюваності. Якщо σ^2_{vos} невідома, доводиться вдаватися до порівняльних критеріям якості для кількох альтернативних поліномів з вибором найбільш точного.

В цьому випадку статистичну значущість відмінності дисперсій альтернативних поліномів проводять за умовою

$$\frac{S^2_{ost-1}}{S^2_{ost-2}} > F_{1-p},$$

де в чисельнику ставиться більша за значенням дисперсія.

Використання S^2_{ost} має місце і при визначенні дисперсії коефіцієнтів регресії за рівняннями (25, 26). Якщо σ^2_{vos} невідома, використовують аналоги цих рівнянь, приймаючи замість σ^2_{vos} її оцінку S^2_{ost}:

$$D\{\bar{b}\} = \left(F^T F\right)^{-1} S^2_{ost} = M^{-1} S^2_{ost} = C S^2_{ost},$$

- для діагональних елементів

$$\sigma^2\{b_j\} = C_{jj} S^2_{ost},$$

- для інших елементів

$$\mu_{11}\{b_j b_q\} = C_{jq} S^2_{ost}.$$

Чим більше за значенням ці величини, тим гірше рівняння. Вони можуть бути використані для порівняння якості альтернативних рівнянь. У граничному випадку - при ідеальної моделі η(x-, b-) вони дорівнюють нулю.

Показник сили стохастичною зв'язку рівняння регресії. Розглянемо дисперсію вектора y_g. Оскільки цей вектор за своїм змістом є вибіркою, дисперсія вектора y_g дорівнюватиме

$$S^2_{yg} = \frac{\sum_{g=1}^{n} \left(y_g - y_{gsr}\right)^2}{n-1}, \quad (2.20)$$

де S^2_{yg} - вибіркова дисперсія,

y_{gsr} - середнє арифметичне за вибіркою величини y_g.

Значення компонента вектора y_g визначається двома чинниками:

- функціональної залежністю $y = \varphi(x_1, x_2, ..., x_k)$,

- впливом функції шуму $\delta(x)$.

Обидва ці чинники визначають і значення дисперсії вектора У.

Конкретний вид аналітичної залежності $y = \varphi(x_1, x_2, ..., x_k)$ невідомий, але її табличний вигляд являє об'єктивно існуючу функцію. У значенні дисперсії S^2_{yg} ця функція представлена складовою yg. Аналогічно суб'єктивна функція $y_{gr} = \eta(x, b)$, якій ми хочемо відобразити об'єктивну функцію $y = \varphi(x_1, x_2, ..., x_k)$, представлена в натуральному виразі (2.18)

$$S^2_{ost} = \frac{SUM_{ost}}{n-(k+1)} = \frac{\sum_{g=1}^{n}(y_g - y_{gr})^2}{n-(k+1)}$$

у вигляді змінної y_{gr}. Таким чином, зіставлення дисперсій S^2_{ost} і σ^2_{vos} може показати, наскільки прийнятий експериментатором вид полінома регресії узгоджується з «об'єктивною реальністю» у вигляді істинного відгуку $\varphi(x)$. Зазначене зіставлення дисперсій проводиться таким чином. Формулу (2.18) представимо у вигляді

$$S^2_{ost} \times [n-(k+1)] = \sum (y_g - y_{gr})^2 . \quad (2.21)$$

Аналогічно рівняння (2.20) представимо у вигляді

$$S^2_{yg} \times (n-1) = \sum (y_g - y_{gsr})^2 . \quad (2.22)$$

Розглянемо відношення рівняння (2.21) до рівняння (2.22):

$$\gamma = \frac{S_{ost}^2 \times [n-(k+1)]}{S_{yg} \times (n-1)} = \frac{\sum (y_g - y_{gr})^2}{\sum (y_g - y_{gsr})^2}. \quad (2.23)$$

Якщо рівняння регресії адекватно ідеальній математичній моделі і функції істинного відгуку, тобто залежність $y = \varphi(x_1, x_2, ..., x_k)$ має не стохастичний, а функціональний характер, то $y_g = y_{gr}$ і $\gamma = 0$. Якщо ж зв'язку між y і x немає і залежність $y = \varphi(x_1, x_2, ..., x_k)$ взагалі відсутня (величини y і x незалежні), то в чисельнику, і в знаменнику рівності (2.23) залишається тільки однакова складова шуму $\delta(w-)$ і $\gamma = 1$. Всі інші значення величини γ, проміжні між кордонами «0» і «1», означають зміну «ступінь функціональності» залежності між y і x. Графічно цю «ступінь функціональності» можна інтерпретувати як тісноту розміщення точок на графіку стохастичної залежності - чим густіше доріжка точок, тим менше значення γ.

На практиці використовують не показник γ, а зворотний йому величину, рівну $\sqrt{(1 - \gamma)}$. Її поведінка аналогічно поведінці коефіцієнта парної кореляції $\rho_{x,\,y}$ - якщо залежність між величинами відсутня, $\rho_{x,\,y}$ дорівнює нулю, якщо залежність функціональна $\rho_{x,\,y}$ дорівнює одиниці. Тому змінну $\sqrt{(1 - \gamma)}$ називають кореляційним відношенням, тоді

$$\theta = \sqrt{1-\gamma} + \sqrt{1 - \frac{\sum_{g=1}^{n} (y_g - y_{gr})^2}{\sum_{g=1}^{n} (y_g - y_{gsr})^2}},$$

2.9 Попередня обробка експериментальних даних

Виключення грубо помилкових даних з варіаційних даних. Попередня обробка експериментальних даних проводиться в основному в двох цілях:

- відсіювання грубих похибок вимірювання, підрахунку або запису цифрового матеріалу;

- оцінка закону розподілу випадкової величини, яка є результатом спостережень і, при необхідності, перехід від цієї величини до іншої, що має нормальний розподіл.

Грубі помилки при фіксуванні значення експериментальних даних - це аномальні, що сильно виділяються значення в варіаційному ряду однорідних даних. Поява таких значень пов'язано або з суб'єктивною помилкою самого експериментатора, або з різким порушенням режиму проведених випробувань (якщо це дійсно помилкові значення!). Такі значення зазвичай носять одиничний характер і проявляються в одному-двох випробуваннях із серії. Не дивлячись на нечисленність, ці значення можуть внести суттєві викривлення в підсумкові результати обробки даних. Тому такі аномальні значення повинні бути безумовно видалені з масиву експериментальних даних, але ...! - аномальні значення не завжди помилкові і іноді ведуть дослідження прямо до нобелівської премії. Бо існує і така причина аномального значення експериментальних даних як стрибкоподібна зміна показників стану об'єкта випробування при зміні параметрів стану середовищ що впливає на нього. Так, наприклад, при монотонній зміні складу або температури металевих сплавів в певному і досить вузькому діапазоні цих змін в сплаві утворюються нові структурні складові (фази), що різко змінюють макроскопічні властивості сплаву. Ще крок в збільшенні чинників впливу - ці фази розчиняються в основі сплаву,

повертаючи вихідний рівень властивостей ... Це і є аномальний «зрив» значень спостережуваних експериментальних даних, виключити які - значить «прогавити» критичний стан матеріалу, здатний в майбутньому стати, наприклад, причиною руйнування якоїсь конструкції.

Найкращим виходом з такої ситуації є повторення серії випробувань, яка містить аномальні результати. Це дозволяє зробити однозначні висновки про те, випадковий аномальний результат чи ні. Але цей вихід не завжди можливий. Найчастіше «аномальність» виявляється на підсумковій обробці експериментального матеріалу. Так чи інакше, визнання результату аномальних спостережень вимагає ретельної професійної експертизи.

Крім питання про причини аномальності результатів даного спостереження є й інше питання - з якого «критичного» значення вважати даний показник аномальним?

У літературі міститься багато рекомендацій для відсіву грубих похибок спостережень. Строго науковий аналіз масиву спостережень в тому відношенні може бути проведений тільки статистичними методами. Кожна груба помилка викликає порушення закону розподілу досліджуваної величини, зміна його параметрів - порушується однорідність спостережень. Тому виявлення грубих помилок можна трактувати як перевірку однорідності випробувань.

Показником помилковості даного спостереження може слугувати лише величина його відхилення від інших спостережень. Сумнівними можуть бути крайні відхилення від інших спостережень. Сумнівними можуть бути крайні відхилення від середнього - як в ту, так і в іншу сторону. Якщо орієнтуватися на закон нормального розподілу, то такі

відхилення симетричні і досліджуються однаково, тобто можна говорити про загальне «крайнє» значення цієї вибірки.

У разі нормального розподілу для одиничного значення даної випадкової величини x при довірчій ймовірності 1-p оцінкою однорідності буде дотримання нерівності

$$|x - M\{x\}| \le U_{1-p} \bullet \sigma, \quad (2.24)$$

де M{x} і σ - відомі параметри розподілу;

U_{1-p} - квантиль стандартного нормального розподілу.

Порушення цієї нерівності, тобто умова

$$|x - M\{x\}| > U_{1-p} \bullet \sigma$$

і буде ознакою грубої помилковості даного значення. Для вибірки обсягом n елементів відповідна довірча ймовірність буде дорівнює $(1 - p)^n$, тобто ймовірність однорідності всіх n подій зменшується з ростом n і при $n \to \infty$ ця ймовірність прагне до нуля.

Якщо x є крайній елемент вибірки, то довірчій оцінці (2.24)

$$(1-p)^n \cong 1 - np.$$

відповідає ймовірність

$$|x - M\{x\}| \le U_{1-p/n} \bullet \sigma,$$

Тоді довірчій ймовірності 1 - p для одного крайнього елемента відповідає оцінка, тобто елемент буде вважатися грубо помилковим, якщо на рівні значущості p

$$|x - M\{x\}| > U_{1-p/n} \bullet \sigma .$$

Все вищевикладене справедливо для випадку, коли відомі параметри розподілу M{x} і σ. якщо ж вони не відомі, то доводиться використовувати їхні вибіркові оцінки x_{sr} і s. Тоді для крайнього елемента робочої статистики

$$t_{раа} = | x - x_{sr} | / s ,$$

буде умова

$$| x - x_{sr} | / s > t_{1-p}$$

Якщо значення робочої статистики потрапляє до лівої зони, крайнє значення не є аномальним. Якщо воно у середній зоні, то необхідний професійний аналіз ситуації та вироблення додаткових аргументів на користь того чи іншого рішення. Якщо $t_{раб}$ у правій зоні, крайнє значення безумовно відкидається.

Приведення розподілу досліджуваної величини до нормального. Причини (умови) процедури регресійного аналізу містять вимоги нормального розподілу відгуку об'єкта дослідження на даному рядку таблиці експериментальних даних. Порушення цієї умови ускладнює проведення другої процедури, оскільки унеможливлює використання

параметрів розподілів, пов'язаних з нормальним: u- та t-розподілів, F-розподілу Фішера та χ^2 розподілу Пірсона. Не можна користуватися квантилами цих розподілів, не можна будувати інтервальні оцінки з допомогою і, відповідно, не можна перевіряти гіпотези про адекватність рівнянь регресії істинної математичної моделі.

Для невеликих вибірок (менше 120 елементів) рекомендується використовувати значення середнього абсолютного відхилення

$$\Delta x = \sum (x_i - x_{sr}) / n .$$

Для вибірки, що має приблизно нормальний розподіл, справедлива умова

$$|\Delta x_i / s - 0{,}7979| < 0.4\sqrt{n} .$$

Для класу вибірок 3 < n < 1000 використовуються значення варіювання розмаху $/x_{max} - x_{min}/$. Для нормального розподілу відношення $/x_{max} - x_{min}/$ до середньоквадратичного вибіркового відхилення повинне лежати у певних межах, що залежать від обсягу вибірки та довірчої ймовірності. Значення нижніх та верхніх меж табульовані.

Перевірка нормальності розподілу може бути проведена за показником асиметрії, яке називається максимальним відносним відхиленням і підпорядковується розподілу Стьюдента. Крайнє значення відкидається як грубо помилкове.

Після виключення аномального значення з варіаційного ряду статистичні характеристики даної вибірки перераховуються для нового обсягу і новий крайній елемент може бути підданий новій перевірці.

Оскільки при використанні вибіркових оцінок виникає їхнє зміщення відносно оцінюваної величини, в робочу статистику повинна бути введена поправка

$$t_{раб} = |x - x_{sr}| / \left(s\sqrt{\frac{n-1}{n}} \right).$$

Межі критичної зони τ_p (де p-відсоткова точка нормованого вибіркового відхилення) виражається через квантелі цієї точки розподілу Стьюдента $t_{p, n-2}$ по співвідношенню

$$\tau_{p,n} = \frac{t_{p,n-2} \cdot \sqrt{n-1}}{\sqrt{(n-2) + \left(t_{p,n-2}\right)^2}}. \qquad (2.25)$$

З урахуванням цього рівняння для вибірок великого обсягу (при n більше 25) рекомендують таку процедуру відсіву аномальних даних:

- вибирають значення x_i з максимальним відхиленням від середнього $/x_i - x_{sr}/$;

- обчислюють значення робочої статистики

$$t_{раб} = |x - x_{sr}| / \left(s\sqrt{\frac{n-1}{n}} \right);$$

- по таблиці t-розподілу знаходять точки $t_{0,05;n-2}$ і $t_{0,001;n-2}$;

- за рівнянням (2.25) знаходять критичні межі $\tau_{0,05;n}$ і $\tau_{0,001;n}$.

Ці точки обмежують три зони:

- ліву до межі $t_{0,05;n-2}$;

- середню між межами $t_{0,05;n-2}$ і $t_{0,001;n-2}$;

- праву від межі $t_{0,001;n-2}$.

Якщо значення робочої статистики потрапляє в ліву зону, крайнє значення не є аномальним. Якщо воно в середній зоні, то є потреба у професійному аналізі ситуації і розробки додаткових аргументів на користь того чи іншого рішення. Якщо $t_{раб}$ в правій зоні, крайнє значення безумовно відкидається.

Приведення розподілу досліджуваної величини до нормального. Передумови (умови) процедури регресійного аналізу містять вимоги нормального розподілу відгуку об'єкта дослідження на цьому рядку таблиці експериментальних даних. Порушення цієї умови ускладнює проведення другої процедури, тому що унеможливлює використання параметрів розподілів, пов'язаних з нормальним: u- і t- розподілів, F-розподілу Фішера і χ^2 розподілу Пірсона. Не можна користуватися квантилями цих розподілів, не можна будувати інтервальні оцінки з їхньою допомогою і, відповідно, не можна перевіряти гіпотези про адекватність рівнянь регресії істинної математичної моделі.

Для невеликих вибірок (менше 120 елементів) рекомендується використовувати значення середнього абсолютного відхилення

$$\Delta x = \sum (x_i - x_{sr}) / n .$$

Для вибірки, що має приблизно нормальний розподіл, справедлива умова

$$\left| \Delta x_i / s - 0.7979 \right| < 0.4\sqrt{n} .$$

Для класу вибірок 3 < n < 1000 використовуються значення варіювання розмаху /x_{max} – x_{min}/. Для нормального розподілу відношення /x_{max} – x_{min}/ до середньоквадратичного вибіркового відхилення повинно лежати в певних межах, що залежать від обсягу вибірки і довірчої ймовірності. Значення нижніх і верхніх меж табульовані.

Перевірка нормальності розподілу може бути проведена по показником асиметрії

$$A_s = \mu_3 / \sigma^3$$

і ексцесу

$$E_k = \left(\mu_4 / \sigma^4\right) - 3$$

(де μ-центральні моменти третього і четвертого порядку). Для перевірки використовуються незмішані оцінки цих показників

$$A_{ns} = \frac{\sqrt{n(n-1)}}{n-2} A_s,$$

$$E_{nk} = \frac{n-1}{(n-2)(n-3)}\left[(n+1)E_k + 6\right].$$

Для наближено нормального закону розподілу ці показники повинні бути близькі до нуля.

Описані методи використовуються для швидкої «приблизної» оцінки нормальності розподілу. Якщо такої оцінки недостатньо, проводять

перевірку гіпотези про нормальність закону розподілу з використанням критерію згоди Пірсона.

Якщо перевірка нормальності розподілу дала негативні результати, слід перетворити дані таким чином, щоб їхній розподіл став нормальним. Такі перетворення проводять, керуючись видом емпіричних поліномів і гістограм частот розподілу досліджуваної випадкової величини.

Після завершення всієї процедури обробки даних для отримання остаточного результату слід виконати зворотні перетворення приведення даних до початкового стану.

Нормування вихідних даних при вирішенні задач регресії. Властивості нормованих величин. Процедуру регресійного аналізу рекомендують вести при нормований-центрованої формі чинників. Вона була введена Гауссом, тому що властивості нормованого-зосереджених величин дозволяють спростити ручні розрахунки. З появою обчислювальної техніки ця обставина втратила своє значення. В даний час цю форму розрахункових величин використовують тоді, коли вона дозволяє проконтролювати правильність проміжних результатів. В даному випадку вона дозволяє проконтролювати правильність розрахунку матриці M.

Різниця між поточним значенням випадкової величини ζ і її середнім (генеральним або вибірковим) називають центрованою випадковою величиною, оскільки вона інтерпретує поточне значення як відрізок від центру (середнє значення), який лежить або зліва від центру (негативне значення) або праворуч - в області позитивних значень. Для обробки даних важливі такі властивості зосереджених величин.

Перша (нульова) властивість: сума зосереджених величин за їхньою сукупністю (вибіркою) дорівнює нулю. Це властивість очевидна, тому що

центрування ділить масив даних на дві рівні частини з протилежними знаками.

Друга (мінімальна) властивість: сума квадратів відхилень поточних значень випадкової величини від їхнього середнього менше, ніж сума квадратів відхилень від будь-якого іншого числа, в тому числі від моди і медіани.

Сума квадратів відхилень s_{otkl} від деякого числа c

$$S_{otkl} = \sum_{i=1}^{n} (z_i - c)^2 = \min .$$

є вимогою найменших квадратів, яка забезпечується отриманням системи нормальних рівнянь. Вона пояснює також, чому величина (2.18) за своєю природою є саме дисперсією: - пояснення в тому, що величина y_{gr} в рівнянні (2.18) є статистична оцінка математичного очікування $M\{y_g\}$ - генерального середнього.

Розділимо центровану величину (ζ_i - $M\zeta$) на середньоквадратичне відхилення σ вихідної величини ζ. Така операція називається нормуванням, тому що середньоквадратичне відхилення тут виступає як міра або норма вимірювання величини (ζ_i - $M\zeta$). Отримана величина Zn називається нормованої:

$$z_{ni} = \frac{z_i - Mz}{\sigma},$$

а сумарна операція центрування і нормування називається стандартизацією масштабу величини ζ.

Фізичний сенс зміною Z_n полягає в тому, що показує, на яке число величин σ відхиляється дане значення ζ_i від свого генерального (або вибіркового) середнього. Таким чином, для нормованої величини початок відліку проводиться від середнього значення ζ_{sr}, а вимір її - в нових одиницях «σ».

Для обробки експериментальних даних важливі два властивості нормованих величин: сума їх по масиву дорівнює нулю в силу першої властивості центрованої величини; сума квадратів нормованих величин дорівнює їхній кількості в масиві. Дійсно, позначаючи нормовано-центровані чинники x як x_n, для вектора розмірності n матимемо

$$\sum xn^2 = \sum\left(\frac{(x_g - xsr}{dx}\right)^2 = \frac{1}{dx^2}\cdot\sum(x_g - xsr)^2 =$$

$$= \frac{1}{\frac{\sum(x_g - xsr)^2}{n}}\cdot\sum(x_g - xsr)^2 = n.$$

Таким чином, Σxn дорівнює нулю, а Σxn^2 дорівнює n. Тоді, замінюючи в матриці моментів відповідні елементи цими результатами, приведемо матрицю M до такого вигляду:

n	0	0	$\sum x_1x_2$	n	n
0	n	$\sum x_1x_2$	$\sum x_1^2x_2$	$\sum x_1^3$	$\sum x_1x_2^2$
0	$\sum x_1x_2$	n	$\sum x_1x_2^2$	$\sum x_1^2x_2$	$\sum x_2^3$
$\sum x_1x_2$	$\sum x_1^2x_2$	$\sum x_1x_2^2$	$\sum x_1^2x_2^2$	$\sum x_1^3x_2$	$\sum x_1x_2^3$
n	$\sum x_1^3$	$\sum x_1^2x_2$	$\sum x_1^3x_2$	$\sum x_1^4$	$\sum x_1^2x_2^2$
n	$\sum x_1x_2^2$	$\sum x_2^3$	$\sum x_1x_2^3$	$\sum x_1^2x_2^2$	$\sum x_2^4$

Такий вид матриці при вирішенні задачі регресії і буде свідченням правильності проміжних розрахунків.

2.10 Приклад виконання роботи

Завдання: дослідити певний процес і визначити коефіцієнти регресійної моделі даного процесу X_1 і X_2, і при відомому діапазоні їхньої зміни і теоретичних даних $У_i$ (табл. 2.1).

Таблиця 2.1

Вихідні дані

Варіант			1	2	3	4	5	6	7	8	9	10
N	X_1	X_2	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
1	-1	+1	400	405	410	415	420	425	430	435	440	445
2	+1	-1	100	110	120	130	140	150	160	170	180	190
3	-1	+1	300	295	290	285	280	275	270	265	260	255
4	+1	-1	100	99	98	97	96	95	94	93	92	91
Варіант			11	12	13	14	15	16	17	18	19	20
N	X_1	X_2	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
1	-1	+1	450	455	400	405	410	415	420	425	430	435
2	+1	-1	200	210	120	130	140	150	160	170	180	190
3	-1	+1	250	245	295	290	285	280	275	270	265	260
4	+1	-1	90	89	97	96	95	94	93	92	91	98
Варіант			21	22	23	24	25					
N	X_1	X_2	Y	Y	Y	Y	Y					
1	-1	+1	455	400	405	410	415					
2	+1	-1	130	140	150	160	170					
3	-1	+1	280	275	270	265	260					
4	+1	-1	90	89	97	96	95					

Вихідні дані

Кількість випробувань m	X_1	X_2	$Y_{теор}$
1	1	1	500
2	-1	1	100
3	1	-1	300
4	-1	-1	90

Рівняння регресії досліджуваного процесу має вигляд:

$$Y = a_0 + a_1 X_1 + a_2 X_2 + a_{12} X_1 X_2$$

Для визначення коефіцієнтів регресійної моделі використовуємо формули

$$a_0 = \frac{1}{m}\sum_{i=1}^{m} Y_i \ , \qquad a_1 = \frac{1}{m}\sum_{i=1}^{m} X_i Y_i \ , \qquad a_2 = \frac{1}{m}\sum_{j,i=1}^{m} X_j Y_i \ , \qquad a_{12} = \frac{1}{m}\sum_{j,i=1}^{m} X_{ij} X_{ji} Y_i$$

В результаті розрахунків отримуємо такі значення коефіцієнтів регресійної моделі:

$a_0 = ¼ (500 + 100 + 300 + 90) = 247,5$

$a_1 = ¼ (500 - 100 + 300 - 90) = 152,5$

$a_2 = ¼ (500 + 100 - 300 - 90) = 52,5$

$a_{12} = ¼ (500 - 100 - 300 + 90) = 47,5$

Поставивши отримані коефіцієнти в вихідне рівняння регресії, знаходимо експериментальні значення $У_i$:

Кількість випробувань m	X_1	X_2	$Y_{теор}$	$Y_{эксп}$
1	1	1	500	500
2	-1	1	100	100
3	1	-1	300	300
4	-1	-1	90	90

Для перевірки адекватності побудованої моделі використовуємо критерій Фішера, який розраховується за формулою

$$K_ф = D_a/D_{ср},$$

де D_a — дисперсія адекватності,

$D_{ср}$ — середня дисперсія.

Дисперсію адекватності розраховуємо за формулою

$$D_a = \frac{\sum_{i=1}^{m} (Y_{iтеор} - Y_{iекс})^2}{m-d},$$

де m - кількість випробувань,

d - кількість параметрів у регресійній моделі, в даному випадку d = 2.

В результаті розрахунків отримуємо:

$$\sum_{i=1}^{m} (Y_{iтеор} - Y_{iексп})^2 =$$

= (500 - 500) 2 + (100 - 100) 2 + (300 - 300) 2 + (90 - 90) 2 = 0.

Внаслідок того, що отримана сума дорівнює нулю, $K_ф = 0$.

Висновок. Результати проведених розрахунків, з використанням критерію Фішера показують, що отриманий закон лінійний і, отже, побудоване рівняння регресії адекватно описує досліджуваний процес.

2.11 Контрольні запитання

1. Що таке регресійний аналіз?
2. Що таке наближення функцій?
3. Які основні елементи регресійного аналізу?
4. Що таке експеримент?
5. Які існують особливості зв'язку між випадковими величинами?
6. Що являє собою таблиця експериментальних даних?
7. Що таке дисперсія відтворюваності?
8. У чому полягає знаходження рівняння регресії?
9. У чому полягає статистичний аналіз якості рівнянь регресії?
10. У чому полягає попередня обробка експериментальних даних?

Список використаної літератури

1. Львовский Е. Н. Статистические методы построения эмпирических формул. - М.: Высшая школа, 1988. 239 с.

2. Бородюк В. П., Вощинин А. П., Иванов А. З. и др. Статистические методы в инженерных исследованиях. - М.: Высшая школа, 1983. - 216 с.

3. Иванов А. З., Круг Г. К., Филаретов Г.Ф. Статистические методы в инженерных исследованиях. Регрессионный анализ. - М.: МЭИ, 1977. - 203 с.

4. Гутер Р. С., Овчинский Б. В. Элементы численного анализа и математической обработки результатов опыта. - М.: Наука, 1970. - 432 с.

5. Пустыльник. Статистические методы анализа и обработки наблюдений. - М.: Наука, 1968. - 288 с.

Список рекомендованої літератури

1. Бахрушин В.Є. Методи аналізу даних: навчальний посібник для студентів – Запоріжжя : КПУ, 2011. – 268 с.

2. Василенко О. А. Математично-статистичні методи аналізу у прикладних дослідженнях: навч. посіб. – Одеса: ОНАЗ ім. О. С. Попова, 2011. – 166 с.

3. Важинський С.Е., Щербак Т І. Методика та організація наукових досліджень : Навч. посіб. – Суми: СумДПУ імені А. С. Макаренка, 2016. – 260 с.

4. Жлухтенко В. І., Наконечний С. І. Теорія ймовірностей і математична статистика: Навч.-метод. посібник. У 2 ч. – Ч. 1. Теорія ймовірностей. – К.: КНЕУ,2000. – 304 с.

5. Жлухтенко В. І., Наконечний С. І. Теорія ймовірностей і математична статистика: Навч.-метод. посібник. У 2 ч. – Ч. 2. Математична статистика. – К.: КНЕУ, 2001.–336 с.

6. Згуровський М.З. Основи системного аналізу: підруч. — К. Видавнича група ВНУ, 2007. — 543с.

7. Конспект лекцій з дисципліни «Дослідження, моделювання та оптимізація зварювальних процесів» для здобувачів вищої освіти другого (магістерського) рівня зі спеціальності 131 «Прикладна механіка» за

освітньопрофесійною програмою «Технології та устаткування зварювання» /Укл.: Носов Д.Г., Перемітько В.В. – Кам'янське, ДДТУ, 2017. – 156 с.

8. Методичні вказівки до проведення практичних занять та до виконання самостійної роботи з курсу «Методи математичної статистики в екології» [Електронний ресурс] / Укл. Сіренко Л.В. – Київ: НТУУ»КПІ», 2012.- http://library.kpi.ua.

9. Мотигін В.В., Павлов С.М. Планування експерименту в інженерних дослідженнях (лабораторний практикум). Навчальний посібник. - Вінниця: ВДТУ, 2001, - 82 с.

10. Пашинський, В. А. Статистичні методи в інженерних дослідженнях : навч. посіб. М-во освіти і науки України, Центральноукраїн. нац. техн. ун-т. - Кропивницький: ЦНТУ, 2020. - 106 с.

11. Савченко О.Г., Валько Н.В., Кавун Г.М., Кузьмич Л.В. Теорія ймовірностей та математична статистика: [базовий курс з прикладами і задачами] – Херсон: РВЦ «Колос», ХДАУ, 2017. – 406 с.

12. Статистичні методи обробки інформації в наукових дослідженнях. Опорний конспект лекцій для здобувачів третього (освітньо-наукового) рівня вищої освіти ступеня доктора філософії (PhD) / Шабатура Т.С. - Одеса : ОДАУ, 2019. - 52 с.

13. Сучасні фізичні та математичні методи досліджень : конспект лекцій для студентів спеціальності 133 «Галузеве машинобудування» / уклад. : В. В. Калініченко. – Краматорськ: ДДМА, 2018. – 74 с.