МІНІСТЕРСТВО ОСВІТИ УКРАЇНИ
Одеський гідрометеорологічний інститут

Школьний Є.П.,
Лоєва І.Д.,
Гончарова Л.Д.

ОБРОБКА ТА АНАЛІЗ ГІДРОМЕТЕОРОЛОГІЧНОЇ ІНФОРМАЦІЇ

Затверджено Міністерством освіти України
як підручник для студентів гідрометеорологічного
напрямку навчання

Одеса 1999 р.

МІНІСТЕРСТВО ОСВІТИ УКРАЇНИ
Одеський гідрометеорологічний інститут

Школьний Є.П., Лоєва І.Д., Гончарова Л.Д.

ОБРОБКА ТА АНАЛІЗ ГІДРОМЕТЕОРОЛОГІЧНОЇ ІНФОРМАЦІЇ

Затверджено Міністерством освіти України
як підручник для студентів гідрометеорологічного
напрямку навчання

Одеса 1999 р.

УДК 551.58

Школьний Є.П., Лоєва І.Д., Гончарова Л.Д.

Обробка та аналіз гідрометеорологічної інформації: - підручник, - К. : Міносвіти України, 1999.- с.

Розглядаються методи статистичного аналізу сукупностей гідрометеорологічних величин, багатовимірного стаистичного аналізу атмосферних процесів та метеорологічних величин. Для судентів, аспірантів гідрометеорологічного напрямку навчання, а також фахівців - гідрометеорологів

Іл. Табл. Бібліогр.: назв.

Рецензенти:

Глушков О.В., д.ф.-м.н., професор, зав. кафедрою вищої та прикладної математики (Одеський гідрометеорологічний інститут).
Відділ метеорології Українського наукового центру екології моря (зав. відділом, д.г.н., професор Тарнопольський А.Г.)

Науковий редактор : д.т.н., професор Школьний Є.П.

ПЕРЕДМОВА

У зв’язку з широким застосуванням методів математичної статистики та теорії випадкових процесів, які є могутнім інструментом досліджень практично у всілякій сфері знання і особливо в гідрометеорологічних науках та практиці обслуговування галузей господарства, в учбові плани вищих учбових закладів, де відбувається підготовка фахівців гідрометеорологічного напрямку, включені спеціальні дисципліни: “Методи обробки та аналізу гідрометеорологічної інформації”, а також “Багатовимірний статистичний аналіз атмосферних процесів та метеорологічних полів”. В цих дисциплінах викладаються статистичні методи аналізу сукупностей гідрометеорологічних величин і метеорологічних полів, методи побудови статистичних моделей гідрометеорологічних прогнозів.

Проходження студентами зазначених учбових курсів пов’язане зі значними труднощами із-за відсутності відповідної учбової літератури. Існуючі численні підручники й учбові посібники з математичної статистики і теорії випадкових функцій ,що створювалися для підготовки фахівців з відповідних галузей науки і техніки, не можуть у повній мірі задовольнити потреби учбового процесу студентів-гідрометеорологів .

Вони, по-перше, утримують специфічний комплекс методів, потрібних для розв’язку відповідних задач, і не відбивають багатьох аспектів теорії, вельми важливих для гідрометеорологічних наук, по-друге, потребують знання специфіки технічних дисциплін.

В основу підручника “Обробка та аналіз гідрометеорологічної інформації” покладені курси лекцій з дисциплін “Методи обробки та аналізу гідрометеорологічної інформації” і “Багатовимірний статистичний аналіз атмосферних процесів та метеорологічних полів”, що викладалися авторами протягом ряду років студентам Одеського гідрометеорологічного інституту. Він розрахований

на студентів спеціальностей : гідрологія, океанологія, метеорологія та агрометеорологія. Більшість розділів підручника може використовуватись студентами екологічного факультету при вивченні дисципліни "Методи обробки інформації", яка міститься у відповідному учбовому плані, та дисциплін, що пов'язані з питаннями охорони природного середовища.

Книга знайде широке використання в науково-дослідній роботі аспірантів та викладачів, а також в практичній діяльності фахівців-гідрометеорологів.

Підручник складається з двох частин. У першій частині зконцентровані методи статистичної обробки та аналізу рядів гідрометеорологічних величин, дослідження їх законів розподілу, кореляційного зв'язку між гідрометеорологічними величинами. Останні розділи цієї частини підручника присвячені методам теорії випадкових функцій, у тому числі й нестаціонарних часових послідовностей.

Велика увага приділяється особливостям практичного застосування цих методів.

Друга частина підручника присвячена методам багатовимірного статистичного аналізу, які покладені в основу задач дослідження статистичної структури метеорологічних полів та побудови статистичних моделей метеорологічних прогнозів. Зокрема, розглядаються основи кореляційного та компонентного аналізів, кластер-аналізу гідрометеорологічних даних, побудови регресійної моделі гідрометеорологічного прогнозу та інші. Кожний розділ цієї частини підручника також ілюструється розв'язками відповідних спеціальних задач.

Таким чином, підручник фактично забезпечує обидві зазначені вище учбові дисципліни. Перелік розділів, які знайшли місце в підручнику, в повній мірі відповідає змісту учбових програм з цих дисциплін.

Оскільки теорія ймовірностей викладається студентам в дисципліні "Вища математика" як одна із окремих її частин, визначення ймовірністних категорій, аксіоми й теореми теорії ймовірностей в підручнику не приводяться, хоча, природно, вони знаходять в ньому широке використання.

При написанні підручника автори намагалися як можна повніше використати попит досліджень, виконаних в зазначеному напрямку школами М.І. Юдіна, Є.П. Борисєнкова, Н.А. Багрова, а також наукових робіт, що проводилися в Одеському гідрометеорологічному інституті.

Підготовку підручника здійснив колектив авторів Одеського гідрометеорологічного інституту. Перший, четвертий, п'ятий розділи першої частини підручника написані кандидатом географічних наук Гончаровою Л.Д., параграф 6.6 першої частини, п'ятий і шостий розділи другої частини підручника підготовлені доктором географічних наук, професором Лоєвою І.Д. Розробка другого, третього розділів і параграфів 6.1-6.5 шостого розділу першої частини, першого, другого, третього і четвертого розділів другої частини підручника, а також загальне наукове редактування здійснені доктором технічних наук, професором Школьним Є.П.

Автори щиро вдячні викладачам кафедри геофізичної гідродинаміки та теорії клімату Одеського гідрометеорологічного інституту за зауваження, що сприяли поліпшенню книги. Особисту подяку автори приносять завідуючому кафедрою вищої математики інституту доктору фізико-математичних наук, професору Глушкову А.В. , який в процесі рецензування підручника висловив корисні зауваження і пропозиції. Вони також висловлюють подяку Серга Є.М. за допомогу у технічній підготовці рукопису.

ВСТУП

Практика ставить перед природними науками задачу пізнання законів, які б дозволили управляти процесами, що відбуваються на Землі. Фізичні та хімічні процеси, що виникають і розвиваються в її сферах, вивчаються багатьма науками. Але гідрометеорологічні науки, які входять до складу наук про Землю, виділяються певними особливостями напрямку розвитку та методологією досліджень. Вони займаються вивченням процесів, що виникають і розвиваються в атмосфері та гідросфері, які є основними ланками кліматичної системи.

Фізичні параметри стану цих двох оболонок Землі складають *гідрометеорологічну інформацію*.

Знання комплексу відповідних статистичних алгоритмів та вміння правильно їх використовувати при аналізі цієї інформації допоможе рішенню актуальних питань утворення, змінення та прогнозування гідрометеорологічних процесів.

Ясно, що емпіричні дослідження в гідрометеорологічних науках мають першорядне значення. На їх основі встановлюються закономірності, які притаманні певним характеристикам атмосфери чи гідросфери. Емпіричні дані є критеріями істинності закономірностей, рівнянь гідродинаміки, особливостей атмосферних чи гідрологічних процесів та тому інше.

Таким чином, *гідрометеорологічна інформація* має важливі *особливості*, які обумовлюються характером процесів, що спостерігаються в перелічених сферах Землі.

Перша з них полягає у тому, що процеси в океані чи атмосфері мають просторові й часові масштаби, які набагато перевищують можливості окремої людини по збиранню та узагальненню інформації про їх стан. Тому дані про процеси в оточуючому середовищі, що збираються з різних регіонів Землі та за тривалі періоди часу, мають надзвичайну цінність для дослідників.

Друга особливість обумовлюється тим, що в науках про Землю, особливо гідрометеорологічних, є дуже обмежені

можливості проведення активного експерименту з природними об'єктами. Отже, аналіз накопичених даних стає головним джерелом досліджень і єдиним засобом перевірки теоретичних висновків та отриманих закономірностей.

Особливості об'єктів, що досліджуються, і методів дослідження підкреслюють важливість систем збирання і накопичення гідрометеорологічної інформації та систем забезпечення доступу до неї багатьох користувачів.

Збирання даних про атмосферу і гідросферу здійснюється, по-перше, з метою оперативного доведення інформації до підрозділів гідрометеорологічної служби, які займаються обслуговуванням різних галузей господарства (прогнози погоди, штормові попередження, тощо) і, по-друге, для накопичення з метою узагальнювання даних про гідрометеорологічний режим та наукових досліджень .

Гідрометеорологічні дані - це кількісні характеристики стану атмосфери і гідросфери. Внаслідок значної мінливості у просторі і за часом фізичних параметрів атмосфери і гідросфери, для спостереження за їх станом з метою вивчення закономірностей процесів, що відбуваються, і, найголовніше, з метою їх прогнозування необхідні численні вимірювання стану цих середовищ. Відомо, що основним джерелом гідрометеорологічної інформації є результати термінових і спеціальних метеорологічних та гідрологічних спостережень і вимірювань, дані аерологічного зондування атмосфери, дані експедиційних досліджень і тому інше.

Значення сукупності гідрометеорологічних величин у даний момент часу визначається станом атмосфери та гідросфери, який обумовлюється дією комплексу фізичних причин. Взагалі кажучи, основні гідрометеорологічні величини є неперервні величини. Це, наприклад, атмосферний тиск, температура і густина повітря, гігрометричні характеристики, швидкість вітру; густина, температура, солоність, швидкість руху води океану тощо. В деяких вимірювальних системах втілюється безперервна регістрація значень тих чи інших фізичних величин. Але в більшості випадків гідрометеорологічні величини вимірюються на світовій мережі

метеорологічних чи гідрологічних станцій та постів через деякі проміжки часу, що встановлюються Всесвітньою Метеорологічною організацією (ВМО) чи особистою програмою досліджень.

Треба зауважити, що і у випадку безперервної реєстрації гідрометеорологічної інформації на тих чи інших носіях перед статистичною обробкою цієї інформації доводиться виконувати її *дискретизацію (квантування)*. Цей процес зводиться до складання рядів значень гідрометеорологічної величини у визначені інтервали часу.

Гідрометеорологічні ряди можуть складатися не тільки з величин безпосередньо вимірних. Їх членами можуть бути і величини, які отримані в результаті узагальнювання первинних вимірювань чи спостережень.

Таким чином, ряди гідрометеорологічних величин складаються з членів, кожний з яких є результатом чи безпосереднього вимірювання або спостереження, чи узагальнювання спостережень за деякий інтервал часу конкретного року.

У подальшому ми познайомимо вас з основними особливостями щодо гідрометеорологічної інформації та статистичних методів її обробки з урахуванням специфіки самої інформації, а також сучасних досягнень в галузі статистичної обробки гідрометеорологічних даних .

ЧАСТИНА І

МЕТОДИ ОБРОБКИ ТА АНАЛІЗУ СУКУПНОСТЕЙ І ЧАСОВИХ ПОСЛІДОВНОСТЕЙ ГІДРОМЕТЕОРОЛОГІЧНИХ ВЕЛИЧИН

1 СТАТИСТИЧНІ ОЦІНКИ ПАРАМЕТРІВ РОЗПОДІЛУ ГІДРОМЕТЕОРОЛОГІЧНИХ ВЕЛИЧИН

1.1 Основні характеристики гідрометеорологічної інформації

Кожний фізичний параметр атмосфери чи гідросфери залежить один від одного, а також від зовнішніх впливів і випадковим чином змінюється за часом та у просторі, утворюючи випадкові поля або послідовності.

Обробка і аналіз систем випадкових величин проводиться за допомогою спеціально розробленого апарату досліджень, що складає методи математичної статистики. Тому гідрометеорологічна інформація повинна задовольняти вимогам, котрі пред'являються до статистичної інформації .

Розглянемо *основні характеристики гідрометеорологічної інформації* .

Однією з важливих ознак рядів є інтервал дискретності. Як правило, ряди гідрометеорологічних величин є еквідистантними, тобто члени рядів визначаються через який-небудь заданий інтервал часу (година, доба, місяць, рік тощо). В деяких випадках при розв'язуванні конкретних задач ряди можуть формуватися із членів, що розташовані на різних відстанях одне від одного .

Ще однією важливою характеристикою ряду гідрометеорологічних величин є його об'єм. Під терміном *об'єм*

сукупності випадкових величин розуміють кількість членів, що складають цю сукупність.

В гідрометеорологічних дослідженнях доводиться мати діло з рядами як великих, так і обмежених об'ємів.

Важливою властивістю ряда гідрометеорологічних величин, що визначає його вид, є характеристика цих величин. Такими характеристиками можуть бути безпосередні значення гідрометеорологічних величин, кількість днів і випадків з атмосферними явищами, їх тривалість, інтенсивність тощо.

Гідрометеорологічні величини можуть бути скалярними або векторними. В останньому випадку ряд являє собою два або більше (в загальному випадку - N) рядів синхронних скалярних характеристик метеорологічної величини.

Отже для гідрометеорологічних досліджень, а також безпосереднього застосування метеорологічної інформації в різних галузях господарства, формується велика множина сукупностей гідрометеорологічних величин, які розрізняються однією або декількома ознаками , а саме :

- інтервалом дискретності ;
- об'ємом сукупності (вибірки) ;
- характеристикою випадкових величин-членів ряду .

Коли кажуть про статистичні сукупності, то мають на увазі дві категорії:

- генеральна сукупність ;
- статистичний ряд (вибірка) .

Термін «*генеральна сукупність»* визначає необмежену кількість незалежних випадкових величин, які підпорядковуються одному закону розподілу. Властивості випадкових величин, які представляються генеральною сукупністю, визначаються параметрами цієї випадкової величини.

Статистичний ряд (вибірка) - обмежена кількість випадкових величин, здобутих випадковим чином із генеральної сукупності. Тому статистичні ряди називають вибірками з генеральної сукупності . Вибірки випадкові та число їх безмежне. Задача дослідника полягає у тому, що б за допомогою вибірки розрахувати деякі оцінки параметрів, котрі б вірогідно характеризували особливості генеральної

сукупності. Із генеральної сукупності, як вже зазначалося, ми можемо мати безмежну кількість вибірок та на основі кожної здобути статистичні оцінки. Значення параметра генеральної сукупності, здобуте на основі вибірки, є *статистичною оцінкою* цього *параметра*, яку позначають символом «^». Якщо взагалі позначити параметр генеральної сукупності як θ, то його оцінка - $\hat{\theta}$.

Оскільки до гідрометеорологічних рядів застосовуються статистичні методи обробки та аналізу, ці ряди повинні задовольняти вимогам, що випливають із умов, покладених в основу цих методів.

Перш за все, кожний *ряд повинний бути однорідним*. Це означає, що всі члени ряду з визначною імовірністю повинні належати до однієї генеральної сукупності, тобто підпорядковуватися визначеному закону розподілу .

В дійсності , в деяких випадках в гідрометеорологічних рядах містяться члени, які не задовольняють сформульованій вимозі. Їх називають «викидами». «Викиди», як правило, виникають тоді , коли спостерігаються аномальні погодні або кліматичні умови .

Наступною вимогою до рядів гідрометеорологічних величин є *незв'язність* їх *членів*. Це означає, що статистична залежність між ними повинна бути відсутньою. Прийняття чи не прийняття цієї вимоги залежить від характеру задачі, що розв'язується. Якщо йдеться про статистичну оцінку моментів випадкових величин, то вихідні ряди повинні бути незв'язними, оскільки методи статистичного оцінювання параметрів спираються на теореми теорії ймовірностей, які, як правило, ставлять вимогу про незалежність випадкових величин .

Зазначену вимогу задовольняють шляхом вибору такого інтервалу дискретності, для якого статистична залежність є незначною (але для цього треба мати апріорну інформацію), або проводити *операцію рандомізації* (від англійського терміну random approximation - *випадкове наближення*). Для останньої може використовуватися відповідна таблиця випадкових чисел або комп'ютерна програма генерації випадкових чисел.

Інша справа, коли ставиться задача дослідження внутрішньої часової статистичної структури гідрометеорологічних величин. Тоді використовуються вихідні часові ряди визначної дискретності, саме такої, щоб статистична залежність між членами ряду проявлялася в тій чи іншій мірі .

Важливе значення при розрахунках оцінок параметрів має об'єм сукупностей .Статистичний *ряд* повинен *володіти представництвом*, тобто бути дійсно вибіркою із генеральної сукупності і мати такий об'єм, який дозволяв би провести оцінку параметрів з заданою точністю, тобто отримати вірогідні оцінки. Ясно, що це можна зробити, коли об'єм вибірок (у згоді з так званим законом великих чисел) є досить великим. Але такі сукупності гідрометеорологічних величин не завжди можна сформувати. В гідрометеорології часто виникає потреба мати справу з малими сукупностями. Це обумовлюється терміном, протягом якого організуються спостереження. У такому випадку потрібно проводити оцінку вірогідності отриманого статистичного параметра .

Метеорологічні (або гідрологічні) ряди необхідно подавати у найбільш зручному для аналізу вигляді в залежності від задачі, що розв'язується.

Найбільш часто сукупності випадкових величин зображаються у двох видах: у виді простого статистичного ряду і у виді згрупованого статистичного ряду .

Первинною формою запису вихідних даних є *простий статистичний ряд*, в якому дані розташовуються в тій послідовності, як вони були отримані в результаті спостережень. Такий ряд об'ємом n має вид :

$$X : x_1, x_2, ..., x_n .$$

Будемо у подальшому позначати випадкові величини великими літерами латинського алфавіту X, Y, Z, а їх конкретні значення - відповідними малими літерами $x, y, z...$

Прикладом простого статистичного ряду є дані, які містяться в табл. 1.1.

Таблиця 1.1 - Середня місячна температура повітря, в º С (березень, м.Одеса).

Рік	0	1	2	3	4	5
1890					3,0	2,3
1900	-	2,9	2,3	3,4	-	1,1
1910	2,5	-0,1	3,5	4,7	5,2	1,3
1920	3,9	4,0	4,9	3,0	-0,3	4,4
1930	4,1	0,9	-2,6	1,1	4,4	1,3
1940	-0,3	1,9	1,9	1,4	3,2	1,9
1950	2,8	2,9	-0,5	1,8	1,3	1,7
1960	0,2	5,5	2,2	-0,3	-0,5	2,3
1970	3,8	1,2	2,6	2,5	2,9	4,3
1980	0,1					

Продовження табл.1.1

6	7	8	9
0,9	4,0	-0,8	2,7
5,4	-0,3	1,8	1,5
3,9	0,5	2,7	2,7
1,0	4,1	-1,8	-2,5
4,4	4,6	4,7	1,5
2,9	3,7	0,6	1,6
-0,6	1,6	1,8	2,9
4,9	2,8	3,7	-0,6
0,9	3,7	4,3	3,0

У деяких випадках простий статистичний ряд ранжирується. *Ранжированим* називають ряд, у якому члени ряду розташовуються у порядку їх збільшення або зменшення.

Розташовувати вихідний матеріал спостережень у вигляді простого статистичного ряду необхідно в тих випадках, коли задача дослідження полягає у вивченні закономірностей змінення випадкової величини за часом. Якщо така задача не ставиться, та особливо, коли об'єм спостережень великий, доцільно вихідний матеріал розташовувати в інакшому, більш компактному, вигляді.

Тоді від простого статистичного ряду переходять до згрупованого.

Побудова згрупованого ряду проводиться таким чином :

- визначається область значень випадкової величини, тобто знаходяться найменше ($x_{\min}$) і найбільше ($x_{\max}$) значення цієї величини;
- розраховується *кількість* k *часткових інтервалів (груп)*, на які треба поділяти область значень випадкової величини. Для цього використовується формула:

$$k = 5 \lg n \,, \qquad (1.1)$$

де n - об'єм ряду (вибірки) .

Результат, отриманий по формулі (1.1), округляється до цілого числа.

Кількість часткових інтервалів, що знаходиться по формулі (1.1), розглядається як верхня границя цієї величини. У деяких випадках кількість груп розраховується по іншій формулі :

$$k = 1 + 3{,}222 \lg n \,; \qquad (1.2)$$

- знаходять *довжину часткових інтервалів градацій* c по очевидному співвідношенню

$$c = \frac{x_{\max} - x_{\min}}{k}\ ; \qquad (1.3)$$

– визначають значення випадкової величини на границях часткових інтервалів і на їх середині. Позначимо останні через $\widetilde{x}_i \quad (i = 1,2,...,k)$;

– підраховують кількість членів ряду m_i, що потрапляють до кожного часткового інтервалу. Величини m_i $(i = 1,2,...,k)$ називають *інтервальними емпіричними частотами*.

Згрупованим статистичним рядом називають сукупність значень випадкової величини на серединах часткових інтервалів (градацій) і відповідних інтервальних частот

$$\begin{array}{l} \widetilde{x}_1; \widetilde{x}_2; ...; \widetilde{x}_{k-1}; \widetilde{x}_k \\ m_1; m_2; ...; m_{k-1}; m_k \ . \end{array} \qquad (1.4)$$

Як правило, окрім інтервальних частот розраховують і інтервальні частості $\hat{p}_i$.

Інтервальні частості - це відносні частоти випадкової величини

$$\hat{p}_i = \frac{m_i}{n}\ . \qquad (1.5)$$

Очевидно , для інтервальних частот виконується рівність

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

а для частостей - рівність

$$\sum_{i=1}^{k} \hat{p}_i = 1 . \qquad (1.6)$$

Використовуючи частості, згрупований ряд можна сформувати таким чином :

$$\begin{array}{l} \widetilde{x}_1 ; \widetilde{x}_2 ; ... ; \widetilde{x}_{k-1} ; \widetilde{x}_k \\ \hat{p}_1 ; \hat{p}_2 ; ... ; \hat{p}_{k-1} ; \hat{p}_k \end{array} . \qquad (1.7)$$

Згруповані ряди часто зображаються за допомогою діаграм. Використовуються дві форми діаграм: гістограма і полігон.

В якості прикладу в таблиці 1.2 приводиться згрупований ряд середніх місячних температур повітря в Одесі у березні.

На рис.1.1 і 1.2 містяться відповідно цьому згрупованому ряду полігон і гістограма.

Гістограма - це система прямокутників, основою яких є довжина часткового інтервалу C, а висота - дорівнює відповідній інтервальній частоті (або частості).

Якщо всі k точок ($\widetilde{x}_i , \hat{p}_i$ або $\widetilde{x}_i , m_i$) нанести в системі координат та з'єднати їх відрізками прямої, то ламана, яка отримана при цьому, називається *полігоном* розподілу.

Таблиця 1.2 - Згрупований ряд середніх місячних температур повітря в Одесі, у березні, в C^0

$$c = 0.8 \quad x_{\min} = -2.6^0 C \quad x_{maz} = 5.5^0 C$$

№ п/п	Градації	Середина градації $\widetilde{x}_i$	Частота m_i	Частість $\hat{p}_i$
1	-2,6 ÷ -1,8	-2,2	3	0,04
2	-1,8 ÷ -1,0	-1,4	0	0,00
3	-1,0 ÷ -0,2	-0,6	9	0,11
4	-0,2 ÷ 0,6	0,2	5	0,06
5	0,6 ÷ 1,4	1,0	11	0,13
6	1,4 ÷ 2,2	1,8	12	0,14
7	2,2 ÷ 3,0	2,6	19	0,22
8	3,0 ÷ 3,8	3,4	7	0,08
9	3,8 ÷ 4,6	4,2	12	0,14
10	4,6 ÷ 5,4	5,0	6	0,07
11	5,4 ÷ 6,2	5,8	1	0,01
		Суми:	85	1,00

2 Статистичні оцінки параметрів

1.2.1 Статистичні оцінки моментів розподілу випадкових величин

З теорії ймовірностей відомо, що властивості випадкових величин можуть характеризуватися початковими (ν), центральними (μ) та основними (r) моментами різних порядків (l) .

В гідрометеорологічних дослідженнях, як правило, використовуються перелічені моменти перших чотирьох порядків, які, як буде показано пізніше, відбивають фізичні властивості процесів, що досліджуються.

Початковий момент l - того порядку для неперервної випадкової величини X визначається таким чином :

$$\nu_l = \int_{-\infty}^{\infty} x^l f(x)dx \ , \qquad (1.8)$$

де $f(x)$ - щільність ймовірності випадкової величини X.

На основі цього визначення отримаємо метод, за допомогою якого можна знайти статистичну оцінку l - того початкового моменту на основі вибірки випадкової величини, яка може бути задана згрупованим рядом виду (1.4) або (1.7).

Як випливає з формули (1.8), випадкова величина X визначена на інтервалі $(-\infty, \infty)$. Інтервал же значень випадкової величини, що визначається вибіркою $X : x_1, x_2, x_3, ..., x_n$, є обмеженим $[x_{\min}, x_{\max}]$.Тому будемо моделювати випадкову величину X таким чином :

$$X = \begin{cases} 0, & якщо \;\; X < x_{\min}, \\ X, & якщо \;\; x_{\min} \le X \le x_{\max}, \\ 0, & якщо \;\; X > x_{\max}. \end{cases} \qquad (1.9)$$

З урахуванням цього , запишемо інтеграл (1.8) у виді :

$$\nu_l = \int_{-\infty}^{x_{\min}} x^l f(x)dx + \int_{x_{\min}}^{x_{\max}} x^l f(x)dx + \int_{x_{\max}}^{\infty} x^l f(x)dx \qquad (1.10)$$

Очевидно, перший та третій інтеграли у формулі (1.10) дорівнюють нулю. Отже,

$$\nu_l = \int_{x_{\min}}^{x_{\max}} x^l f(x)dx. \qquad (1.11)$$

При побудові згрупованого ряду інтервал $[x_{\min}, x_{\max}]$ роздрібнювався на k часткових інтервалів довжиною С. Ураховуючи це , розіб'ємо інтеграл (1.11) на k інтегралів

$$\nu_l = \sum_{i=1}^{k} \int_{x_{\min}+(i-1)c}^{x_{\min}+ic} x^l f(x)dx \qquad (1.12)$$

Розглянемо проміжок

$$\left[x_{\min} + (i-1)c;\ \ x_{\min} + ic \right] \text{ (рис.1.3)}$$

Очевидно при апроксимації криволінійної трапеції $AB'C'Д$ прямокутником $ABCД$ середнє значення випадкової величини X в цьому інтервалі є її значення X_i на середині i - того часткового інтервалу, а середнє значення щільності імовірності є її значення в точці X_i. З урахуванням цього, застосуємо до інтегралу у формулі (1.12) теорему про середнє. Будемо мати

$$\nu_l \approx \sum_{i=1}^{k} \tilde{x}_i^{\,l} f(x_i) \int_{x_{\min}+(i-1)c}^{x_{\min}+ic} dx = \hat{\nu}_l \qquad (1.13)$$

або

$$\hat{\nu}_l \approx \sum_{i=1}^{k} \tilde{x}_i^{\,l} f(x_i)c \qquad (1.14)$$

Добуток

$$f(\tilde{x}_i)c = \hat{p}_i \qquad (1.15)$$

є площа прямокутника $ABCД$, яка приблизно дорівнює площі криволінійної трапеції $AB'C'Д$, що відображає інтервальну імовірність $\hat{p}_i$ випадкової величини X. Тому можна вважати, що величина $\hat{p}_i$, яка визначається формулою (1.15), є оцінкою цієї імовірності, тобто інтервальною частістю. Таким чином, статистична оцінка l - того початкового моменту дорівнює

$$\hat{\nu}_l = \sum_{i=1}^{k} \tilde{x}_i^{\,l} \hat{p}_i \qquad (1.16)$$

або, оскільки $\hat{p}_i = \frac{m_i}{n}$

$$\hat{\nu}_l = \frac{1}{n} \sum_{i=1}^{k} \tilde{x}_i^{\,l} m_i\,, \qquad (1.17)$$

де m_i - емпірична частота i-того інтервалу, n- об’єм вибірки.

Із теорії ймовірностей відомо, що

$$\hat{\nu}_l = \int_{-\infty}^{\infty} x f(x) dx = m_x \qquad (1.18)$$

є математичне сподівання випадкової величини X. Знайдемо оцінку першого початкового моменту.

$$\hat{\nu}_1 = \frac{1}{n} \sum_{i=1}^{k} \widetilde{x}_i m_i = \overline{x} \qquad (1.19)$$

Очевидно, вона є середнім значенням величини. Отже, *середнє значення є статистичною оцінкою математичного сподівання випадкової величини* X.

Оцінка другого початкового моменту дорівнює:

$$\hat{\nu}_2 = \frac{1}{n} \sum_{i=1}^{k} \widetilde{x}_i^{\,2} m_i = \overline{x^2}, \qquad (1.20)$$

а третього має вид :

$$\hat{\nu}_3 = \frac{1}{n} \sum_{i=1}^{k} \widetilde{x}_i^{\,3} m_i = \overline{x^3} \qquad (1.21)$$

і т. д .

За означенням *центральний момент* l - того порядку визначається рівнянням

$$\mu_l = \int_{-\infty}^{\infty} (x - m_x)^l f(x) dx. \qquad (1.22)$$

Аналогічним чином можна прийти до формули, яка дає змогу отримати на основі вибірки випадкової величини X статистичні оцінки центрального моменту l - того порядку

$$\hat{\mu}_l = \sum_{i=1}^{k} (\tilde{x}_i - \bar{x})^l \hat{p}_i \qquad (1.23)$$

або

$$\hat{\mu}_l = \frac{1}{n} \sum_{i=1}^{k} (\tilde{x}_i - \bar{x})^l m_i \ . \qquad (1.24)$$

Очевидно, центральний момент першого порядку дорівнює нулю. Таке ж значення має і його оцінка $\hat{\mu}_1 = 0$. Як відомо ,

$$\mu_2 = \int_{-\infty}^{\infty} (x - m_x)^2 f(x) dx = \sigma_x{}^2 \qquad (1.25)$$

є дисперсією випадкової величини X. Отже оцінка його

$$\hat{\mu}_2 = \sum_{i=1}^{k} (\tilde{x}_i - \bar{x})^2 \hat{p}_i \qquad (1.26)$$

або

$$\hat{\mu}_2 = \frac{1}{n} \sum_{i=1}^{k} (\tilde{x}_i - \bar{x})^2 m_i \qquad (1.27)$$

є *оцінкою дисперсії:* $\hat{\mu}_2 = \hat{\sigma}_x^{\,2}$; а $\hat{\sigma}_x = \sqrt{\hat{\sigma}_x^{\,2}}$ називається *оцінкою середнього квадратичного відхилу*.

Центральні моменти можна розрахувати по формулам зв'язку їх з початковими моментами. А саме:

$$\hat{\mu}_2 = \hat{\nu}_2 - \hat{\nu}_1^{\,2}, \qquad (1.28)$$

$$\hat{\mu}_3 = \hat{\nu}_3 - 3\hat{\nu}_2\hat{\nu}_1 + 2\hat{\nu}_1^{\,3}, \qquad (1.29)$$

$$\hat{\mu}_4 = \hat{\nu}_4 - 4\hat{\nu}_1\hat{\nu}_3 + 6\hat{\nu}_1^{\,2}\hat{\nu}_2 - 3\hat{\nu}_1^{\,4}. \qquad (1.30)$$

Крім початкових і центральних моментів при статистичному аналізі гідрометеорологічних процесів знаходять широке використання основні (нормовані) моменти випадкових величин. За означенням *основним моментом* l -

того порядку r_l називається відношення l-того центрального моменту до l - того степеня середнього квадратичного відхилу.

$$r_l = \frac{\mu_l}{\sigma_x^{\,l}}. \qquad (1.31)$$

Як правило, оскільки $r_1 = 0$, а $r_2 = 1$, використання основних моментів обмежується лише третім та четвертим r_3 і r_4. Ці моменти дають важливу інформацію про характер розподілу випадкових величин. Третій основний момент відбиває характер асиметрії кривої розподілу. Тому його називають *коефіцієнтом асиметрії* $r_3 = A_S$. При $r_3 = 0$, крива розподілу є симетричною відносно центру розподілу. Як відомо, гаусовий (нормальний) розподіл є симетричним відносно математичного сподівання і для нього $r = 0$. Крім асиметрії крива розподілу характеризується сплюснутостю або витягнутостю відносно кривої нормального розподілу. Цю міру називають *коефіцієнтом ексцесу* E. Коефіцієнт ексцесу має такий зв'язок з четвертим основним моментом :

$$E = r_4 - 3. \qquad (1.32)$$

Як буде показано у розділі 2.3.2, для нормального розподілу $r_4 = 3$ і $E = 0$. При $E > 0$ крива розподілу є витянутою, при $E < 0$ - сплюснутою. Приклади форм кривих розподілу при різних значеннях r_3 і E приводяться на рис.1.4 і 1.5.

Для розрахування статистичних оцінок третього та четвертого основних моментів використовуються формули:

$$\hat{r}_3 = \frac{\hat{\mu}_3}{\hat{\sigma}_x{}^3}, \tag{1.33}$$

$$\hat{r}_4 = \frac{\hat{\mu}_4}{\hat{\sigma}_x{}^4}. \tag{1.34}$$

Початкові та центральні моменти можна розраховувати без попереднього групування вихідної інформації . У такому разі використовуються виборочні формули :

$$\hat{\nu}_l = \frac{1}{n}\sum_{i=1}^{n} x_i{}^l, \tag{1.35}$$

$$\hat{\mu}_l = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^l. \tag{1.36}$$

Ці формули випливають із рівностей (1.16) і (1.24), якщо уявити собі, що кожний член вибірки x_i являє собою групу, яка складається з одного значення випадкової величини з частотою $m_i = 1$.

В якості прикладу в табл.1.3 приводяться результати статистичного оцінювання першого початкового та другого, третього, четвертого центральних моментів середньої місячної температури повітря в Одесі у березні. Всі розрахунки зроблені на основі згрупованого ряду, що міститься в табл.1.2.

Нагадаємо, що оцінювання параметрів статистичних сукупностей можна провести як по згрупованому, так і по простому статистичному ряду.

Результат оцінювання ж основних моментів буде приведено в підрозділі 1.2.2 тому, що потребує додаткових пояснень, пов'язаних з вимогами, щодо оцінок моментів.

Таблиця 1.3 - Приклад розрахунків оцінок статистичних моментів ряду середньої місячної температури (березень, м.Одеса)

№ п/п	Температура (градації)		$\tilde{x}_i$	m_i	$\hat{p}_i$	$\tilde{x}_i m_i$
1	-2.6	-1.8	-2.2	3	0.04	-6.6
2	-1.8	-1.0	-1.4	0	-	-
3	-1.0	-0.2	-0.6	9	0.11	-5.4
4	-0.2	0.6	0.2	5	0.06	1.0
5	0.6	1.4	1.0	11	0.13	11.0
6	1.4	2.2	1.8	12	0.14	21.6
7	2.2	3.0	2.6	19	0.22	49.4
8	3.0	3.8	3.4	7	0.08	23.8
9	3.8	4.6	4.2	12	0.14	50.4
10	4.6	5.4	5.0	6	0.07	30.0
11	5.4	6.2	5.8	1	0.01	5.8
СУМИ :				85	1.00	181.00

Продовження табл.1.3

$\tilde{x}_i - \bar{x}$	$(\tilde{x}_i - \bar{x})^2$	$(\tilde{x}_i - \bar{x})^2 m_i$	$(\tilde{x}_i - \bar{x})^3$
-4.3	18.49	55.47	-79.51
-	-	-	-
-2.7	7.29	65.61	-19.68
-1.9	3.61	18.05	-6.86
-1.1	1.21	13.31	-1.33
-0.3	0.09	1.08	-0.03
0.5	0.25	4.75	0.13

1.3	1.69	11.83	2.2
2.1	4.41	52.92	9.26
2.9	8.41	50.46	24.39
3.7	13.69	13.69	50.65
		287.17	

Продовження табл.1.3

$(\widetilde{x}_i - \bar{x})^3 m_i$	$(\widetilde{x}_i - \bar{x})^4$	$(\widetilde{x}_i - \bar{x})^4 m_i$
-238.53	341.88	1025.64
-	-	-
-177.12	53.14	478.26
-34.3	13.03	65.15
-14.63	1.46	16.06
-0.36	0.01	0.12
2.47	0.06	1.14
15.40	2.86	20.02
111.12	19.45	233.4
146.34	70.73	424.38
50.65	187.42	187.42
-138.96		2451.59

Тоді маємо: $\bar{x} = 2.1 \;\; \hat{\mu}_2 = 3.38 \;\; \hat{\mu}_3 = -1.64$

$$\hat{\mu}_4 = 28.84$$

1.2.2 Властивості статистичних оцінок параметрів

Нехай з генеральної сукупності здобуто випадковим чином N вибірок. Тоді будемо мати N оцінок параметру. Якими повинні бути оцінки $\hat{\theta}_N$, щоб достатньо вірогідно

характеризувати параметр θ? Вони повинні задовольняти *трьом умовам: незсуненості, ефективності* та *умотивованості.*

Оцінка параметра називається *незсуненою*, якщо її математичне сподівання дорівнює параметру, який оцінюється, тобто

$$M\left[\hat{\theta}_N\right] = \theta. \qquad (1.37)$$

Якщо ця рівність не виконується, то оцінка $\hat{\theta}_N$ може або завищувати значення θ (тобто $M[\hat{\theta}_N] > \theta$), або занижувати його ($M[\hat{\theta}_N] < \theta$). В обох випадках це приводить до систематичних (одного знаку) похибок в оцінці параметра θ. Отже, вимога незсуненості гарантує відсутність систематичних похибок при оцінках параметрів.

Оскільки $\hat{\theta}_N$ - випадкова величина, значення якої змінюється від вибірки до вибірки (нагадаємо, що вибірки добуваються із генеральної сукупності випадковим чином), то міра її розсіювання відносно математичного сподівання параметра θ характеризується дисперсією $Д[\hat{\theta}_N]$. Нехай $\hat{\theta}_N$ і $\hat{T}_N$ - дві незсунені оцінки параметра θ, тобто $M[\hat{\theta}_N] = \theta$; $M[\hat{T}_N] = \theta$. Із двох оцінок $\hat{\theta}_N$ і $\hat{T}_N$ слід віддати перевагу тій, котра має менше розсіювання навколо параметра, що оцінюється, тому що у цьому випадку значення оцінки, знайдене по конкретній вибірці, менше всього відхиляється від істинного значення параметра θ. Якщо $Д[\hat{\theta}_N] < Д[\hat{T}_N]$, то за оцінку треба прийняти $\hat{\theta}_N$. Незсунена оцінка, котра має найменшу дисперсію серед усіх

можливих незсунених оцінок параметра, розрахованих по вибірках одного й того ж об'єму, називається *ефективною оцінкою*. Іншими словами, оцінка $\hat{\theta}_N$ більш ефективна ніж $\hat{T}_N$, якщо $M[(\hat{\theta}_N - \theta)^2] < M[(\hat{T}_N - \theta)^2]$.

Розглянемо нижню межу $\inf M[(\hat{\theta}_N - \theta)^2]$ множини величин $M[(\hat{\theta} - \theta)^2]$ по всіх можливих $\hat{\theta}_N$. Оцінка $\hat{\theta}_N$, для якої досягається нижня межа, називається ефективною. Оцінка $\hat{\theta}$, для якої міра розкиду $M[(\hat{\theta} - \theta)^2]$ при $N \to \infty$ всюди поводиться так, як і для ефективної оцінки, тобто

$$\lim_{N\to\infty} \frac{M[(\hat{\theta}_N - \theta)^2]}{M[(\hat{\theta} - \theta)^2]} = 1$$

називається *асимптотичною ефективною оцінкою*.

Оцінка $\hat{\theta}_N$ параметра θ називається *умотивованою*, якщо вона по ймовірності збігається до параметра θ:

$$\lim_{n\to\infty}(P\left|\hat{\theta} - \theta\right| < \varepsilon) = 1. \qquad (1.38)$$

Рівність (1.38) позначає, що чим більший об'єм n вибірки, тим більша ймовірність того, що похибка оцінки не перевищує скільки завгодно малого додатного числа. Отже, у випадку використання умотивованих оцінок виправдовується збільшення числа одиниць вибірки, оскільки при цьому все менше імовірною стає можливість значної помилки в оцінці невідомого параметра .

Ясно, що коли $Д[\hat{\theta}_N] \to 0$ при $n \to \infty$, то незсунена оцінка є й ефективною, й умотивованою.

Покажемо, що *середнє значення випадкової величини є незсунена, ефективна та умотивована оцінка математичного сподівання*. Для цього знайдемо математичне сподівання середнього значення випадкової величини X, використовуючи формулу (1.35) при $l = 1$:

$$M[\bar{x}] = M\{\frac{1}{n}\sum_{i=1}^{n} x_i\}. \qquad (1.39)$$

За відомими властивостями математичного сподівання маємо

$$M[\bar{x}] = \frac{1}{n}\sum_{i=1}^{n} M(x_i) = \frac{1}{n}\sum_{i=1}^{n} m_x = m_x. \qquad (1.40)$$

Рівність (1.40) і є визначення незсуненості оцінки $\bar{x}$ математичного сподівання m_x.

Знайдемо тепер дисперсію середнього значення

$$Д[\bar{x}] = Д\{\frac{1}{n}\sum_{i=1}^{n} x_i\}. \qquad (1.41)$$

У відповідності до властивостей дисперсії можна записати

$$Д[\bar{x}] = \frac{1}{n^2}\sum_{i=1}^{n} Д[x_i] = \frac{1}{n^2}\sum_{i=1}^{n} \sigma_x{}^2 = \frac{\sigma_x{}^2}{n}.$$

(1.42)

Оскільки при $n \to \infty \quad Д[\bar{x}] \to 0,$ то середнє значення є ефективною та умотивованою оцінкою математичного сподівання.

Очевидно

$$\sqrt{Д[\bar{x}]} = \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} \,. \tag{1.43}$$

Формула (1.43) характеризує середній квадратичний відхил середнього значення .

Оцінка дисперсії випадкової величини, котра отримується за допомогою формул (1.26), (1.27) і (1.36) при $l = 2$ не є незсуненою. Для того, щоб отримати незсунену оцінку дисперсії треба помножити оцінку другого центрального моменту $\hat{\mu}_2$ на множник Бесселя $\frac{n}{n-1}$. Тобто *незсунена оцінка дисперсії*, позначимо її $S_x{}^2$, дорівнює :

$$S_x{}^2 = \frac{n}{n-1}\hat{\mu}_2 = \frac{1}{n-1}\sum_{i=1}^{k}(\tilde{x}_i - \bar{x})^2 m_i \tag{1.44}$$

або

$$S_x{}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{1.45}$$

якщо оцінка дисперсії знаходиться без попереднього групування інформації.

Тепер, ураховуючи позначені вище властивості щодо статистичних оцінок моментів на прикладі їх розрахунків (табл.1.3) згрупованого ряду середньої місячної температури березня в м. Одеса (табл.1.2), приведемо результати розрахунків за формулою (1.44) незсуненої оцінки дисперсії $S_x^{\,2}$ та середнього квадратичного відхилу $S_x = \sqrt{S_x^{\,2}}$. Тоді маємо:

$$S_x^{\,2} = 3.42; S_x = 1.85.$$

Відповідні оцінки третього та четвертого основних моментів за формулами (1.33) і (1.34) складають:

$$\hat{r}_3 = -0.26; \hat{r}_4 = 2.46; E = -0.54 .$$

Покажемо далі, що *частість є незсунена, ефективна та умотивована оцінка ймовірності.*

Нехай проводяться експерименти з деяким явищем А, яке може з'явитися, а може не з'явитися. Знайдемо z_ν - число появлень події A в одній ν-тій спробі. Очевидно, $z_\nu = 1$, якщо подія А відбулася з імовірністю p і $z_\nu = 0$, якщо подія А не відбулася з імовірністю $q = 1 - p$.

Найдемо математичне сподівання числа появлень події А

$$M[z_\nu] = 1 \cdot p + 0 \cdot q = p, \qquad (1.46)$$

а також дисперсію

$$Д[z_\nu] = (1-p)^2 p + (0-p)^2 \cdot q = \\ + q^2 p + p^2 q = pq(p+q) = pq \quad . \qquad (1.47)$$

Легко зрозуміти, що частоту m можна виразити через число появлень події А таким чином :

$$m = \sum_{\nu=1}^{n} z_\nu \qquad (1.48)$$

де n - загальне число незалежних експериментів .
Тоді частість $\hat{p}$ дорівнює

$$\hat{p} = \frac{\sum_{\nu=1}^{n} z_\nu}{n} . \qquad (1.49)$$

Знайдемо математичне сподівання частості

$$M[\hat{p}] = M\left[\frac{\sum_{\nu=1}^{n} z_\nu}{n}\right] = \frac{1}{n}\sum_{\nu=1}^{n} M[z_\nu] . \qquad (1.50)$$

Урахування співвідношення (1.46) приводить до результату

$$M[\hat{p}] = \frac{1}{n}\sum_{\nu=1}^{n} p = p \quad . \tag{1.51}$$

Рівність (1.51) свідчить про те , що частість є незсуненою оцінкою імовірності.

Розрахуємо тепер дисперсію частості. Будемо мати

$$Д[\hat{p}] = Д\left[\frac{1}{n}\sum_{\nu=1}^{n} z_\nu\right] = \frac{1}{n^2}\sum_{\nu=1}^{n} Д\left[z_\nu\right]. \tag{1.52}$$

Підставимо (1.52) в результат (1.47) .Маємо:

$$Д[\hat{p}] == \frac{1}{n^2}\sum_{i=1}^{n} pq = \frac{pq}{n} \tag{1.53}$$

Очевидно при $n \to \infty \quad Д\left[\hat{p}\right] \to 0$. Отже частість є незсуненою, ефективною та умотивованою оцінкою ймовірності .

2 ЗАКОНИ РОЗПОДІЛУ ГІДРОМЕТЕОРОЛОГІЧНИХ ВЕЛИЧИН

2.1 Поняття про закон розподілу

Дослідження законів розподілу гідрометеорологічних величин має велике практичне значення. Як уже зазначалося неодноразово, в основних ланках кліматичної системи атмосфері та гідросфері постійно відбуваються зміни їх фізичного стану, а кількісні характеристики цього стану такі, наприклад, як температура повітря чи води, атмосферний тиск, хмарність, вологість повітря, кількість опадів, річний стік та інші розглядаються як випадкові величини.

Задача дослідника полягає у тому, щоб серед множини випадкових подій чи явищ виявити закономірності, відкинувши несуттєві події. А це можна зробити шляхом побудови моделей фізичних параметрів, які висвітлюють властивості цих випадкових величин, які, як відомо, містяться у законі розподілу.

Законом розподілу випадкової величини називають всіляку відповідність між можливими значеннями випадкової величини та її ймовірностями.

Таким чином, вичерпною характеристикою будь-якого випадкового процесу є закон розподілу, знання якого дає можливість правильно протулмачити смисл того чи іншого статистичного моменту та на основі цього методично правильно організувати вивчення гідрометеорологічних особливостей регіону, що досліджується. Підібравши закон розподілу до статистичного ряду (вибірки), можна розрахувати імовірність того, що випадкова величина із генеральної сукупності, яка вивчається, знаходиться у заданому інтервалі або ймовірність того, що випадкова величина прийме значення менше (більше) деякого конкретного числа.

У більшості випадків закони розподілу гідрометеорологічних величин неможливо визначити апріорно тільки шляхом аналізу відомих фізичних властивостей. Тип розподілу та його параметри визначаються шляхом

статистичної обробки експериментальних даних. Найбільш поширеним є метод групування даних, при якому вся множина значень даної випадкової величини розділяється на ряд неперетинних часткових інтервалів, а потім підраховується число даних, що потрапили до кожного часткового інтервалу. Але таким шляхом, як було показано в параграфі 1.1, будується і згрупований ряд. Отже *згрупований ряд має сенс емпіричного розподілу випадкової величини,* графічним зображенням якого є гістограма або полігон.

Отже вивчення особливостей статистичної структури гідрометеорологічних величин базується на інформації, у якості якої виступають статистичні ряди (вибірки), що сформовані по результатах вимірювань та спостережень.

Згрупований ряд, як емпіричний розподіл, апроксимують аналітичним виразом, який відбиває властивості генеральної сукупності. У зв'язку з цим, основним етапом статистичного аналізу гідрометеорологічної інформації є підбір закону розподілу по даних статистичної сукупності. Ця задача розв'язується шляхом апроксимації емпіричного розподілу таким теоретичним законом, який би у визначеному смислі найкращим чином відповідав би емпіричному розподілу. Але, як би добре, на підставі відомих властивостей закону розподілу, не була підібрана теоретична крива чи то нормального розподілу, чи то розподілів Пірсона, Пуассона або інших відомих, про властивості яких піде мова у цьому розділі пізніше, між нею і емпіричним (статистичним) розподілом неминучі деякі розбіжності. Тому обов'язково після розрахунків теоретичних частот (по відповідних формулах, в залежності від закону розподілу) проводять перевірку гіпотези про міру розбіжності між емпіричними та теоретичними частотами. Розбіжності між цими частотами можуть носити як випадковий характер, так і бути статистично значущими. Останнє вказує на те, що підібрана теоретична крива не відповідає даному емпіричному розподілу. Щоб з'ясувати ці питання використовують так звані "*критерії згоди"*. І тільки після використання таких критеріїв можна зробити висновок про успішність апроксимації статистичного розподілу теоретичним законом.

Тому *процес дослідження закону розподілу складається з таких етапів:*

- на основі зовнішнього вигляду емпіричного розподілу, який зображається гістограмою чи полігоном, та з урахуванням статистичних оцінок моментів та допоміжних статистик, формулюють гіпотезу про закон розподілу;
- на основі статистичної сукупності знаходять оцінки параметрів вибраного теоретичного розподілу та відповідні для них статистики (*етап поновлювання закону розподілу*);
- розраховують теоретичні інтервальні частоти для випадкової величини яка досліджується;
- роблять оцінку розбіжності між емпіричними та теоретичними частотами.

Теоретичні основи та впровадження на практиці перших трьох етапів дослідження закону розподілу буде послідовно розглянемо у цьому (другому) розділі, а з останнім – ви познайомитесь у розділі 4 : “Перевірка статистичних гіпотез”.

А, щоб виконати перелічені етапи дослідження, треба, перш за все, успішно підібрати теоретичний закон, аналітичний вираз якого може бути у вигляді функції розподілу чи щільності ймовірності. Необхідно також знати властивості законів розподілу, які найбільш часто використовуються при дослідженнях гідрометеорологічних величин.

2.2 Функція розподілу і щільність імовірності

Функція розподілу є найбільш загальною формою визначення закону розподілу. Вона зазначає ймовірність того, що випадкова величина X приймає значення, менше фіксованого дійсного числа x. Для неперервної випадкової величини це визначення має вигляд:

$$F(x) = P(X < x). \qquad (2.1)$$

Імовірність того, що $X < x$ залежить від x, отже $F(x)$ є функцією від x. Тому $F(x)$ і називається *функцією розподілу.*

Для дискретної випадкової величини X, яка може приймати значення $x_1, x_2, \ldots, x_i, \ldots x_n$, функція розподілу має вигляд :

$$F(x) = \sum_{x_i < X} P(X - x_i), \tag{2.2}$$

де нерівність $x_i < x$ під знаком суми означає, що підсумування поширюється на всі ті значення x_i, котрі менші x.

Побудуємо функцію розподілу для дискретної випадкової величини, розподіл імовірностей для якої визначається такою таблицею :

$$X : x_1 x_2 x_3 \ldots x_i \ldots x_{n-1} x_n,$$

$$P : p_1 p_2 p_3 \ldots p_i \ldots p_{n-1} p_n.$$

При $x \le x_1$:

$$F(x) = P(X < x_1) = 0;$$

при $x_1 < x \le x_2$:

$$F(x) = P(X < x_2) = P(X = x_1) = p_1;$$

при $x_1 < x \le x_2$:

$$F(x) = P(X < x_2) = P(X = x_1) + \\ + P(X = x_2) = p_1 + p_2;$$

при $x_3 < x \le x_4$:

$$F(x) = P(X < x_4) = P(X = x_1) + \\ + P(X = x_2) + P(X = x_3) = p_1 + p_2 + p_3;$$

при $x_{n-1} < x \le x_n$:

$$F(x) = P(X < x_n) = P(X = x_1) + \\ + P(X = x_2) + P(X = x_3) + ... + \\ + P(X = x_{n-1}) = p_1 + p_2 + p_3 + ... + p_{n-1};$$

при $x > x_n$:

$$F(x) = P(X = x_1) + P(X = x_2) +$$
$$+ P(X = x_3) + ... + P(X = x_{n-1}) +$$
$$+ P(X = x_n) = p_1 + p_2 +$$
$$+ p_3 + ... + p_{n-1} + p_n = 1$$

тобто $\sum_i p_i = 1.$

Функція розподілу має стрибок у тих точках, де випадкова величина приймає конкретні значення, що утримуються в таблиці. В інтервалах між значеннями випадкової величини функція $F(x)$ постійна. Сума всіх стрибків функції розподілу дорівнює одиниці. Графік функції розподілу дискретної випадкової величини є розривна ступінчата ламана лінія (рис.2.1).

Неперервна випадкова величина має неперервну чи кусочно-неперервну функцію розподілу; графік цієї функції має форму плавної кривої (рис.2.2).

Розглянемо *загальні властивості функції розподілу*:

1. Функція розподілу $F(x)$ є невід'ємною функцією з областю значень

$$0 \le F(x) \le 1.$$

Ця властивість випливає з визначення функції розподілу як імовірності здійснювання події $X < x$.

Ясно, що на мінус нескінченності функція розподілу $F(x)$ дорівнює нулю, а на плюс нескінченності функція розподілу дорівнює одиниці, тобто

$$F(-\infty)=0;\quad F(+\infty)=1.$$

Ця властивість стає очевидною при геометричній інтерпретації функції розподілу (див. рис. 2.2). Якщо точка x необмежено пересувається ліворуч, то попадання X ліворуч від x y границі стає неможливою подією. Тому можна вважати, що ймовірність цієї події прагне до нуля, тобто

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

При необмеженому пересуванні точки x праворуч, попадання випадкової точки X праворуч від x у границі стає вірогідною подією. Тому можна вважати, що ймовірність цієї події прагне до одиниці, тобто

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

Очевидно,

$$P(X>x)=1-F(x).$$

Функція розподілу, як і всіляка імовірність, є безрозмірною.

2. Імовірність попадання випадкової величини в інтервал $[\alpha,\beta]$ дорівнює різниці значень функції розподілу на кінцях цього інтервалу

$$P(\alpha < X < \beta) = F(\beta) - F(\alpha).$$

3. Функція розподілу випадкової величини є неспадною функцією

$$F(\beta) \geq F(a), \text{ якщо } \beta > \alpha.$$

Неперервну випадкову величину можна задати не тільки функцією розподілу, але й щільністю ймовірностей. Розглянемо цю форму задавання випадкової величини.

Нехай випадкова величина X визначається функцією розподілу $F(x)$. Згідно з другою властивістю функцією розподілу імовірність попадання цієї величини в елементарний інтервал $[x; x + \Delta x]$ дорівнює

$$P(x < X < x + \Delta x) =$$

$$= F(x + \Delta x) - F(x) \quad .$$

Розділимо обидві частини на Δx:

$$\frac{P(x < X < x + \Delta x)}{\Delta x} = \frac{F(x + \Delta x) - F(x)}{\Delta x} .$$

Вважаючи, що функція розподілу неперервна, перейдемо до границі при $\Delta x \to 0$. Тоді отримаємо похідну від функції розподілу, яка й називається *щільністю імовірності*

$$f(x) = \frac{dF(x)}{dx} = \lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x}.$$

(2.3)

Крива, що зображує щільність імовірності $f(x)$ випадкової величини, називається *кривою розподілу*. Чисельні приклади кривих розподілу будуть розглядатися пізніше.

Тепер можна привести визначення *неперервної випадкової величини*: випадкова величина X називається неперервною, якщо її функція розподілу $F(x)$ неперервна на всій осі OX, а щільність розподілу $f(x)$ існує всюди, за винятком, може бути, скінченого числа точок.

Розглянемо *властивості щільності імовірності.*

1. Щільність імовірності є функцією невід'ємною :

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

Ця властивість випливає безпосередньо із визначення цієї функції як похідної від неспадної функції $F(x)$.

2. Імовірність попадання неперервної випадкової величини X до інтервалу $[\alpha; \beta]$ дорівнює :

$$P(\alpha < X < \beta) = \int_{\alpha}^{\beta} f(x)dx.$$

3. Функцію розподілу можна виразити через щільність імовірності за формулою:

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

4. Інтеграл у нескінченних межах від щільності імовірності дорівнює одиниці:

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

Геометрично це означає, що площа між кривою розподілу та осью абцис дорівнює одиниці.

На основі властивості 2 щільності ймовірностей при апроксимації емпіричного розподілу проводяться розрахунки інтервальних теоретичних імовірностей. Це означає, що для кожного інтервалу змінної x

$$x_1 + (i-1)c < x < x_1 + ic,$$

де x_1 – початкове значення змінної x,

$i = 1,2,...,k$ – номер часткового інтервалу, а

c – його довжина, розраховують імовірність p_i того, що випадкова величина X належить до цього інтервалу:

$$p_i = P[x_1+(i-1)c < x < x_1 + ic] =$$

$$= \int_{x_1+(i-1)c}^{x_1+ic} f(x)dx \quad .$$

В цьому рівнянні $f(x)$ щільність імовірності випадкової величини X.

2.3 Розподіл Пірсона

2.3.1 Елементи загальної теорії

Нехай ми маємо статистичний ряд деякої випадкової величини X об'ємом n. Побудуємо згрупований ряд $\widetilde{x}_i, m_i$ $(i = \overline{1,k})$, де

$\widetilde{x}_i$ - середина часткового інтервалу, а

m_i- емпірична інтервальна частота, і перейдемо до безрозмірної випадкової величини

$$z = \frac{x - \hat{x}}{c}, \qquad (2.4)$$

де

$\hat{x}$ - деяка статистика, що буде визначена пізніше;

c - величина часткового інтервалу.

Розглянемо диференціальне рівняння

$$\frac{dy}{dz} = y\frac{z+b}{c_2 z^2 + c_1 z + c_0}, \tag{2.5}$$

де

b, c_0, c_1, c_2 - дійсні числа.

Якщо змінна z у ньому є випадковою величиною, пов'язаною з вихідною випадковою величиною X рівністю (2.4), і якщо коефіцієнти c_0, c_1, c_2 є статистиками випадкової величини X такими, що

$$c_0 = -\sigma^2 \frac{S+1}{S-2}; \tag{2.6}$$

$$c_1 = -b = -\frac{r_3\sigma}{2}\frac{(S+2)}{(S-2)}; \tag{2.7}$$

$$c_2 = \frac{1}{S-2}; \tag{2.8}$$

$$\sigma = \frac{\sigma_x}{c}; \tag{2.9}$$

$$S = \frac{6(r_4 - r_3^{\,2} - 1)}{3r_3^{\,2} - 2r_4 + 6}, \qquad (2.10)$$

σ_x - середній квадратичний відхил величини X,

r_3 та r_4 - її третій та четвертий основні моменти, то розв'язок диференціального рівняння (2.5) являє собою сукупність законів розподілу Пірсона. Тобто, *розподіл Пірсона* – це сім'я розподілів імовірностей, щільність яких $y = P(z)$ задовольняє диференціальному рівнянню (2.5). Відповідні графіки $y = P(z)$, що відображають залежність щільності ймовірності від z, називають кривими Пірсона взагалі.

Розподіли Пірсона класифікують в залежності від значення коефіцієнтів b, c_0, c_1, c_2 та області змінення z. Сім'я розподілів Пірсона складається із 12 типів. Будь-який розподіл Пірсона однозначно визначається своїми першими чотирма початковими моментами

$$\nu_l = \int_{-\infty}^{\infty} z f(z) dz, \quad l = 1 \div 4.$$

Розглянемо теорію розподілів Пірсона більш детально.
Розв'язок рівняння (2.5), очевидно, має такий вигляд :

$$y = y_0 e^{\int \frac{z+b}{c_2 z^2 + c_1 z + c_0} dz}, \qquad (2.11)$$

де y_0 - довільна стала.

Як відомо, інтеграл у правій частині рівності (2.11) залежить від характеру коефіцієнтів квадратного трьохчлена

$$Q(z) = c_2 z^2 + c_1 z + c_0 \,. \tag{2.12}$$

Підставимо до нього значення коефіцієнтів (2.6)-(2.8) і перетворимо його у рівняння. Розв'язок квадратного рівняння дає такі значення коренів трьохчлена (2.12)

$$z_{1,2} = \sigma \frac{r_3 (S+2) \pm t}{4}, \tag{2.13}$$

де

$$t = 4\sqrt{(S+1)(1-\chi)}, \tag{2.14}$$

а

$$\chi = -\frac{{r_3}^2 (S+2)^2}{16(S+1)} \,. \tag{2.15}$$

Після елементарних перетворень рівність (2.14) приймає вигляд:

$$t = \frac{16|S+1|}{|r_3||S+2|}\sqrt{\chi(\chi-1)}\ . \qquad (2.16)$$

Із рівності (2.16) випливає, що значення коренів квадратного трьохчлена (2.12) залежать від величини

$$u = \chi(\chi-1). \qquad (2.17)$$

Рівняння (2.17), якщо його доповнити до повного квадрата

$$u + \frac{1}{4} = (\chi - \frac{1}{2})^2, \qquad (2.18)$$

є рівнянням параболи з вершиною в точці з координатами $(\frac{1}{2};-\frac{1}{4})$. Графік функції (2.17) приводиться на рис.2.3.

Як виходить з (2.17) і графіка функції $u = f(\chi)$, область значень χ може бути розподіленою на три підобласті, в кожній з котрих функція u має один і той же знак. У відповідності до цього, статистика (2.16) відноситься до множини дійсних або уявних чисел. У залежності від цього, приймає визначний вид інтеграл в експоненті розв'язку (2.11) і, таким чином, й сам розв'язок. Оскільки кожний із розв'язків диференціального рівняння (2.5) співвідноситься з визначеним типом розподілу Пірсона, кожному із підобластей значень χ теж відповідає той чи інший *тип розподілів Пірсона* , а саме:

– області $\chi < 0$ відповідає I тип;

– області $0 < \chi < 1$ відповідає IV тип;

– області $\chi > 1$ відповідає VI тип;

– значенню $\chi = 0$ відповідають:

$$\begin{cases} \text{II тип при } r_4 < 3\ ; \\ \text{нормальний розподіл при } r_4 = 3; \\ \text{VII тип при } r_4 > 3\ ; \end{cases}$$

– значенню $\chi = 1$ відповідає V тип;

– області $-\infty < \chi < \infty$ відповідає III тип.

Отже, III тип розподілів Пірсона може спостерігатися при будь-яких значеннях статистики χ.

Якщо для кожної з областей значень χ отримати щільність імовірності $y = f(z)$ для того чи іншого типу розподілів Пірсона і проінтегрувати її в межах i-го часткового інтервалу, то, як було показано вище, будемо мати *інтервальну імовірність*

$$p_i = P[z_1 + (i-1)c < z < z_1 + ic] = \\ = \int_{z_1+(i-1)c}^{z_1+ic} y(z)dz \quad , \qquad (2.19)$$

де $z_1 = \dfrac{x_1 - \hat{x}}{c}$ (2.20)

значення лівої межі області значень змінної x у новій системі координат. З другого боку

$$p_i = \frac{\widetilde{m}_i}{n}, \qquad (2.21)$$

де $\widetilde{m}_i$ - *теоретична частота* для і-того часткового інтервалу; n - загальний об'єм сукупності випадкової величини.

Таким чином, розподіли Пірсона можуть розглядатися і в термінах теоретичних частот $\widetilde{m}_i$. Статистика $\hat{x}$ визначається формулою:

$$\hat{x} = \overline{x} - \frac{r_3 \sigma_x}{2} \frac{(S+2)}{(S-2)}, \qquad (2.22)$$

2.3.2 Нормальний розподіл як частинний випадок розподілів Пірсона. Властивості нормального розподілу.

Як було зазначено , при $\chi = 0$, коли $r_4 = 3$ (в точці $\chi = 0$, очевидно, $r_3 = 0$), розв'язок диференціального рівняння (2.5) відповідає нормальному розподілу. Обгрунтуємо цей факт. Для цього знайдемо значення статистик S, c_2, c_1, c_0 при зазначених вище основних моментах r_3 і r_4

$$S = \frac{6(3-1)}{-2 \cdot 3 + 6} = \infty;$$

$$c_2 = \lim_{S\to\infty} \frac{1}{S-2} = 0;$$

$$c_1 = -b = \lim_{S\to\infty}\left[-\frac{r_3\sigma}{2}\frac{(S+2)}{(S-2)}\right] = -\frac{r_3\sigma}{2} = 0;$$

$$c_0 = -\lim_{S\to\infty}\left[\sigma^2 \frac{S+1}{S-2}\right] = -\sigma^2.$$

Отже, інтеграл у показнику степені розв'язку (2.11) приймає значення:

$$\int \frac{z+b}{c_2 z^2 + c_1 z + c_0}\, dz = -\int \frac{z}{\sigma^2}\, dz = -\frac{z^2}{2\sigma^2}.$$

(2.23)

Оскільки, як видно з рівності (2.22), у нашому випадку $\hat{x} = \overline{x}$, то, враховуючи формули (2.9) і (2.4) і те, що $\overline{x}$ є оцінкою m_x, маємо:

$$\int \frac{z+b}{c_2 z^2 + c_1 z + c_0}\, dz = -\frac{(x-m_x)^2}{2\sigma_x{}^2}.$$

(2.24)

Будемо вважати, що довільна стала y_0 дорівнює

$$y_0 = \frac{1}{\sqrt{2\pi}\sigma_x}.$$

Тоді остаточно маємо

$$y = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left[- \frac{(x - m_x)^2}{2\sigma_x^2} \right], \qquad (2.25)$$

тобто нормальний розподіл випадкової величини X, де m_x та σ_x - параметри цього розподілу.

Нормальному закону підпорядковуються, наприклад, температура повітря, парціальний тиск водяної пари, атмосферний тиск як біля земної поверхні, так і в вільній атмосфері, компоненти швидкості вітру у вільній атмосфері, а також деякі інші гідрометеорологічні величини. Тому нормальний закон (закон Гаусса) у гідрометеорологічних дослідженнях часто використовується.

Визначимо *основні властивості нормального розподілу.*

1. Крива розподілу симетрична відносно ординати, що проходить через точку m_x.

2. Крива має один максимум при $x = m_x$. В точці максимуму щільність розподілу дорівнює

$$f(x)_{\max} = \frac{1}{\sigma_x \sqrt{2\pi}}. \qquad (2.26)$$

3. При $|x| \to \infty$ вітки кривої асимптотично наближаються до осі OX.

4. Згідно з симетричністю кривої розподілу, математичне сподівання, мода й медіана нормального розподілу співпадають: $m_x = M_0 = M_e$ (*мода* – це значення випадкової величини, на яку припадає максимум імовірності; *медіана* – це таке значення випадкової величини, для якої $P(x < M_e) = P(x > M_e)$.

5. Крива розподілу має дві точки перегину з координатами

$$(m_x - \sigma_x; \frac{1}{\sigma_x \sqrt{2\pi e}}) \text{ і } (m_x + \sigma_x; \frac{1}{\sigma_x \sqrt{2\pi e}}).$$

6. Непарні центральні моменти нормального розподілу дорівнюють нулю.

Згідно з визначенням центрального моменту q-го порядку маємо:

$$\mu_q = \int_{-\infty}^{\infty} (x - m_x)^q f(x) dx = \\ = \frac{1}{\sigma_x \sqrt{2\pi}} \int_{-\infty}^{\infty} (x - m_x)^q e^{-\frac{(x-m_x)^2}{2\sigma_x{}^2}} dx \quad . \qquad (2.27)$$

Введемо змінну

$$t = \frac{x - m_x}{\sigma_x} \,. \tag{2.28}$$

Тоді вираз (2.27) має вигляд :

$$\mu_q = \frac{\sigma_x^{\,q}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^q e^{-\frac{t^2}{2}} dt \,, \tag{2.29}$$

Інтегруючи цей вираз частинами, отримаємо:

$$\mu_q = \frac{(q-1)\sigma_x^{\,q}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^{q-2} e^{-\frac{t^2}{2}} dt \tag{2.30}$$

З формули (2.29) можна знайти μ_{q-2}

$$\mu_{q-2} = \frac{\sigma_x^{\,q-2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^{q-2} e^{-\frac{t^2}{2}} dt \,. \tag{2.31}$$

Звідси

$$\frac{\mu_q}{\mu_{q=2}} = (q-1)\sigma_x^{\,2} \,. \tag{2.32}$$

Отже маємо просте рекурентне співвідношення, яке дає можливість знайти моменти вищих порядків через моменти нижчих порядків

$$\mu_q = (q-1)\sigma_x{}^2 \mu_{q-2}\,, \qquad q = 2,3,4,.... \quad (2.33)$$

З формули (2.33) випливає, що оскільки $\mu_1 = 0$, всі моменти непарного порядку дорівнюють нулю.

Крім того маємо:

$$\mu_2 = \sigma_x{}^2; \quad \mu_4 = 3\sigma_x{}^4; \quad \mu_6 = 15\sigma_x{}^6 \text{ і т.д.}$$

7. Коефіцієнти асиметрії та ексцесу нормального розподілу дорівнюють нулю. Дійсно,

$$r_3 = A_s = \frac{\mu_3}{\sigma_x{}^3} = 0\,; \qquad (2.34)$$

$$E = r_4 - 3 = \frac{\mu_4}{\sigma_x{}^4} - 3 = \frac{3\sigma_x{}^4}{\sigma_x{}^4} - 3 = 0\,. \qquad (2.35)$$

Стає зрозумілою важливість обчислювання цих коефіцієнтів для емпіричних рядів, оскільки вони характеризують скісінсть та сплющеність (чи витягнутість) даного розподілу порівняно з нормальним.

8. Форма кривої нормального розподілу не змінюється із зміною математичного сподівання. При зменшенні чи

збільшенні математичного сподівання графік щільності ймовірності зсувається ліворуч чи праворуч.

При змінюванні дисперсії змінюється форма кривої розподілу. Зі збільшенням $\sigma_x^{\,2}$ максимальна ордината, що визначається рівністю (2.26), зменшується, а із зменшенням $\sigma_x^{\,2}$ - збільшується. Згідно з властивістю щільності розподілу, площа, що обмежена кривою розподілу і осью абцис, дорівнює одиниці. Тому при зростанні σ_x крива розподілу розтягується вздовж осі ординат.

9. У виразі (2.25) застосуємо нову змінну за допомогою формули (2.28). Будемо мати:

$$f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} . \qquad (2.36)$$

Легко бачити, що $m_t = 0$, а $\sigma_t^{\,2} = 1$.

Функція (2.36) називається *нормованою щільністю нормального розподілу*, а її графік – *нормованою нормальною кривою*, або *кривою нормального розподілу нормованої випадкової величини* (рис.2.4). Значення нормованої щільності ймовірності для різних значень t проводяться в табл.1 Додатку . Нормована функція розподілу є функцією парною, тобто $f(-t) = f(t)$. Вона на практиці часто застосовується при розрахунках імовірності того, що випадкова величина знаходиться у визначених межах, тобто при розрахунках інтервальної імовірності. Останню і потрібно обчислювати при апроксимації емпіричної ймовірності аналітичним виразом. Обчислювання легко провести і за допомогою спеціальної функції, яка носить назву *інтеграла ймовірності:*

$$\Phi(t) = \frac{2}{\sqrt{2\pi}} \int_0^t e^{-\frac{t^2}{2}} dt \ . \qquad (2.37)$$

Ця функція затабульована і її значення при різних t приводяться в табл.2. Додатку .

Інтеграл імовірності – функція непарна, тобто

$$\Phi(-t) = -\Phi(t).$$

Згідно із зазначеною властивістю функції розподілу

$$P(\alpha < x < \beta) = F(\beta) - F(\alpha). \qquad (2.38)$$

Перейдемо від x до нормованої випадкової величини (2.28). Ясно, що нерівності

$$\alpha < x < \beta \quad \text{і} \quad \frac{\alpha - m_x}{\sigma_x} < t < \frac{\beta - m_x}{\sigma_x}$$

рівносильні. Тому рівними є ймовірності

$$P(\alpha < x < \beta) = P\left\{ \frac{\alpha - m_x}{\sigma_x} < t < \frac{\beta - m_x}{\sigma_x} \right\}. \qquad (2.39)$$

Але

$$P(t_1 < t < t_2) = F(t_2) - F(t_1), \qquad (2.40)$$

де

$$t_1 = \frac{\alpha - m_x}{\sigma_x},\ t_2 \frac{\beta - m_x}{\sigma_x}\ . \qquad (2.41)$$

Згідно з зазначеною вище властивістю функції розподілу, маємо

$$F(t) = \int_{-\infty}^{t} f(t)dt =$$

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-\frac{t^2}{2}} dt + . \qquad (2.42)$$

$$+ \frac{1}{\sqrt{2\pi}} \int_{0}^{t} e^{-\frac{t^2}{2}} dt$$

Перший інтеграл правої частини рівності (2.42) дорівнює, очевидно, $\frac{1}{2}$, а другий $\frac{1}{2}\Phi(t)$.

Отже

$$P(\alpha < x < \beta) = P(t_1 < t < t_2) =$$

$$= \frac{1}{2} + \frac{1}{2}\Phi(t_2) - \frac{1}{2} - \frac{1}{2}\Phi(t_1) =$$

$$= \frac{1}{2}\left[\Phi(t_2) - \Phi(t_1)\right]. \qquad (2.43)$$

Рівність (2.43) поряд з формулами (2.19) і (2.21) використовується при розрахунках інтервальних теоретичних частот $\widetilde{m}_i$ нормального розподілу. При цьому α і β є границями i - го часткового інтервалу довжиною c.

Можливо використовувати і інший шлях. Дійсно, оскільки

$$p_i = P(x_{i-1} < x < x_{i+1}) = \int_{x_{i-1}}^{x_{i+1}} f(x)dx$$

$$= \int_{t_{i-1}}^{t_{i+1}} f(x)dt \quad ,$$

де $t_{i-1} = \frac{x_{i-1} - \overline{x}}{\sigma_x}$; $t_{i+1} = \frac{x_{i+1} - \overline{x}}{\sigma_x}$, а x_{i-1}, x_{i+1} - межі i-го часткового інтервалу випадкової величини X.

За теоремою про середнє маємо

$$p_i = f(t_i)\left[t_{i+1} - t_{i-1}\right] = f(t_i)\frac{x_{i+1} - x_{i-1}}{\sigma_x}$$

$$= \frac{f(t_i)c}{\sigma_x},$$

де $t_i = \frac{t_{i-1} + t_{i+1}}{2}$.

Отже, використовуючи рівність (2.21), маємо

$$\widetilde{m}_i = n\frac{f(t_i)c}{\sigma_x}. \qquad (2.44)$$

Значення щільності *нормованного нормального розподілу* в точці t_i беруться з таблиці Додатку . В формулі (2.44) при конкретних розрахунках σ_x замінюється на його статистичну оцінку S_x.

10. Знайдемо ймовірність того, що нормально розподілена випадкова величина X відхиляється від свого математичного сподівання m_x на величину меншу ніж ε, тобто

$$P\left\{\left|x - m_x\right| \le \varepsilon\right\}. \qquad (2.45)$$

Перейдемо до нормованої випадкової величини. Оскільки $x = m_x \pm \varepsilon$, маємо:

$$t = \frac{m_x \pm \varepsilon - m_x}{\sigma_x} = \pm \frac{\varepsilon}{\sigma_x}. \qquad (2.46)$$

Тоді

$$P(t_1 < t < t_2) = \frac{1}{2}\Phi(t_2) - \frac{1}{2}\Phi(t_1) =$$

$$= \frac{1}{2}\Phi(\frac{\varepsilon}{\sigma_x}) - \frac{1}{2}\Phi(-\frac{\varepsilon}{\sigma_x}) = \Phi(\frac{\varepsilon}{\sigma_x})$$

Отже, маємо:

$$P\{|x - m_x| < \varepsilon\} = \Phi(\frac{\varepsilon}{\sigma_x}). \qquad (2.47)$$

Будемо тепер вважати, що ε послідовно дорівнює σ_x,$2\sigma_x$ і $3\sigma_x$. Тоді за допомогою формули (2.47) прийдемо до таких співвідношень :

При $\varepsilon = \sigma_x, P(|x - m_x| < \sigma_x) = \Phi(1) = 0{,}6827,$;

при $\varepsilon = 2\sigma_x, P(|x - m_x| < 2\sigma_x) = \\ = \Phi(2) = 0{,}9545,$;

при $\varepsilon = 3\sigma_x, P(|x - m_x| < 3\sigma_x) = \\ = \Phi(3) = 0{,}9973$.

Із останньої рівності виходить, що практично розсіяння нормально розподіленої випадкової величини X укладається на інтервалі $m_x \pm 3\sigma_x$ (рис.2.5). Імовірність того, що X виходить за цей інтервал, дуже мала і дорівнює 0,0027. Така подія може вважатися практично неможливою. На проведенному міркуванні основується *правило трьох сигм*, яке формулюється таким чином: якщо випадкова величина має нормальний розподіл, то відхил цієї величини від математичного сподівання по абсолютній величині не перебільшує трьох середніх квадратичних відхилів. Але, якщо остання властивість виконується, це ще не означає, що випадкова величина X підпорядковується нормальному розподілу.

Термін “нормальний розподіл” застосовується в умовному смислі, як загальноприйнятий в літературі термін. Так, твердження, що якась ознака підпорядковується нормальному закону розподілу, зовсім не позначає наявність будь-яких непорушних норм, ніби-то таких, що полягають в основі явища відбиттям якого є ознака, яка розглядається, а підпорядкування іншим видам законів не означає якусь-то анормальність цього явища. Головна особливість нормального закону полягає у тому, що він є граничним законом, до якого наближаються інші закони розподілу.

В табл.2.1 наводиться приклад розрахунків теоретичних частот нормального розподілу на основі емпіричного ряду середніх місячних температур повітря в червні в Одесі. Підставою для гіпотези про те, що емпіричні інтервальні частоти відповідають частотам нормального розподілу є близькість до нуля третього і до трійки – четвертого основних моментів. Розрахунки теоретичних частот в табл.2.1 проводяться на основі двох розглянутих вище методів: за допомогою нормованого нормального розподілу (з 5 до 9 стовпця) і на основі інтеграла ймовірностей (10-12 стовпці табл.2.1).

Як видно із таблиці, і перший , і другій методи дають результати, що , по-перше, дуже близькі між собою, а по-друге, мало відрізняються від відповідних емпіричних частот. Результат же апроксимації цього статистичного розподілу (табл.2.1) нормальним розподілом приводиться в розділі 4 (§4.5).

Таблиця 2.1- Розрахунки теоретичних частот нормального розподілу по ряду середніх місячних температур повітря червень, Одеса)

Статистики ряду та оцінки параметрів

n	$c,^0C$	$\bar{x},^0C$	$S_x,^0C$	r_3	r_4
100	1.5	17.3	2.9	0.16	2.57

Розрахунки теоретичних частот

№	Границі градації	m_i	Нові градації	t_i

п/п	ліва, x_{i-1}	права x_{i+1}		t_{i-1}	t_{i+1}	
1.	2.	3.	4.	5.	6.	7.
1	10,0	11,5	1	-2,52	-2,00	-2,26
2	11,5	13,0	6	-2,00	-1,48	-1,74
3	13,0	14,5	11	-1,48	-0,97	-1,23
4	14,5	16,0	16	-0,97	-0,45	-0,71
5	16,0	17,5	20	-0,45	0,07	-0,19
6	17,5	19,0	19	0,07	0,59	0,33
7	19,0	20,5	14	0,59	1,10	0,85
8	20,5	22,0	7	1,10	1,62	1,36
9	22,0	23,5	4	1,62	2,14	1,88
10	23,5	25,0	2	2,14	2,66	2,40

Продовження табл.2.1 (Розрахунки теоретичних частот)

$f(t_i)$	$\widetilde{m}_i$	$\Phi(t_{i+1})$	$\Phi(t_{i-1})$	$\widetilde{m}_i$
9.	10.	11.	12.	13.
0,0310	1,6	-0,95450	-0,98826	1,7
0,0878	4,5	-0,86113	-0,95450	4,7
0,1872	9,7	-0,66795	-0,86113	9,7
0,3101	16,0	-0,34729	-0,66795	16,0
0,3918	20,3	0,05581	-0,34729	20,2
0,3778	19,5	0,44481	0,05581	19,5
0,2780	14,4	0,72867	0,44481	14,2
0,1582	8,2	0,89477	0,72867	8,3
0,0681	3,5	0,96765	0,89477	3,6
0,0224	1,2	0,99219	0,96765	1,2

Розглянемо тепер деякі інші, найбільш популярні в гідрометеорологічних дослідженнях, типи розподілів Пірсона.

2.3.3 Перший тип розподілів Пірсона

I тип спостерігається при $\chi < 0$, а його основне рівняння має вид :

$$\widetilde{m}_i = \widetilde{m}_0 (1 + \frac{z_i}{l_1})^{q_1} (1 - \frac{z_i}{l_2})^{q_2} . \qquad (2.48)$$

В ньому $q_1, q_2, l_1, l_2, \widetilde{m}_0$- параметри, які визначаються формулами :

$$\left.\begin{matrix} q_1 \\ \\ q_2 \end{matrix}\right\} = \frac{1}{2}\left[S - 2 \mp S(S+2)\frac{r_3}{t} \right]; \qquad (2.49)$$

$$l_1 = \frac{q_1 l}{S-2}; \qquad (2.50)$$

$$l_2 = \frac{q_2 l}{S-2}; \qquad (2.51)$$

$$l = \frac{\sigma t}{2}. \qquad (2.52)$$

Статистикам σ і t відповідають рівності (2.9) і (2.14), або (2.16).

$$\tilde{m}_0 = \frac{n}{l}\frac{{q_1}^{q_1}{q_2}^{q_2}}{(S-2)^{S-2}}\frac{\Gamma(q_1+q_2+2)}{\Gamma(q_1+1)\Gamma(q_2+1)}. \qquad (2.53)$$

У рівності (2.53) $\Gamma(a)$ - гамма-функція відповідного аргументу.

$$\Gamma(a) = \int_0^{\infty} t^{a-1} e^{-t} dt. \qquad (2.54)$$

Вигляд кривих розподілу Пірсона I типу залежить від параметрів форми q_1 і q_2, а параметри l_1 і l_2 визначають межі області визначення функції (2.48).

Розглянемо *різні значення параметрів форми.*

1) Параметри форми $q_1 > 0;\ \ q_2 > 0$. В цьому разі $\widetilde{m}_i = 0$ при $z_i = -l_1$ і $z_i = l_2$. При $z_i = 0$, $\widetilde{m}_i = \widetilde{m}_0$. Отже, крива розподілу починається в точці $z_i = -l_1$ і закінчується в точці $z_i = l_2$, досягаючи максимального значення на частковому інтервалі, центром якого є точка $z_i = 0$.

а) $q_1 > 1;\ \ q_2 > 1$ - крива розподілу має дві точки перегину (рис. 2.6) ;

б) $0 < q_1 < 1;\ \ q_2 > 1$ - крива розподілу має лише одну точку перегину праворуч відносно початку координат. Ліворуч відбувається вертикальний дотик (рис.2.7);

в) $q_1 > 1;\ \ 0 < q_2 < 1$ - крива розподілу має вертикальний дотик праворуч, а ліворуч – точку перегину (рис.2.8);

г) $0 < q_1 < 1;\ \ 0 < q_2 < 1$ - крива розподілу має тільки вертикальні дотики, а точки перегину відсутні (рис.2.9).

2) Параметр $q_1 < 0$, а параметр $q_2 > 0$.

Тоді основне рівняння I типу можна записати таким чином :

$$\widetilde{m}_i = \frac{\widetilde{m}_0 (1 - \frac{z_i}{l_2})^{q_2}}{(1 + \frac{z_i}{l_1})^{-q_1}}, \quad (-q_1 > 0). \qquad (2.55)$$

Звідси видно, що при $z_i \to -l_1, \widetilde{m}_i \to \infty$. При $z_i = l_2, \widetilde{m}_i = 0$. Отже, пряма $z_i = -l_1$ є асимптотою графіка функції (2.55)

а) $-1 < q_1 < 0; \quad q_2 > 1$ - крива розподілу праворуч має горизонтальний дотик в точці $z_i = l_2$ (рис. 2.10);

б) $-1 < q_1 < 0; \quad 0 < q_2 < 1$ - крива розподілу праворуч має вертикальний дотик (рис.2.11).

3) Параметр $q_1 > 0$, а параметр $q_2 < 0$.

У цьому випадку рівняння (2.48) має вид:

$$\widetilde{m}_i = \frac{\widetilde{m}_0 (1 + \frac{z_i}{l_1})^{q_1}}{(1 - \frac{z_i}{l_2})^{-q_2}}, \qquad (-q_2 > 0). \tag{2.56}$$

Видно, що $\widetilde{m}_i \to \infty$, коли $z_i \to l_2$. Отже пряма $z_i = l_2$ є асимптотою графіка функції (2.56) :

а) Коли $q_1 > 1$, а $-1 < q_2 < 0$, то ліворуч маємо горизонтальний дотик (рис.2.12);

б) Якщо $0 < q_1 < 1$, а $-1 < q_2 < 0$, то ліворуч спостерігається вертикальний дотик (рис. 2.13).

4) Обидва параметри форми менші нуля ($q_1 < 0; \quad q_2 < 0$).

Тоді рівняння (2.48) перетворюється в рівняння

$$\widetilde{m}_i = \frac{\widetilde{m}_0}{(1 + \frac{z_i}{l_1})^{-q_1} (1 - \frac{z_i}{l_2})^{-q_2}}, \tag{2.57}$$

$$(-q_1 > 0; \ -q_2 > 0)$$

У цьому випадку $\widetilde{m}_i \to \infty$ і при $z_i \to -l_1$, і при $z_i \to l_2$. Маємо асиметричний U - видний розподіл (рис. 2.14).

Таким чином, I тип розподілу Пірсона включає 9 форм кривих розподілу.

Для I типу розподілу статистика $\hat{x}$ визначається формулою

$$\hat{x} = \bar{x} - \frac{r_3 \sigma_x}{2} \frac{(S+2)}{(S-2)}. \qquad (2.58)$$

В табл. 2.2 наводяться результати розрахунків теоретичних частот I типу розподілу Пірсона на основі статистичного ряду модуля швидкості вітру на висоті 0.1 км.

Як випливає з табл. 2.2, емпіричні інтервальні частоти m_i розрізняються з відповідними теоретичними частотами $\widetilde{m}_i$ незначно. Крім того, можна зробити висновок про те, що на модальний інтервал частот, центром якого є швидкість вітру $\hat{x}$=6.5 м/с, припадає біля 26% випадків.

Таблиця 2.2 - Результати розрахунків теоретичних частот I типу Пірсона на основі ряду швидкості вітру на висоті 0,1 км.

Статистики ряду та оцінки параметрів

n	c	$\bar{x}$	S_x	$\hat{r}_3$	$\hat{r}_4$	σ	S	χ	t
220	2,0	6,3	2,78	0,07	2,32	1,39	5,74	-0,002	10,40

Продовження табл.2.2 (Статистики ряду та оцінки параметрів)

q_1	q_2	l	l_1	l_2	$\hat{x}$	$\widetilde{m}_0$
1,73	2,01	7,22	3,34	3,88	6,54	56,03

Розрахунки теоретичних частот

$$\widetilde{m}_i = \widetilde{m}_0 (1 + \frac{z_i}{l_1})^{-q_1} (1 - \frac{z_i}{l_2})^{-q_2}$$

№ п/п	$\widetilde{x}_i$	m_i	z_i	$1 + \frac{z_i}{l_1}$	$(1 + \frac{z_i}{l_1})^{-q_1}$
1.	2.	3.	4.	5.	6.
1	1,5	13	-2,52	0,25	0,09
2	3,5	38	-1,52	0,54	0,35
3	5,5	53	-0,52	0,84	0,75
4	7,5	56	0,48	1,14	1,26
5	9,5	39	1,48	1,44	1,89
6	11,5	19	2,48	1,74	2,62
7	13,5	2	3,48	2,04	3,44

Продовження табл.2.2 (Розрахунки теоретичних частот)

$(1-\frac{z_i}{l_2})$	$(1-\frac{z_i}{l_2})^{-q_2}$	$\widetilde{m}_i$
7.	8.	9.
1,65	2,74	13,5
1,39	1,94	38,1
1,13	1,29	53,8
0,88	0,77	54,2
0,62	0,38	40,2
0,36	0,13	18,8
0,10	0,01	2,0

2.3.4 Другий тип розподілів Пірсона

II тип спостерігається, коли $\chi = 0; r_3 = 0; r_4 < 3.$

Основне рівняння для інтервальних частот цього розподілу має вид:

$$\widetilde{m}_i = \widetilde{m}_0 \left(1 - \frac{z_i^{\,2}}{\widetilde{l}^{\,2}}\right)^q, \qquad (2.59)$$

де $q, \widetilde{l}, \widetilde{m}_0$ - параметри цього розподілу.

Порівнюючи рівняння (2.48) та (2.59), легко побачити, що II тип породжується I типом при $q_1 = q_2 = q$ і

$l_1 = l_2 = \tilde{l}$. З формули (2.49) при цих умовах випливає (оскільки $r_3 = 0$):

$$q_{1,2} = \frac{1}{2}(S-2) = q \ . \tag{2.60}$$

Але при умові $r_3 = 0,$ статистика S дорівнює

$$S = \frac{6(r_4 - 1)}{6 - 2r_4}. \tag{2.61}$$

З урахуванням цього маємо формулу для параметра форми q:

$$q = \frac{5r_4 - 9}{2(3 - r_4)} \ . \tag{2.62}$$

При цих же умовах, статистика t дорівнює

$$t = 4\sqrt{\frac{2r_4}{3 - r_4}} \ . \tag{2.63}$$

Оскільки $l_1 = l \ = \tilde{l}$, то

$$\tilde{l} = \frac{l}{2} = \frac{\sigma t}{4} = \sigma\sqrt{\frac{2r_4}{3 - r_4}}. \qquad (2.64)$$

Таким же чином можна показати, що

$$\tilde{m}_0 = \frac{n}{\tilde{l}\,2^{2q+1}} \frac{\Gamma(2q+2)}{\{\Gamma(q+1)\}^2}. \qquad (2.65)$$

Проаналізуємо формулу (2.62). Очевидно, що $q = 0,$ коли $5r_4 - 9 = 0.$ Звідси $r_4 = 1{,}8.$

Отже, доцільно розглянути три ситуації:

а) $r_4 = 1{,}8; \quad q = 0.$ З формули (2.59) випливає, що в інтервалі $-\tilde{l} < z_i < \tilde{l} \qquad \tilde{m}_i = \tilde{m}_0 = const.$ Таким чином, у цьому випадку ми маємо *рівномірний розподіл*. Отже рівномірний розподіл є частинним випадком II типу розподілів Пірсона (рис. 2.15)

б) $1{,}8 < r_4 < 3.$ Тоді $q > 0.$ З формули (2.59) виходить, що $\tilde{m}_i = 0$ при $z_i = \pm\tilde{l}$ і $\tilde{m}_i = \tilde{m}_0$ при $z_i = 0.$ Отже ми маємо симетричну відносно початку координат криву розподілу, яка починається в точці $z_i = -\tilde{l}$ і закінчується в точці $z_i = \tilde{l}$ (рис. 2.16)

в) $r_4 < 1{,}8; \quad q < 0.$ Тоді формулу (2.59) можна перетворити так :

$$\widetilde{m}_i = \frac{\widetilde{m}_0}{\left(1 - \frac{z_i^{\,2}}{\widetilde{l}^{\,2}}\right)^{-q}}, \quad (-q > 0). \tag{2.66}$$

З рівняння (2.66) видно, що при $z_i \to \pm\widetilde{l}$, $\widetilde{m}_i \to \infty$, а при $z_i = 0$, $\widetilde{m}_i = \widetilde{m}_0$. Ми отримаємо симетричний U - видний розподіл (рис. 2.17).

Треба мати на увазі, що оскільки $r_3 = 0$, статистика $\hat{x}$, як видно із загального співвідношення (2.22), дорівнює середньому значенню, тобто

$$\hat{x} = \overline{x}.$$

Крива розподілів Пірсона II типу при $q > 0$ суттєво відрізняється від кривої нормального розподілу. По-перше, область існування розподілу другого типу Пірсона є обмеженою $(-\widetilde{l} \le z \le \widetilde{l})$, а нормального розподілу – необмеженою, по друге, вона є більш сплюснутою $(E < 0)$, ніж крива нормального розподілу.

В табл. 2.3 наводяться результати розрахунків теоретичних частот II типу розподілу Пірсона на основі ряду меридіональної складової швидкості вітру на висоті 50 км. Підставою для прийняття рішення про можливість використання для цього II типу кривих Пірсона є той факт, що

$r_3 \approx 0,$ тобто ми маємо діло з симетричним розподілом. Крім того, параметр $\chi \approx 0,$ а $\hat{r}_4 < 3.$

Як випливає з табл.2.3, де містяться результати розрахунків, отримані інтервальні теоретичні частоти $\widetilde{m}_i$ дуже близькі до відповідних емпіричних частот m_i, а найбільша частота, що дорівнює $\widetilde{m}_0 = 13$ припадає на частковий інтервал, центром якого є швидкість меридіонального вітру $\hat{x} = \bar{x} = 7$ м/с. Результат же апроксимації цього емпіричного розподілу розподілом Пірсона II типу приводиться в розділі 4 (§ 4.5).

Таблиця 2.3 - Результати розрахунків теоретичних частот II типу Пірсона для меридіональної складової швидкості вітру на висоті 50 км.

Статистики ряду та оцінки параметрів

n	c	$\bar{x}$	S_x	$\hat{r}_3$	$\hat{r}_4$	σ	S
86	3,8	7,2	8,22	0,07	2,05	2,16	3,25

Продовження табл.2.3 (Статистики ряду та оцінки параметрів)

χ	t	q	$\widetilde{l}$	$\widetilde{m}_0$
-0,0019	8,29	0,63	4,48	12,92

Розрахунки теоретичних частот $\widetilde{m}_i = \widetilde{m}_0 (1 - \frac{z_i^2}{\widetilde{l}^2})^q$

№ п/п	$\widetilde{x}_i$	m_i	z_i	z_i^2	$\frac{z_i^2}{\widetilde{l}^2}$
1.	2.	3.	4.	5.	6.
1	-9,1	2	-4,29	18,40	0,92
2	-5,3	7	-3,29	10,82	0,54
3	-1,5	10	-2,29	5,24	0,26
4	2,3	16	-1,29	1,66	0,08
5	6,1	11	-0,29	0,08	0,004
6	9,9	14	0,71	0,50	0,03
7	13,7	10	1,71	2,92	0,15
8	17,5	8	2,71	7,34	0,37
9	21,3	8	3,71	13,76	0,69

Продовження табл.2.3 (Розрахунки теоретичних частот)

$(1 - \frac{z_i^2}{\widetilde{l}^2})$	$(1 - \frac{z_i^2}{\widetilde{l}^2})^q$	$\widetilde{m}_i$
7.	8.	9.
0,08	0,20	2,6
0,46	0,61	7,9
0,74	0,83	10,7
0,92	0,95	12,3
1,00	1,00	12,9
0,97	0,98	12,7
0,85	0,90	11,6
0,63	0,75	9,7
0,31	0,48	6,2

2.3.5 Третій тип розподілів Пірсона

Як вже зазначалось, III тип може спостерігатися при будь-яких значеннях статистики χ. Але, як показує попит, найбільш часто це відбувається при $|\chi| > 4$.

Основне рівняння III типу має вид:

$$\tilde{m}_i = \tilde{m}_0 (1 + \frac{z_i}{l'})^p e^{-\frac{pz_i}{l'}}. \qquad (2.67)$$

Параметри розподілу визначаються формулами :

$$p = \frac{4}{r_3^{\,2}} - 1; \qquad (2.68)$$

$$l' = \sigma\left(\frac{2}{r_3} - \frac{r_3}{2}\right); \qquad (2.69)$$

$$\tilde{m}_0 = \frac{n}{|l'|} \frac{p^{p+1}}{e^p \Gamma(p+1)}, \qquad (2.70)$$

при цьому :

$$\hat{x} = \bar{x} - \frac{r_3 \sigma_x}{2}. \qquad (2.71)$$

Криві, що відносяться до III типу розподілів Пірсона, можна поділити на дві групи .

1) $l' > 0$. При додатних значеннях параметру масштабу l' спостерігається три різновиди кривих розподілу. Загальними для них є дві властивості: при $z_i \to \infty,\ \widetilde{m}_i \to 0$ (експоненціальна функція спадає швидше, ніж зростає степенева функція), при $z_i = -l',\ \widetilde{m}_i = 0,$ тобто точка $z_i = -l'$ визначає початок розподілу.

а) якщо $p > 1$, то крива розподілу має дві точки перегину, які розташовуються праворуч та ліворуч відносно початку координат (рис. 2.18);

б) при $0 < p < 1$ маємо тільки одну точку перегину графіка функції. При наближенні до $z_i = -l'$, відбувається вертикальний дотик (рис. 2.19);

в) при $-1 < p < 0$ основне рівняння III типу Пірсона можна записати таким чином :

$$\widetilde{m}_i = \frac{\widetilde{m}_0 e^{-\frac{p z_i}{l'}}}{(1 + \frac{z_i}{l'})^{-p}}, \quad (-p > 0). \qquad (2.72)$$

Коли $z_i \to -l'$, знаменник рівняння (2.72) наближається до нуля, а $\widetilde{m}_i \to \infty$ крива розподілу зліва має асимптоту $z_i = -l'$ (рис. 2.20).

2) $l' < 0.$ Таким значенням цього параметру теж відповідає три типа кривих розподілу, які відрізняються від попередніх кривих протилежною асиметрією (при $l' > 0$ розподіли мають правосторонню асиметрію, а при $l' < 0$ - лівосторонню).

Ясно, що тепер $\widetilde{m}_i \to 0$ при $z_i \to -\infty$, тобто асимптотою графіка функції виявляється від'ємна піввісь осі OZ. Крім того, $\widetilde{m}_i = 0$ ($\widetilde{m}_i \to \infty$, якщо $p < 0$) при $z_i = -l'$ (оскільки $l' < 0$, то точка початку кривої розподілу знаходиться на додатній півосі, тобто праворуч від початку координат). Криві розподілу, що відносяться до цієї групи, зображені на рис. 2.21- 2.23.

В табл. 2.4 у якості прикладу знаходяться результати розрахунків теоретичних частот розподілу Пірсона III типу на основі статистичного ряду модуля швидкості вітру на висоті 0,1 км.

Порівняння емпіричних m_i та теоретичних $\widetilde{m}_i$ інтервальних частот дає можливість прийти до висновку про невелику розбіжність між цими частотами. Найбільшу ж частоту $\widetilde{m}_0 = 54$ при $n = 179$ має швидкість вітру 5 м/с.

Про результат апроксимації цього (табл. 2.4) емпіричного розподілу розподілом Пірсона III типу дивись у розділі 4 (§ 4.5).

Ми розглянули у попередніх параграфах цього розділу тільки ті типи розподілів Пірсона, які найбільш часто використовуються при вивченні статистичної структури гідрометеорологічних величин.

Інші типи досить докладно викладені в книжці А.К. Мітропольского "Техника статистических вычислений ".

Таблиця 2.4 - Результати розрахунків теоретичних частот $\widetilde{m}_i$ III типу Пірсона на основі ряду швидкості вітру на висоті 0,1 км.

Статистики ряду та оцінки параметрів

n	c	$\bar{x}$	S_x	$\hat{r}_3$	$\hat{r}_4$	σ	S
179	2,0	5,7	2,86	0,57	2,88	1,43	7,8

Продовження табл. 2.4 (Статистики ряду та оцінки параметрів)

χ	p	l'	$\hat{x}$	$\widetilde{m}_0$
-0,22	11,12	4,61	4,88	53,7

Розрахунки теоретичних частот $\widetilde{m}_i = \widetilde{m}_0 (1 + \frac{z_i}{l'})^p e^{-\frac{pz_i}{l'}}$

№ п/п	$\widetilde{x}_i$	m_i	z_i	$1 + \frac{z_i}{l'}$	$(1 + \frac{z_i}{l'})^p$
1.	2.	3.	4.	5.	6.
1	1,5	20	-1,69	0,63	0,006
2	3,5	52	-0,69	0,85	0,16
3	5,5	43	0,31	1,07	2,12
4	7,5	31	1,31	1,28	15,57
5	9,5	23	2,31	1,50	90,81
6	11,5	8	3,31	1,72	415,98
7	13,5	1	4,31	1,93	1497,6
8	15,5	1	5,31	2,15	4974,1

Продовження табл. 2.4 (Розрахунки теоретичних частот)

$-\frac{pz_i}{l'}$	$e^{-\frac{pz_i}{l'}}$	$\tilde{m}_i$
7.	8.	9.
4,08	58,94	19,0
1,66	5,28	45,4
-0,75	0,47	53,9
-3,16	0,04	35,5
-5,57	0,004	18,5
-7,98	0,0003	7,6
-10,4	0,00003	2,5
-12,81	$0,3\ 10^{-5}$	0,7

2.4 Гамма-розподіл

Гамма розподілом називається розподіл, щільність імовірности для котрого визначається формулою

$$f(x) = \frac{\alpha^{\lambda} x^{\lambda-1} e^{-\lambda x}}{\Gamma(\lambda)}, \qquad (2.73)$$

де λ - параметр форми, α - параметр масштабу.

Гамма розподіл може мати іншу форму, якщо застосувати перетворення

$$u = \alpha x \qquad (2.74)$$

Запишемо щільність імовірності через похідну від функції розподілу, і розділимо обидві частини рівності на α. З урахуванням перетворення (2.74) будемо мати:

$$f(u) = \frac{u^{\lambda-1}e^{-u}}{\Gamma(\lambda)} . \qquad (2.75)$$

Розглянемо властивості гамма-розподілу.

Крива розподілу починається в точці $u = o$. При $u \to \infty, f(u) \to 0$. Таким чином, щільність імовірності гамма-розподілу існує в області $[0, \infty)$.

Знайдемо похідну функції (2.75):

$$f'(u) = e^{-u}u^{\lambda-2}(\lambda - 1 - u) . \qquad (2.76)$$

Точку екстремуму отримаємо, прирівнюючи похідну до нуля. Вона визначається рівністю

$$u = \lambda - 1 . \qquad (2.77)$$

З рівності (2.77) видно, що екстремум функція (2.75) може мати лише у випадку, коли $\lambda > 1$.

Друга похідна функції (2.75) дорівнює

$$f''(u) = e^{-u}u^{\lambda-3}\begin{bmatrix} u^2 - 2(\lambda-1)u + \\ +(\lambda-1)(\lambda-2) \end{bmatrix}. \qquad (2.78)$$

Можна легко показати, що в точці екстремуму квадратний трьохчлен дорівнює мінус одиниці. Але саме він визначає знак другої похідної y цій точці. Таким чином, $f''(u=\lambda-1)<0$ і в точці екстремуму ми маємо максимум функції (2.75). Стосовно до щільності імовірності це означає, що точка (2.77) є модальним значенням випадкової величини u.

Крім того, видно, що крива розподілу має моду тільки у випадку, коли $\lambda>1.$

Умови наявності точок перегину отримаємо, прирівнюючи другу похідну (2.78) до нуля. Приходимо до квадратного рівняння

$$u^2-2(\lambda-1)u+(\lambda-1)(\lambda-2)=0, \qquad (2.79)$$

розв'язок якого має вид

$$u=\lambda-1\pm\sqrt{\lambda-1}. \qquad (2.80)$$

Рівності (2.80) свідчать про те, що при $1<\lambda\le 2.$ Крива розподілу має лише одну точку перегину праворуч від початку координат (рис. 2.24). При $\lambda>2$ точок перегину дві (рис. 2.25).

Якщо параметр форми $\lambda<1$, то як випливає з формули (2.75), $f(u)\to\infty$ при $u\to 0$ (рис. 2.26).

Частинним випадком гамма-розподілу є експоненціальний розподіл (рис. 2.27). Ми його, очевидно, отримаємо при $\lambda=1.$ Оскільки $\Gamma(1)=1,$ рівняння (2.75) приймає вид:

$$f(u) = e^{-u} \ . \qquad (2.81)$$

Знайдемо моменти випадкових величин, що підпорядковуються гамма-розподілу.

Початкові моменти гамма-розподілу дорівнюють

$$\nu_l = \mu[u^l] = \frac{1}{\Gamma(\lambda)} \int_0^{\infty} u^{\lambda+l-1} e^{-u} du = \frac{\Gamma(\lambda+l)}{\Gamma(\lambda)} =$$

$$= \frac{(\lambda+l-1)!}{(\lambda-1)!} = \lambda(\lambda+1)...(\lambda+l-1) \qquad (2.82)$$

Звідси $\nu_1 = \lambda; \quad \nu_2 = \lambda(\lambda+1)$.

Тому математичне сподівання випадкової величини u і її дисперсія дорівнюють

$$m_u = \lambda; \qquad (2.83)$$

$$\sigma_u{}^2 = \lambda(\lambda+1) - \lambda^2 = \lambda. \qquad (2.84)$$

Таким же чином можна показати, що

$$\mu_3 = 2\lambda \text{ і } \mu_4 = 3\lambda(\lambda+2), \qquad (2.85)$$

що дає змогу легко визначити третій та четвертий основні моменти випадкової величини u:

$$r_3 = \frac{\mu_3}{\sigma_u{}^3} = \frac{2\lambda}{\sqrt{\lambda^3}} = \frac{2}{\sqrt{\lambda}}; \qquad (2.86)$$

$$r_4 = \frac{\mu_u}{\sigma_u{}^4} = \frac{3\lambda^2 + 6\lambda}{\lambda^2} = \frac{6}{\lambda} + 3. \qquad (2.87)$$

Очевидно, коефіцієнт ексцесу визначається формулою

$$E = \frac{6}{\lambda}. \qquad (2.88)$$

Отже, міри косості та крутості (асиметрія та ексцес) залежать тільки від параметра форми λ гамма-розподілу. Оскільки $\lambda > 0$, то крива розподілу має правосторонню асиметрію і більшу крутість, ніж крива нормального розподілу.

На основі отриманих рівностей для моментів розподілу, а також рівності (2.74), можна легко визначити формули статистичних оцінок параметрів гамма-розподілу.

Із формули (2.84) випливає, що оцінкою параметра форми $\hat{\lambda}$ є $S_u{}^2$- оцінка дисперсії випадкової величини u. Але враховуючи рівність (2.74), маємо:

$$S_u{}^2 = \overline{[\hat{\alpha}^2 (x - \bar{x})^2]} = \hat{\alpha}^2 \overline{(x - \bar{x})^2} = \hat{\alpha}^2 S_x{}^2.$$

Отже

$$\hat{\lambda} = \hat{\alpha}^2 S_x{}^2. \qquad (2.89)$$

З іншого боку

$$\bar{u} = \hat{\lambda}, \qquad (2.90)$$

тобто

$$\hat{\lambda} = \hat{\alpha}\bar{x}. \qquad (2.91)$$

Якщо рівність (2.91) підставити до рівняння (2.89), то отримаємо після скорочення

$$\hat{\alpha} = \frac{\bar{x}}{S_x{}^2}. \qquad (2.92)$$

Враховуючи тепер рівність (2.92) в формулі (2.89), будемо мати:

$$\hat{\lambda} = \frac{(\bar{x})^2}{S_x{}^2}. \qquad (2.93)$$

Теоретичні частоти $\tilde{m}_i$ для гамма-розподілу можна знайти за допомогою формули

$$\tilde{m}_i = np_i = nP(u_{i-1} < u < u_{i+1}), \qquad (2.94)$$

де u_{i-1} і u_{i+1} - межі і-того часткового інтервалу. В свою чергу, інтервальну імовірність можна знайти, застосовуючи відповідну властивість щільності ймовірності

$$P(u_{i-1} < u < u_{i+1}) = \frac{1}{\Gamma(\lambda)} \int_{u_{i-1}}^{u_{i+1}} u^{\lambda-1} e^{-u} du$$

або

$$P(u_{i-1} < u < u_{i+1}) =$$
$$= \frac{1}{\Gamma(\lambda)} \left[\int_0^{u_{i+1}} u^{\lambda-1} e^{-u} du - \int_0^{u_{i-1}} u^{\lambda-1} e^{-u} du \right]. \qquad (2.95)$$

Функція

$$\gamma(\lambda,t)=\frac{1}{\Gamma(\lambda)}\int_0^t u^{\lambda-1}e^{-u}du \qquad (2.96)$$

називається *неповною гамма-функцією*.

Таким чином,

$$\widetilde{m}_i = n\left[\gamma(\lambda,u_{i+1})-\gamma(\lambda,u_{i-1})\right]. \qquad (2.97)$$

Існують таблиці цих функцій (наприклад, “Таблиці неповних гамма-функцій” Є.Слуцького), за допомогою яких можна обчислити інтервальні імовірності й, після цього, інтервальні теоретичні частоти.

Розрахунки теоретичних частот можна втілити й іншим шляхом.

Покажемо, що при $l'>0$ до гамма-розподілу зводиться III тип розподілу Пірсона.

Для цього від інтервальних частот $\widetilde{m}_i$ у формулі (2.94) перейдемо до інтервальних імовірностей за формулою

$$p_i=\frac{\widetilde{m}_i}{n}, \qquad (2.98)$$

і будемо вважати, що довжина часткового інтервалу $c\to 0$. Тоді отримаємо відповідну формулу для щільності ймовірності, яка при умові, що $\widetilde{m}_0$ визначається рівністю (2.70), має вид

$$f(z)=\frac{1}{l'}\frac{p^{p+1}}{e^{p}\Gamma(p+1)}\left(1+\frac{z}{l'}\right)^{p}e^{-\frac{pz}{l'}}. \qquad (2.99)$$

Розділимо обидві частини рівності (2.99) на $\frac{p}{l'}$. Будемо мати в лівій частині (2.99):

$$\frac{1}{\frac{p}{l'}}f(z)=\frac{dF}{d\left(\frac{zp}{l'}+p\right)}=f\left(\frac{zp}{l'}+p\right), \qquad (2.100)$$

а рівняння (2.99) перетворюється на таку формулу:

$$f\left(\frac{zp}{l'}+p\right)=\frac{1}{e^{p}\Gamma(p+1)}\left[p+\frac{zp}{l'}\right]^{p}e^{-\frac{pz}{l'}}$$

або, якщо позначити $p+\frac{zp}{l'}=u$, то

$$f(u)=\frac{1}{e^{p}\Gamma(p+1)}u^{p}e^{p-u}. \qquad (2.101)$$

Тепер залишилося ввести позначення

$$p+1=\lambda, \tag{2.102}$$

щоб отримати рівняння

$$f(u)=\frac{u^{\lambda-1}e^{-u}}{\Gamma(\lambda)}. \tag{2.103}$$

Порівняння формул (2.75) і (2.103) показує, що вони є тотожними. Аналогічний вид мають й криві гамма-розподілу і розподілу Пірсона III типу при $l'>0$ (рис.2.18-2.20 і 2.24-2.26). Отже, інтервальні теоретичні частоти гамма розподіленої випадкової величини можна розрахувати за допомогою розподілу Пірсона III типу при умові, коли $r_3<2$ $(l'>0)$.

2.5 Логарифмічно нормальний розподіл

Логарифмічно нормальним розподілом називається розподіл такої випадкової величини $x>0$, логарифм якої $u=\ln x$ має нормальний розподіл.

$$f(u)=\frac{1}{\sigma_u\sqrt{2\pi}}e^{-\frac{(u-m_u)^2}{2\sigma_u{}^2}}. \tag{2.104}$$

Щільність імовірності випадкової величини x, яку можна подати як показникові функцію $x = e^u$, зв'язана з щільністю розподілу величини u таким чином:

$$f(x) = \frac{dF(x)}{dx} = \frac{dF[x(u)]}{du}\frac{du}{dx} = \frac{1}{x} f(u).$$

(2.105)

Звідси маємо

$$f(x) = \frac{1}{x\sigma_u\sqrt{2\pi}} e^{-[\ln x - m_u]^2 / 2\sigma_u^2}. \qquad (2.106)$$

Очевидно,

$$m_x = \int_0^\infty x f(x)dx = e^{\frac{\sigma_u^2 + 2m_u}{2}}, \qquad (2.107)$$

$$\sigma_x^2 = \int_0^\infty [x - m_x]^2 f(x)dx = \\ = e^{\sigma_u^2 + 2m_u}\left[e^{\sigma_u^2} - 1\right]. \qquad (2.108)$$

Знайдемо математичне сподівання й дисперсію випадкової величини u. Поділивши σ_x^2 на m_x^2, будемо мати:

$$\frac{\sigma_x^2}{m_x^2} = e^{\sigma_u^2} - 1, \tag{2.109}$$

звідки

$$\sigma_u^2 = \ln\left[\frac{\sigma_x^2}{m_x^2} + 1\right]. \tag{2.110}$$

Прологарифмуємо рівність (2.107)

$$\ln m_x = \frac{\sigma_u^2}{2} + m_u. \tag{2.111}$$

Отже,

$$m_u = \ln m_x - \frac{\sigma_u^2}{2} = \ln m_x - \frac{1}{2}\ln\left\{\frac{\sigma_x^2}{m_x^2} + 1\right\}. \tag{2.112}$$

Таким чином щільність логарифмічно нормального розподілу можна записати у вигляді:

$$f(x)=\frac{1}{x\sqrt{2\pi\ln\left\{\frac{\sigma_x^{\,2}}{m_x^{\,2}}+1\right\}}}\times$$

$$\times e^{-\frac{\left[\ln x-\ln m_x+\frac{1}{2}\ln\left\{\frac{\sigma_x^{\,2}}{m_x^{\,2}}-1\right\}\right]^2}{2\ln\left\{\frac{\sigma_x^{\,2}}{m_x^{\,2}}+1\right\}}}. \qquad (2.113)$$

Логарифмічно нормальним розподілом добре апроксимуються емпіричні розподіли випадкових величин зі значною правосторонньою асиметрією.

Приведемо приклад розрахунків теоретичних частот логарифмічно нормального закону розподілу для ряду швидкості вітру біля земної поверхні.

Статистичні характеристики випадкової величини x ,а саме $\overline{x}$ і $S_x^{\,2}$ приймають відповідно за математичне сподівання m_x і дисперсію $\sigma_x^{\,2}$ і по формулах (2.110) і (2.112) знаходять m_u і $\sigma_u^{\,2}$, які є вихідними характеристиками для обчислювання теоретичних частот.

Таблиця 2.5 - Результати розрахунків теоретичних частот логарифмічно нормального розподілу для ряду швидкості вітру біля земної поверхні

Інтервали $a-b$	$\widetilde{x}_i$	m_i	$x_i'=\frac{\widetilde{x}_i-x_0}{c}$	$x_i'm_i$	$x_i'^2m_i$

0-3,5	1,75	9	-5	-45	225
3,5-7,0	5,25	101	-4	-404	1616
7,0-10,5	8,75	185	-3	-555	1665
10,5-14,0	12,25	112	-2	-224	448
14,0-17,5	15,75	49	-1	-49	49
17,5-21,0	19,25	24	0	0	0
21,0-24,5	22,75	12	1	12	12
24,5-28,0	26,25	5	2	10	20
28,0-31,5	29,75	2	3	6	18
31,5-35,0	33,25	1	4	4	16

Продовження табл.2.5

$\ln a$	$\ln b$	$t_1 = [\ln a - - m_u] / \sigma_u$	$t_2 = [\ln b - - m_u] / \sigma_u$
$-\infty$	1,2528	$-\infty$	-2,28
1,2528	1,9459	-2,28	-0,71
1,9459	2,3514	-0,71	0,21
2,3514	2,6391	0,21	0,87
2,6391	2,8622	0,87	1,37
2,8622	3,0445	1,37	1,79
3,0445	3,1987	1,79	2,14
3,1987	3,3322	2,14	2,44
3,3322	3,4500	2,44	2,71
3,4500	3,5553	2,71	2,95

Продовження табл.2.5

$\frac{1}{2}\Phi(t_1)$	$\frac{1}{2}\Phi(t_2)$	$\frac{1}{2}\Phi(t_2) - \frac{1}{2}\Phi(t_1)$	$\tilde{m}_i$
11.	12.	13.	14.
-0,5000	-0,4885	0,0115	6
-0,4884	-0,2610	0,2275	114
-0,2610	0,0830	0,3440	172

0,0830	0,3080	0,2250	113
0,3080	0,4145	0,1065	53
0,4145	0,4635	0,0490	25
0,4635	0,4840	0,0205	10
0,4840	0,4925	0,0085	4
0,4925	0,4965	0,0040	2
0,4965	0,4985	0,0020	1

$$x_0 = 19{,}25; \qquad \bar{x} = 10{,}5350; \qquad \sigma_u^2 = 0{,}1937;$$
$$m_u = 2.2578; \qquad c = 3.5; \qquad S_x^2 = 23.7392;$$
$$\sigma_u = 0.4402.$$

В табл.2.5 для зручності проводиться перенесення початку координат в точку $x_0 = 19.25$, яка поділяє область значень змінної приблизно на дві рівні частини. У подальшому нові значення змінної $(\widetilde{x}_i - x_0)$ зменшується в c разів, де c – довжина часткового інтервалу. Це приводить до того, що у якості випадкових величин, які моделюються, виступають прості числа (x_i'), що мають порядок 10^0.

З табл. 2.5 випливає, що емпіричні та теоретичні частоти добре співпадають.

2.6 Біномний розподіл.

Розглянемо випадок повторення однієї і тієї ж події при постійних умовах тільки з двома можливими наслідками, при чому у якості елементарних наслідків кожного елементарного випробування ми будемо розрізняти тільки два наслідки: появу деякої події A та не появу її $\overline{A}$ (тобто появу події, протилежної події A). При такому формулюванні задачі $u = A + \overline{A}$.

Будемо вважати, що імовірність появи події для кожного випробування постійна і дорівнює $P(A) = p$, де $0 < p < 1$. Для події $\overline{A}$ будемо мати:

$$p(\overline{A}) = 1 - P(A) = 1 - p = q, \ p + q = 1.$$

У якості події A може виступати день з грозою у літні місяці. Гроза може спостерігатися з імовірністю p, а може не спостерігатися з імовірністю $q = 1 - p.$

Нехай проведено n незалежних випробувань, які ми будемо розглядати як одне складне випробування. Результат кожного випробування ми будемо відзначати, ставлючи літеру A або $\overline{A}$ на відповідному місці. Ясно, що при двох випробуваннях можливі такі 2^2=4 наслідки: $\overline{A}\,\overline{A}$, $\overline{A}A$ $A\overline{A}$, AA (подія A два рази не з'явилася , подія A не з'явилася у першому і з'явилася у другому випробуванні, подія A з'явилася у першому і не з'явилася у другому випробуванні, подія A з'явилася два рази). При трьох випробуваннях можливі такі 2^3 =8 наслідки:

$$\overline{A}\,\overline{A}\,\overline{A}, A\overline{A}\,\overline{A}, AA\overline{A}, A\overline{A}A, \overline{A}AA, AAA, A\overline{A}A, \overline{A}A\overline{A}.$$

Кожному можливому результату n випробувань (це 2^n результатів) буде відповідати послідовність n літерів A і $\overline{A}$, що чергуються у тому порядку, у якому з'являються ці події у n випробуваннях, наприклад $A\overline{A}\,\overline{A}A...\overline{A}.$

Оскільки випробування незалежні, то імовірність кожного такого результату можна знайти шляхом добутку імовірностей подій A і $\overline{A}$ у відповідних випробуваннях. Так, наприклад, для записаного вище результату знайдемо:

$$P(A)P(\overline{A})P(\overline{A})P(A)...P(A) = pqqp...q,$$

оскільки $P(A) = p$ і $P(\overline{A}) = q$.

Ясно, що коли у записаної послідовності літера A зустрічається x разів, літера $\overline{A}$ зустрічається $n - x$ разів, то імовірність такого результату буде $p^x q^{n-x}$, незалежно від того, у якому порядку чергуються ці x літерів A, та $n - x$ літерів $\overline{A}$. Імовірність для можливих восьми наслідків трьох випробувань приводиться в табл.2.6

Таблиця 2.6 - Імовірності восьми наслідків трьох незалежних випробувань

Наслідки	$\overline{A}\overline{A}\overline{A}$	$\overline{A}\overline{A}A$	$\overline{A}A\overline{A}$	$A\overline{A}\overline{A}$
Імовір-ність	$q \cdot q \cdot q = q^3$	$q \cdot q \cdot p = q^2 p$	$q \cdot p \cdot q = q^2 p$	$p \cdot q \cdot q = pq^2$

Продовження табл. 2.6

$AA\overline{A}$	$A\overline{A}A$	$\overline{A}AA$	AAA
$p \cdot p \cdot q = p^2 q$	$p \cdot q \cdot p = p^2 q$	$q \cdot p \cdot p = qp^2$	$p \cdot p \cdot p = p^3$

Непоявлення події A у всіх трьох випробуваннях має імовірність $P_3(0) = q^3$, відповідаючу тільки за один наслідок

$\overline{A}\,\overline{A}\,\overline{A}$. Нехай $p_3(1)$ є імовірністю появи події A тільки один раз протягом трьох випробувань. Це може відбутися, якщо здійсниться який-небудь з трьох варіантів: $\overline{A}\,\overline{A}A$ або $\overline{A}A\overline{A}$, або $A\overline{A}\,\overline{A}$, кожний з котрих має імовірність $q^2 p$. Тому

$$P_3(1) = P(\overline{A}\,\overline{A}A) + P(\overline{A}A\overline{A}) + P(A\overline{A}\,\overline{A}) = \\ = 3q^2 p$$

Аналогічно цьому, з табл. 2.6 знаходимо, що імовірність появи A рівно два рази при трьох випробуваннях дорівнює:

$$P_3(2) = P(AA\overline{A}) + P(A\overline{A}A) + P(\overline{A}AA) = 3p^2 q$$

і , нарешті,

$$P_3(3) = P(AAA) = p^3.$$

Сума

$$P_3(0) + P_3(1) + P_3(2) + P_3(3) = q^3 + 3q^2 p + \\ + 3p^2 q + p^3 = (p+q)^3 = 1$$

Це і треба було чекати, оскільки ми розглядали суму імовірностей подій, утворюючих повну групу. Вірогідно, що

подія A при трьох випробуваннях трапиться або 0, або 1, або 2, або 3 рази.

Якщо ми розглянемо імовірність $P_n(x)$ x разів спостерігати подію A протягом n випробувань, то міркуючи аналогічно попередньому, прийдемо до рівняння:

$$P_n(x) = C_n^x p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x q^{n-x} =$$

$$= \frac{n(n-1)...(n-x+1)}{x!} p^x q^{n-x} \quad , \tag{2.114}$$

де $C_n^x = \frac{n!}{x!(n-x)!}$ - число сполучень із n елементів по x.

Сукупність імовірностей $p_n(x)$ при $x = 0,1,2,...,n$ тобто $p_n(0), p_n(1),..., p_n(n)$, називається *біномним розподілом імовірності*. Оскільки ці імовірності відповідають несумісним подіям, утворюючим повну групу, то

$$\sum_{x=0}^{n} P_n(x) = 1. \tag{2.115}$$

Це легко перевірити, оскільки імовірності $P_n(x)$ відповідно (2.114) утворюють члени бінома $(p+q)^n$, за що вони і отримали свою назву.

Сталі n і p у формулі (2.114) називаються *параметрами біномного закону.*

Розглянемо спочатку зовнішній вигляд біномного закону. Для цього знайдемо відношення $P_n(x)$ до $P_n(x-1)$. Очевидно,

$$\frac{P_n(x)}{P_n(x-1)} = \frac{C_n^x p^x q^{n-x}}{C_n^{x-1} p^{x-1} q^{n-x+1}} = \frac{(n-x+1)}{x}\frac{p}{q} =$$

$$1+\left[\frac{(n-x+1)p}{xq}-1\right] =$$

$$=1+\frac{(n-x+1)p-xq}{xq} = 1+\frac{(n+1)p-x(p+q)}{xq}$$

$$=1+\frac{(n+1)p-x}{xq}$$

(2.116)

У залежності від того, чи буде x меншим, чи більшим від $(n+1)p$, права частина (2.116) буде більша чи менша від одиниці і, у відповідності до цього, $P_n(x) > P_n(x-1)$ або $P_n(x) < P_n(x-1)$. Отже, доки $x < (n+1)p$, кожний член $P_n(x)$ більше попереднього $P_n(x-1)$, тобто послідовність імовірностей збільшується і, навпаки, коли буде $x > (n+1)p$, то послідовність імовірностей буде зменшуватися. Приведемо такий приклад. Нехай відомо, що число днів з грозою влітку відповідає біномному розподілу і в

червні, липні та вересні $(n = 30 + 31 + 31)$ імовірність грози $P(A) = p = \frac{1}{5}$. Тоді

$$(n+1)p = 92 \cdot 0{,}2 = 18.4 \approx 18$$ і тому $x_0 = 18$.

Отже при змінюванні x від 0 до n імовірність $P_x(x)$ (імовірність того, що за цей період гроза буде спостерігатися x днів) спочатку буде збільшуватися, досягаючи максимуму при $x_0 = 18$, потім стане зменшуватися. Отже $x_0 = \mathrm{M}_0$, де M_0 - модальне значення дискретної випадкової величини x, що має біномний розподіл. Як правило, *біномній розподіл має одну моду і відноситься до одномодальних розподілів.* При цьому, найбільшу імовірність можуть мати які-небудь з крайніх значень 0 чи n. Тоді імовірності в ряду $P_n(x)$ будуть весь час зменшуватися або збільшуватися. Постійне зменшування $P_n(x)$ спостерігається при малих p, якщо n не дуже велике. У цьому випадку ціла частина $(n+1)p$ дорівнює нулю. Протилежний випадок постійного збільшення $P_n(x)$ може спостерігатися при p, близьких до одиниці, та невеликих n. При великих n завжди розподіл буде мати моду в центральній частині розподілу. Два значення моди мають місце при $n = 4$ і $p = \frac{2}{5}$, коли $(n+1)p = 2$. У цьому випадку з формули (2.114) випливає:

$$P_n(0) = \frac{81}{625}; \quad P_n(1) = P_n(2) = \frac{81}{625} = \frac{216}{625};$$

$$P_n(3) = \frac{91}{625}; \quad P_n(4) = \frac{16}{625}.$$

При $n = 10$ і при тому ж значенні імовірності $p = \frac{2}{5}$ отримаємо лише одну моду $(n+1)p = \frac{22}{5} = 4{,}4$, тобто $x_0 = 4$ модальне значення x_0 розташовується на інтервалі $np - q \le x_0 \le np + q$.

Початкові моменти для біномного розподілу визначаються формулами :

$$\gamma_0 = 1, \tag{2.117}$$

$$\gamma_1 = np, \tag{2.118}$$

$$\gamma_2 = npq + n^2 p^2, \tag{2.119}$$

$$\gamma_3 = npq(q - p) + 3n^2 p^2 q + n^3 p^3, \tag{2.120}$$

$$\gamma_4 = npq(1 - 6pq) + n^2 p^2 q(7q - 4p) + 6n^3 p^3 q + n^4 p^4. \tag{2.121}$$

Запишемо також формули і для центральних моментів:

$$\mu_0 = 1, \tag{2.122}$$

$$\mu_1 = 0, \tag{2.123}$$

$$\mu_2 = npq, \tag{2.124}$$

$$\mu_3 = npq(q-p), \tag{2.125}$$

$$\mu_4 = npq[3(n-2)pq+1]. \tag{2.126}$$

Від центральних моментів дуже просто перейти до основних моментів біномного розподілу

$$r_0 = 1, \tag{2.127}$$

$$r_1 = 0, \tag{2.128}$$

$$r_2 = 1, \tag{2.129}$$

$$r_3 = \frac{q-p}{\sqrt{npq}}, \tag{2.130}$$

$$r_4 = 3 + \frac{1-6pq}{npq}. \tag{2.131}$$

При $n \to \infty$, маємо $r_3 = 0,\ r_4 = 3$, що відповідає нормальному розподілу.

Якщо np - ціле число, то математичне сподівання і мода збігаються .

Розподіл (2.114), як правило асиметричний за винятком $p = 0{,}5$. При $p < 0{,}5$ асиметрія додатня, при $p > 0{,}5$ - від'ємна.

При збільшенні числа випробувань n форма полігона розподілу наближається до симетричної. Вона стає такою й при крайніх p. Це випливає з формул (2.130), (2.131). Практично графік біномного розподілу можна вважати симетричним при $np \geq 4$. На рис.2.28 зображується графік розподілу при $p = 0{,}05$і $n = 5$. Видно, що при збільшенні n і p графік біномного розподілу стає майже симетричним відносно моди , що спостерігається в точці $x = 3$ (рис.2.29).

При великих значеннях n біномний розподіл з хорошим наближенням можна описати за допомогою нормального розподілу з тим же центром і з тією ж дисперсією, що і у біномного розподілу. В основі переходу від перервного біномного розподілу до неперервного нормального розподілу є теорема Муавра-Лапласа: *якщо відбувається необмежена кількість випробувань, при кожному з яких імовірність появи події* A *дорівнює* p*, то при безмежній кількості випробувань імовірність* $P(n, t_1, t_2)$ *того, що число* m *появи події* A *задовольняє нерівності*

$$t_1\sqrt{npq} < m - np < t_2\sqrt{npq}$$

наближається до границі

$$\frac{1}{\sqrt{2\pi}}\int_{t_1}^{t_2} e^{-t^2/2}\,dt \ ,$$

де t_1 *і* t_2 *- які-небудь два числа* $(t_1 < t_2)$.

Ця важлива теорема дає можливість приблизно розрахувати імовірності подій, зв’язаних з біномним розподілом,

безпосереднє обчислювання яких при великих n викликає великі труднощі із-за потреби знаходження факторіалів великих чисел.

2.6 Розподіл Пуассона

У багатьох випадках, коли мають діло з подіями, які рідко відбуваються, розглядаються дискретні випадкові величини, що підпорядковуються розподілу Пуассона. У метеорологічній практиці такими величинами є число днів з грозою, хуртовиною, ожеледдю, тощо.

Нехай дискретна випадкова величина χ може прийняти значення із зліченої множини цілих чисел (0,1,2,…). Тоді, якщо імовірність того, що $\chi = m$, визначається формулою

$$P(\chi = m) = \frac{\lambda^m}{m!} e^{-\lambda} \quad (m = 0,1,2,...), \qquad (2.132)$$

то говорять, що така випадкова величина має розподіл Пуассона. В ньому $\lambda > 0$ - параметр розподілу Пуассона.

Визначимо математичне сподівання і дисперсію випадкової величини χ.

$$m_\chi = \sum_{m=0}^{\infty} mP(\chi = m) = \sum_{m=0}^{\infty} m \frac{\lambda^m}{m!} e^{-\lambda} =$$
$$= e^{-\lambda} \sum_{m=0}^{\infty} m \frac{\lambda^m}{m!} = e^{-\lambda} \lambda \sum_{m=0} \frac{\lambda^{m-1}}{(m-1)!} \qquad (2.133)$$

Як відомо, для нескінченного ряду

$$\sum_{m=0}^{\infty} \frac{\lambda^{m-1}}{(m-1)!} = e^{\lambda}. \qquad (2.134)$$

Отже,

$$m_x = e^{-\lambda} \lambda e^{\lambda} = \lambda. \qquad (2.135)$$

Постійна величина λ є математичним сподіванням випадкової величини, яка має розподіл Пуассона. Це означає, що розподіл Пуассона однозначно визначається математичним сподіванням випадкової величини.

Для визначення дисперсії використаємо відоме співвідношення

$$\sigma_x^2 = \nu_2 - \nu_1^2. \qquad (2.136)$$

Знайдемо спочатку другий початковий момент

$$\nu_2 = \sum_{m=0}^{\infty} m^2 \frac{\lambda^m}{m!} e^{-\lambda} = \lambda \sum_{m=0}^{\infty} m \frac{\lambda^{m-1}}{(m-1)!} e^{-\lambda} =$$

$$\lambda \sum_{m=1}^{\infty} [(m-1)+1] \frac{\lambda^{m-1}}{(m-1)!} e^{-\lambda} =$$

$$= \lambda \left[\sum_{m=1}^{\infty} (m-1) \frac{\lambda^{m-1}}{(m-1)!} e^{-\lambda} + \sum_{m=1} \frac{\lambda^{m-1}}{(m-1)!} \right] =$$

$$= \lambda \left[\lambda e^{-\lambda} \sum_{m=2}^{\infty} \frac{\lambda^{m-2}}{(m-2)!} + 1 \right] =$$

$$= \lambda(\lambda + 1).$$

(2.137)

Оскільки $\nu_1 = m_x = \lambda$, то

$$\sigma_x^2 = \lambda^2 + \lambda - \lambda^2 = \lambda. \qquad (2.138)$$

Таким чином,

$$\sigma_x^2 = m_x = \lambda. \qquad (2.139)$$

Отже, дисперсія випадкової величини, що має розподіл Пуассона, чисельно дорівнює її математичному сподіванню.

Третій та четвертий центральні моменти розподілу Пуассона визначаються формулами

$$\mu_3 = \lambda\ , \tag{2.140}$$

$$\mu_4 = \lambda + 3\lambda^2. \tag{2.141}$$

Тому для коефіцієнтів асиметрії та ексцесу можна знайти такі формули:

$$A_s = r_3 = \frac{\mu_3}{\sigma_x^3} = \frac{\lambda}{(\sqrt{\lambda})^3} = \frac{1}{\sqrt{\lambda}}, \tag{2.142}$$

$$E = \frac{\mu_4}{\sigma_x^4} - 3 = \frac{\lambda + 3\lambda^2}{\lambda^2} - 3 = \frac{1}{\lambda}. \tag{2.143}$$

У Додатку приводиться значення функції

$$P(\chi = m).$$

В якості прикладу, в табл. 2.7 наводяться результати розрахунків теоретичних частот розподілу Пуассона на основі статистичного ряду числа днів з опадами ≥ 10 мм.

Таблиця 2.7 - Результати розрахунків теоретичних частот розподілу Пуассона на основі статистичного ряду числа днів з опадами ≥ 10 мм у листопаді, Одеса.

$$\widetilde{m}_i = nP(x)$$

x_i	m_i	$x_i m_i$	$P(x)$	$\widetilde{m}_i$
0	16	0	0,27	10,8
1	10	10	0,36	14,4
2	8	16	0,23	9,2
3	1	3	0,10	4,0
4	3	12	0,03	1,2
5	2	10	0,01	0,4
$\sum$	**40**	**51**	**1,00**	**40**

По даних табл.2.7 отримаємо, що $\overline{\chi} \cong 1{,}3$, а $S_x^2 \cong 2{,}0$, тобто оцінка середнього числа днів з опадами ≥ 10 мм і дисперсія цієї величини мають близькі значення, що дає підставу вибрати для апроксимації цієї кліматичної характеристики розподіл Пуассона.

3 КОРРЕЛЯЦІЙНИЙ ЗВ’ЯЗОК МІЖ ДВОМА ВИПАДКОВИМИ ВЕЛИЧИНАМ.

3.1 Функціональні, стахостичні та корреляційні залежності між випадковими величинами

Зв’язки між різними явищами у природі складні та різноманітні. Однак їх можна відповідним чином класифікувати. Будемо далі вести мову про залежність між випадковими величинами, які підлягали вивченню і раніше.

Є всі підстави розглядати гідрометеорологічні величини як випадкові величини, незважаючи на те, що їх змінення обумовлені певними фізичними причинами. Справа у тому, що зв’язки у гідрометеорологічних процесах виявляються залежними від багатьох факторів і у нас нема, як правило, достатніх знань про те, який саме зв’язок, або які саме зв’язки, обумовили змінення цієї величини в той чи інший час.Тим більше, що у багатьох випадках у нас ще нема повних уявлень про причинно-наслідкові залежності у природних явищах, які спричиняють гідрометеорологічні процеси. Та чи інша міра невизначеності завжди має місце, коли йдеться про майбутній стан атмосфери чи об’єктів гідросфери, тобто про гідрометеорологічне прогнозування .

Існують два погляди відносно цієї невизначеності. По-перше, вважають, що вона пов’язана, як вже мовилося, з неповнотою наших знань про закони еволюції процесів в атмосфері і гідросфері, або з похибками чи нечисленністю вихідних вимірювань гідрометеорологічних величин. Але, по-друге, існує думка й про те, що невизначеність майбутнього стану є внутрішньою властивістю атмосфери і океану та зв’язана з незалежністю розвитку деяких процесів від початкових умов. Це приводить до принципового обмеження передбаченості. Інакше кажучи, стохастичність гідрометеорологічних процесів є внутрішньою властивістю атмосфери та гідросфери.

У багатьох випадках, коли йдеться про статистичну обробку результатів гідрометеорологічних спостережень, ми взагалі абстрагуємося від розглядання будь-яких фізичних закономірностей, а приймаємо сукупність гідрометеорологічних величин у якості вибірки випадкової величини із деякої генеральної сукупності.

У цьому розділі ми обмежимося розгляданням залежності тільки між двома випадковими величинами. Ці залежності можуть бути функціональними та стохастичними.

Функціональна залежність між різними фізичними величинами вивчається фізикою. У математиці вона формалізується шляхом впровадження поняття функції. Але якщо у функціональний залежності $y = f(x)$ аргумент x є випадкова величина, то випадковою величиною є і залежна змінна y. Отже, *функціональною залежністю* між двома випадковими величинами називається така залежність, коли можливому значенню однієї випадковою величини відповідає тільки одне значення другої.

Між випадковими величинами може існувати зв'язок і іншого роду. Він виявляється у тому, що одна з них реагує на змінення іншої зміненням свого закону розподілу. Такий *зв'язок* називається *стохастичним.* Стохастичний зв'язок між двома випадковими величинами спостерігається, наприклад, коли існують загальні випадкові фактори, впливаючі як на одну, так і на другу величину поряд з іншими неоднаковими для обох величин випадковими факторами. Наприклад, якщо у є деяка функція від випадкових величин $z_1, z_2, \ldots, z_m, \upsilon_1, \upsilon_2, \ldots, \upsilon_k$

$$y = f(z_1, z_2, \ldots, z_m, \upsilon_1, \upsilon_2, \ldots, \upsilon_k),$$

а x - функція від тих же випадкових величин $z_1, z_2, \ldots, z_m$ і деякої сукупності інших випадкових величин $u_1, u_2, \ldots, u_l$

$$x = \varphi(z_1, z_2, ..., z_m, u_1, u_2, ..., u_k)$$

то залежність між випадковими величинами X і Y буде стохастичною.

Найбільш важливі особливості стохастичного зв'язку виявляються у тих змінеснях, які зазнає центр *умовного розподілу* однієї величини при зміненні другої. Якщо припустити, що умовний розподіл є нормальним, то центром його є *умовне математичне сподівання* $m_{y/x}$. Отже розглядається залежність умовного математичного сподівання однієї випадкової величини від другої. Таку *залежність* називають *корреляційною (статистичною)* залежністю між двома випадковими величинами.

Будемо, як і раніше, позначати випадкові величини великими латинськими літерами $X, Y, Z, ...$, а конкретні їх значення - малими літерами $x, y, z, ...$. Отже, *корреляційну залежність* між двома випадковими величинами можна визначити як функціональну залежність умовного математичного сподівання однієї з них від значення другої, тобто

$$M\left[Y/X = x\right] = m_{Y/X} = f(x) \qquad (3.1)$$

Функцію $f(x)$ називають *функцією регресії* випадкової величини Y по (на) X. Рівняння (3.1) називається рівнянням регресії.

Розглянемо два приклади.

Приклад 1. Нехай двовимірна випадкова величина (X,Y) визначена на прямокутнику $0 \le X \le 1;\ 0 \le Y \le 2$ і має сумісну щільність ймовірності

$$P(x,y) = \frac{1}{x} + \frac{y}{3}. \qquad (3.2)$$

Треба знайти функцію регресії змінної Y для випадку, коли $X = x$.Очевидно

$$m_{y/x} = \int_0^2 yP\left(Y/X = x\right)dy, \qquad (3.3)$$

де $P\left(Y/X = x\right)$ - є *умовна щільність імовірності випадкової величини* Y..

Відомо, що

$$P\left(Y/X = x\right) = \frac{P(x,y)}{P(x)}. \qquad (3.4)$$

У рівності (3.4) $P(x)$ - *безумовний розподіл випадкової величини* X, котрий визначається формулою:

$$P(x)=\int_0^2 P(x,y)dy\ . \qquad (3.5)$$

Отже,

$$P(x)=\int_0^2\left(\frac{1}{x}+\frac{y}{3}\right)dy=\frac{2}{x}+\frac{2}{3}\ . \qquad (3.6)$$

Таким чином,

$$P\left(Y/X=x\right)=\frac{\frac{1}{x}+\frac{y}{3}}{\frac{2}{x}+\frac{2}{3}}=\frac{y+3}{2x+6}. \qquad (3.7)$$

Знайдемо умовний розподіл випадкової величини Y при двох значеннях $X:\ x=\frac{1}{2}$ і $x=1$, які належать до області значень цієї випадкової величини.

При $x=\frac{1}{2}$,

$$P\left(\left.Y\right/X=\frac{1}{2}\right)=\frac{y+3}{7}, \tag{3.8}$$

а при $x=1$,

$$P\left(\left.Y\right/X=1\right)=\frac{y+3}{8}\ . \tag{3.9}$$

Як свідчать рівності (3.8) і (3.9), змінення значення випадкової величини X приводить до змінення закону розподілу випадкової величини Y.Отже, між випадковими величинами X і Y існує стохастична залежність. Знайдемо тепер *рівняння регресії*, яке, як було зазначено вище, є функціональною залежністю умовного математичного сподівання випадкової величини Y від значення величини X. Для цього використуємо рівняння (3.3) при умові (3.7) :

$$\begin{aligned} m_{\left.y\right/x} &= \int_0^2 y\frac{y+3}{2x+6}dy = \\ &= \frac{1}{2x+6}\int_0^2 (y^2+3y)dy = \frac{26}{3\left(2x+6\right)} \end{aligned}\ . \tag{3.10}$$

Отже, рівняння регресії Y по X визначається функціональною залежністю

$$m_{y/x} = \frac{26}{3(2x+6)}. \qquad (3.11)$$

В області значень випадкової величини X графік цієї функції являє собою відрізок гіперболи. Він зображається на рис.3.1.

Приклад 2. Нехай в результаті експерименту отримані такі значення випадкових величин X і Y:

при $x = 2; y: 4;5;6,$

при $x = 3; y: 10;6;4;8,$

при $x = 5; y: 12;16;9;7.$

Знайдемо *оцінки умовного математичного сподівання* , тобто *умовні середні значення випадкової величини* Y . Очевидно,

$$\bar{y}_{X=2} = 5; \quad \bar{y}_{X=3} = 7; \quad \bar{y}_{X=5} = 11.$$

Графік цієї залежності $\bar{y}_{X} = f(x)$ зображений на рис 3.2

Легко переконатися, що ми отримали рівняння прямої

$$\bar{y}_{/X} = 2x + 1 \qquad (3.12)$$

Отже, випадкові величини X і Y зв'язані корреляційною залежністю, яка відбивається рівнянням регресії (3.12).

Останній приклад можна узагальнити. Нехай корреляційна залежність характеризується рівнянням :

$$m_{y/x} = \alpha x + \beta \ , \qquad (3.13)$$

а

$$P\left(Y/X = x\right) = \frac{1}{\sqrt{2\pi}\sigma_{y/x}} \exp\left[- \frac{\left(y - m_{y/x} \right)^2}{2\sigma^2_{y/x}} \right]$$

(3.14)

Це означає, що випадкова величина Y має *умовний нормальний розподіл* з *умовною дисперсією* $\sigma^2_{y/x}$. Ці факти графічно відображаються на рис 3.3.

Як видно з рис 3.3, зі зміненням значення випадкової величини X змінюються на осі OY положення центрів розподілів випадкової величини Y, тобто умовне математичне сподівання, оскільки розподіл є нормальним. Це означає, по-перше, що ці випадкові величини зв'язані корреляційною

залежністю і, по-друге, те, що при зміненні значень випадкової величини X змінюються параметри закону розподілу (3.14) (як видно на графіку, розмах кривої і, таким чином і умовна дисперсія теж міняються). Тобто змінюється закон розподілу випадкової величини Y. А це означає, що випадкові величини зв'язані стохастичною залежністю. Отже, *корреляційна залежність є частинним випадком стохастичної залежності* .

3.2 Тіснота та форма корреляційного зв'язку

Незважаючи на важливість поняття функції регресії, можливості її практичного вживання дуже обмежені. Як видно з прикладу 1, для оцінки функції регресії потрібно знати аналітичний вид двовимірного розподілу (X, Y). Тільки знаючи вид цього закону розподілу, можна точно визначити форму функції регресії та її параметри. Однак, ми найчастіше маємо лише вибірку обмеженого об'єму, на основі якої треба знайти вид двовимірного розподілу (X, Y), а потім вид функції регресії. Це може привести до значних помилок, оскільки одну і ту ж сукупність точок (x_i, y_i) на площині можна з однаковою мірою успішності описати за допомогою різних функцій розподілу.

Для характеристики форми зв'язку при вивченні корреляційної залежності використовують поняття лінії регресії.

Лінію регресії Y по X (або Y на X) називають умовне середнє значення випадкової змінної Y, яка розглядається як функція від X, тобто

$$\bar{y}(x) = f(x).$$

Саме така лінія регресії була побудована у прикладі 2 попереднього розділу.

Аналогічно, умовне середнє значення випадкової величини X, тобто $\bar{x}(y)$, що розглядається як функція y, називається *лінією регресії X по Y*.

Виникає питання, чому для визначення лінії регресії користуються саме умовним середнім $\bar{y}(x)$? Функція $\bar{y}(x)$ має одну надзвичайну властивість: вона дає найменшу середню похибку оцінки прогнозу.

Припустимо, що лінія регресії є довільною функцією. Середня похибка прогнозу по кривій регресії визначається математичним сподіванням квадрата різниці між експериментальною величиною та величиною, розрахованою по рівнянню лінійної регресії, тобто $M[Y - f(x)]^2$. Звичайною є вимога знайти таку лінію регресії, середня похибка прогнозу на основі якої була б найменшою. Такою й є $f(x) = \bar{y}(x)$. Це випливає з властивості мінімальності розсіювання навколо центра розподілу $\bar{y}(x)$. Якщо розсіювання визначається відносно $f(x) \neq \bar{y}(x)$, то середній квадрат відхилу збільшується. Тому можна сказати, що *лінія регресії $\bar{y}(x)$ мінімізує середню квадратичну похибку прогнозу величин Y по X*.

Визначення корреляційної залежності було сформульовано безвідносно до сумісного закону розподілу випадкових величин (X, Y), тобто корреляційну залежність можна досліджувати при будь-якому законі розподілу (X, Y). Однак у теорії корреляції важливе місце займає двовимірний нормальний закон розподілу випадкових змінних (X, Y) (дивитись розділ 3.7). У цьому випадку умовні середні $\bar{y}(x)$ і $\bar{x}(y)$ є оцінками центрів умовного розподілів

$M(Y/X=x)$ і $M(X/Y=y)$ випадкових величин Y і X, що випливає із властивостей нормального закону розподілу. Отже, крива регресії є оцінкою функції регресії, і, таким чином, середня похибка прогнозу по рівнянню регресії є мінімальною.

Пару випадкових чисел $(x;y)$ можна зобразити графічно як точку з координатами $(x;y)$. Таким же чином можна зобразити весь набір пар випадкових чисел, тобто всю вибірку двох випадкових величин. Таке предоставлення корреляційної залежності називають *корреляційним графіком, або полем корреляції.*

По розташуванню точок на ньому можна зробити попередні висновки про форму і тісноту корреляційного зв'язку. По-перше, якщо зі збільшенням x випадкова величина y зростає, то ми маємо *прямий зв'язок* між ними. У протилежному випадку корреляційний *зв'язок* є *зворотнім*. По-друге, точки на корреляційному графіку можуть грунтуватися біля деякої середньої прямої лінії. У такому випадку корреляційний *зв'язок є лінійним і функція регресії має вид рівняння прямої*:

$$\bar{y}(x)=ax+b\ . \qquad (3.15)$$

Коефіцієнти a і b є оцінками параметрів α і β генерального рівняння регресії

$$m_{y/x}=\alpha x+\beta \qquad (3.16)$$

Крім лінійної зустрічаються *нелінійні корреляційні зв'язки*, які відбиваються відповідними рівняннями регресії. Серед них для гідрометеорологічних випадкових величин часто реалізуються:

параболічне рівняння регресії:

$$\bar{y}(x) = a_0 + a_1 x + a_2 x^2 \; ; \qquad (3.17)$$

кубічне рівняння:

$$\bar{y}(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 \; ; \qquad (3.18)$$

показникове рівняння регресії:

$$\bar{y}(x) = a b^{cx} \; ; \qquad (3.19)$$

частинним випадком якого є *експоненціальне рівняння регресії:*

$$\bar{y}(x) = a e^{bx} \; ; \qquad (3.20)$$

гіперболічне рівняння регресії:

$$\bar{y}(x) = \frac{a}{x^b} \; . \qquad (3.21)$$

По-третє, по розкиду точок на корреляційному графіку можна зробити попередній висновок про тісноту корреляційного зв'язку. Звичайно, він є більш тісним, коли точки тісніше групуються навколо лінії регресії, і навпаки.

На рис.3.4 зображаються корреляційні графіки при різних характерах корреляційного зв'язку.

3.3 Корреляційне відношення

Розкид точок на корреляційному графіку може дати лише якісне уявлення про тісноту корреляційного зв'язку. Але у більшості дослідницьких задач треба мати кількісну міру зв'язку.

Характер розкиду точок (x, y) на корреляційному полі визначається не тільки впливом випадкової величини X на випадкову величину Y, а і впливом на неї й інших випадкових величин. Якщо на величину Y чинить вплив тільки величина X, то всі точки будуть розміщуватися на лінії регресії. Чим більший вплив інших випадкових величин на випадкову величину Y, тим більшим виявляється розкид точок на корреляційному графіку. Розглянемо, яким показником можна вимірювати тісноту корреляційного зв'язку.

Основною характеристикою є *загальний показник мінливості - повна дисперсія* σ_y^2. Умовимося під повною дисперсією розуміти дисперсію випадкової величини Y відносно умовного математичного сподівання $m_{Y/X}$. Повну дисперсію можна розкласти на дві складові, кожна з яких характеризує дію окремих факторів. Одна з них характеризує вплив фактора X на Y, друга - вплив інших факторів. Очевидно, чим менший вплив інших факторів, тим тіснішим є зв'язок між X і Y. Отже, повну дисперсію можна виразити таким чином:

$$\sigma_y^2 = M\{\left(Y - m_{Y/X}\right)^2\} = M\{(Y - \bar{y}(x) +$$
$$\bar{y}(x) - m_{Y/X})^2\} = M\{[Y - \bar{y}(x)]^2 + [\bar{y}(x) -$$
$$- m_{Y/X}]^2 + 2[(Y - \bar{y}(x))(\bar{y}(x) - m_{Y/X})]\} =$$
$$= M\{[Y - \bar{y}(x)]^2\} + M\{[\bar{y}(x) - m_{Y/X}]^2 +$$
$$+ 2M\{[Y - \bar{y}(x)][\bar{y}(x) - m_{Y/X}]\}.$$

Розглянемо останній член отриманої рівності. Очевидно,

$$M\{[Y - \bar{y}(x)][\bar{y}(x) - m_{Y/X}]\} =$$
$$= M\{Y \cdot \bar{y}(x) - Y \cdot m_{Y/X} - [\bar{y}(x)]^2 +$$
$$+ \bar{y}(x) \cdot m_{Y/X}\} = [\bar{y}(x)]^2 - \bar{y}(x) \cdot m_{Y/X} -$$
$$- [\bar{y}(x)]^2 + \bar{y}(x) \cdot m_{Y/X} = 0.$$

Отже

$$\sigma_y^2 = M\{[Y - \bar{y}(x)]^2\} + M\{[\bar{y}(x) - \\ - m_{Y/X}]^2\} \quad . \qquad (3.22)$$

Член

$$\sigma^2_{y/x} = M\{[Y - \bar{y}(x)]^2\} \qquad (3.23)$$

характеризує розсіювання точок на корреляційному полі відносно лінії регресії. Тому дисперсія (3.23) описує вплив інших діючих факторів на випадкову величину Y. Другий член

$$\delta^2_{y/x} = M\{[\bar{y}(x) - m_{Y/X}]^2\} \qquad (3.24)$$

має смисл міри розсіювання вибіркових ліній регресії відносно генеральної лінії регресії і, таким чином, характеризує вплив випадкової величини X на Y.

Підставимо формули (3.23) і (3.24) до рівняння (3.22) і поділимо його на повну дисперсію. Будемо мати:

$$\frac{\sigma^2_{y/x}}{\sigma^2_y} + \frac{\delta^2_{y/x}}{\sigma^2_y} = 1\,. \qquad (3.25)$$

Величину

$$\eta^2_{y/x} = \frac{\delta^2_{y/x}}{\sigma^2_y}\,, \qquad (3.26)$$

яка має смисл відносного внеску випадкової величини X у повне розсіювання випадкової величини Y, тобто є мірою впливу випадкової величини X на Y, або мірою корреляційного зв'язку між ними, називають *корреляційним відношенням*. Очевидно,

$$\eta^2_{y/x} = 1 - \frac{\sigma^2_{y/x}}{\sigma^2_y} \ . \qquad (3.27)$$

Дисперсія $\sigma^2_{y/x}$ є частиною повної дисперсії σ^2_y, а тому $\sigma^2_{y/x} \le \sigma^2_y$.

Тоді, як свідчить рівність (3.27),

$$0 \le \eta^2_{y/x} \le 1 \ . \qquad (3.28)$$

Корреляційне відношення дорівнює нулю тоді, коли на випадкову величину Y величина X не впливає, а діють на неї тільки інші випадкові величини. У цьому разі $\sigma^2_{y/x} = \sigma^2_y$. Навпаки, $\eta^2_{y/x} = 1$, коли $\sigma^2_{y/x} = 0$, тобто коли на величину Y інші випадкові величини, окрім X, не чинять впливу. У цьому випадку між Y і X існує функціональна залежність.

Відсутність корреляційного зв'язку не можна ототожнювати з незалежністю випадкових величин. Якщо $\eta^2_{y/x} = 0$, то змінна Y може бути залежною від X, але так,

що центри умовних розподілів не змінюються при зміненні X, а змінюються лише умовні дисперсії.

Замість величини $\eta^2_{y/x}$ використовується величина

$$\eta_{y/x} = \sqrt{\eta^2_{y/x}}\ , \tag{3.29}$$

Границі її змінення такі ж, як і у $\eta^2_{y/x}$, тобто

$$0 \le \eta^2_{y/x} \le 1\ . \tag{3.30}$$

Звичайно, на основі вибірок ми знаходимо оцінки розглянутих дисперсій

$$S^2_y = S^2_{y/x} + \hat{\delta}^2_{y/x}, \tag{3.31}$$

де

$$S^2_{y/x} = \frac{1}{n-1}\sum_{i=1}^{n}[y_i - \bar{y}(x_i)]^2 \tag{3.32}$$

називається *внутрішньогруповою дисперсією* і характеризує відхил ординат точок від умовної середньої i -того інтервалу, а

$$\hat{\delta}^2_{y/x} = \frac{1}{k}\sum_{j=1}^{k}[\bar{y}_j(x) - \bar{y}]^2 \qquad (3.33)$$

міжгрупова дисперсія характеризує відхил групових умовних середніх від загального значення випадкової величини Y. У формулах (3.32) і (3.33) $\bar{y}(x)$ є *емпіричною лінією регресії*.

Емпіричним корреляційним відношенням є величина $\hat{\eta}^2_{y/x}$, що розраховується за формулою:

$$\hat{\eta}^2_{y/x} = 1 - \frac{S^2_{y/x}}{S^2_y}. \qquad (3.34)$$

З формули (3.34) випливає, що

$$0 \le \hat{\eta}^2_{y/x} \le 1. \qquad (3.35)$$

При $\hat{\eta}^2_{y/x} = 0$ мінливість середніх $\bar{y}_i(x)$ є відсутньою, і лінія регресії паралельна до осі абсцис, що означає відсутність корреляційного зв'язку між випадковими величинами X і Y. При $\hat{\eta}^2_{y/x} = 1$ всі точки корреляційного поля розташовуються на лінії регресії. Це свідчить про наявність функціональної залежності між випадковими величинами.

Треба мати на увазі, що емпіричне корреляційне відношення у визначній мірі завищує тісноту зв'язку і тим більше, чим менше число спостережень.

3.4 Рівняння регресії між двома випадковими величинами

3.4.1 Метод найменших квадратів

Рівняння регресії, що розглядалися у попередньому розділі, являють собою не що інше, *як моделі процесу взаємозв'язку між випадковими величинами* Y *і* X. Звичайно, модель повинна бути адекватною процесу, що моделюється.

Можуть бути різні ступені адекватності моделі процесу, що досліджується. Тому при моделюванні треба визначити кількісну міру адекватності. Таку *кількісну міру називають критерієм якості або функцією цілі.*

Критерій якості вибирають у залежності від характеру задачі, яка ставиться. При побудові статистичних моделей найбільш часто використовують такий критерій:

$$\Delta^2 = \sum_{i=1}^{n} [y_i - f(x_i)]^2 = \min, \qquad (3.36)$$

де y_i - ординати точок корреляційного поля, $f(x_i)$ - ординати точок на лінії регресії при значенні незалежної змінної x_i (рис.3.5). Смисл критерію Δ^2 полягає у тому, що для опису взаємозв'язку між випадковими величинами вибирають такі параметри рівняння регресії $\bar{y}(x)$, які дають мінімум суми квадратів різниць між ординатами експериментальних точок і точок лінії регресії при відповідних значеннях змінної x. Тому метод, основою якого є критерій якості у формі (3.36), носить назву *метода найменших квадратів.*

Рівняння (3.36) є функцією параметрів моделі, які підлягають визначенню. Як відомо, умовний екстремум функції декількох змінних визначається шляхом зрівнювання до нуля частинних похідних цієї функції по незалежних змінних. Нехай, наприклад,

$$\bar{y}(x) = f(a, b, x), \tag{3.37}$$

тобто функція регресії має два невідомих параметра a і b. Тоді необхідно записати рівняння, що визначають умови екстремуму

$$\frac{\partial \Delta^2}{\partial a} = 0; \quad \frac{\partial \Delta^2}{\partial b} = 0 . \tag{3.38}$$

Можна показати, що визначені за допомогою рівнянь (3.38) параметри a і b надають дійсно мінімум критерія якості Δ^2. Як буде показано далі, операції (3.38) приводять до системи алгебраїчних рівнянь, розв'язок якої дає шукані значення параметрів моделей. Ці *алгебраїчні рівняння називають нормальними*, а відповідну систему *системою нормальних рівнянь*.

Параметри вибраної регресійної моделі знаходять на основі вибірок випадкових величин X і Y. Тому вони є оцінками параметрів генерального рівняння регресії. Якщо вони отримані за методом найменших квадратів, то їх називають оцінками метода найменших квадратів.

3.4.2 Оцінювання параметрів лінійного рівняння регресії

Як відомо, лінійне рівняння регресії має вид:

$$\bar{y}(x) = ax + b, \tag{3.39}$$

де a і b - коефіцієнти регресії, які треба визначити на основі вибірок випадкових величин X і Y. Для цього використаємо метод найменших квадратів. Відповідний критерій якості має таку форму:

$$\Delta^2 = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2 = \min. \tag{3.40}$$

Для отримання системи нормальних рівнянь прирівняємо до нуля частинні похідні по a і b від критерію Δ^2:

$$\frac{\partial \Delta^2}{\partial a} = -2\sum_{i=1}^{n} [y_i - ax_i - b]x_i = 0 ; \tag{3.41}$$

$$\frac{\partial \Delta^2}{\partial a} = -2\sum_{i=1}^{n} [y_i - ax_i - b] = 0 . \tag{3.42}$$

Після простих перетворень ми приходимо до системи алгебраїчних рівнянь

$$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i\, y_i\,, \tag{3.43}$$

$$a\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i\,, \tag{3.44}$$

з них отримаємо, що a дорівнює:

$$a = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}\,. \tag{3.45}$$

Індекси сумування в формулі (3.45) тимчасово опущені.

Поділимо чисельник і знаменник рівняння (3.45) на n^2. Будемо мати :

$$a = \frac{\frac{1}{n}\sum x_i y_i - \frac{1}{n}\sum x_i \frac{1}{n}\sum y_i}{\frac{1}{n}\sum x_i^2 - \left(\frac{\sum x_i}{n}\right)^2}$$

або

$$a = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\overline{x^2} - \left(\bar{x}\right)^2}\,. \tag{3.46}$$

Виконуючи ряд перетворень приходимо до рівнянь:

$$\overline{xy} - \bar{x}\bar{y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = K_{xy}, \qquad (3.47)$$

$$\overline{x^2} - (\bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = S_x^2. \qquad (3.48)$$

K_{xy} - *носить назву коваріацій випадкових величин* X і Y. Відомо, що

$$K_{xy} = r_{xy} S_x S_y, \qquad (3.49)$$

де r_{xy} - *статистичний параметр, який має назву коефіцієнта корреляції*. Ураховуючи рівності (3.47) - (3.49), приходимо до такого виразу для кутового коефіцієнта рівняння регресії:

$$a = r_{xy}\frac{S_y}{S_x}. \qquad (3.50)$$

Перейдемо тепер до обчислювання вільного члена рівняння регресії (3.39). Аналогічно, для коефіцієнта регресії b маємо:

$$b = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} . \qquad (3.51)$$

Формулу (3.51) можна переписати таким чином:

$$b = \frac{n\sum x_i^2 \frac{\sum y_i}{n} - \frac{\sum x_i}{n} n\sum x_i y_i + \frac{1}{n}\left(\sum x_i\right)^2 \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} -$$

$$- \frac{\frac{1}{n}\left(\sum x_i\right)^2 \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \bar{y}\frac{n\sum x_i^2 - \left(\sum x_i\right)^2}{n\sum x_i^2 - \left(\sum x_i\right)^2} - \qquad ,$$

$$- \bar{x}\frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

або , після скорочення і врахування формули (3.45)

$$b = \bar{y} - a\bar{x} . \qquad (3.52)$$

Рівняння (3.50) і (3.52) є оцінками параметрів лінійної регресійної моделі метода найменших квадратів. Вони утримують важливий статистичний параметр - коефіцієнт корреляції r_{xy}. Розглянемо його.

3.5 Коефіцієнт корреляції як міра тісноти лінійного

корреляційного зв'язку

Як було показано в розділі 3.3, мірою тісноти корреляційного зв'язку між випадковими величинами X і Y є корреляційне відношення:

$$\eta^2_{y/x} = \frac{M[\bar{y}(x) - m_y]^2}{\sigma_y^2} = \frac{\sum_{i=1}^{n}[\bar{y}(x_i) - \bar{y}]^2}{n\sigma_y^2}. \quad (3.53)$$

Замість $\bar{y}(x)$ і $\bar{y}$ підставимо в (3.53) їх значення з рівнянь (3.39) і (3.52). Будемо мати:

$$\eta^2_{y/x} = \frac{\sum_{i=1}^{n}[ax_i + b - a\bar{x} - b]^2}{n\sigma_y^2} =$$

$$= \frac{\sum_{i=1}^{n} a^2 (x_i - \bar{x})^2}{n\sigma_y^2} = \frac{a^2}{\sigma_y^2}\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = a^2 \frac{\sigma_x^2}{\sigma_y^2}.$$

(3.54)

Підставимо тепер в (3.54) замість a його значення із рівності (3.50). Отримаємо:

$$\eta^2_{y/x} = r^2_{xy} \,. \tag{3.55}$$

Отже, коефіцієнт *корреляції є мірою тісноти лінійного корреляційного зв'язку*. Оскільки $\eta^2_{y/x} \le 1$, то

$$\left| r_{xy} \right| \le 1 \tag{3.56}$$

або

$$-1 \le r_{xy} \le 1 \,. \tag{3.57}$$

При $r_{xy} = 0$, лінійний корреляційний зв'язок відсутній, тобто *величина* X *і* Y *- некоррельовані* ; при $0 < r_{xy} < 1$ він є *прямим*, при $-1 < r_{xy} < 0$ - *зворотнім*.

Як випливає з формули (3.49),

$$r_{xy} = \frac{K_{xy}}{S_x S_y}, \tag{3.53}$$

або, якщо врахувати рівняння (3.47),

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{nS_x S_y} \; . \qquad (3.54)$$

Коефіцієнт корреляції характеризує ступінь наближення корреляційного зв'язку між випадковими величинами X *і* Y *до лінійної функціональної залежності.*

Зв'язок між випадковими величинами тим тісніший, чим більшим по абсолютній величині є коефіцієнт корреляції.

3.6 Оцінювання коефіцієнтів нелінійних рівнянь регресії

Як було зазначено вище, до нелінійних рівнянь регресії відносяться показникове, гіперболічне, та параболічне рівняння.

До *показникового рівняння регресії* відноситься рівняння:

$$\bar{y}(x) = ab^{cx} \; . \qquad (3.55)$$

Його частинний випадок є *експоненціальне рівняння регресії*

$$\bar{y}(x) = ae^{bx} \; . \qquad (3.56)$$

Обчислення коефіцієнтів цих рівнянь здійснюється на основі метода найменших квадратів. Але зручно застосувати відповідне нелінійне перетворення, яке привело б нелінійні

рівняння регресії до лінійних. Покажемо це на прикладі експоненціального рівняння регресії (3.56).

Якщо метрика

$$\Delta^2 = \sum_{i=1}^{n} [y_i - \bar{y}(x_i)]^2 = \min, \qquad (3.57)$$

то, очевидно, досягає мінімуму і метрика

$$\Delta^2 = \sum_{i=1}^{n} [\ln y_i - \ln \bar{y}(x_i)]^2 = \min. \qquad (3.58)$$

Стосовно до рівняння (3.56), метрика (3.58) має вид

$$\Delta^2 = \sum_{i=1}^{n} [\ln y_i - \ln a - bx_i]^2 = \min . \qquad (3.59)$$

Позначимо $z_i = \ln y_i$; $c = \ln a$. Тоді будемо мати:

$$\Delta^2 = \sum_{i=1}^{n} [z_i - bx_i - c]^2 = \min, \qquad (3.60)$$

тобто такий же критерій якості, як і для лінійного рівняння регресії.

Тоді коефіцієнти регресії розраховуються по формулах:

$$b = r_{zx} \frac{S_z}{S_x}, \tag{3.61}$$

$$c = \bar{z} - b\bar{x} \tag{3.62}$$

або

$$b = r_{\ln y, x} \frac{S_{\ln y}}{S_x}, \tag{3.63}$$

$$c = \overline{\ln y} - r_{x, \ln y} \frac{S_{\ln y}}{S_x} \bar{x}, \tag{3.64}$$

$$a = e^c \ . \tag{3.65}$$

Аналогічно, логарифмічне перетворення приводить *гіперболічне рівняння регресії*

$$\bar{y}(x) = \frac{a}{x^b} \tag{3.66}$$

до лінійного

$$\ln \bar{y}(x) = \ln a - b \ln x, \tag{3.67}$$

а критерій якості для нього до форми

$$\Delta^2 = \sum_{i=1}^{n} [z_i - c + bu_i]^2 = \min, \tag{3.68}$$

де $z_i = \ln y_i$; $c = \ln a$; $u_i = \ln x_i$.

Зупинимося тепер на *параболічному рівнянні*

$$\bar{y}(x) = a_0 + a_1 x + a_2 x^2. \tag{3.69}$$

Відповідно до метода найменших квадратів критерій якості такої моделі дорівнює

$$\Delta^2 = \sum_{i=1}^{n} [y_i - (a_0 + a_1 x_i + a_2 x_i^2)]^2 = \min . \tag{3.70}$$

Систему нормальних рівнянь ми одержимо за допомогою операцій

$$\frac{\partial \Delta^2}{\partial a_0} = 0; \quad \frac{\partial \Delta^2}{\partial a_1} = 0; \quad \frac{\partial \Delta^2}{\partial a_2} = 0, \tag{3.71}$$

які дають

$$\begin{cases} a_0 n + a_1 \sum_{i=1}^{n} x_i + a_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i \\ a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 + a_2 \sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} y_i x_i \\ a_0 \sum_{i=1}^{n} x_i^2 + a_1 \sum_{i=1}^{n} x_i^3 + a_2 \sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} y_i x_i^2 \end{cases} \qquad (3.72)$$

Розв'язок цієї системи лінійних неоднорідних алгебраїчних рівнянь дає шукані коефіцієнти регресії a_0, a_1 і a_2.

Аналогічним чином знаходяться і коефіцієнти кубічного рівняння регресії.

3.7 Двовимірний нормальний розподіл системи двох випадкових величин

Система двовимірних випадкових величин X і Y характеризується функцією щільності сумісного розподілу $f(x, y)$. Однієї з них є щільність двовимірного розподілу. Щільність нормального розподілу двох залежних неперервних випадкових величин має вид:

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{xy}^2}}\times$$

$$\times\exp\left\{-\frac{1}{2\left(1-\rho_{xy}^2\right)}\left[\begin{array}{l}\dfrac{(x-m_x)^2}{\sigma_x^2}-2\rho_{xy}\times\\ \times\dfrac{(x-m_x)(y-m_y)}{\sigma_x\sigma_y}+\\ +\dfrac{(y-m_y)^2}{\sigma_y^2}\end{array}\right]\right\} \tag{3.73}$$

Розглянемо сенс параметрів $m_x,\sigma_x^2,m_y,\sigma_y^2,\rho_{xy}$ цього закону розподілу. Як відомо з курсу "Теорія ймовірностей", якщо відомою є щільність сумісного розподілу двох випадкових величин $f(x,y)$, то

$$\left.\begin{array}{l}f(x)=\int\limits_{-\infty}^{\infty}f(x,y)dy\\ f(y)=\int\limits_{-\infty}^{\infty}f(x,y)dx\end{array}\right\}. \tag{3.74}$$

Для випадкових величин, визначених в області $(-\infty,\infty)$, на підставі цього з урахуванням (3.73) маємо :

$$f(x)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{xy}^2}}\times$$
$$\times\exp\left\{-\frac{(x-m_x)^2}{2\left(1-\rho_{xy}^2\right)\sigma_x^2}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2\left(1-\rho_{xy}^2\right)}\times\right.\right.$$
$$\left.\left.\times\left[\frac{(y-m_y)^2}{2\sigma_y^2}-2\rho_{xy}\frac{(x-m_x)(y-m_y)}{\sigma_x\sigma_y}\right]\right\}dy\right\}$$

(3.75)

Запровадимо позначення $\upsilon=\dfrac{y-m_y}{\sigma_y}$ і $u=\dfrac{x-m_x}{\sigma_x}$.

Тоді

$$f(x)=\frac{1}{2\pi\sqrt{1-\rho_{xy}^2}\sigma_x}e^{-\frac{u^2}{2\left(1-\rho_{xy}^2\right)}}\times$$
$$\times\int_{-\infty}^{\infty}e^{-\frac{\upsilon^2-2\rho_{xy}u\upsilon}{2\left(1-\rho_{xy}^2\right)}}d\upsilon.$$

(3.76)

Якщо показник експоненти доповнити до повного квадрату

$$\upsilon^2 - 2\rho_{xy} u\upsilon = \left(\upsilon - \rho_{xy} u\upsilon\right)^2 - \rho_{xy}^2 u^2, \qquad (3.77)$$

то будемо мати

$$f(x) = \frac{1}{2\pi\sqrt{1-\rho_{xy}^2}\sigma_x} e^{-\left[\frac{u^2}{2\left(1-\rho_{xy}^2\right)} - \frac{\rho_{xy}^2 u^2}{2\left(1-\rho_{xy}^2\right)}\right]} \times$$

$$\times \int_{-\infty}^{\infty} e^{-\frac{\left(\upsilon-\rho_{xy}u\upsilon\right)^2}{2\left(1-\rho_{xy}^2\right)}} d\upsilon \quad . \qquad (3.78)$$

Позначимо

$$\frac{\upsilon - \rho_{xy}\upsilon u}{\sqrt{2}\sqrt{1-\rho_{xy}^2}} = t \ . \qquad (3.79)$$

Тоді

$$d\upsilon = \sqrt{2}\sqrt{1-\rho_{xy}^2}dt \qquad (3.80)$$

і

$$\int_{-\infty}^{\infty} e^{-\frac{(\upsilon-\rho_{xy}u\upsilon)^2}{2\left(1-\rho_{xy}^2\right)}} d\upsilon = \sqrt{2}\sqrt{1-\rho_{xy}^2}\int_{-\infty}^{\infty} e^{-t^2} dt =$$

$$= \sqrt{2}\sqrt{1-\rho_{xy}^2}\sqrt{\pi} \quad .$$

(3.81)

Отже маємо:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{u^2}{2}}$$

або повертаючись до вихідної змінної,

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{(x-m_x)^2}{2\sigma_x^2}} \quad . \qquad (3.82)$$

Таким чином, параметри m_x і σ_x^2 двовимірного нормального розподілу (3.73) мають сенс математичного сподівання і дисперсії випадкової величини X.

Застосовуючи до рівняння (3.73) другу формулу системи (3.74), шляхом аналогічних перетворень можна показати, що

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{(y-m_y)^2}{2\sigma_y^2}} \quad . \qquad (3.83)$$

Отже, параметри m_y і σ_y^2 також є математичним сподіванням і дисперсією нормально розподіленої випадкової величини Y.

Зупинимося тепер на поясненні смислу параметра ρ_{xy}. Для цього знайдемо коваріацію K_{xy} випадкових величин X і Y. Маємо за визначенням:

$$K_{xy} = M\{(x - m_x)(y - m_y)\} = \\ = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x - m_x)(y - m_y)f(x,y)dxdy \quad . \tag{3.84}$$

Введемо позначення $x - m_x = \Delta x; \quad y - m_y = \Delta y.$

Тоді з урахуванням рівняння (3.73) отримаємо:

$$K_{xy} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Delta x\Delta y\frac{1}{2\pi\sqrt{1-\rho_{xy}^2}}\times \\ \times e^{-\frac{1}{2(1-\rho_{xy}^2)}\left[\frac{\Delta x^2}{\sigma_x^2}-2\rho_{xy}\frac{\Delta x\Delta y}{\sigma_x\sigma_y}+\frac{\Delta y^2}{\sigma_y^2}\right]}d\Delta xd\Delta y\,. \tag{3.85}$$

Позначимо

$$\frac{\Delta x}{\sigma_x \sqrt{2}} = \xi \quad ; \tag{3.86}$$

$$\frac{1}{\sqrt{2\left(1-\rho_{xy}^2\right)}}\left(\frac{\Delta y}{\sigma_y} - \rho_{xy}\frac{\Delta x}{\sigma_x}\right) = \eta. \tag{3.87}$$

Тоді

$$\Delta x = \sigma_x \sqrt{2}\Delta\xi, \tag{3.88}$$

$$\Delta y = \sigma_y \sqrt{2}\sqrt{1-\rho_{xy}^2}\left(\eta + \frac{\rho_{xy}\xi}{\sqrt{1-\rho_{xy}^2}}\right). \tag{3.89}$$

Возводячи до квадрату рівняння (3.87) і доповнюючи ліву частину його до повного квадрату, маємо:

$$\frac{1}{2\left(1-\rho_{xy}^2\right)}\left[\frac{\Delta x^2}{\sigma_x^2} - 2\rho_{xy}\frac{\Delta x \Delta y}{\sigma_x \sigma_y} + \frac{\Delta y^2}{\sigma_y^2}\right] = \eta^2 + \xi^2 \quad . \tag{3.90}$$

Отже з урахуванням рівностей (3.88) - (3.90) отримаємо:

$$K_{xy} = \frac{2\sigma_x \sigma_y \sqrt{1-\rho_{xy}^2}}{\pi} \times$$
$$\times \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \xi \left(\eta + \frac{\rho_{xy} \xi}{\sqrt{1-\rho_{xy}^2}} \right) e^{-\xi^2 - \eta^2} d\xi d\eta \quad . \tag{3.91}$$

Останнє рівняння можна переписати так:

$$K_{xy} = \frac{2\sigma_x \sigma_y \sqrt{1-\rho_{xy}^2}}{\pi} \int_{-\infty}^{\infty} \xi e^{-\xi^2} d\xi \int_{-\infty}^{\infty} \eta e^{-\eta^2} d\eta +$$
$$+ \frac{2\sigma_x \sigma_y \sqrt{1-\rho_{xy}^2}}{\pi} \int_{-\infty}^{\infty} \xi^2 e^{-\xi^2} d\xi \int_{-\infty}^{\infty} e^{-\eta^2} d\eta \quad . \tag{3.92}$$

Але

$\int_{-\infty}^{\infty} \xi e^{-\xi^2} d\xi = \int_{-\infty}^{\infty} \eta e^{-\eta^2} d\eta = 0$ як інтеграл від

непарної функції, а

$$\int_{-\infty}^{\infty} \xi^2 e^{-\xi^2} d\xi = \frac{\sqrt{\pi}}{2}, \tag{3.93}$$

$$\int_{-\infty}^{\infty} e^{-\eta^2} d\eta = \sqrt{\pi} \ . \tag{3.94}$$

Враховуючи ці результати, маємо:

$$K_{xy} = \sigma_x \sigma_y \rho_{xy} \tag{3.95}$$

і

$$\rho_{xy} = \frac{K_{xy}}{\sigma_x \sigma_y} \tag{3.96}$$

Таким чином, параметр ρ_{xy} двовимірного нормального розподілу є коефіцієнтом корреляції між випадковими величинами X і Y.

У випадку, коли випадкові величини X і Y - некоррельовані $\left(\rho_{xy} = 0\right)$, то

$$f(x,y)=\frac{1}{\sqrt{2\pi}\sigma_x\sigma_y}e^{-\frac{1}{2}\left[\frac{(x-m_x)^2}{\sigma_x^2}-\frac{(y-m_y)^2}{\sigma_y^2}\right]}=$$

$$=\frac{1}{\sqrt{2\pi}\sigma_x}e^{-\frac{(x-m_x)^2}{2\sigma_x^2}}\frac{1}{\sqrt{2\pi}\sigma_y}e^{-\frac{(y-m_y)^2}{2\sigma_y^2}}=$$

$$=f(x)f(y)\ .$$

(3.97)

Рівняння (3.97) є необхідною і достатньою умовою незалежності випадкових величин X і Y. Отже, некоррельовані нормально розподілені випадкові величини є незалежними. Але такий висновок не може бути прийнятним для випадкових величин, які мають інші розподіли ймовірностей. Якщо коефіцієнт корреляції дорівнює нулю, то в загальному випадку випадкові величини X і Y можуть бути і незалежними і залежними. Нульове значення коефіцієнта корреляції є умовою необхідною, але недостатньою для незалежності випадкових величин. Такі випадкові величини можуть бути зв'язаними стохастичним або функціональним зв'язком.

Розглянемо рівняння:

$$\frac{(x-m_x)^2}{\sigma_x^2} - 2\rho_{xy}\frac{(x-m_x)(y-m_y)}{\sigma_x\sigma_y} + \\ + \frac{(y-m_y)^2}{\sigma_y^2} = K^2 \quad , \qquad (3.98)$$

що є показником експоненти рівняння (3.73).

Рівняння (3.98) є рівнянням сім'ї еліпсів рівної щільності розподілу, центр яких розташовується в точці з координатами (m_x, m_y), а головні осі розсіювання (осі симетрії еліпсів) не паралельні координатним осям (рис.3.6). Він відповідає умові $m_x = m_y = 0$. Вони складають з осю OX кути, що визначаються рівнянням :

$$tg2\varepsilon = \frac{2\rho_{xy}\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2} \; . \qquad (3.99)$$

Рівняння (3.99) дає значення кутів ε_1 і ε_2, які розрізняються на величину $\frac{\pi}{2}$.

Якщо нормально розподілені величини є незалежними $(\rho_{xy} = 0)$, то, очевидно, рівняння (3.98) придбає форму:

$$\frac{(x-m_x)^2}{\sigma_x^2}+\frac{(y-m_y)^2}{\sigma_y^2}=K^2\ . \qquad (3.100)$$

У цьому випадку $\varepsilon=0$, тобто осі симетрії еліпса паралельні осям координат (рис.3.7). Параметр K відповідає вибраному значенню рівня z, на якому відбувається перетин поверхні двовимірного нормального розподілу площиною, паралельною координатній площині OXY (рис.3.6). При $z_0=K=1$ півосі еліпса дорівнюють середнім квадратичним відхилам σ_x і σ_y.

Запровадимо позначення :

$$\frac{x-m_x}{\sigma_x}=u;\quad \frac{y-m_y}{\sigma_y}=\upsilon$$

і, враховуючи формулу (3.73) знайдемо умовний розподіл $f\left(y/x\right)$:

$$f\left(\frac{y}{x}\right)=\frac{f(x,y)}{f(x)}=$$

$$=\frac{e^{-\frac{1}{2\left(1-\rho_{xy}^2\right)}\left(u^2-2\rho_{xy}u\upsilon+\upsilon^2\right)}\sqrt{2\pi}\sigma_x}{2\pi\sigma_x\sigma_y e^{-\frac{u^2}{2}}\sqrt{1-\rho_{xy}^2}}=$$

$$=\frac{e^{-\frac{1}{2}\left(\frac{\upsilon-\rho_{xy}u}{\sqrt{1-\rho_{xy}^2}}\right)^2}}{\sqrt{2\pi}\sigma_y\sqrt{1-\rho_{xy}^2}}\quad .$$

Переходячи до попередніх змінних, маємо:

$$f\left(\frac{y}{x}\right)=\frac{1}{2\pi\sigma_y\sqrt{1-\rho_{xy}^2}}e^{-\frac{1}{2}\left(\frac{y-m_y-\rho_{xy}\frac{\sigma_y}{\sigma_x}(x-m_x)}{\sigma_y\sqrt{1-\rho_{xy}^2}}\right)^2}.$$

(3.101)

Із рівняння (3.101) видно, що умовний розподіл є нормальним з центром розсіювання

$$m_{y/x} = m_y + \rho_{xy}\frac{\sigma_y}{\sigma_x}(x - m_x) \qquad (3.102)$$

і середнім квадратичним відхилом

$$\sigma_{y/x} = \sigma_y\sqrt{1-\rho_{xy}^2}\ . \qquad (3.103)$$

Ясно, що рівність (3.102) є рівняння лінії регресії Y по X, яка має форму прямої лінії (3.16), де

$$\alpha = \rho_{xy}\frac{\sigma_y}{\sigma_x}, \qquad (3.104)$$

а

$$\beta = m_y - \alpha m_x\ . \qquad (3.105)$$

Отже, аналіз двовимірного нормального розподілу дає змогу обгрунтувати основні засади теорії корреляційного зв'язку між двома випадковими величинами, які розглядалися у попередніх параграфах цього розділу.

4 ПЕРЕВІРКА СТАТИСТИЧНИХ ГІПОТЕЗ

4.1 Загальна постановка задачі про перевірку статистичних гіпотез

Перевірка статистичних гіпотез тісно поєднується з теорією оцінювання параметрів. У природознавстві, різних галузях техніки і економіки часто для з'ясування того чи іншого випадкового факту звертаються до висловлювання гіпотез, які можна перевірити статистично, тобто опираючись на результати спостережень у випадкових вибірках. Під *статистичними гіпотезами* розуміють такі гіпотези, котрі відносяться або до виду, або до окремих параметрів розподілу випадкової величини. Наприклад, статистичною є гіпотеза про те, що середні добові температури повітря мають нормальний закон розподілу. Статистичною буде також гіпотеза про те, що місячні кількості опадів на сусідніх метеорологічних станціях суттєво не розрізнюються.

Сформулюємо задачу статистичної перевірки гіпотези у загальному виді. Нехай $f(x,\Theta)$ - закон розподілу випадкової величини x, який залежить від одного параметра Θ. Припустимо, що необхідно перевірити гіпотезу про те, що $\Theta = \Theta_0$. Будемо називати цю *гіпотезу нульовою* і позначимо її через H_0. Гіпотезу про те, що $\Theta = \Theta_1$ назвемо *конкуруючою* і позначимо її через H_1. Гіпотезу H_0 іноді називають *основною*, а гіпотезу H_1 - *альтернативною*. Отже, стоїть задача перевірки гіпотези H_0 відносно конкуруючої гіпотези H_1 на основі вибірки, що складається з n незалежних спостережень $x_1, x_2, x_3, \ldots, x_n$ над випадковою величиною X.

Як вже відомо, статистична сукупність розглядається як випадкова вибірка з генеральної сукупності. Отже, всю можливу

множину N вибірок об'ємом n можна розділити на дві неперетинних підмножини (позначимо їх через u_1 і u_2), таких, що гіпотеза H_0 повинна бути відкинутою, якщо вибірка, яка розглядається, потрапляє до підмножини u_1 , і прийнятою, якщо вибірка належить до підмножини u_2.

Більш зручно, проте, мати діло не з вибірками, а з деякими статистичними параметрами k, одержаними на основі вибірок за визначним правилом. Оскільки ці параметри є числами, а останні зображаються точками на числовій осі, підмножини u_1 і u_2 вибірок зводяться до двох підмножин точок числової осі або до двох одномірних областей W_1 і W_2. Область W_1 параметрів k називають *критичною областю*, а область W_2 - *областю припустимих значень*. Оскільки область W_2 складається з точок, які не увійшли до області W_1, то область W_1 однозначно визначає область W_2 , і навпаки.

Виникає питання про те, якими принципами треба керуватися при будуванні критичної області W_1. Ці принципи полягають у тому, що приймаючи чи відкидаючи гіпотезу H_0 можна припустити помилку двох видів.

Помилка першого роду полягає у тому, що нульова гіпотеза H_0 відкидається, тобто приймається гіпотеза H_1 тоді, коли в дійсності все ж таки вірною є гіпотеза H_0.

Помилка другого роду допускається тоді, коли приймається гіпотеза H_0 , у той час, коли вірною є гіпотеза H_1.

Смисл помилок першого і другого роду ілюструє табл. 4.1.

Таблиця 4.1 - Помилки першого і другого роду

Гіпотеза H_0	є вірною	є невірною
Відкидається	Помилка I роду	Правильне рішення
Приймається	Правильне рішення	Помилка II роду

Імовірності помилок першого і другого роду однозначно визначаються вибором критичної області W_1. Умовимося для будь-якої критичної області W_1 позначати через α імовірність помилки першого роду. Її називають *рівнем значущості*. Критичну область W_1 відокремлює від області прийняття гіпотези H_0 критична точка $k_{кр}$. Можуть розглядатися правостороння, лівостороння, двохстороння і симетрична двохстороння критичні області. Вони позначені на рис. 4.1.

Як вже позначалося, параметр k формується за визначеним правилом в залежності від характеру задачі, яка розв'язується, та властивостей випадкових величин. Критичне значення параметра $k_{кр}$ знаходиться із такої умови. Припустимо йдеться про правосторонню критичну область (рис.4.1а). Нехай гіпотеза H_0 є вірною. Тоді імовірність того, що гіпотеза H_0 відкидається тобто, що робиться помилка I роду, дорівнює

$$P(k > k_{кр}) = \alpha.$$

Для лівосторонньої критичної області (рис.4.1б)

$$P(k < k_{кр}) = \alpha,$$

для двохсторонньої критичної області (рис.4.1в),

$$P(k < k_{кр2}) + P(k > k_{кр1}) = \alpha$$

і для двохсторонньої симетричної області (рис.4.1г)

$$P(k < -k_{кр}) + P(k > k_{кр}) = \alpha$$

або

$$P(k > k_{кр}) = \frac{\alpha}{2}.$$

Отже, можна сказати, що при великій кількості вибірок доля помилкових рішень дорівнює α , якщо гіпотеза H_0 є вірною.

Для знаходження параметра k будемо використовувати такі теореми:

1. Нехай незалежні величини $z : z_1, z_2, \ldots, z_u$ підпорядковуються нормальному закону розподілу. Тоді сума квадратів цих величин

$$\chi^2 = \sum_{i=1}^{n} z_i^2 \qquad (4.1)$$

підпорядковується закону розподілу, який визначається щільністю імовірності

$$f(\chi^2) = \begin{cases} \dfrac{1}{2^{\nu/2}\Gamma(\dfrac{\nu}{2})} (\chi^2)^{\frac{\nu}{2}-1} e^{-\chi^2/2} & \text{при } \chi^2 \geq 0 \\ 0 & \text{при } \chi^2 < 0 \end{cases} \qquad (4.2)$$

з ν - числом ступенів волі. Число ступенів волі є параметром розподілу. Закон розподілу (4.2) називають χ^2 - *розподілом.*

2. Нехай ми маємо дві незалежні випадкові величини u і υ, які підпорядковуються χ^2 - розподілу з числами ступенів волі ν_1 і ν_2 відповідно. Тоді випадкова величина

$$F = \frac{u/\nu_1}{\upsilon/\nu_2} \qquad (4.3)$$

підпорядковується розподілу

$$f(F)=\begin{cases}\dfrac{1}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})}\left(\dfrac{\nu_1}{\nu_2}\right)^{\nu_1/2}\dfrac{F^{\frac{\nu_1-2}{2}}}{\left(1+\dfrac{\nu_1}{\nu_2}F\right)^{\frac{\nu_1+\nu_2}{2}}} & \text{при } F\geq 0 \\ 0, & \text{при } F<0\end{cases}. \qquad (4.4)$$

Формула (4.4) носить назву *закона Фішера-Снедекора*.

Розподіл Фішера-Снедекора двохпараметричний з параметрами ν_1 і ν_2.

3. Нехай маємо дві незалежні випадкові величини u і υ такі, що u - підпорядковується нормальному закону, а $\upsilon - \chi^2$ розподілу з числом ступенів волі ν.

Тоді випадкова величина t

$$t=\frac{u}{\sqrt{\dfrac{\upsilon}{\nu}}} \qquad (4.5)$$

підпорядковується розподілу

$$f(t)=\frac{1}{\sqrt{\pi\nu}}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},-\infty<t<\infty$$

(4.6)

Рівняння (4.6) називається *розподілом Стьюдента*. Як видно, цей закон є однопараметричним, з параметром ν.

4.2 Перевірка статистичної гіпотези про однорідність членів статистичної сукупності.

Процеси, що відбуваються в атмосфері і гідросфері, мають різну інтенсивність. У залежності від цього гідрометеорологічні величини, які характеризують інтенсивність того чи іншого процесу, можуть приймати значення, котрі відрізняються одне від одного. В окремих випадках значення деяких метеорологічних величин може дуже відрізнятися від загального рівня, який характеризується середнім значенням. Наприклад, у Одесі середня температура повітря у листопаді дорівнює 5,8 ^{0}C . В 1994 році цього місяця відбулося вторгнення дуже холодного континентального арктичного повітря на Україну і середні добові температури протягом декількох днів коливалися у границях - 10 ^{0}C…-12 ^{0}C, тобто різко відрізнялися від звичних середніх добових температур. Такі значення метеорологічних величин прийнято називати "*випадіннями*" або "*викидами*". Випадіння чинять вплив на середнє значення, але особливо на значення центральних моментів, оскільки при їх оцінці в суму входять доданки, які являють собою великі різниці між випадіннями і середніми значеннями, котрі підносяться до другої, третьої чи четвертої степені у залежності від того, оцінку якого моменту треба знайти.

Виникає питання, що робити з означенними випадіннями: чи зберігати їх у вихідній статистичній сукупності, чи вилучити? На це питання треба відповісти після перевірки статистичної гіпотези H_0 про однорідність членів статистичної сукупності.

Розглянемо принцип формування критерія k для перевірки цієї гіпотези на основі вибірки випадкової величини $X : x_1, x_2, ..., x_n$.

Відомо, що середнє значення, як випадкова величина, підпорядковується нормальному закону. Вибиремо із вибірки $x_{\min}$ і $x_{\max}$. Ясно, що випадкові величини $\bar{x} - x_{\min}$ та $x_{\max} - \bar{x}$ також будуть підпорядковуватися нормальному закону. Тому можна позначити

$$u = \left| x_{екстр} - \bar{x} \right| , \qquad (4.7)$$

де $x_{екстр}$ об'єднує мінімальну і максимальну величини. Випадкова величина u відповідає одній з умов теореми 3.

Будемо вважати, що випадкова величина X має нормальний розподіл. Тоді на основі теореми 1

$$\upsilon = \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 = \chi^2 \qquad (4.8)$$

підпорядковується розподілу χ^2 з числом ступенів волі $\nu = n - 1$. Отже, ця величина задовольняє другій умові теореми 3. Таким чином, на основі цієї теореми отримаємо випадкову величину

$$t = \frac{\left| x_{екстр} - \bar{x} \right|}{\sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}} = \frac{\left| x_{екстр} - \bar{x} \right|}{S_x}, \qquad (4.9)$$

що підпорядковується розподілу Стьюдента. Далі на рівні значущості α, який установлюється дослідником, знаходять по відповідній таблиці Додатку $t_{кр}(\alpha, \nu)$ з умови (для правосторонньої критичної області):

$$P(t > t_{кр}) = \alpha. \qquad (4.10)$$

Отже, якщо $t < t_{кр}$, то гіпотеза H_0 про однорідність членів вибірки не відхиляється, у протилежному випадку, тобто коли $t > t_{кр}$, гіпотеза H_0 відхиляється й приймається альтернативна гіпотеза H_1 про те, що $x_{екстр}$, інакше кажучи "випадіння", не належать до тієї ж генеральної сукупності, що і решта членів цієї сукупності. У такому разі неоднорідні члени вилучаються з вибірки, і знову проводиться оцінка відповідних параметрів. Але "випадіння" не відкидаються зовсім. Їх треба пильно вивчати, оскільки вони відбивають ті чи інші аномальні особливості атмосферних процесів, наслідком яких вони є.

Наведемо приклад перевірки такої гіпотези.

Нехай в січні в Одесі отримані такі оцінки математичного сподівання і середнього квадратичного відхилу середньої місячної температури повітря на основі вибірки об'єму $n = 35: \bar{x} = -1.8^\circ C$, $S_x = 3{,}1^\circ C$. Однак вибірка утримує

низькі значення температури, які значно відрізняються від загального їх рівня: $x_1 = -8{,}8^\circ C; x_2 = -7{,}8^\circ C;$ $x_3 = -6{,}0^\circ C$. Необхідно визначити, чи не відносяться перелічені значення температури до "випадінь"? Для цього сформулюємо основну гіпотезу H_0 таким чином: середня місячна температура $x_1 = -8{,}8^\circ C$ належить до тієї ж генеральної сукупності, що й інші члени вибірки. Для її перевірки розрахуємо емпіричний критерій Стьюдента

$$t = \frac{-1{,}8 + 8{,}8}{3{,}1} = 2{,}26.$$

При рівні значущості $\alpha = 0{,}05$ і числа ступенів волі $\nu = n - 1 = 34,\quad t_{кр}(\alpha, \nu) = 2{,}03$. Отже, оскільки $t > t_{кр}(\alpha, \nu)$, гіпотеза H_0 відхиляється. Тепер треба вилучити x_1 з вибірки і знову знайти значення $\bar{x}$ і S_x. Розрахунки показують, що $\bar{x} = -1{,}5^\circ C; S_x = 3{,}0^\circ C$ при $n = 34$.

Далі формулюємо гіпотезу H_0, як і у попередньому випадку, відносно $x_2 = -7{,}8^\circ C$. Маємо:

$$t = \frac{-1{,}5 + 7{,}8}{3{,}0} = 2{,}1;\ t_{кр}(\alpha, \nu) = 2{,}04 \text{ при } \alpha = 0{,}05$$

і $\nu = 33$. Таким чином, гіпотеза H_0 знову відхиляється, а x_2 вилучається з вибірки. Після цього знову розраховуємо $\bar{x}$ і S_x

при $n = 33$. Будемо мати: $\bar{x} = -1{,}3^{\circ}C; S_x = 2{,}9^{\circ}C$. Формулюємо тепер гіпотезу H_0 відносно члена $x_3 = -6{,}0^{\circ}C$. Для нього критерій Стьюдента дорівнює:

$$t = \frac{-1{,}3 + 6{,}0}{2{,}9} = 1{,}62; \quad t_{кр}(\alpha, \nu) = 2{,}04$$

при $\alpha = 0{,}05$ і $\nu = 32$. Отже, оскільки $t < t_{кр}(\alpha, \nu)$, гіпотеза H_0, не відхиляється. Таким чином, $x_1 = -8{,}8^{\circ}C$ і $x_2 = -7{,}8^{\circ}C$ є "випадіннями", тобто не належать до тієї ж генеральної сукупності, що інші її члени на рівні значущості $\alpha = 0{,}05$. *Статистично обгрунтованими є оцінки, які отримані на основі сукупності однорідних членів*, тобто $\bar{x} = -1{,}3^{\circ}C$ і $S_x = 2{,}9^{\circ}C$. Значення середньомісячної температури $x_1 = -8{,}8^{\circ}C$ і $x_2 = -7{,}8^{\circ}C$ є результатом дії аномальних синоптичних процесів і підлягають окремому розгляданню.

4.3 Перевірка статистичної гіпотези про однорідність двох нормально розподілених рядів.

Два ряду випадкових величин називають *однорідними,* коли вони на рівні значущості α належать до одної і тієї ж генеральної сукупності (підпорядковуються одному і тому ж закону розподілу).

Нехай ми маємо дві статистичні сукупності:

$$X : x_1, x_2, ..., x_n; \tag{4.11}$$

$$Y : y_1, y_2, ..., y_m \quad , \tag{4.12}$$

які мають об'єми відповідно n і m . Відомо, що ці сукупності підпорядковуються нормальному закону розподілу, тобто

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left[- \frac{(x - m_x)^2}{2\sigma_x^2} \right], \tag{4.13}$$

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma_y} \exp\left[- \frac{(y - m_y)^2}{2\sigma_y^2} \right]. \tag{4.14}$$

За визначенням ряди X і Y будуть однорідними, якщо $m_x = m_y$ і $\sigma_x^2 = \sigma_y^2$. Але ми не маємо математичних сподівань і дисперсій, які є параметрами генеральних сукупностей, але на основі вибірок (4.11) і (4.12) можна отримати їх оцінки $\bar{x}, \bar{y}, S_x^2, S_y^2$. Оскільки оцінки - це випадкові величини, вони відрізняються один від одного. Можливі ситуації, що зображені на рис. 4.2.

На першому з них дуже відрізняються одне від одного математичні сподівання m_x і m_y випадкових величин X і Y. Між їх оцінками $\bar{x}$ і $\bar{y}$ різниці будуть значущими. Тому

статистичні сукупності, на основі котрих отримані ці оцінки, не будуть належати до однієї генеральної сукупності.

У другому випадку (рис. 4.2б) випадкові величини мають різні дисперсії σ_x^2 і σ_y^2 . Це приведе до значущих різниць між оцінками дисперсій S_x^2 і S_y^2. І знову сукупності випадкових величин не будуть належити до однієї генеральної сукупності. *Ряди випадкових величин* X *і* Y *будуть однорідними*, тобто будуть на рівні значущості α належати до однієї генеральної сукупності, *якщо різниці між оцінками параметрів їх розподілу* S_x^2 і S_y^2 *з однієї сторони і* $\bar{x}$ *і* $\bar{y}$ *- з другої носять випадковий характер*, або, інакше кажучи, *не будуть значущими* (рис.4.2в). Звідси видно, що дослідження однорідності двох випадкових рядів полягають у перевірці двох статистичних гіпотез: гіпотези $H_0^{'}$ про незначущість різниць оцінок дисперсій та гіпотези $H_0^{''}$ про незначущість різниць середніх значень.

а) Перевірка гіпотези $H_0^{'}$ *про незначущість різниць між оцінками дисперсій.*

Нам відомо, що ряди випадкових величин X і Y підпорядковуються нормальному розподілу. Тоді на основі першої теореми сума їх квадратів підпорядковується χ^2 розподілу. Отже випадкові величини

$$u = \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (4.15)$$

та

$$v = \sum_{i=1}^{m} (y_i - \bar{y})^2 \qquad (4.16)$$

підпорядковуються розподілу χ^2 з числами ступенів волі відповідно $\nu_1 = n-1$ та $\nu_2 = m-1$. Результати ж (4.15) і (4.16) є умовою теореми 2. Тому

$$\frac{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}{\dfrac{\sum_{i=1}^{m}(y_i - \bar{y})^2}{m-1}} = F \qquad (4.17)$$

є випадковою величиною, що підпорядковується розподілу Фішера-Снедекора (4.4). Очевидно, рівність (4.17) можна записати

$$F = \frac{S_x^2}{S_y^2} \quad . \qquad (4.18)$$

Нам залишилося визначити рівень значущості α і за допомогою параметрів щільності ймовірності Фішера-Снедекора $\nu_1 = n-1$ і $\nu_2 = m-1$ знайти по відповідних таблицях $F_{кр}(\alpha, \nu_1, \nu_2)$ (Додаток)

Якщо

$$F < F_{кр}(\alpha, \nu_1, \nu_2),$$

то гіпотеза $H_0^{'}$ не відхиляється, і навпаки.

З формул (4.17) - (4.18) виходить, що оцінки дисперсій S_x^2 і S_y^2 як випадкові величини, підпорядковуються χ^2 розподілу.

б) Перевірка статистичної гіпотези $H_0^{''}$ про незначущість різниць між середніми значеннями.

Як відомо, середні значення - випадкові величини, що підпорядковуються нормальному закону розподілу. Тому випадкова величина $u = \bar{x} - \bar{y}$ теж підпорядковується цьому закону. Оскільки S_x^2 і S_y^2, як було відзначено вище, підпорядковуються χ^2 розподілу, то $(n-1)S_x^2$ і $(m-1)S_y^2$ теж йому підпорядковуються. Тому випадкова величина

$$\upsilon = (n-1)S_x^2 + (m-1)S_y^2 \qquad (4.19)$$

має закон розподілу χ^2 з числом ступенів волі $v = m + n - 2$. Але такі величини u і υ задовольняють умовам теореми 3. Тому маємо

$$t = \frac{\left|\bar{x} - \bar{y}\right|}{\sqrt{\dfrac{(n-1)S_x^2 + (m-1)S_y^2}{m+n-2}\left(\dfrac{1}{n} + \dfrac{1}{m}\right)}}. \qquad (4.20)$$

У формулі (4.20) модуль у чисельнику застосовується для того, щоб мати діло лише з додатними значеннями параметра

Стьюдента t, крім того під знаком радікалу вводиться нормуючий множник $\left(\frac{1}{n}+\frac{1}{m}\right)$.

Тепер за рівнем значущості α та числом ступенів волі $v = m + n - 2$ за допомогою відповідної таблиці знаходиться значення $t_{кр}(\alpha, v)$. Якщо

$$t < t_{кр}(\alpha, v),$$

то гіпотеза $H_0^{''}$ не відхиляється.
Коли

$$t > t_{кр}(\alpha, v)$$

ми потрапляємо до критичної області і гіпотеза $H_0^{''}$ відхиляється та приймається альтернативна гіпотеза $H_1^{''}$.

Гіпотеза H_0 про однорідність рядів (4.11) і (4.12) не відхиляється лише у тому випадку, коли не відхиляються і $H_0^{'}$, і $H_0^{''}$. Якщо яка-небудь з них відхиляється, то відхиляється гіпотеза H_0 на даному рівні значущості.

Перевірку розглянутих гіпотез треба виконувати і тоді, коли проводиться, наприклад, дослідження просторових чи часових особливостей випадкової величини, наприклад кліматичних характеристик якої-небудь метеорологічної величини на сусідніх метеорологічних станціях або в різні місяці (чи роки). Тільки тоді треба шукати фізичні причини розбіжностей між відповідними середніми значеннями та дисперсіями, коли шляхом перевірки зазначених вище гіпотез установлюється, що ці розбіжності є значущими.

Нижче наводиться приклад перевірки статистичної гіпотези про однорідність двох рядів середніх місячних температур у січні в Одесі. Перший з них відноситься до періоду 1906-1940 р.р., другий – до періоду 1941-1975 р.р. Ці періоди 20-го сторіччя відрізняються тим, що у першому з них відбувалось загальне потепління глобального клімату, а у другому - його похолодання.

Отже треба визначити, чи не відбулося при цьому порушення однорідності загального ряду спостережень температури повітря.

Позначимо літерою X середні місячні температури першого з періодів, а літерою Y - другого. Розрахунки середніх значень і дисперсій, що відносяться до зазначених рядів, дали такі результати:

Ряд X			Ряд Y		
n	$\overline{x}$	S_x^2	m	$\overline{y}$	S_y^2
35	- 1,9 ^{0}C	8,1(^{0}C)^2	35	- 2,4^{0}C	10,3(^{0}C)^2

Відомо, що середні місячні температури повітря підпорядковуються нормальному закону розподілу. Тому перевірка гіпотези H_0 про однорідність цих рядів середньої місячної температури проводиться за допомогою параметричних критеріїв.

По-перше, сформулюємо гіпотезу H_0': дисперсії рядів середніх місячних температур повітря в січні в Одесі значуще не розрізняються. Як відзначалося вище, ця гіпотеза перевіряється за допомогою критерія Фішера-Снедекора. Очевидно,

$$F = \frac{S_y^2}{S_x^2} = \frac{10,3}{8,1} = 1,28 .$$

Числа ступенів волі, при цьому, дорівнюють $v_1 = v_2 = n - 1 = 34$. Якщо визначити рівень значущості $\alpha = 0{,}05$, то $F_{кр}(\alpha, v_1, v_2) = 1{,}77$. Оскільки $F < F_{кр}(\alpha, v_1, v_2)$ гіпотеза $H_0^{'}$ не відхиляється. Таким чином, можна зробити висновок, що розбіжність в оцінках дисперсій є випадковою, тобто обумовлюється особливостями вибірок.

Сформулюємо тепер гіпотезу відносно середніх значень: середні значення середніх місячних температур в різні періоди розрізняються незначуще. Як відзначалося, така гіпотеза перевіряється шляхом використання критерія Стьюдента. Очевидно,

$$t = \frac{\left|-1{,}9 + 2{,}4\right|}{\sqrt{\frac{34 \cdot 8{,}1 + 3{,}4 \cdot 10{,}3}{68}\left(\frac{1}{35} + \frac{1}{35}\right)}} = 0{,}86$$

Критичне значення критерія Стьюдента при $\alpha = 0{,}05$ і $v = n + m - 2 = 68$ дорівнює: $t_{кр}(\alpha, v) = 2{,}00$. Отже, оскільки $t < t_{кр}(\alpha, v)$, гіпотеза $H_0^{''}$ також не відхиляється.

Таким чином, ми приходимо до висновку, що справедливою є гіпотеза H_0 про те, що з імовірністю 95% ряди середніх місячних температур у періоди 1906-1940 рр. і 1941-1975 рр. є однорідними. Це означає, що процеси потепління і похолодання клімату, що відбувалися в ці періоди, до порушення однорідності загального ряду середніх місячних температур повітря в січні в Одесі не привели.

4.4 Перевірка статистичної гіпотези про однорідність рядів випадкових величин за допомогою критерія Вілкоксона.

Критерії однорідності, що розглядалися вище, називають *параметричними*, тому, що їх використовування пов'язано з необхідністю прийняття для вибірок, які розглядаються умови щодо нормального закону розподілу. Але випадкові величини, у тому числі й гідрометеорологічні величини, не завжди підпорядковуються цьому закону. Більш того, нормальний закон може бути використованим для характеристики властивостей гідрометеорологічних величин лише у деяких випадках. *Якщо випадкові величини не підлягають нормальному розподілу або якщо невідомо до якого закону розподілу відноситься випадкова величина, вживаються непараметричні критерії*. Одним з таких критеріїв є критерій Вілкоксона.

Критерій Вілкоксона буває двох видів: *інверсійний* та *ранговий*. Розглянемо перший з них.

Суть критерія Вілкоксона з'ясуємо на такому прикладі. Нехай ми маємо дві вибірки

$$X : x_1, x_2, x_3, x_4, x_5$$
$$Y : y_1, y_2, y_3, y_4, y_5, y_6$$

Випадкові величини, що належать до вибірок X і Y розташовують у загальній послідовності у порядку збільшення (або зменшення) їх значень, наприклад у виді:

$$y_1 x_1 y_2 y_3 x_2 x_3 y_4 y_5 y_6 x_4 x_5$$

Якщо якому-небудь значенню x *попереджує деяке значення* y*, то кажуть, що ця пара утворює інверсію*. В загальній послідовності число інверсій u дорівнює

$$u = 1 + 3 + 3 + 6 + 6 = 19.$$

Відомо, що в однорідних рядах, кожний з котрих має не менше 10 членів, число інверсій розподіляється приблизно за нормальним законом з математичним сподіванням

$$m_u = \frac{m \cdot n}{2} \tag{4.21}$$

і дисперсією

$$\sigma_u^2 = \frac{m \cdot n}{12}(m + n + 1), \tag{4.22}$$

де n і m - число членів у першій та другій вибірках.

У якості нульової гіпотези H_0 приймемо гіпотезу про належність вибірок X і Y до однієї генеральної сукупності. Тепер необхідно знайти границі допустимих значень u, що відділяють область прийняття гіпотези від критичної області. Це роблять, як показувалося вище, шляхом установлення рівня значущості α або довірчої ймовірності $p = 1 - \alpha$. Якщо значення критерію, яке отримане за даними спостережень, попаде до критичної області, то нульова гіпотеза відхиляється й з імовірністю p приймається альтернативна гіпотеза.

Область прийняття гіпотези H_0 визначається нерівністю:

$$m_u - t_{кр}(\alpha, v)\sigma_u \le u \le m_u + t_{кр}(\alpha, v)\sigma_u, \quad (4.23)$$

а критична область - нерівностями

$$u < m_u - t_{кр}(\alpha, v)\sigma_u, \quad (4.24)$$

$$u > m_u + t_{кр}(\alpha, v)\sigma_u. \quad (4.25)$$

У нерівностях (4.23)-(4.25) $\sigma_u = \sqrt{\sigma_u^2}$ - середній квадратичний відхил числа інверсій, $t_{кр}(\alpha, v)$ - критерій Стьюдента для рівня значущості α і числа ступенів волі $v = m + n - 2$.

Критерій однорідності Вілкоксона відповідає задачі порівняння тільки двох вибірок. Але він може вживатися для попарного порівняння вибірок в S пунктах спостережень деякого регіона, який вважається однорідним.

Приведемо приклад використання інверсійного критерія Вілкоксона.

Маємо вибірки значень температури повітря біля земної поверхні для двох метеорологічних станцій

$$t_1{}^0c: 11{,}1;\ 12{,}2;\ 13{,}7;\ 27{,}0;\ 18{,}1;\ 14{,}3;$$

$$11{,}4;\ 11{,}3;\ 10{,}8;\ 11{,}6;\ 10{,}0.$$

$$t_2{}^0c: 11{,}4;\ \ 12{,}6;\ \ 18{,}2;\ \ 20{,}6;\ \ 9{,}1;\ \ 11{,}5;$$

$$13{,}6;\ \ 19{,}1;\ \ 12{,}3;\ \ 11{,}7;\ \ 10{,}1;\ \ 9{,}0.$$

Об'єми вибірок дорівнюють відповідно $n = 11; m = 12$.

Об'єднаємо обидва ряди і розташуємо елементи загального ряду в порядку збільшення їх значень. Отримаємо ранжированний ряд

(9,0) (9,1) 10,0 (10,1) 10,8 11,1 11,3 (11,4) 11,4 (11,5) 11,6 (11,7) 12,2 (12,3) (12,6) (13,6) 13,7 14,3 18,1 (18,2) (19,1) (20,6) 27,0

В дужках ранжированого ряду приводяться елементи ряду t_2. Тепер підрахуємо число інверсій відносно величин першого ряду:

$$u_1 = 2 + 3 + 3 + 3 + 4 + 5 + 6 + 9 + 9 +$$
$$+ 9 + 12 = 65 \ \ .$$

Розрахуємо по формулах (4.21) і (4.22) математичне сподівання та дисперсію числа інверсій. Вони дорівнюють

$m_u = 66;\ \ \sigma_u^2 = 264$. Отже $\sigma_u = 16{,}2$

Встановимо рівень значущості $\alpha = 0.05$. Як видно, число ступенів волі $v = 21$. Це дає $t_{кр}(0{,}05;21) = 2{,}08$. Границями критичної області є точки

$$u_{кр1} = m_u - t_{кр}(\alpha, v)\sigma_u = 32{,}3$$

і

$$u_{кр2} = m_u - t_{кр}(\alpha, v)\sigma_u = 99{,}7,$$

Видно, що емпіричне число інверсій $(u = 65)$ не виходить за границі області прийняття гіпотези [32,3; 99,7] і гіпотеза H_0 про однорідність рядів температури повітря на цих метеорологічних станціях не відхиляється.

4.5 Перевірка статистичної гіпотези про відповідність емпіричного розподілу теоретичному

Як вже неодноразово зазначалося, емпіричним законом розподілу є згрупований ряд випадкової величини

$$\begin{matrix} \widetilde{x}_1; \widetilde{x}_2; ...; \widetilde{x}_k \\ m_1; m_2; ...; m_k. \end{matrix} \qquad (4.26)$$

Реалізація алгоритму апроксимації емпіричного розподілу визначеним теоретичним законом приводить до отримання

інтервальних теоретичних частот. Отже, в результаті розрахунків будемо мати

$$\widetilde{x}_1;\widetilde{x}_2;...;\widetilde{x}_k$$

$$\widetilde{m}_1;\widetilde{m}_2;...;\widetilde{m}_k. \qquad (4.27)$$

Виникає питання, яка міра розбіжності між емпіричними m_i і теоретичними $\widetilde{m}_i$ інтервальними частотами нас задовольняє. Відповідь на нього дає перевірка статистичної гіпотези про відповідність емпіричного розподілу теоретичному закону на заданому рівні значущості.

Перш за все, як і при перевірці всілякої статистичної гіпотези, треба визначити критерій k, за допомогою якого втілюється перевірка гіпотези H_0 відносно альтернативної гіпотези H_1. Цей критерій формується на такій підставі. Відомо, що емпіричні частоти, як випадкові величини, мають нормальний розподіл. Тому, як зазначалося в розділі 4.1, сума їх квадратів підлягає χ^2 розподілу. Оскільки теоретичні інтервальні частоти $\widetilde{m}_i$ - величини невипадкові , то розподілу χ^2 підлягає і така сума квадратів

$$\sum_{i=1}^{k}\frac{(m_i-\widetilde{m}_i)^2}{\widetilde{m}_i}=\chi^2. \qquad (4.28)$$

Випадкова величина χ^2 і використовується у якості зазначенного критерія, який носить назву критерія Пірсона.

Застосування цього критерія для перевірки узгодженості між емпіричними та теоретичними частотами є вельми зручним, тому що, як випливає з рівності (4.28), він складається із різниць між цими частотами, що і визначає міру розбіжності між ними.

Гіпотеза H_0 формулюється таким чином: розбіжності між емпіричними і теоретичними частотами є незначущими на рівні значущості α. Тому

$$P\left[\chi^2 > \chi^2{}_{кр}(\nu,\alpha)\right] = \alpha\ , \qquad (4.29)$$

що визначає правосторонню критичну область. Вона зображається на рис. 4.3. Як видно, область прийняття гіпотези H_0 визначається нерівністю $\chi^2 < \chi^2{}_{кр}(\nu,\alpha)$, а критична область - нерівністю $\chi^2 > \chi^2{}_{кр}(\nu,\alpha)$. Отже, коли величина χ^2, що розрахована за формулою (4.28), відповідає першій з нерівностей, то гіпотеза H_0 не відхиляється, а якщо другій, то приймається альтернативна гіпотеза H_1.

Рівень значущості α, як відомо, визначається дослідником. Він дорівнює площі, що показується на рис.4.3. Величина ν є числом ступенів волі, котре відіграє роль параметра χ^2 розподілу. Цей параметр залежить від виду теоретичного розподілу, яким апроксимується емпіричний розподіл, і визначається за таким правилом:

$$\nu = k - s\,, \qquad (4.30)$$

де k - кількість часткових інтервалів.

Треба пам'ятати, що при використанні критерія χ^2 необхідно, щоб частота (емпірична та теоретична) у кожній градації була не менше 5. Для цього градації, які мають частоти менші за 5, необхідно об'єднати з сусідніми. S - кількість лінійних зв'язків, відносно частот m_i, що реалізуються при статистичному оцінюванні моментів, які беруть участь при розрахунках параметрів теоретичного розподілу.
Треба мати на увазі, що лінійний зв'язок

$$\sum_{i=1}^{k} m_i = n, \tag{4.31}$$

де n - загальний об'єм статистичної сукупності, реалізується завжди.

Наведемо приклади визначення v для деяких розподілів.

а) Нехай емпіричний розподіл апроксимується нормальним розподілом, як відомо, нормальний розподіл має два параметри- m_x і σ_x^2, оцінками яких є

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} m_i \tilde{x}_i \tag{4.32}$$

і

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{k} m_i (\tilde{x}_i - \bar{x})^2 . \tag{4.33}$$

Рівності (4.32) і (4.33) являють собою лінійні форми відносно інтервальних частот m_i. Таким чином, з урахуванням лінійного

зв'язку (4.31) для нормального розподілу маємо $S = 3$, а $\nu = k - 3$.

б) Якщо проводиться апроксимація емпіричного розподілу розподілом Пірсона I типу, то для визначення параметрів цього розподілу, окрім статистичних характеристик $\bar{x}$ і S_x^2, необхідно використовувати оцінки основних моментів $\hat{r}_3$ і $\hat{r}_4$, які безпосередньо пов'язані з оцінками третього

$$\hat{\mu}_3 = \frac{1}{n}\sum_{i=1}^{k} m_i (\tilde{x}_i - \bar{x})^3 \qquad (4.34)$$

і четвертого

$$\hat{\mu}_4 = \frac{1}{n}\sum_{i=1}^{k} m_i (\tilde{x}_i - \bar{x})^4 \qquad (4.35)$$

центральних моментів. Ясно, що вони відносно інтервальних частот m_i також є лінійними залежностями. Отже, при визначенні параметрів цього розподілу треба використовувати лінійні залежності (4.32) - (4.35), що з урахуванням рівності (4.31) дає $S = 5$, а $\nu = k - 5$.

в) Для розподілу Пірсона III типу параметри розраховуються за допомогою $\bar{x}, S_x^2$ і $\hat{r}_3$. Тому для III типу розподілів Пірсона маємо $S = 4$, а $\nu = k - 4$.

У якості прикладів в табл. 4.1-4.3 наводяться результати розрахунків "критерія згоди" χ^2 для розподілів Пірсона.

Таблиця 4.1 - Розрахунки критерія χ^2 на основі вибірки місячних температур повітря у червні в Одесі для нормального розподілу

№ п/п	m_i	$\tilde{m}_i$	$m_i - \tilde{m}_i$	$(m_i - \tilde{m}_i)^2$	$\frac{(m_i - \tilde{m}_i)^2}{\tilde{m}_i}$
1	1 } 7	1,6 } 6,3	0,9	0,81	0,13
2	6	4,5			
3	11	9,7	1,3	1,69	0,17
4	16	16,0	0,0	0,00	0,00
5	20	20,3	-0,3	0,09	0,004
6	19	19,5	-0,5	0,25	0,01
7	14	14,4	-0,4	0,16	0,01
8	7	8,2	-1,2	1,44	0,18
9	4 } 6	3,5 } 4,7	1,3	1,69	0,36
10	2	1,2			
				Сума:	0,864

Як видно із табл.4.1, розрахунки критерія Пірсона χ^2 дають таку величину: $\chi^2 = 0{,}864$. Значення $\chi^2_{кр}(\alpha, \nu)$, яке потім порівнюють з χ^2, знайдемо по таблиці Додатку ... при заданому рівні значущості $\alpha = 0{,}05$ та числа ступенів волі ν. Треба розрахувати ν.

$$\nu = k - 3 = 8 - 3 = 5.$$

Тоді

$$\chi^2_{кр}(0{,}05;5) = 11{,}1.$$

Отже $\chi^2 < \chi^2_{кр}(\alpha, v)$ і гіпотеза H_0 про те, що статистичний розподіл апроксимується нормальним законом розподілу з імовірністю $p = 0{,}95$ не відхиляється.

Розрахунки критерія χ^2 при апроксимації емпіричного розподілу меридіональної складової швидкості вітру на висоті 50 км розподілом Пірсона II типу дають такі результати (табл.4.2). Значення $\chi^2 = 2{,}82$. Відповідність емпіричних і теоретичних частот (гіпотеза H_0) перевіримо на рівні значущості $\alpha = 0{,}10$. Число ступенів волі для цієї задачі дорівнює $v = k - 4 = 8 - 4 = 4$.
Тоді $\chi^2_{кр}(0{,}10;4) = 7{,}78$.

Таблиця 4.2 - Результат розрахунків критерія χ^2 на основі вибірки меридіональної складової швидкості вітру на висоті 50 км

№ п/п	m_i	$\tilde{m}_i$	$m_i - \tilde{m}_i$	$(m_i - \tilde{m}_i)^2$	$\frac{(m_i - \tilde{m}_i)^2}{\tilde{m}_i}$
1	2 } 9	2,6 } 10,5			
2	7	7,9	- 1,5	2,25	0,21
3	10	10,7	- 0,7	0,49	0,05
4	16	12,3	3,7	13,69	1,11
5	11	12,9	- 1,9	3,61	0,28
6	14	12,7	1,3	1,69	0,13
7	10	11,6	- 1,6	2,56	0,22
8	8	9,7	- 1,7	2,89	0,30
9	8	6,2	1,8	3,24	0,52
				Сума:	2,82

Оскільки $\chi^2 < \chi^2_{кр}(\alpha, v)$, то з імовірністю 90% меридіональну складову швидкості вітру на висоті 50 км можна апроксимувати розподілом Пірсона II типу.

Приклад розрахунків критерія χ^2 при апроксимації емпіричного розподілу швидкості вітру на висоті 0,1 км розподілом Пірсона III типу наводиться в табл. 4.3.

Таблиця 4.3 - Розрахунки критерія χ^2 на основі статистичного ряду швидкості вітру на висоті 0,1 км.

№ п/п	m_i	$\tilde{m}_i$	$m_i - \tilde{m}_i$	$(m_i - \tilde{m}_i)^2$	$\frac{(m_i - \tilde{m}_i)^2}{\tilde{m}_i}$
1	20	19,0	1,0	1	0,05
2	52	45,4	6,6	43,56	0,96
3	43	53,9	-10,9	118,81	2,20
4	31	35,5	-4,5	20,25	0,57
5	23	18,5	4,5	20,25	1,09
6	8	7,6	-0,8	0,64	0,06
7	1 (10)	2,5 (10,8)			
8	1	0,7			
				Сума:	4,93

Розрахунки дали: $\chi^2 = 4{,}93$. Якщо рівень значущості прийняти $\alpha = 0{,}025$, то враховуючи те, що $v = k - 4 = 6 - 4 = 2$, здобудемо по таблиці Додатку $\chi^2_{кр}(0{,}025;2) = 7{,}38$.

В даному випадку емпіричні m_i та теоретичні $\tilde{m}_i$ частоти ряду швидкості вітру на висоті 0,1 км з ймовірністю 97,5% не

відрізняються статистично значуще і емпіричний розподіл можна апроксимувати розподілом Пірсона III типу.

Окрім критерія χ^2 для перевірки зазначеної гіпотези можуть використовуватись й інші критерії, наприклад, критерій Колмогорова, критерій Романовського. Про їх властивості і правила застосування можна прочитати у відповідних довідниках з математичної статистики.

До перевірки статистичних гіпотез ми ще повернемося при розв'язках інших задач статистичного аналізу гідрометеорологічної інформації. Але принципи побудови відповідних критеріїв будуть засновуватися на сформульованих вище теоремах.

5 ІНТЕРВАЛЬНІ ОЦІНКИ ПАРАМЕТРІВ

5.1 Уявлення про довірчий інтервал

Вище ми мали діло з оцінками параметрів, які характеризуються дійсними числами. Кожне число, як відомо, зображається точкою на числовій осі. Тому такі оцінки параметрів називають точечними.

Точечні оцінки розраховуються на основі випадкової вибірки з генеральної сукупності. Тому неможливо отримати уявлення про те, в якій мірі ця оцінка відрізняється від параметра генеральної сукупності, що оцінюється. Таку інформацію можна отримати шляхом інтервального оцінювання.

Задачу інтервального оцінювання у самому загальному виді можна сформулювати таким чином: по даних вибірки побудувати числовий інтервал, відносно якого із наперед вибраною імовірністю можна мовити, що параметр, який оцінюється, знаходиться усередині цього інтервалу. Інтервальне оцінювання особливо необхідне при невеликій кількості спостережень, коли точечна оцінка у значній мірі є випадковою, отже мало надійною.

Довірчим інтервалом $[\theta_1;\theta_2]$ для параметра θ називається, такий інтервал, відносно котрого можна із наперед вибраною імовірністю $p=1-\alpha$ близькою до одиниці, твердити, що він утримує невідоме значення параметра θ. Це визначення можна записати так:

$$P[\theta_1<\theta<\theta_2]=1-\alpha.$$

Чим меншим для вибраної імовірності є $[\theta_1,\theta_2]$, тим точнішу оцінку невідомого параметра θ ми маємо, і навпаки, якщо цей інтервал великий, то оцінка є мало придатною для

практики. Оскільки границі довірчого інтервалу θ_1 і θ_2 залежать від елементів вибірки, тобто ними вони визначаються, то значення θ_1 і θ_2 можуть змінюватися від вибірки до вибірки. Ймовірність $p = 1 - \alpha$ умовилися називати довірчою імовірністю.

5.2 Довірчий інтервал для математичного сподівання

Будемо вважати, що випадкова величина x має нормальний розподіл. Треба знайти інтервальну оцінку математичного сподівання m_x. Як відомо, середнє значення $\overline{x}$ підпорядковується нормальному закону, а S_x^2 - закону розподілу χ^2. Тому, як зазначалося вище, випадкова величина

$$t = \frac{\overline{x} - m_x}{\sqrt{\frac{S_x^2}{n}}} \tag{5.1}$$

або

$$t = \frac{\overline{x} - m_x}{S_x}\sqrt{n} \tag{5.2}$$

підпорядковується розподілу Стьюдента. Вибираючи імовірність $p = 1 - \alpha$ і знаючи об'єм вибірки n, можна знайти $t_{кр}(\alpha, n)$ $(v = n)$ таке, що

$$P[|t| < t_{кр}(\alpha, n)] = 1 - \alpha \ . \qquad (5.3)$$

Підставимо до формули (5.3) співвідношення (5.2). Будемо мати:

$$P\left[\frac{|\bar{x} - m_x|}{S_x}\sqrt{n} < t_{кр}(\alpha, n)\right] = 1 - \alpha \qquad (5.4)$$

або

$$P\left[|\bar{x} - m_x| < t_{кр}(\alpha, n)\frac{S_x}{\sqrt{n}}\right] = 1 - \alpha \ . \qquad (5.5)$$

Нерівність, що розташовується у квадратних дужках, еквівалентна двохсторонній нерівності і тому

$$P\left[-t_{кр}(\alpha, n)\frac{S_x}{\sqrt{n}} < m_x - \bar{x} < t_{кр}(\alpha, n)\frac{S_x}{\sqrt{n}}\right] = 1 - \alpha \qquad (5.6)$$

або

$$P\left[\bar{x} - t_{кр}(\alpha, n)\frac{S_x}{\sqrt{n}} < m_x < \bar{x} + t_{кр}(\alpha, n)\frac{S_x}{\sqrt{n}}\right] = 1 - \alpha \tag{5.7}$$

Але формула (5.7) є визначенням довірчого інтервалу. Отже довірчий інтервал для математичного сподівання визначається нерівністю

$$\bar{x} - t_{кр}(\alpha, n)\frac{S_x}{\sqrt{n}} < m_x < \bar{x} + t_{кр}(\alpha, n)\frac{S_x}{\sqrt{n}}. \tag{5.8}$$

Із співвідношення (5.8) видно, що значення границь довірчого інтервалу залежить, по-перше, від об'єму вибірки n і, по-друге, від мінливості випадкової величини x, яка характеризується оцінкою її середнього квадратичного відхилу S_x. Чим більшим є S_x і чим меншим об'єм вибірки, тим більший довірчий інтервал, і навпаки.

Розглянемо такий приклад. Нехай на метеорологічній станції на основі п'ятидесятирічної вибірки $(n = 50)$ отримані точечні оцінки математичного сподівання середньої місячної температури травня $\bar{T} = 8^0 C$ і середнього квадратичного відхилу $S_T = 6^0 C$. Побудуємо довірчий інтервал для математичного сподівання температури повітря із довірчою імовірністю $p = 0{,}95 \quad (\alpha = 0{,}05)$. За допомогою відповідної таблиці (Додаток 5) знайдемо, що $t_{кр}(0{,}05;50) = 2{,}009$. Отже, довірчий інтервал є

$$\left(8-2{,}009\frac{6}{\sqrt{50}}\right)^0 C < m_T <$$

$$< \left(8+2{,}009\frac{6}{\sqrt{50}}\right)^0 C$$

або

$$6{,}3^0 C < m_T < 9{,}7^0 C.$$

Якщо мінливість температури збільшити у 2 рази $(S_T = 12^0 C)$ при такому ж об'ємі вибірки, то довірчий інтервал буде:

$$4{,}6^0 C < m_T < 11{,}4^0 C$$

Нехай попередні значення $\overline{T}$ і S_T отримані на основі вибірки об'єму $n = 25$. Тоді, оскільки $t_{кр}(0{,}05;25) = 2{,}064$, довірчий інтервал має вид

$$\left(8-2{,}064\frac{6}{5}\right)^0 C < m_T < \left(8+2{,}064\frac{6}{5}\right)^0 C$$

або

$$5{,}5^0 C < m_T < 10{,}5^0 C.$$

5.3 Довірчий інтервал для дисперсії

Нехай випадкова величина x має нормальний розподіл. Треба побудувати довірчий інтервал для дисперсії генеральної сукупності σ_x^2, якщо відома її статистична оцінка S_x^2.

Як було показано в розділі 4, оцінки дисперсії S_x^2 мають χ^2 розподіл з $\nu = n-1$ ступенями волі. Тоді величина $\dfrac{nS_x^2}{\sigma_x^2}$ також підпорядковується χ^2 розподілу з $\nu = n-1$ ступенями волі. За визначенням довірчого інтервалу маємо

$$P\left(\chi_1^2 < \frac{nS_x^2}{\sigma_x^2} < \chi_2^2 \right) = 1-\alpha. \qquad (5.9)$$

Інтервальна імовірність (5.9), як відомо, характеризується площею під кривою $f(\chi^2)$, обмеженною ординатами точок χ_1^2 та χ_2^2. Ця площа, очевидно, дорівнює $1-\alpha$. Ясно, що існує багато варіантів побудови площі такого розміру шляхом пересування точок χ_1^2 і χ_2^2, тобто задача в такій формуліровці характеризується невизначеністю. Тому домовимося вибрати точки χ_1^2 і χ_2^2 так, щоб виконувалася умова

$$P\left(\chi^2 < \chi_1^2\right) = P\left(\chi^2 > \chi_2^2\right) = \frac{\alpha}{2} \qquad (5.10)$$

У таблицях, що містяться у Додатку , отримуються імовірності

$$P\left[\chi^2 > \chi_{кр}^2(v,\alpha)\right] = 1 - \int_0^{\chi_{v,\alpha}^2} f(\chi^2)d(\chi^2). \qquad (5.11)$$

Отже, по цих таблицях при рівні значущості $\frac{\alpha}{2}$ і числу ступенів волі v визначається χ_2^2. Очевидно, оскільки вся площа, що розташована під кривою $f(\chi^2)$, дорівнює одиниці,

$$P(\chi^2 > \chi_1^2) = 1 - P(\chi^2 < \chi_1^2) = 1 - \frac{\alpha}{2}. \qquad (5.12)$$

Отже в таблиці Додатку по числу ступенів волі $v = n - 1$ та рівню значущості $1 - \frac{\alpha}{2}$ знайдемо χ_1^2.

Перетворимо подвійну нерівність

$$\chi_1^2 < \frac{nS_x^2}{\sigma_x^2} < \chi_2^2 . \qquad (5.13)$$

Розглянемо дві нерівності, еквівалентні (5.13):

$$\chi_1^2 < \frac{nS_x^{2\,2}}{\sigma_{x\ \ 2}^2} \quad \text{і} \quad \frac{nS_x^2}{\sigma_x^2} < \chi_2^2 . \qquad (5.14)$$

З них маємо

$$\frac{nS_x^2}{\chi_2^2} < \sigma_x^2 \quad \text{і} \quad \sigma_x^2 < \frac{nS_x^2}{\chi_1^2} . \qquad (5.15)$$

Об’єднуючи ці результати, отримаємо таку подвійну нерівність:

$$\frac{nS_x^2}{\chi_2^2} < \sigma_x^2 < \frac{nS_x^2}{\chi_1^2} \qquad (5.16)$$

Нерівність (5.13) та (5.16) є еквівалентними. Тому рівність (5.9) з
врахуванням цього факту дає

$$P\left(\frac{nS_x^2}{\chi_2^2} < \sigma_x^2 < \frac{nS_x^2}{\chi_1^2}\right) = 1-\alpha, \qquad (5.17)$$

а це є не що інше, як визначення довірчого інтервалу для дисперсії σ_x^2 випадкової величини x. Отже, довірчий інтервал для дисперсії визначається подвійною нерівністю

$$\frac{nS_x^2}{\chi_2^2} < \sigma_x^2 < \frac{nS_x^2}{\chi_1^2}. \qquad (5.18)$$

Для середнього квадратичного відхилу довірчий інтервал має вид:

$$\sqrt{\frac{nS_x^2}{\chi_2^2}} < \sigma_x < \sqrt{\frac{nS_x^2}{\chi_1^2}}.$$

При великому $\nu = n-1$, розподіл $f(\chi^2)$ наближається до нормального розподілу. Отже, при великому об'ємі вибірки n оцінки дисперсії S_x^2 мають розподіл, близький до нормального. Оскільки S_x^2 - незсунена оцінка σ_x^2, то $M[S_x^2] = \sigma_x^2$ і $M(S_x) = \sigma_x$. Вище зазначалося, що

$$\sigma_{S_x} = \frac{S_x}{\sqrt{2(n-1)}}. \qquad (5.19)$$

Тобто йдеться про будову довірчого інтервалу для математичного сподівання σ_x нормально розподільної випадкової величини S_x. Як вже відомо, довірчий інтервал у такому випадку має вид:

$$S_x - t_{кр}(\alpha, v)\sigma_{S_x} < \sigma_x < S_x + t_{кр}(\alpha, v)\sigma_{S_x}, \quad (5.20)$$

де $t_{кр}$ - критерій Стьюдента при рівні значущості α і числі ступенів волі $v = n - 1$.

У якості прикладу розглянемо такі прості задачі.

На основі даних $(n = 31)$ про середню місячну температуру у квітні в Одесі отримана оцінка дисперсії $S_T^2 = 0{,}9(^0C)^2$. Знайти довірчий інтервал для дисперсії температури σ_T^2 при $\alpha = 0{,}05$. Оскільки $\frac{\alpha}{2} = 0{,}025$, а $1 - \frac{\alpha}{2} = 0{,}975$, то маємо $\chi_2^2(0{,}025;30) = 47{,}0$; $\chi_1^2(0{,}975;30) = 16{,}8$. Таким чином,

$$\frac{31 \cdot 0{,}9}{47{,}0} < \sigma_T^2 < \frac{31 \cdot 0{,}9}{16{,}8}$$

і

$$0{,}6(^0C)^2 < \sigma_T^2 < 1{,}7(^0C)^2.$$

Довірчий інтервал для середнього квадратичного відхилу температури має, очевидно, такий вид:

$$0{,}8^0 C < \sigma_T < 1{,}3^0 C.$$

Нехай тепер при таких же вихідних даних збільшується об'єм вибірки $(n = 100)$. Тоді

$$\chi_2^2(0{,}025;99) = 128{,}4;$$
$$\chi_1^2(0{,}975;99) = 73{,}3.$$

Це дає

$$0{,}7(^0C)^2 < \sigma_T^2 < 1{,}2(^0C)^2$$

і

$$0{,}8^0 C < \sigma_T < 1{,}1^0 C$$

Отже, при збільшенні об'єму вибірки довірчий інтервал зменшується.

Побудуємо тепер довірчий інтервал для σ_T по формулі (5.20) маємо

$$\sigma_{S_T} = \frac{S_T}{\sqrt{2(n-1)}} = \frac{\sqrt{0,9}}{\sqrt{2(100-1)}} =$$

$$= \frac{0,95}{14,07} = 0,07$$

$$t_{кр}(0,05;99) = 1,98 \quad .$$

Отже

$$0,95 - 1,98 \cdot 0,07 < \sigma_T < 0,95 + 1,98 \cdot 0,07 \ ,$$

що приводить до інтервалу

$$0,8^0 C < \sigma_T < 1,1^0 C .$$

Порівняння довірчих інтервалів, отриманих по двох різних методах при великих об'ємах вибірок показує, що вони не розрізнюються.

5.4 Інтервальне оцінювання коефіцієнта корреляції та перевірка гіпотези про його значущість

Як і для інших статистичних характеристик, оцінка коефіцієнта корреляції, що отримана по формулі (3.54), є точечною оцінкою генерального коефіцієнта корреляції ρ_{xy}.

Для обгрунтування точності такої оцінки потрібно побудувати довірчий інтервал.

Коефіцієнт корреляції, особливо *коли об'єм вибірки невеликий, не підпорядковується нормальному закону*. Але нелінійне логарифмічне перетворення Фішера

$$\hat{z} = \frac{1}{2}\ln\frac{1+r_{xy}}{1-r_{xy}} \tag{5.21}$$

при невеликих n має розподіл, близький до нормального з параметрами

$$z = \frac{1}{2}\ln\frac{1+\rho_{xy}}{1-\rho_{xy}} + \frac{\rho_{xy}}{2(n-1)} \tag{5.22}$$

та

$$\sigma_z^2 = \frac{1}{n-3}\ , \tag{5.23}$$

де n - об'єм вибірки.

Цей факт дає можливість просто побудувати довірчий інтервал для перетворення z , а від нього за допомогою зворотнього перетворення

$$r_{xy} = \frac{e^{2z} - 1}{e^{2z} + 1} \qquad (5.24)$$

отримати довірчий інтервал для ρ_{xy}. Дійсно, для перетворення z довірчий інтервал має вид

$$\hat{z} - t_{кр}(\alpha, v)\sigma_z < z < \hat{z} + t_{кр}(\alpha, v)\sigma_z, \qquad (5.25)$$

де $t_{кр}(\alpha, v)$ - статистика Стьюдента
або

$$z_1 < z < z_2 , \qquad (5.26)$$

де

$$z_1 = \hat{z} - t_{кр}(\alpha, v)\sigma_z, \qquad (5.27)$$

$$z_2 = \hat{z} + t_{кр}(\alpha, v)\sigma_z . \qquad (5.28)$$

Після обчислення z_1 і z_2 при заданих значеннях рівня значущості α і числа ступенів волі $v = n$ за допомогою формули (5.24) знайдемо відповідні значення r_{xy1} і r_{xy2} - границь довірчого інтервалу для коефіцієнта корреляції. Отже, довірчий інтервал для коефіцієнта корреляції має вид:

$$r_{xy1} < \rho_{xy} < r_{xy2} \ . \qquad (5.29)$$

Якщо величина x і y мають розподіл, близький до нормального, *а об'єм вибірки - великий, то розподіл коефіцієнта кореляції - близький до нормального з параметрами* ρ_{xy} *і* σ_r^2. Тому довірчий інтервал для коефіцієнта кореляції має вид:

$$r_{xy} - t_{кр}(\alpha, v)\sigma_r < \rho_{xy} < r_{xy} + t_{кр}(\alpha, v)\sigma_r . \qquad (5.30)$$

Стандартний відхил коефіцієнта кореляції σ_r визначається формулою:

$$\sigma_r = \frac{1 - r_{xy}^2}{\sqrt{n-1}} \ . \qquad (5.31)$$

Після отримання точечної оцінки коефіцієнта кореляції необхідно оцінити вірогідність статистичного зв'язку між випадковими величинами. Це здійснюється шляхом перевірки відповідної гіпотези. Нульова гіпотеза H_0 формулюється так: коефіцієнт кореляції не є вірогідним $(\rho_{xy} = 0)$ на заданому рівні значущості. Альтернативна гіпотеза H_1 - коефіцієнт кореляції вірогідний. Зазначена гіпотеза перевіряється по t критерію Стьюдента. Підставою для його використання є

теорема 3, що сформульована у розділі 4. На її основі емпіричне значення t критерія визначається формулою :

$$t = \frac{|\hat{z}|}{\sigma_z}, \qquad (5.32)$$

якщо об'єм вибірки малий, або

$$t = \frac{|r_{xy}|}{\sigma_r} \qquad (5.33)$$

при великому об'ємі вибірки.

Гіпотеза H_0 не відхиляється, якщо $t < t_{кр}(\alpha, v)$, де α - рівень значущості, а $v = n - 1$ - число ступенів волі. При $t > t_{кр}(\alpha, v)$ приймається альтернативна гіпотеза H_1, про те, що коефіцієнт корреляції, отриманий по вибірках випадкових величин x і y є вірогідним.

5.5 Довірчий інтервал для коефіцієнтів лінійної регресії та перевірка гіпотези про їх значущість

Коефіцієнти a і b рівняння лінійної регресії

$$\bar{y}(x) = ax + b, \qquad (5.34)$$

як зазначалося, є точечними оцінками параметрів генерального рівняння регресії

$$m_{x/y} = \alpha x + \beta \ . \qquad (5.35)$$

Міру адекватності моделі (5.34) *процесу взаємозв'язку випадкових величин* y *і* x *можна оцінити за допомогою інтервальних оцінок вибіркових параметрів регресії* a *і* b. Будемо вважати, що випадкова величина y має умовний нормальний розподіл з параметрами $m_{x/y}$ та σ_y^2 , причому дисперсія y не залежить від x. У протилежному випадку ми мали б діло із стохастичним зв'язком між y і x.

Покажемо, що коефіцієнт регресії a можна представити у виді лінійної комбінації величин y_i, що дозволе у подальшому використати теорему про те, що лінійна комбінація нормально розподілених випадкових величин підпорядковується нормальному розподілу.
Тобто,

$$a = \sum_{i=1}^{n} k_i y_i \ . \qquad (5.36)$$

Для цього проведемо очевидні перетворення в формулі (3.45), що дає оцінку метода найменших квадратів кутового коефіцієнта рівняння (5.34)

$$a = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2} = \frac{\Sigma x_i y_i - \Sigma y_i \dfrac{\Sigma x_i}{n}}{\Sigma x_i^2 - n\left(\dfrac{\Sigma x_i}{n}\right)^2} =$$

$$= \frac{\Sigma x_i y_i - \bar{x}\Sigma y_i}{\Sigma x_i^2 - n(\bar{x})^2} = \frac{\Sigma x_i y_i - \Sigma \bar{x} y_i}{n[\overline{x^2} - (\bar{x})^2]} = \frac{\Sigma (x_i - \bar{x}) y_i}{nS_x^2} =$$

$$= \sum_{i=1}^{n} \frac{x_i - \bar{x}}{nS_x^2} y_i = \sum_{i=1}^{n} k_i y_i,$$

де

$$k_i = \frac{x_i - \bar{x}}{nS_x^2}. \qquad (5.37)$$

Оскільки випадкові величини y_i мають нормальний розподіл, то їх лінійна комбінація (5.36) також підпорядковується нормальному закону. Тоді довірчий інтервал для її математичного сподівання m_a має вид:

$$a - t_{кр}(p,v)\sigma_a \le m_a \le a + t_{кр}(p,v)\sigma_a, \qquad (5.38)$$

де σ_a - середній квадратичний відхил величини a, $t_{кр}(p,v)$ - параметр Стьюдента.

Знайдемо на основі формули (5.36) математичне сподівання m_a та дисперсію σ_a^2, використовуючи відомі властивості математичного сподівання й дисперсії. Маємо

$$m_a = M\left[\sum_{i=1}^{n} k_i y_i\right] = \sum_{i=1}^{n} M[k_i y_i] = \sum_{i=1}^{n} k_i M[y_i] =$$

$$= \sum_{i=1}^{n} k_i m_{y/x_i} = \sum_{i=1}^{n} k_i(\alpha x_i + \beta) = \frac{1}{nS_x^2} \sum_{i=1}^{n} (x_i - \bar{x}) \times$$

$$\times (\alpha x_i + \beta) = \frac{1}{nS_x^2} \sum_{i=1}^{n} (\alpha x_i^2 - \alpha \bar{x} x_i + \beta x_i - \beta \bar{x}) \quad .$$

(5.39)

Після сумування рівність (5.39) приймає вид :

$$m_a = \frac{1}{S_x^2} \left\{ \alpha \left[\overline{x^2} - (\bar{x})^2 \right] + \beta \bar{x} - \beta \bar{x} \right\} = \alpha \quad . \quad (5.40)$$

Здобутий результат свідчить про те, що *коефіцієнт* a *є незсуненою оцінкою параметра* α.

Знайдемо тепер дисперсію коефіцієнта a

$$\sigma_a^2 = Д(a) = Д(\sum_{i=1}^{n} k_i y_i) = \sum_{i=1}^{n} Д(k_i y_i) =$$

$$= \sum_{i=1}^{n} k_i^2 Д(y_i) = \sum_{i=1}^{n} k_i^2 \sigma_y^2 = \sigma_y^2 \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n^2 S_x^4} =$$

$$= \frac{\sigma_y^2}{nS_x^4} \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n} = \frac{\sigma_y^2}{nS_x^2}$$

або, якщо дисперсію σ_y^2 замінити на її оцінку S_y^2,

$$\sigma_a^2 = \frac{S_y^2}{nS_x^2} \ . \qquad (5.41)$$

Із формули (5.41) видно, *що оцінка* a *є не тільки незсуненою, але й ефективною і умотивованою оцінкою.*

Тепер, повертаючись до нерівності (5.38), ми можемо побудувати довірчий інтервал для коефіцієнта регресії α:

$$a - t_{кр}(p,v)\frac{S_y}{\sqrt{n}S_x} \le \alpha \le a + t_{кр}(p,v)\frac{S_y}{\sqrt{n}S_x} \ . \qquad (5.42)$$

Довірчий інтервал для вільного члена рівняння регресії можна, очевидно, записати так:

$$b - t_{кр}(p,v)\sigma_b \le m_b < b + t_{кр}(p,v)\sigma_b \ . \qquad (5.43)$$

Отже, задача полягає у тому, щоб знайти m_b і σ_b. Для цього використуємо формулу для оцінки метода найменших квадратів коефіцієнта b

$$m_b = M[b] = M[\bar{y} - a\bar{x}] = M[\bar{y}] - M[a\bar{x}] =$$

$$= M\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] - M\left[\frac{1}{n}\sum_{i=1}^{n} ax_i\right] = \frac{1}{n}M\left[\sum_{i=1}^{n} y_i\right] -$$

$$- \frac{1}{n}M\left[\sum_{i=1}^{n} ax_i\right] = \frac{1}{n}\sum_{i=1}^{n} M[y_i] - \frac{1}{n}\sum_{i=1}^{n} M[ax_i] =$$

$$= \frac{1}{n}\sum_{i=1}^{n} m_{y/x=x_i} - \frac{1}{n}\sum_{i=1}^{n} x_i M[a] =$$

$$= \frac{1}{n}\left[\sum_{i=1}^{n}(\alpha x_i + \beta) - \sum_{i=1}^{n}\alpha x_i\right] = \beta$$

Таким чином

$$m_b = \beta. \qquad (5.44)$$

Це означає, що *метод найменших квадратів дає не тільки незсунену оцінку параметра* a*, але і параметра* b *рівняння лінійної регресії.*

Оскільки оцінки a і $\bar{y}$ є некоррельованими, то

$$\sigma_b^2 = Д(b) = Д\left[\bar{y} - a\bar{x}\right] = Д\left[\bar{y}\right] + Д\left[a\bar{x}\right] =$$

$$= Д\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] + Д\left[a\bar{x}\right] = \frac{1}{n^2}\sum_{i=1}^{n} Д\left[y_i\right] +$$

$$+ (\bar{x})^2 Д(a) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_y^2 + (\bar{x})^2\frac{\sigma_y^2}{n\sigma_x^2} =$$

$$= \frac{\sigma_y^2}{n}\left(1 + \frac{(\bar{x})^2}{\sigma_x^2}\right) \quad .$$

(5.45)

Отриманий результат свідчить про те, що b - *не тільки незсунена, а й ефективна і умотивована оцінка коефіцієнта регресії* β, оскільки $\sigma_b^2 \to 0$ при $n \to \infty$. Із формули (5.45) випливає, крім того, що

$$\sigma_b = \frac{\sigma_y}{\sqrt{n}}\left[1 + \frac{1}{c_{v_x}^2}\right]^{1/2}, \qquad (5.46)$$

де

$$c_{v_x} = \frac{\sigma_x}{|\bar{x}|} \quad - \qquad (5.47)$$

коефіцієнт варіації (відносна мінливість) випадкової величини x.

Тепер можна записати довірчий інтервал для коефіцієнта β. Він є

$$b - t_{кр}(p,v)\frac{S_y}{\sqrt{n}}\left[1+\frac{1}{c_{v_x}^2}\right]^{1/2} \le \beta \le$$
$$\le b + t_{кр}(p,v)\frac{S_y}{\sqrt{n}}\left[1+\frac{1}{c_{v_x}^2}\right]^{1/2} \qquad (5.48)$$

Перевірка гіпотези про статистичну значущість коефіцієнтів регресії a і b здійснюється за допомогою критерія Стьюдента

$$t = \frac{|a|}{\sigma_a} \qquad (5.49)$$

і

$$t = \frac{|b|}{\sigma_b}, \qquad (5.50)$$

у якому σ_a і σ_b визначаються, відповідно, рівностями (5.41) і (5.46).

Нульова гіпотеза H_0 формулюється так: с довірчою ймовірністю p коефіцієнт регресії a (чи b) не є вірогідним $\alpha = 0$ (чи $\beta = 0$). Гіпотеза H_0 відхиляється, тобто приймається протилежна гіпотеза H_1, що коефіцієнт a (чи b) є вірогідним, якщо $t > t_{кр}(p, v)$. Число ступенів волі $v = n - 1$.

Можна побудувати *довірчий коридор, у якому з імовірністю* p *розташовується генеральне рівняння регресії* (5.35). Він створюється на такій підставі.

Як вже неодноразово зазначалося, оцінкою цієї лінії є лінійне рівняння

$$\bar{y}(x) = ax + b \ .$$

Оскільки

$$b = \bar{y} + a(x - \bar{x}) \ ,$$

то, очевидно,

$$\bar{y}(x) = \bar{y} + a(x - \bar{x}) \ . \qquad (5.51)$$

Рівність (5.51) - лінійна комбінація нормально розподілених випадкових величин $\bar{y}$ і a. Оскільки вони некоррельовані, то дисперсія

$$Д[\bar{y}(x)] = Д[ax + b] = \frac{\sigma_y^2}{n} + (x - \bar{x})^2 \times$$

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

(5.52)

Звідси,

$$\sigma[\bar{y}(x)] = \sigma(ax + b) = \sigma_y \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{nS_x^2} \right]^{1/2}$$

або, якщо замість σ_y поставити його оцінку S_y,

$$\sigma[\bar{y}(x)] = \sigma(ax + b) = \frac{S_y}{\sqrt{n}} \left[1 + \frac{(x - \bar{x})^2}{S_x^2} \right]^{1/2}.$$

(5.53)

Запишемо тепер довірчий інтервал для умовного математичного сподівання

$$\bar{y}(x) - t_{кр}(p,v)\sigma[\bar{y}(x)] < m_{y/x} < \\ < \bar{y}(x) + t_{кр}(p,v)\sigma[\bar{y}(x)] \quad . \qquad (5.54)$$

З урахуванням рівностей (5.35) і (5.53) остаточно маємо:

$$ax + b - t_{кр}(p,v)\frac{S_y}{\sqrt{n}}\left[1 + \frac{(x-\bar{x})^2}{S_x^2}\right]^{1/2} < \\ < \alpha x + \beta < \qquad , \qquad (5.55) \\ < ax + b + t_{кр}(p,v)\frac{S_y}{\sqrt{n}}\left[1 + \frac{(x-\bar{x})^2}{S_x^2}\right]^{1/2}$$

де p в дужках біля критерія Стьюдента позначає довірчу ймовірність.

На рис. 5.1 у якості прикладу міститься рівняння регресії між меридіональними компонентами швидкості вітру на висоті 20 км в пунктах Уайт Сендз u_{ws} і Канаверал u_k

$$u_{ws} = 0{,}72u_k + 2{,}75 \qquad (5.56)$$

й довірчий коридор з довірчою ймовірністю $p = 0{,}95$. Коефіцієнт корреляції між цими метеорологічними величинами дорівнює 0,71.

6 ЕЛЕМЕНТИ ТЕОРІЇ ВИПАДКОВИХ ПРОЦЕСІВ, ЧАСОВІ ПОСЛІДОВНОСТІ ГІДРОМЕТЕОРОЛОГІЧНИХ ВЕЛИЧИН

6.1 Поняття про випадкову функцію

У попередніх розділах розглядалися закономірності розподілу випадкових величин, тобто величин, можливі значення яких є числа. Однак часто доводиться мати діло з експериментами, результати яких характеризуються не числами, а функцією деякого аргументу. Наприклад, швидкість вітру в деякій точці простору випадковим чином змінюється за часом. Результат вимірювання швидкості вітру за деякий інтервал часу є не число, а функція часу, причому ця функція від досліду до досліду змінюється випадковим чином. Такі функції називаються випадковими функціями.

Випадковій функції можна дати таке визначення: *випадковою функцією* $X(t)$ *аргументу* t *називається функція, ординати якої для будь-яких фіксованих значень аргументу є випадковими величинами.*

Графічний приклад випадкової функції зображається на рис. 6.1. Криві, що відбивають результати окремих експериментів $x_1(t), x_2(t), ..., x_k(t)$ називають реалізаціями випадкової функції. Лінії, що проведені паралельно осі ординат, дають перерізи випадкової функції. Переріз випадкової функції $X(t)$ при значенні аргументу $t = t_i$ дає ряд точок, які відповідають випадковій величині $x_1(t_1), x_2(t_1), ..., x_k(t_1)$, при значенні $t = t_i$: $x_1(t_i), x_2(t_i), ..., x_k(t_i)$. Отже визначення випадкової функції можна сформулювати і таким чином: *випадковою функцією* $X(t)$ *аргументу* t *називається*

функція, перерізи якої при будь-яких значеннях аргументу t_i *дають випадкову величину* $X(t_i)$.

З випадковими функціями доводиться зустрічатися при розв'язуванні задач із самих різноманітних гілок науки та техніки. Взагалі кажучи, кожний параметр стану атмосфери або гідросфери є випадковою функцією координат простору та часу. Різні шуми в радіотехнічних системах є випадковими функціями часу. Велике значення мають методи теорії випадкових функцій у зв'язку з широким застосуванням різних систем автоматичного регулювання, розвитком систем зв'язку, зростаючими вимогами до точності роботи різних систем і приладів. При цьому, методи теорії випадкових функцій повинні давати можливість оцінювати вплив різних випадкових факторів за часом на цікавлючі нас характеристики.

У цьому розділі розглядаються способи опису і аналізу різних випадкових функцій, методи визначення характеристик випадкових функцій, які дають можливість отримати уявлення про фізичні властивості процесів і явищ, що розглядаються дослідником.

Випадкові функції можуть бути скалярними і векторними, у загальному випадку тензорними. Прикладом скалярної випадкової функції може бути температура повітря в деякій точці простору, а векторної - швидкість вітру. Очевидно, векторна випадкова функція може розглядатися як система скалярних випадкових функцій. Взагалі кажучи, швидкість вітру можна розкласти на три складові: зональну, меридіональну та вертикальну. Отже швидкість вітру як векторна випадкова функція складається з трьох скалярних випадкових функцій. Множина T аргументу часто є підмножиною дійсної прямої, при цьому аргумент t може приймати або будь-які дійсні значення у заданому скінченому або нескінченному інтервалі, чи тільки визначені дискретні значення. У першому випадку *функцію* $X(t)$ *називають випадковим процесом,* у другому випадку - *випадковою послідовністю.* У якості аргументу t не обов'язково розглядають час. Можна, наприклад, розглядати атмосферний тиск як випадкову функцію висоти.

У випадку, коли множина T є деяка область в n - вимірному векторному просторі, випадкова функція буде залежати вже не від скалярного, а від векторного аргументу $T(t_1, t_2, ..., t_k)$, у якому $t_1, t_2, ..., t_k$ - координати вектора T. У такому разі її можна розглядати як функцію від k скалярних аргументів $t_1, t_2, ..., t_k$. *Випадкову функцію декількох аргументів називають випадковим полем.*

В метеорології, наприклад, розглядаються поля температури, атмосферного тиску, швидкості вітру, геопотенціалу, тощо. Кожне з цих полів є випадковою функцією трьох координат (x, y, z) і часу t. При цьому, випадкове поле може бути скалярним, наприклад, поле тиску чи поле вітру. В останньому випадку реалізація є векторною функцією.

6.2 Закон розподілу випадкової функції і її імовірнісні характеристики

Розглянемо, яким чином можливо описати випадкову функцію з імовірнісної точки зору.

Випадкова функція вважається повністю визначеною, якщо відомий її закон розподілу, який характеризується або функцією розподілу, або щільністю імовірності.

Зафіксуємо аргументи $t = t_j$ і розглянемо переріз випадкової функції $X(t_j)$. Ми отримаємо випадкову величину $X(t_j) = \{x_1(t_j), x_2(t_j), ..., x_n(t_j)\}$. Вона вважається *повністю визначеною*, якщо знанна її функція розподілу

$$F(x) = P(X < x) \qquad (6.1)$$

або щільність імовірності.

Якщо розглянути систему перерізів випадкової функції, відповідаючих значенням аргументу $t_1, t_2, ..., t_k$, то отримаємо систему випадкових величин.

Система випадкових величин визначена, якщо відома функція розподілу

$$F(x_1, x_2, ..., x_k) = \\ = P(X_1 < x_1; X_2 < x_2; ...; X_k < x_k) \quad (6.2)$$

Ця функція розподілу буде приблизно характеризувати випадковий процес $X(t)$.

Виходячи з цього, *випадковий процес* $X(t)$ *вважається заданим*, якщо для кожного значення t визначена функція розподілу випадкової величини $X(t)$

$$F_1(x,t) = P[X(t) < x], \quad (6.3)$$

для кожної пари значень t_1 і t_2 аргументу t визначена функція розподілу системи випадкових величин $X_1 = X(t_1)$ і $X_2 = X(t_2)$

$$F_2(x_1, x_2, t_1, t_2) = P(X_1 < x_1; X_2 < x_2) \quad (6.4)$$

і взагалі для будь-яких n значень $t_1, t_2, ..., t_n$ аргументу t визначена n - вимірна функція розподілу системи випадкових величин

$$X_1 = X(t_1); X_2 = X(t_2); ...; X_n = X(t_n);$$

$$F(x_1, x_2, ..., x_n; t_1, t_2, ..., t_n) = \\ = P(X_1 < x_1; X_2 < x_2; ...; X_n < x_n) \quad (6.5)$$

Функція $F_1(x,t)$ називається одновимірною функцією розподілу випадкового процесу. Вона характеризує закон розподілу кожного її перерізу, але не відповідає на питання про взаємну залежність між різними перерізами.

Функція $F_2(x_1, x_2; t_1, t_2)$, яка має назву двовимірної функції розподілу випадкового процесу, також не є його вичерпною характеристикою. Для повної характеристики випадкового процесу треба задавати всі багатовимірні функції розподілу.

Для неперервних випадкових функцій, таких, кожний переріз котрих є неперервною випадковою величиною, можна користуватися багатовимірними щільностями імовірностей. Якщо $F_1(x,t)$ має частинну похідну по x

$$\frac{\partial F_1(x,t)}{\partial x} = f_1(x,t), \quad (6.6)$$

то вона називається *одновимірною щільністю розподілу.*

Аналогічно визначаються багатовимірні щільності розподілу випадкових функцій. Якщо існує мішана частинна похідна від n - вимірної функції розподілу

$$\frac{\partial^n F_n(x_1, x_2, ..., x_n; t_1, t_2, ..., t_n)}{\partial x_1 \partial x_2 ... \partial x_n} = \\ = f_n(x_1, x_2, ..., x_n; t_1, t_2, ..., t_n) \ , \tag{6.7}$$

то вона називається n- *вимірною щільністю розподілу* випадкового процесу.

Функції розподілу і щільності розподілу повинні задовольняти вимогам симетрії, тобто повинні бути незмінними при будь-якому вибиранні значень аргументу $t_1, t_2, ..., t_n$.

Із функцій розподілу і щільності розподілу системи n випадкових величин можна отримати функції розподілу будь-якої її підсистеми. Тому, якщо відома n - вимірна функція розподілу або щільність розподілу, то заданими є і всі функції і щільності розподілу більш низького порядку.

Одержання і використання на практиці багатовимірних щільностей і функцій розподілу для описування випадкових процесів при розв'язаннях практичних задач у великій мірі утруднюється і в більшості випадків спряжене з дуже громіздкими математичними перетвореннями. Тому на практиці частіше користуються імовірносними характеристиками випадкових функцій, які аналогічні числовим (статистичним) характеристикам розподілу випадкової величини: математичному сподіванню дисперсії та коваріації (або корреляції). На відміну від числових характеристик випадкових величин, які є постійними числами, характеристики випадкових функцій є функціями її аргументу.

Математичним сподіванням випадкової функції $X(t)$ *називається невипадкова функція* $m_x(t)$ *аргументу* t *, яка*

дорівнює математичному сподіванню значень випадкової функції для кожного фіксованого значення аргументу t.

У відповідності до визначення математичне сподівання випадкової функції має вид:

$$m_x(t) = M[X(t)] = \int_{-\infty}^{\infty} xf(x,t)dx \quad . \tag{6.8}$$

Математичне сподівання випадкової функції має характер *"середньої" функції*, біля якої групуються конкретні реалізації випадкової функції. На рис.6.1 зображені реалізації випадкової функції і її математичне сподівання.

Дисперсією випадкової функції $X(t)$ називається невипадкова функція $Д_x(t)$ аргументу t, яка дорівнює дисперсії значень випадкової функції для кожного фіксованого значення аргументу t. Дисперсія випадкової функції визначається формулою:

$$Д_x(t) = Д[X(t)] = \int_{-\infty}^{\infty} [x - m_x(t)]^2 f(x,t)dx \quad . \tag{6.9}$$

Дисперсія випадкової функції характеризує розкид реалізацій випадкової функції відносно її математичного сподівання. Замість дисперсії випадкової функції часто використовується середній квадратичний відхил випадкової функції, що дорівнює

$$\sigma_x(t) = \sqrt{Д_x(t)} \quad . \tag{6.10}$$

Математичне сподівання і дисперсія є дуже важливими характеристиками випадкової функції, але вони не дозволяють міркувати про характер зв'язку між значеннями випадкової функції при різних значеннях аргументу, оскільки вони повністю визначаються одновимірними щільностями розподілу.

Для характеристики зв'язку між значеннями випадкової функції при різних значеннях аргументу використовується коваріаційна функція або автоковаріаційна функція.

$$K_x(t_i,t_j) = M\{[X(t_i) - m_x(t_i)][X(t_j) -$$

$$- m_x(t_j)]\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[x_i - m_x(t_i)][x_j - m_x(t_j)\times$$

$$\times f(x_i,x_j;t_i,t_j)dx_i dx_j \quad .$$

(6.11)

На рис.6.2 зображається вид коваріаційної функції деякої випадкової функції $X(t)$. Очевидно, графік цієї функції двох аргументів є поверхня. Пояснимо характер цієї поверхні. При значеннях аргументів $t_i = t_j$ коваріаційна функція $K_x(t_i,t_j)$ приймає максимальне значення, оскільки у цьому випадку розглядається зв'язок значення випадкової функції з собою, а такий зв’язок є максимальним. При збільшенні різниці $t_i - t_j$ зав'язок між значеннями випадкової функції для аргументів t_i і

t_j зменшується, відповідно до цього зменшується й коваріаційна функція $K_x(t_i, t_j)$.

Розглядаючи вирази (6.9) і (6.11), можна побачити, що коли аргументи коваріаційної функції дорівнюють одне одному, тобто $t_i = t_j = t$, то

$$K_x(t,t) = Д_x(t) \ . \qquad (6.12)$$

На графіку коваріаційної поверхні ця умова характеризується лінією, що утворюється при діагональному перерізі коваріаційної поверхні.

Коваріаційна функція не змінюється від переставлення аргументів, тобто вона є симетричною функцією своїх аргументів.

На практиці часто використовується *нормована коваріаційна функція,* яка має назву *корреляційної функції*

$$r_x(t_i, t_j) = \frac{K_x(t_i, t_j)}{\sigma_x(t_i)\sigma_x(t_j)} \ . \qquad (6.13)$$

Нормована коваріаційна функція по суті аналогічна коефіцієнту корреляції між випадковими величинами, але залежить від двох аргументів t_i й t_j. Як і коефіцієнт корреляції, корреляційна функція може приймати значення у границях від -1 до +1. Якщо обидва аргументи корреляційної функції дорівнюють одне одному, то $r_x(t,t) = 1$. При збільшенні різниці $t_i - t_j$ корреляційна функція, як правило, зменшується. При сумісному розгляданні декількох випадкових функцій

необхідно враховувати можливий зв'язок між окремими випадковими функціями. Цей зв'язок характеризується коваріаційною функцією зв'язку $K_{xy}(t_i, t_j)$, або взаємною коваріаційною функцією. Щоб підкреслити факт, що розглядається зв'язок між різними перерізами однієї й тієї ж випадкової функції, коваріаційну $K_x(t_i, t_j)$ (а також корреляційну функцію) часто називають *автоковаріаційною* функцією.

Поряд з взаємною коваріаційною функцією використовується взаємна корреляційна функція

$$r_{xy}(t_i, t_j) = \frac{K_{xy}(t_i, t_j)}{\sigma_x(t_i)\sigma_y(t_j)} \quad . \tag{6.14}$$

На відміну від автокорреляційної функції ця функція при $t_i = t_j = t$ може не дорівнювати одиниці.

Імовірнісні характеристики випадкової функції, що розглядалися вище, не є вичерпними характеристиками у загальному випадку, хоча знання цих характеристик, як правило, достатньо для розв'язання практичних задач. Ці характеристики є вичерпними для нормально розподілених випадкових функцій, що часто зустрічаються.

Без доведення приведемо основні властивості випадкової функції.

1. Математичне сподівання суми випадкових функцій дорівнює сумі математичних сподівань функцій

$$M[\sum_{i=1}^{n} X_i(t)] = \sum_{i=1}^{n} m_{x_i}(t) \quad . \tag{6.15}$$

2. Якщо випадкова функція $Z(t) = X(t) + Y(t)$ дорівнює сумі двох випадкових функцій, то

$$K_z(t_i, t_j) = K_x(t_i, t_j) + K_y(t_i, t_j) + K_{xy}(t_i, t_j) + \\ + K_{yx}(t_i, t_j) \quad .$$

(6.16)

Якщо випадкові функції $X(t)$ і $Y(t)$ - некоррельовані між собою, тобто $K_{xy}(t_i, t_j) = 0$, то

$$K_z(t_i, t_j) = K_x(t_i, t_j) + K_y(t_i, t_j) \ . \qquad (6.17)$$

Як частинний випадок, розглянемо характеристики суми випадкової функції і випадкової величини:

$$z(t) = X(t) + Y \quad . \qquad (6.18)$$

За визначенням математичного сподівання маємо:

$$m_z(t) = m_x(t) + m_y \quad . \qquad (6.19)$$

Крім того, формула (6.17) дає

$$K_z(t_i, t_j) = K_x(t_i, t_j) + Д_y \ . \qquad (6.20)$$

3. Нехай випадкова функція $Z(t)$ є сумою випадкової функції $X(t)$ і невипадкової функції $f(t)$, тобто

$$Z(t) = X(t) + f(t) \ , \qquad (6.21)$$

тоді

$$m_z(t) = m_x(t) + f(t) \qquad (6.22)$$

і

$$K_z(t_i, t_j) = K_x(t_i, t_j) \ . \qquad (6.23)$$

4. При множенні випадкової функції на невипадкову функцію математичне сподівання випадкової функції помножується на невипадкову функцію, тобто, якщо

$$z(t) = f(t)X(t) \ , \qquad (6.24)$$

то

$$m_z(t) = M[f(t)X(t)] = f(t)m_x(t) \ . \qquad (6.25)$$

Визначимо коваріаційну функцію такої функції:

$$K_z(t_i,t_j) = M\{[X(t_i)f(t_i) - m_x(t_i)f(t_i)]\times$$
$$\times[X(t_j)f(t_j) - m_x(t_j)f(t_j)]\} = f(t_i)f(t_j)\times$$
$$\times M\{[X(t_i) - m_x(t_i)][X(t_j)m_x(t_j)]\} =$$
$$= f(t_i)f(t_j)K_x(t_i,t_j) \ . \qquad (6.26)$$

Звідси випливає, що при *помноженні випадкової функції на постійну величину* математичне сподівання випадкової функції треба помножити на постійну величину, а коваріаційну функцію на квадрат постійної. Отже, коли

$$Z(t) = aX(t) \ , \qquad (6.27)$$

то

$$m_z(t) = am_x(t) \ ,$$
$$K_z(t_i,t_j) = a^2K_x(t_i,t_j) \ . \qquad (6.28)$$

5. Нехай

$$Y(t) = \frac{dX(t)}{dt} \quad , \tag{6.29}$$

то

$$m_y(t) = \frac{dm_x(t)}{dt} \quad , \tag{6.30}$$

$$K_y(t_i, t_j) = \frac{\partial^2 K_x(t_i, t_j)}{\partial t_i \partial t_j} \; . \tag{6.31}$$

При існуванні похідної необхідна неперервність випадкової функції, тобто наявність статистичного зв'язку між значеннями випадкової функції при достатньо малому приросту аргументу Δt і, крім того, цей зв'язок повинен бути таким, щоб виконувалася рівність:

$$\frac{\partial K_x(t_i, t_j)}{\partial t_i} = \frac{\partial K_x(t_i, t_j)}{\partial t_j} = 0, \text{ при } i = j. \tag{6.32}$$

У всіх практичних випадках виконання умови (6.32) буває достатнім для існування похідної від випадкової функції.

6. Нехай тепер випадкова функція $Y(t)$ дорівнює

$$Y(t) = \int_{\alpha}^{t} X(t)dt \ , \qquad (6.33)$$

де a - довільне число. Тоді

$$m_y(t) = \int_{a}^{t} m_x(t)dt \ , \qquad (6.34)$$

$$K_y(t_i, t_j) = \int_{\alpha}^{t}\int_{\alpha}^{t} K_x(t_i, t_j)dt_i dt_j \ . \qquad (6.35)$$

6.3 Стаціонарні випадкові функції

6.3.1 Поняття про стаціонарну випадкову функцію

Серед многостатності випадкових функцій важливе значення має особливий клас випадкових функцій - стаціонарні випадкові функції.

Стаціонарною випадковою функцією називається така випадкова функція, математичне сподівання якої постійне, а коваріаційна функція залежить тільки від різниці аргументів t_i і t_j, тобто

$$m_x = const \quad , \tag{6.36}$$

$$K_x(t_i, t_j) = K_x(\tau) \quad , \tag{6.37}$$

де

$$\tau = t_j - t_i \ .$$

Це так зване *визначення* стаціонарної функції *у широкому смислі.*

Випадкову функцію називають *стаціонарною у вузькому смислі,* якщо її n - вимірний закон розподілу залежить тільки від різниць аргументів $t_1, t_2, ..., t_n$ і не залежить від самих значень цих аргументів.

Оскільки в більшості випадків ми будемо цікавитися тільки характеристиками випадкової функції $m_x(t)$ і $K_x(\tau)$, то далі будемо розглядати тільки випадкові функції, стаціонарні в широкому смислі.

Стаціонарність випадкової функції *означає*, що випадковий процес протікає однорідне зі змінюванням аргументу t, при цьому стаціонарна випадкова функція повинна існувати в області аргументу $(-\infty, \infty)$.

Як випливає із визначення, *коваріаційна функція* стаціонарної випадкової функції *є функцією тільки одного аргументу* τ, а не двох, як це було у загальному випадку. Це значно спрощує всі операції з характеристиками стаціонарних випадкових функцій.

Дисперсію стаціонарної випадкової функції отримаємо, коли у виразі для коваріаційної функції випадкової функції припустимо, що $t_i = t_j = t$. Тоді маємо:

$$Д_x(t) = K_x(t,t) = K_x(t-t) = K_x(0) = const \tag{6.38}$$

Отже, *дисперсія* стаціонарної випадкової функції є *постійною величиною*, яка дорівнює значенню коваріаційної функції при нульовому значенні аргументу. Оскільки у загальному випадку коваріаційна функція симетрична відносно її аргументів t_i і t_j, то *коваріаційна функція* стаціонарної випадкової функції *є парною функцією*, тобто

$$K(\tau) = K(-\tau) \ . \tag{6.39}$$

На рис.6.3 зображується коваріаційна поверхня стаціонарної випадкової функції.

Як видно, на відміну від коваріаційної нестаціонарної випадкової функції *коваріаційна функція* стаціонарного процесу *постійна* для всіх $t_i - t_j = const$.

Стаціонарні випадкові функції в чистому виді зустрічаються дуже рідко, оскільки для цього необхідно існування однорідності випадкового процесу при нескінченному зміненні аргументу t. Але нас, як правило, цікавить протікання випадкового процесу на невеликому інтервалі значень t і, якщо умови стаціонарності виконуються для цього інтервалу, то ми можемо вважати таку випадкову функцію стаціонарною.

Зустрічаються такі випадкові функції, у яких коваріаційна функція задовольняє умові стаціонарності, але математичне сподівання її не є постійним. Такі властивості мають випадкові гідрометеорологічні процеси, що формуються під дією взбурюючих факторів не тільки випадкового характеру, але й

факторів детермінованих. Останні можуть мати різну періодичність. Наприклад, добовий та річний хід. Зустрічаються періодичності і з іншими часовими характеристиками. Очевидно, такі випадкові функції ми можемо розглядати також як стаціонарні після віднімання із вихідного процесу його математичного сподівання $m(t)$, що є деякою функцією часу. Статистичний аналіз таких нестаціонарних гідрометеорологічних процесів буде розглядатися пізніше.

6.3.2 Ергодична властивість стаціонарної випадкової функції

Більшість стаціонарних випадкових функцій має дуже важливу для практики ергодичну властивість. *Суть ергодичної властивості* полягає у тому, що на основі однієї, достатньо довгої окремої реалізації, можна міркувати про всі властивості випадкової функції, як і по будь-якій кількості реалізацій.

Оскільки розподіл випадкової функції може бути охарактеризовано за допомогою достатньо великої кількості моментів розподілу, а аргументом випадкової функції в більшості випадів є час, то ергодичну властивість стаціонарної випадкової функції можна визначити таким чином: *осереднення за часом від будь-якої реалізації дорівнює осередненню по множині реалізації*. Математично ергодична властивість для моментів першого та другого порядку записується у такому виді:

$$m_x = \int_{-\infty}^{\infty} xf(x,t)dx = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t)dt = \overline{X(t)}\,, \tag{6.40}$$

$$Д_x = \int_{-\infty}^{\infty}(x - m_x)^2 f(x,t)dx =$$

$$= \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}[x(t) - m_x]^2 dt = \overline{[x(t) - m_x]^2},$$

(6.41)

$$K_x(\tau) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_i - m_x)(x_j - m_x)\times$$

$$\times f(x_i, x_j; t_i, t_j)dx_i dx_j =$$

$$= \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}[x(t) - m_x][x(t+\tau) - m_x]dt =$$

$$= \{\overline{[x(t) - m_x][x(t+\tau) - m_x]} \ ,$$

(6.42)

де рискою зверху позначені середні за часом; $x(t)$ - будь-яка реалізація випадкової функції $X(t)$.

Випадкові функції можуть мати властивість ергодичності по відношенню до моментів не всіх порядків. Так, наприклад, випадкова функція може бути ергодичною по відношенню до математичного сподівання і неоргодичною по відношенню до коваріаційної функції. Можуть бути випадкові функції ергодичні по відношенню тільки до двох перших моментів. Оскільки для опису випадкової функції ми користуємося тільки

математичним сподіванням та коваріаційною функцією, то будемо вважати функцію ергодичною у тому випадку, коли умова ергодичності виконується для цих характеристик.

Однак не всі стаціонарні функції мають ергодичну властивість. На рис.6.4 приводяться реалізації деякої стаціонарної випадкової функції $X(t)$. Ясно, що коли ми для находження математичного сподівання будемо користуватися осередненням значень однієї реалізації по аргументу t, то отриманий результат буде залежати від вибраної реалізації і взагалі мати значення, відмінне від математичного сподівання випадкової функції, тобто функція не підпорядковується ергодичній властивості. Можна показати, що випадкова функція має ергодичну властивість, якщо її коваріаційна функція $K_x(\tau)$ прагне до нуля при безмежному зростанні τ. Ця умова є достатньою для того, щоб випадкова функція мала ергодичну властивість відносно моментів другого порядку (тобто дисперсії та коваріаційної функції). Випадкова функція буде мати ергодичну властивість відносно математичного сподівання не тільки при виконанні умови $K_x(\tau) \to 0$ при $\tau \to \infty$, але і у тому випадку, коли $K_x(\tau)$ при великих τ має коливальний характер, але при цьому середнє значення $K_x(\tau)$ дорівнює нулю.

Для прикладу покажемо, що випадкова функція, яка дорівнює сумі випадкової функції і незалежної від неї випадкової величини, вже не має ергодичної властивості.

Дійсно, у цьому разі маємо:

$$K_z(\tau) = M[\dot{z}(t)\dot{z}(t+\tau)] = M\{[\dot{x}(t) + \dot{y}] \times$$
$$\times (\dot{x}(t+\tau) + \dot{y}]\} = K_x(\tau) + Д_y \quad ,$$

де

$$\dot{z}(t) = \dot{x}(t) + \dot{y}, \quad \dot{x}(t) = x(t) - m_x .$$

Коваріаційна функція $K_z(\tau)$ при збільшенні τ (якщо $\dot{x}(t)$ ергодична функція), прагне не до нуля, а до дисперсії випадкової величини y. Отже, відповідно до зазначеної вище умови ергодичності, випадкова функція $z(t)$ вже не має ергодичної властивості.

6.3.3 Спектральне розкладання стаціонарної випадкової функції

6.3.3.1 Лінійчатий спектр стаціонарної випадкової функції

Спектральним розкладанням деякої функції називається зображення її у виді суми гармонічних коливань, які мають різні амплітуди гармонік. Залежність амплітуд від частоти називається *спектром функції*. Для втілення спектрального розкладання використовується перетворення Фурьє.

Розглянемо яким чином можна отримати спектральне розкладання стаціонарної випадкової функції.

На рис.6.5 умовно зображається випадкова функція $X(t)$ на інтервалі $(-T;T)$, яку будемо позначати $X_T(t)$. Ця функція може бути розкладеною у ряд Фурьє.

$$X_T(t) = \sum_{i=0}^{\infty} (X_{1i} \cos \omega_i t + X_{2i} \sin \omega_i t) , \qquad (6.43)$$

де $\omega_i = i\Delta\omega; \Delta\omega = \dfrac{2\pi}{2T} = \dfrac{\pi}{T};$ X_{1i} і X_{2i} - випадкові коефіцієнти ряду Фурьє.

Визначимо дисперсію випадкової функції $X_T(t)$:

$$
\begin{aligned}
Д_{X_T} &= M\{\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}(X_{1i}\cos\omega_i t + X_{2i}\sin\omega_i t)\times \\
&\times(X_{1j}\cos\omega_i t + X_{2j}\sin\omega_j t)\} = \\
&= \sum_{i=0}^{\infty}(M[X_{1i}{}^2]\cos^2\omega_i t + M[X_{2i}{}^2]\sin^2\omega_i t) + \\
&+ \{\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}(M[X_{1i}X_{1j}]\cos\omega_i t\cos\omega_j t + \\
&+ M[X_{2i}X_{2j}]\sin\omega_i t\sin\omega_j t) + \\
&+ \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}(M[X_{1i}X_{2j}]\cos\omega_i t\sin\omega_j t + \\
&+ M[X_{2i}X_{1j}]\sin\omega_i t\cos\omega_j t)\} \quad i \neq j \quad .
\end{aligned}
\tag{6.44}
$$

Оскільки функції $\cos\omega t$ і $\sin\omega t$ - є ортогональними, останній член рівності (6.44) дорівнює нулю. Крім того, функція $X(t)$ - стаціонарна. Тому дисперсія її повинна бути постійною, тобто незалежною від аргументу t. Як видно, це можливо, коли виконуються умови:

$$M[X_{1i}{}^2] = M[X_{2i}{}^2] = Д_i \ , \tag{6.45}$$

$$M[X_{1i}X_{1j}]=M[X_{2i}X_{2j}]=0 \quad . \qquad (6.46)$$
$$i\neq j$$

Якщо запровадити ці умови до рівності (6.44), то будемо мати

$$Д_{X_T} = \sum_{i=0}^{\infty} Д_i \quad . \qquad (6.47)$$

Умова (6.46) є не що інше, як умова незв'язності коефіцієнтів розкладення (6.43).

Отримаємо тепер коваріаційну функцію випадкової функції $X_T(t)$. Очевидно, з урахуванням умов (6.45), (6.46) і (6.47) маємо:

$$K_{X_T}(t_v - t_\mu) = \sum_{i=0}^{\infty} \{M[X_i{}^2 \cos\omega_i t_v \cos\omega_i t_\mu] +$$
$$+ M[X_{2i}{}^2 \sin\omega_i t_v \sin\omega_i t_\mu) =$$
$$= \sum_{i=0}^{\infty} Д_i(\cos\omega_i t_v \cos\omega_i t_\mu + \sin\omega_i t_v \sin\omega_i t_\mu) =$$
$$= \sum_{i=0}^{\infty} Д_i \cos\omega_i(t_v - t_\mu) \quad .$$

(6.48)

Позначимо $t_v - t_\mu = \tau$. Тоді приходимо до рівняння:

$$K_{X_T}(\tau) = \sum_{i=0}^{\infty} Д_i \cos \omega_i \tau \ . \qquad (6.49)$$

Воно є не що інше, як розкладання парної функції $K_{X_T}(\tau)$ у ряд Фурьє на інтервалі $[-T, T]$. Як відомо, коефіцієнти розкладання визначаються формулами:

$$Д_0 = \frac{1}{T} \int_0^T K_{X_T}(\tau) d\tau \ , \qquad (6.50)$$

$$Д_i = \frac{1}{2T} \int_{-2T}^{2T} K_{X_T}(\tau) \cos \omega_i \tau d\tau \ ,$$

$$i = 1,2,3,... \ . \qquad (6.51)$$

Таким чином, для опису стаціонарної випадкової функції можна використовувати залежність дисперсії коефіцієнтів розкладення випадкової функції в ряд Фурьє $Д_i$ від частоти ω_i замість коваріаційної функції $K_{X_T}(\tau)$. Залежність коефіцієнтів $Д_i$ від частоти ω_i називається спектром випадкової функції. Якщо випадкова функція розглядається на обмеженному інтервалі аргументів, то частота приймає дискретні значення з проміжком $\Delta\omega$. У цьому випадку *спектр випадкової функції* називається *дискретним* або *лінійчатим*. Приклад лінійчатого спектру приводиться на рис.6.6. Лінійчатий спектр отримується у разі розкладання в ряд Фурьє l - періодичної випадкової функції.

6.3.3.2 Спектральна щільність стаціонарної випадкової функції

Спектральне розкладання випадкової функції на обмеженому інтервалі $[-T;T]$ дає лише наближений опис випадкової функції. Більш повне уявлення про випадкову функцію при її спектральному розкладанні можна отримати при збільшенні T, тобто при $T \to \infty$. У цьому випадку частота із дискретної величини перетворюється на безперервну, а замість дисперсій амплітуд для кожної частоти необхідно розглядати щільність дисперсій амплітуд на одиницю частоти. Отже дискретний спектр перетворюється у безперервний.

Знайдемо відношення $\frac{Д_i}{\Delta\omega}$, яке має сенс середньої щільності дисперсій на одиницю частоти. Позначимо її через $S_{X_T}(\omega_i)$ Таким чином,

$$S_{X_T}(\omega_i) = \frac{Д_i}{\Delta\omega} \qquad (6.52)$$

або з урахуванням рівняння (6.51), а також залежності між інтервалом дискретності частоти $\Delta\omega$ і інтервалом задання процесу T,

$$S_{X_T}(\omega_i) = \frac{1}{2\pi}\int_{-2T}^{2T} K_{X_T}(\tau)\cos\omega_i\tau d\tau \ . \qquad (6.53)$$

Звідси, при $T \to \infty$ маємо:

$$S_X(\omega) = \lim_{\substack{T\to\infty \\ (\Delta w\to 0)}} S_{X_T}(\omega_i) = \\ = \frac{1}{2\pi}\int_{-\infty}^{\infty} K_X(\tau)\cos\omega\tau d\tau \quad . \tag{6.54}$$

Функція $S_X(\omega)$ визначає щільність розподілу дисперсії гармонічних коливань у залежності від частоти і тому називається *спектральною щільністю випадкової функції.*

За допомогою формули (6.54) спектральна щільність стаціонарної випадкової функції однозначно визначається її коваріаційною функцією. Справедливим виявляється й зворотне перетворення Фурьє, яке дає залежність між коваріаційною функцією й спектральною щільністю випадкової функції:

$$K_X(\tau) = \int_{-\infty}^{\infty} S_X(\omega)\cos\omega\tau d\omega \; . \tag{6.55}$$

Рівності (6.54) і (6.55) є частинними випадками перетворення Фурьє - "косинус перетвореннями Фурьє". У загальному випадку для функцій $S_X(\omega)$ і $K_X(\tau)$ визначається такі співвідношення:

$$S_X(\omega) == \frac{1}{2\pi}\int_{-\infty}^{\infty} K_X(\tau)e^{-i\omega\tau} d\tau, \tag{6.56}$$

$$K_X(\tau) = \int_{-\infty}^{\infty} S_X(\omega) e^{i\omega\tau} d\tau \ . \qquad (6.57)$$

Оскільки за відомою формулою Ейлера

$$e^{-i\omega\tau} = \cos\omega t - i\sin\omega t \ ,$$

$$e^{i\omega\tau} = \cos\omega t + i\sin\omega t \ ,$$

а функції $K_x(\tau)$ і $S_x(\omega)$ - парні, то формули (6.54), (6.55) і (6.56), (6.57) еквівалентні.

Розглянемо основні властивості спектральної щільності стаціонарної випадкової функції.

1) Спектральна щільність є функція парна, тобто

$$S_x(\omega) = S_x(-\omega) \ . \qquad (6.58)$$

2) Інтеграл від спектральної щільності по всіх значеннях частоти дорівнює дисперсії випадкової функції. Дійсно, вважаючи що $\tau = 0$, із (6.55) маємо:

$$Д_x = \int_{-\infty}^{\infty} S_x(\omega) d\omega \ . \qquad (6.59)$$

3) Спектральну щільність випадкової функції можна розглядати як *"енергетичний спектр"* випадкової функції. Це пояснюється тим, що в якості випадкової функції часто розглядаються такі величини, як швидкість вітру, швидкість течії в океані і напруження. Тоді розподіл дисперсій, які мають квадрат розмірності амплітуди, пропорційний щільності розподілу енергії сигналу по частотах.

4) Подібно нормованій коваріаційній функції (корреляційній функції) може використовуватися і нормована спектральна щільність, яка визначається таким чином:

$$S(\omega) = \frac{S_x(\omega)}{Д_X} \quad . \tag{6.60}$$

Нормована спектральна щільність зв'язана з корреляційною функцією такими ж співвідношеннями, як і спектральна щільність $S_x(\omega)$ з коваріаційною функцією $K_x(\tau)$.

Нормована спектральна щільність має всі властивості, як і спектральна щільність випадкового процесу. Крім того

$$\int_{-\infty}^{\infty} S(\omega)d\omega = \int_{-\infty}^{\infty} \frac{S_x(\omega)d\omega}{Д_x} =$$

$$= \frac{1}{Д_x}\int_{-\infty}^{\infty} S_x(\omega)d\omega = \frac{Д_x}{Д_x} = 1$$

5) При збільшенні масштабу аргументу коваріаційної функції масштаб аргументу спектральної щільності і сама спектральна щільність зменшується у таке ж саме число разів. Іншими словами, якщо коваріаційній функції $K_X(\tau)$

відповідає спектральна щільність $S_X(\omega)$, то коваріаційній функції з аргументом $K_X(\alpha\tau)$ відповідає спектральна щільність $\frac{1}{\alpha}S_X\left(\frac{\omega}{\alpha}\right)$.

Дійсно, якщо позначити $z = \alpha\tau$, то будемо мати

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} K(\alpha\tau)\cos\omega\tau d\tau =$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} K(z)\cos\frac{\omega\tau}{\alpha}d\frac{z}{\alpha} = \frac{1}{\alpha}\times$$

$$\times\frac{1}{2\pi}\int_{-\infty}^{\infty} K_X(z)\cos\frac{\omega}{\alpha}zdz = \frac{1}{\alpha}S_X\left(\frac{\omega}{\alpha}\right) \tag{6.61}$$

Ця властивість значить, що при стискуванні $(\alpha < 1)$ коваріаційної функції вздовж осі τ спектр випадкової функції розширюється і навпаки (при $\alpha > 1$). Така поведінка функції $K_X(\tau)$ і $S_X(\omega)$ фізично пояснюється тим, що при переваженні у сигналі високих частот коваріаційна функція швидко зменшується, а спектральна щільність розтягується в сторону високих частот; якщо у сигналі переважують низькі частоти, то маємо обернену картину.

Використовуючи формулу (6.54) або (6.56), можна отримати аналітичний вираз для спектральної щільності, якщо удається апроксимувати коваріаційну функцію прийнятною формулою. Приведемо деякі важливі приклади.

1. Нехай стаціонарний випадковий процес має коваріаційну функцію

$$K(\tau) = Д_x e^{-\alpha|\tau|}, \quad \alpha > 0 . \tag{6.62}$$

Відповідно до рівності (6.56), спектральна щільність дорівнює:

$$\begin{aligned} S_x(\omega) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} Д_x e^{-\alpha|\tau|} e^{-i\omega\tau} = \\ &= \frac{Д_x}{2\pi}\left\{\int_{-\infty}^{0} e^{(\alpha - i\omega)\tau} d\tau + \int_{0}^{\infty} e^{-(\alpha + i\omega)\tau} d\tau\right\} = \\ &= \frac{Д_x}{2\pi}\left[\frac{1}{\alpha - i\omega} + \frac{1}{\alpha + i\omega}\right] = \frac{Д_x \alpha}{\pi(\alpha^2 + \omega^2)} . \end{aligned} \tag{6.63}$$

Як видно з (6.63), це парна функція, яка досягає найбільшого значення $\frac{Д_x}{\pi\alpha}$ при $\omega = 0$. Але, як і коваріаційна функція (6.62) у точці $\tau = 0$, спектральна щільність (6.63) не має похідної у точці $\omega = 0$.

Величина α, як випливає з рівності (6.62), має сенс масштабу аргументу, при якому коваріаційна функція зменшується в e разів. Тому її називають *декрементом затухання коваріаційної функції.*

На рис.6.7 приводяться графіки корреляційних функцій що відповідають коваріаційній функції (6.62) і відповідні спектральні щільності. Порівнювання кривих показує, що при малих α корреляційна функція зменшується відносно повільно, ніж при великих, а спектральна щільність зменшується зі збільшенням частоти ω швидко. Це свідчить про те, що у

спектрі випадкового процесу мають місце малі частоти (або великі масштаби флуктуацій фізичної величини). Процеси такого типу мають назву *вузькосмужного*, оскільки енергія такого процесу зосереджена у вузькій смузі частот. Отже, вузькосмужному процесу відповідає великий час корреляції, тобто наявність корреляційного зв'язку між перерізами випадкового процесу, який повільно зменшується зі збільшенням інтервалу між перерізами.

При збільшенні α, тобто із зменшенням часу корреляції, щільність змінюється більш плавно. Для великих α спектральна щільність змінюється при зростанні ω дуже повільно. Такі процеси називають *широкосмужними*. Вони характеризуються ш в и д к и м падінням корреляційного зв'язку між перерізами випадкового процесу.

Розглядається випадковий процес, для якого спектральна щільність є величиною постійною на всьому інтервалі частот: $S_x(\omega) = S_x(0) = const$. Його називають *"білим шумом"* по аналогії з білим світлом, який утворюється при рівномірному змішуванні всіх кольорів видимого спектру. Такий випадковий процес відзначається рівномірним розподілом енергії по всіх частотах. Реально такі процеси не існують, оскільки енергія такого процесу (інакше кажучи його дисперсія) повинна бути нескінченною, але вони є зручною абстракцією для тих випадків, коли спектральна щільність приблизно постійна на діапазоні частот, що нас цікавить.

Коваріаційна функція "білого шуму" дорівнює нулю всюди, крім точки $\tau = 0$, де, як вже відомо, вона дорівнює дисперсії $Д_x$. Це означає, що для "білого шуму" цілком є відсутнім корреляційний зв'язок між значеннями випадкової функції при будь-яких значеннях аргументу t.

2. Як відзначалося, коваріаційна функція (6.62) не має похідної в точці $\tau = 0$. При розв'язуванні деяких задач це є суттєвою завадою. Тому кращою апроксимацією (якщо це можливо) є апроксимація для коваріаційної функції

$$K_x(\tau) = Д_x e^{-\alpha^2 \tau^2}, \quad \alpha > 0 \tag{6.64}$$

Тоді

$$\begin{aligned} S_x(\omega) &= \frac{Д_x}{2\pi} \int_{-\infty}^{\infty} e^{-\alpha^2 \tau^2} e^{-i\omega\tau} d\tau = \\ &= \frac{Д_x}{2\pi} \int_{-\infty}^{\infty} e^{-(\alpha^2 \tau^2 + i\omega\tau)} d\tau \quad . \end{aligned} \tag{6.65}$$

Доповнимо показник степеня до повного квадрата:

$$\alpha^2 \tau^2 + i\omega\tau = (\alpha\tau + i\frac{\omega}{2\alpha})^2 + \frac{\omega^2}{4\alpha^2} \tag{6.66}$$

і підставимо результат (6.66) до рівності (6.65). Будемо мати:

$$S_x(\omega) = \frac{Д_x}{2\pi} \left(\int_{-\infty}^{\infty} e^{-(\alpha\tau + i\frac{\omega\tau}{2\alpha})^2} d\tau \right) e^{-\frac{\omega^2}{4\alpha^2}}. \tag{6.67}$$

Інтеграл у дужках формули (6.67) є інтеграл Пуассона і дорівнює $\frac{\sqrt{\pi}}{\alpha}$;

Отже,

$$S_x(\omega) = \frac{Д_x}{2\sqrt{\pi}\alpha} e^{-\frac{\omega^2}{4\alpha^2}} . \qquad (6.68)$$

На рис.6.8 містяться корреляційна функція й спектральна щільність такого випадкового процесу при різних α.

3. Розглянемо тепер випадкову функцію з лінійною коваріаційною функцією. Нехай коваріаційна функція описується на інтервалі від 0 до τ_0 лінійною функцією

$$K_x(\tau) = \begin{cases} Д_x\left(1 - \frac{\tau}{\tau_0}\right), & при \quad 0 < \tau < \tau_0, \\ 0, & при \quad \tau \geq \tau_0 . \end{cases} \qquad (6.69)$$

Графік такої функції приводиться на рис.6.9. Вона може використовуватися для апроксимації швидко згасаючої монотонної коваріаціоної функції.

Розрахуємо спектральну щільність цієї випадкової функції

$$S_x(\omega) = \frac{1}{\pi}\int_0^{\infty} K_x(\tau)\cos\omega\tau d\tau = \frac{Д_x}{\pi}\int_0^{\tau_0}(1-\frac{\tau}{\tau_0})\times$$

$$\times\cos\omega\tau d\tau = \frac{Д_x}{\pi\omega}\int_0^{\tau_0}(1-\frac{\tau}{\tau_0})d\sin\omega\tau =$$

$$= \frac{Д_x}{\pi\omega}\left[\int_0^{\tau_0} d\sin\omega\tau - \frac{1}{\tau_0}\int_0^{\tau_0}\tau\cos\omega\tau d\tau\right] =$$

$$= \frac{Д_x}{\pi\omega^2\tau_0}(1-\cos\omega\tau_0) \quad . \tag{6.70}$$

Графік спектральної щільності (6.70) приводиться на рис.6.9. Аналізуючи функції (6.69) і (6.70) разом, можна побачити, що коли коваріаційна функція стискується (τ_0 зменшується), то спектральна щільність розтягується, і навпаки.

4. Дуже часто при дослідженні гідрометеорологічних процесів ми отримуємо коваріаційну функцію, яка колихаючись згасає по експоненціальному закону. Прикладом такої коваріаційної функції може бути така функція:

$$K_x(\tau) = Д_x e^{-\alpha|\tau|}\cos\omega_0\tau \ . \tag{6.71}$$

Знайдемо спектральну щільність такої випадкової функції:

$$S_x(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} Д_x e^{-\alpha|\tau|} \cos\omega_0 \tau e^{-i\omega\tau} d\tau \ . \quad (6.72)$$

Як видно, ми використали загальну форму перетворення Фурьє для коваріаційної функції. Використовуючи відомі формули Ейлера, будемо мати:

$$\cos\omega_0\tau = \frac{e^{i\omega_0\tau} + e^{-i\omega_0\tau}}{2} \ . \quad (6.73)$$

Тоді

$$S_x(\omega) = \frac{Д_x}{4\pi} \int_{-\infty}^{\infty} \left(\begin{array}{l} e^{-\alpha|\tau| - i\omega\tau + i\omega_0\tau} + \\ + e^{-\alpha|\tau| - i\omega\tau - i\omega_0\tau} \end{array} \right) d\tau =$$

$$= \frac{Д_x}{4\pi} \left[\begin{array}{l} \int_{-\infty}^{0} \left(\begin{array}{l} e^{\alpha\tau - i(\omega_0 - \omega)\tau} + \\ + e^{\alpha\tau - i(\omega_0 + \omega)\tau} \end{array} \right) d\tau + \\ + \int_{0}^{\infty} \left(\begin{array}{l} e^{-\alpha\tau - i(\omega_0 - \omega)\tau} + \\ + e^{-\alpha\tau - i(\omega_0 + \omega)\tau} \end{array} \right) d\tau \end{array} \right] =$$

$$= \frac{Д_x\alpha}{2\pi} \left[\frac{1}{\alpha^2 + (\omega_0 - \omega)^2} + \frac{1}{\alpha^2 + (\omega_0 + \omega)^2} \right] =$$

$$= \frac{Д_x\alpha}{\pi} \frac{\alpha^2 + \omega_0^2 + \omega^2}{(\alpha^2 + \omega_0^2 + \omega^2) - 4\omega_0^2\omega^2} \ . \quad (6.74)$$

Характер цієї функції залежить від співвідношення між параметрами α і ω_0. Якщо коваріаційна функція має різко виражений коливальний характер, тобто ω_0 - велике, то у спектральної щільності $S_x(\omega)$ буде спостерігатися максимум приблизно в точці $\omega = \omega_0$ (рис.6.10). Якщо основну роль в коваріаційній функції $K_x(\tau)$ грає показникова функція, тобто α - велике, то функція $S_x(\omega)$ буде мати максимальне значення в точці $\omega = 0$ (рис.6.10).

На практиці, одначе, зустрічаються випадкові функції, ковариаційні функції яких не можуть бути з достатньою точністю апроксимованими аналітичними виразами. Тому для розрахування спектральних щільностей використовуються чисельні методи, основні риси яких будуть розглянуті нижче.

6.4 Взаємний спектральний аналіз випадкових процесів

Для системи стаціонарних випадкових процесів $X_1(t), X_2(t), ..., X_n(t)$, крім спектральних щільностей $S_{x_i}(\omega)$ кожного процесу, розглядаються й взаємні спектральні щільності $S_{x_i x_j}(\omega)$, які є перетвореннями Фурьє від відповідних взаємних коваріаційних функцій:

$$S_{x_i x_j}(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} K_{x_i x_j}(\tau) e^{-i\omega\tau} d\tau \qquad (6.75)$$

Навпаки, взаємні коваріаційні функції є зворотними перетвореннями Фурьє від взаємних спектральних щільностей:

$$K_{x_i x_j}(\tau) = \int_{-\infty}^{\infty} S_{x_i x_j}(\omega) e^{i\omega\tau} d\omega. \qquad (6.76)$$

Взаємна коваріаційна функція не має властивостей парності. Позначимо через $K^{(+)}_{x_i x_j}(\tau)$ - *парну частину взаємної коваріаційної функції* $K_{x_i x_j}(\tau)$, а через $K^{(-)}_{x_i x_j}(\tau)$ - її *непарну частину*. Отже,

$$K_{x_i x_j}(\tau) = K^{(+)}_{x_i x_j}(\tau) + K^{(-)}_{x_i x_j}(\tau), \qquad (6.77)$$

де

$$K^{(+)}_{x_i x_j}(\tau) = \frac{1}{2}\left[K_{x_i x_j}(\tau) + K_{x_i x_j}(-\tau)\right], \qquad (6.78)$$

$$K^{(-)}_{x_i x_j}(\tau) = \frac{1}{2}\left[K_{x_i x_j}(\tau) - K_{x_i x_j}(-\tau)\right]. \qquad (6.79)$$

Підставляючи (6.77) в (6.75) та використовуючи відому формулу Ейлера, отримаємо:

$$S_{x_i x_j}(\omega) = C_{x_i x_j}(\omega) - iQ_{x_i x_j}(\omega), \tag{6.80}$$

де

$$C_{x_i x_j}(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} K^{(+)}_{x_i x_j}(\tau)\cos\omega\tau d\tau \tag{6.81}$$

називається *ко-спектром*,

$$Q_{x_i x_j}(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} K^{(-)}_{x_i x_j}(\tau)\sin\omega\tau d\tau \tag{6.82}$$

квадратурним спектром випадкових процесів $X_i(t)$ і $X_j(t)$. Ко-спектр, як косинус - перетворення Фурьє від парної функції $K^{(+)}_{x_i x_j}(\tau)$, є парною функцією. Якщо підставити (6.80) у (6.76) і застосувати для функції $e^{i\omega t}$ формулу Ейлера, то отримаємо формулу для *взаємної коваріаційної функції :*

$$K_{x_i x_j}(\tau) = \int_{-\infty}^{\infty} C_{x_i x_j}(\omega)\cos\omega\tau d\omega + \int_{-\infty}^{\infty} Q_{x_i x_j}(\omega)\sin\omega\tau d\omega \quad . \tag{6.83}$$

Нехай у формулі (6.83) $\tau = 0$. Тоді :

$$K_{x_i x_j}(0) = \int_{-\infty}^{\infty} C_{x_i x_j}(\omega) d\omega \ . \qquad (6.84)$$

Формула (6.42) свідчить про те, що ко-спектр дає розкладання по різних частотах взаємної коваріаційної функції двох випадкових процесів при нульовому зсуві аргументу і має смисл середнього добутку процесів X_i та X_j у вузькому інтервалі частот $\omega + d\omega$, поділеному на частотний інтервал.

На рис.6.11 приводиться ко-спектр зональних індексів вітру на рівні 500 гПа на широтах 40^0 і 60^0 ш. Можна бачити, що самий великий внесок у взаємну коваріацію між вітрами на широтах 40^0 і 60^0 с.ш. дають періоди біля 25 днів. Як відомо, коваріації між швидкостями вітру на цих широтах від'ємні, тому всі значення ко-спектру у цьому випадку від'ємні або близькі до нуля.

Квадратурний спектр $Q_{x_i x_j}(\omega)$ дає внесок різних гармонік у сумарну коваріацію у випадку, коли всі гармоніки часової послідовності $X_i(t)$ зсунуті по фазі на чверть періоду назад, а послідовність $X_j(t)$ залишається незмінною. Дійсно, коли $\tau = \frac{\pi}{2\omega} = \frac{T}{4}$, то формула (6.83) дає :

$$K_{x_i x_j}(\frac{T}{4}) = \int_{-\infty}^{\infty} Q_{x_i x_j}(\omega) d\omega \ . \qquad (6.85)$$

Отже ми отримали такий результат:
Якщо у формулі (6.83) $\tau = 0$. Тоді

$$K_{x_i x_j}(0) = \int_{-\infty}^{\infty} C_{x_i x_j}(\omega) d\omega. \qquad (6.86)$$

Звідси бачимо, що ко-спектр дає розкладення по різних частотах взаємної коваріаційної функції двох випадкових процесів при нульовому зсуві аргументу. Аналогічно, вважаючи $\tau = \frac{\pi}{2\omega} = \frac{T}{4}$, отримаємо:

$$K_{x_i x_j}(\frac{T}{4}) = \int_{-\infty}^{\infty} Q_{x_i x_j}(\omega) d\omega. \qquad (6.87)$$

Отже, квадратурний спектр характеризує внесок у загальну взаємну корреляцію двох випадкових процесів гармонік, що в них утримуються, при зсуві фаз цих гармонік на чверть періоду.

Комплексну функцію (6.80) можна записати у показниковій формі

$$S_{x_i x_j}(\omega) = \left| S_{x_i x_j}(\omega) \right| e^{\psi_{x_i x_j}(\omega)}. \qquad (6.88)$$

Модуль взаємної спектральної щільності :

$$\left| S_{x_i x_j}(\omega) \right| = \sqrt{C^2_{x_i x_j}(\omega) + Q^2_{x_i x_j}(\omega)} \qquad (6.89)$$

називають *амплітудним спектром*, а функцію

$$\psi_{x_i x_j}(\omega) = arctg\left[-\frac{Q_{x_i x_j}(\omega)}{C_{x_i x_j}(\omega)}\right] \qquad (6.90)$$

фазовим спектром.

При частотному зображенні процесів з'являється можливість порівняння взаємної енергії на фіксованій частоті з енергіями кожного з процесів на цій же частоті шляхом визначення співвідношення:

$$\Gamma(\omega) = \frac{C^2_{x_i x_j}(\omega) + Q^2_{x_i x_j}(\omega)}{S_{x_i}(\omega) S_{x_j}(\omega)}. \qquad (6.91)$$

Величина

$$\gamma(\omega) = \sqrt{\Gamma(\omega)} \qquad (6.92)$$

має *сенс спектрального коефіцієнта взаємної корреляції* процесів $X_i(t)$ і $X_j(t)$, який визначає тісноту корреляційного зв'язку між цими процесами на фіксованих частотах. Вона має назву *когерентності* й може приймати значення від 0 до 1.

На рис.6.12 міститься когерентність між меридіональною складовою швидкості вітру на висотах 10 і 30 км в пункті острова Воллон (Північна Америка). Видно, що спостерігається високий корреляційний зв'язок $\gamma(n) > 0{,}9$ між періодичними коливаннями меридіонального вітру з періодами 12 неділь (хвилі Маддена-Джуліана) і 48 неділь (річне коливання) на цих висотах.

При визначенні міри взаємозв'язку спектральних компонентів двох процесів важливо з'ясувати, яким є співвідношення між взаємною енергією синхронної і несинхронної взаємодії, оскільки саме від характеру цієї взаємодії залежить різниця фаз коливань на фіксованій частоті. Із рівняння (6.90) видно, що коли $C_{x_i x_j}(\omega) \neq 0$, а $Q_{x_i x_j}(\omega) = 0$, різниця фаз коливань повинна дорівнювати нулю, оскільки взаємозв'язок процесів буде існувати за рахунок синхронної їх взаємодії. При $C_{x_i x_j}(\omega) = 0$ і $Q_{x_i x_j}(\omega) \neq 0$ різниця фаз спектральних компонент дорівнює $\frac{\pi}{2}$ (чверть періоду). Це означає, що взаємозв'язок коливань відбувається тільки в результаті несинхронної взаємодії процесів $X_i(t)$ і $X_j(t)$. У всіх інших випадках, тобто коли $C_{x_i x_j}(\omega) \neq 0$ і $Q_{x_i x_j}(\omega) \neq 0$, різниця фаз спектральних компонентів (фазовий спектр) визначається рівнянням (6.90). Фазовий спектр визначає відставання по фазі процесу $X_j(t)$ від процесу $X_i(t)$ при умові, що величину $\psi_{x_i x_j}(\omega)$ вважають додатною від 0^0 до 180^0 і від'ємною від 180^0 до 360^0. Зсув фаз в 0 відповідає додатній корреляції між процесами ("*у фазі*"), а зсув фаз в 180 - від'ємній корреляції ("*у протифазі*").

Корегентність одночасно являє собою міру стійкості різниці фаз. При постійності різниці фаз процесів $\gamma(\omega) \to 1$, якщо різниця фаз нестійка, то $\gamma(\omega) \to 0$.

6.5 Визначення спектральної щільності по експериментальних даних

Спектральну щільність стаціонарного процесу, як вже було показано вище, можна визначити як перетворення Фурьє від відповідної коваріаційної функції. Але при цьому треба знати коваріаційну функцію на нескінченному інтервалі значень її аргументу. При визначенні характеристик випадкової функції по експериментальних даних інтервал значень аргументу коваріаційної функції є обмеженим. Крім того, ми маємо не саму коваріаційну функцію, а її оцінку. Виникає питання про отримання на її основі оцінки спектральної щільності, котра задовольняла б основним вимогам до статистичних оцінок, тобто була б незсуненою, ефективною та умотивованою.

При визначенні оцінки спектральної щільності здавалося б простіше всього використати перетворення Фурьє, змінивши нескінчені границі інтегрування кінцевими значеннями, які відповідають найбільшому значенню аргументу $\tau = \tau_m$ оцінки коваріаційної функції. Це рівнозначно тому, що ми замінили справжню коваріаційну функцію $K(\tau)$ її оцінкою $\hat{K}(\tau)$ на інтервалі $[-\tau_m, \tau_m]$, а поза цього інтервалу запропонували $\hat{K}(\tau) = 0$.

Але оцінка спектральної щільності виду :

$$\hat{S}_1(\omega) = \frac{1}{2\pi} \int_{-\tau_m}^{\tau_m} \hat{K}(\tau) e^{-i\omega\tau} d\tau \qquad (6.93)$$

не є ефективною і умотивованою, тому, що дисперсія цієї оцінки не збігається до нуля при $\tau_m \to \infty$. Отже, така оцінка не може нас задовольнити. Кращі оцінки спектральної щільності дають методи, що основані на попередньому згладжуванні коваріаційної функції.

Нехай ми маємо функцію $\hat{K}(\tau)$, яка дорівнює справжньому значенню коваріаційної функції $K(\tau)$ при $|\tau| \le \tau_m$ і нулю при $|\tau| > \tau_m$. Цю функцію можна розглядати як добуток функції $K(\tau)$ на функцію $\lambda(\tau)$:

$$\hat{K}(\tau) = K(\tau)\lambda(\tau)\ , \tag{6.94}$$

де

$$\lambda(\tau) = \begin{cases} 1\ , & при \quad |\tau| \le \tau_m\ ; \\ \\ 0\ , & при \quad |\tau| > \tau_m\ . \end{cases} \tag{6.95}$$

Тепер функція $\hat{K}(\tau)$ визначена на всій дійсній осі. Знайдемо від неї перетворення Фурьє та будемо вважати його за оцінку $\hat{S}(\omega)$ спектральної щільності $S(\omega)$:

$$\hat{S}(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-j\omega\tau}\hat{K}(\tau)d\tau = \\ = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-j\omega\tau}\lambda(\tau)K(\tau)d\tau \quad . \tag{6.96}$$

Позначимо через $Q(\omega)$ спектр функції $\lambda(\tau)$, тобто

$$Q(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\omega\tau}\lambda(\tau)d\tau \tag{6.97}$$

Тоді

$$\lambda(\tau) = \int_{-\infty}^{\infty} e^{-i\omega\tau}Q(\omega)d\omega \tag{6.98}$$

і, таким чином,

$$\lambda(\tau)K(\tau) = \int_{-\infty}^{\infty} e^{i\omega_1\tau}S(\omega_1)d\omega_1 \times \\ \times \int_{-\infty}^{\infty} e^{i\omega_2\tau}Q(\omega_2)d\omega_2 = \\ = \int_{-\infty}^{\infty} S(\omega_1)\left(\int_{-\infty}^{\infty} e^{i(\omega_1+\omega_2)}Q(\omega_2)d\omega_2\right)d\omega_1 \quad .$$

Якщо в інтегралі, що міститься в дужках, зробити заміну змінної $\omega_1 + \omega_2 = \omega$, тоді будемо мати

$$\lambda(\tau)K(\tau) = \\ = \int_{-\infty}^{\infty} e^{i\omega\tau} \left(\int_{-\infty}^{\infty} S(\omega_1) Q(\omega - \omega_1) d\omega_1 \right) d\omega_2 \quad . \tag{6.99}$$

Але із (6.96) випливає, що

$$\lambda(\tau)K(\tau) = \int_{-\infty}^{\infty} e^{i\omega\tau} \hat{S}(\omega) d\omega \tag{6.100}$$

Порівняння формул (6.99) і (6.100) приводить до рівності

$$\hat{S}(\omega) = \int_{-\infty}^{\infty} S(\omega_1) Q(\omega - \omega_1) d\omega_1 \, . \tag{6.101}$$

Таким чином, $\hat{S}(\omega)$ є значення справжньої спектральної щільності, осередненної з ваговою функцією по всьому інтервалу частот. Для функції (6.95) спектр $Q(\omega)$ має вид:

$$Q(\omega) = \frac{1}{2\pi} \int_{-\tau_m}^{\tau_m} e^{i\omega\tau} d\tau = \frac{\sin \omega\tau_m}{\pi\omega} . \tag{6.102}$$

Отже, користуючись при визначенні спектральної щільності у якості оцінки коваріаційної функції добутком виду (6.94), ми отримаємо не справжню спектральну щільність $S(\omega)$, а її значення, яке згладжене за допомогою вагової функції, яка є спектром функції $\lambda(\tau)$. Звідси виникає задача вибору оптимальної функції $\lambda(\tau)$, тобто такої, щоб згладжування (6.101) було найкращим у смислі близкості функції $\hat{S}(\omega)$ і $S(\omega)$. Такими функціями є:

1. Функція Бартлетта, що визначається формулою (6.95);
2. Модифікована функція Бартлетта:

$$\lambda(\tau) = \begin{cases} 1 - \dfrac{|\tau|}{\tau_m}, & при \quad |\tau| \le \tau_m \ ; \\ 0, & при \quad |\tau| > \tau_m \ . \end{cases} \tag{6.103}$$

3. Функція Тьюкі

$$\lambda(\tau) = \begin{cases} 1 - 2a + 2a\cos\dfrac{\pi\tau}{\tau_m}, & при \quad |\tau| \le \tau_m \ ; \\ 0, & при \quad |\tau| > \tau_m \ . \end{cases} \tag{6.104}$$

Тьюкі запропонував брати $a = 0{,}23$, не обгрунтувавши вибір такого значення. Парзен показав, що оптимальним є значення $a = 0{,}25$.

4. Функція Хеннінга

$$\lambda(\tau) = \begin{cases} 0{,}5\left(1 - \cos\dfrac{\pi\tau}{\tau_m}\right), & при \quad |\tau| \le \tau_m \quad ; \\ 0, & при \quad |\tau| > \tau_m \quad . \end{cases}$$

(6.105)

5. Функція Парзена

$$\lambda(\tau) = \begin{cases} 1 - \left(\dfrac{|\tau|}{\tau_m}\right)^q, & при \quad |\tau| \le \tau_m \quad ; \\ 0, & при \quad |\tau| > \tau_m \quad . \end{cases}$$

(6.106)

Парзен розглядав цю функцію при $q = 2$. Існують й інші види згладжуючих функцій $\lambda(\tau)$. Таким чином, задачу визначення спектральної щільності можна сформулювати так: нехай маємо оцінку коваріаційної функції $\hat{K}(\tau)$ при $|\tau| < T$, де T - права границя інтервалу визначення випадкової функції. Будемо шукати оцінку спектральної щільності $\hat{S}(\omega)$ за формулою

$$\hat{S}(\omega) = \frac{1}{2\pi} \int_{-\tau_m}^{\tau_m} e^{-i\omega\tau} \lambda(\tau) \hat{K}(\tau) d\tau , \qquad (6.107)$$

підбираючи функцію $\lambda(\tau)$ і значення τ_m так, щоб задовольнити деякому критерію оптимальності.

При визначенні оцінки спектральної щільності за формулою (6.107) з вибраною згладжуючою функцією $\lambda(\tau)$, результат буде залежати від вибору точки зрізу коваріаційної функції τ_m. При малих τ_m буде відбуватися зсунення оцінки спектральної щільності, при великих τ_m - збільшення дисперсії оцінки. Прагнення вибрати τ_m таким, щоб мінімізувати як зсунення оцінки спектральної щільності, так і її дисперсії, потребує необхідності задовільнення двох суперечливих вимог.

Для отримання оцінки спектральної щільності може бути використаний і інший метод. Нехай маємо реалізацію $x(t)$ ергодичного стаціонарного процесу $X(t)$, визначену на інтервалі $[0, T]$. Тоді коваріаційну функцію можна одержати за формулою:

$$K_x(\tau) = \frac{1}{T-\tau}\int_0^{T-\tau}[x(t) - m_x][x(t+\tau) - m_x]dt$$

(6.108)

Якщо підставити (6.108) у формулу (6.93), то після відокремлювання змінних, будемо мати

$$\hat{S}_1(\omega) = \frac{1}{2\pi T}\left|\int_0^T x(t)e^{-i\omega t}dt\right|^2 \quad .$$

(6.109)

Величину (6.109) називають *вибірковим спектром*, або *періодограмою*. Вона отримується за допомогою перетворення Фурьє самої реалізації. Періодограма не є умотивованою оцінкою спектральної щільності. Для того, щоб задовольнити вимогу умотивованості, необхідно провести згладжування періодограми за формулою:

$$\hat{S}(\omega) = \int_{-\infty}^{\infty}\hat{S}_1(\omega)Q(\omega - \omega_1)d\omega \ , \qquad (6.110)$$

вибираючи згладжуючу функцію такою, щоб оцінка була умотивованою.

Порівнюючи (6.110) та (6.101), бачимо, що коли у якості функції $Q(\omega)$ виступає спектр вагової функції $\lambda(\tau)$, за допомогою якого виконувалось згладження коваріаційної

функції при першому методі оцінювання спектральної щільності, то при другому методі отримуємо ту ж саму оцінку.

Згладжуючу функцію $\lambda(\tau)$ часто називають *корреляційним вікном*, а її перетворення Фурьє $Q(\omega)$ - *спектральним вікном*. Обидва методи оцінювання спектральної щільності дають однаковий результат, але супроводжуються різними обчислювальними труднощами. Другий метод при застосуванні так званого швидкого перетворення Фурьє є більш раціональним і у цей час користується широкою популярністю.

У зв'язку з наявністю у спектрі гідрометеорологічних процесів шумової компоненти при аналізі спектрограм важливе значення приїдається оцінці невипадковості знайдених періодичностей. Як відомо, у гідрометеорологічних процесах має місце широкий спектр коливань: від випадкових незв'язних ("білий шум"), до таких, які характеризуються значними зв'язками ("червоний шум").

При використанні спектрального аналізу конкретних гідрометеорологічних процесів треба враховувати співвідношення між часовим кроком вибірки Δt, довжиною вибірки N, максимальним зрізом корреляційної функції $\tau_{\max}$ (йому відповідає число інтервалів Δt), числом ступенів волі l та нормованою стандартною похибкою оцінок спектральної щільності $\hat{S}_x(\omega)$. З одного боку, число зсувів $\tau_{\max}$ повинно бути малим порівняно з довжиною вибірки (наприклад $\tau_{\max} = 0{,}1N$), число ступенів волі - по можливості більшим. Це дасть визначну міру статистичної надійності оцінок. З іншого боку, число зсувів повинно бути досить великим, щоб отримати велике розділення по смузі частот. При цьому надійність статистичних оцінок у границях частотної смуги зменшується. Часто вибір максимального часового зсуву $\tau_{\max}$ корреляційної функції основується на можливій точності розрахунків спектральної щільності.

Спектральні щільності, що розраховані по вибіркових даних, будуть відрізнятися від спектра генеральної сукупності.

Як і у випадкову оцінок одновимірних розподілів, для оцінок значущості спектра використовується перевірка статистичної гіпотези на заданому рівні значущості. Нульова гіпотеза H_0 полягає у тому, що в спектрі часової послідовності відсутні гармонічні коливання на фоні спектра реалізації "білого шуму" (його спектр характеризується умовою $S(\omega) = S_x(0) = const$), або "червоного шуму" (спектр якого є спадаюча експоненціально крива). Вважається, що вихідна вибірка має нормальний розподіл. Тоді за відомою теоремою, що була сформульована у попередньому розділі, значення спектральної щільності, які характеризують розподіл дисперсій по спектру частот, мають χ^2 розподіл з числом ступенів волі l. Перевірка нульової гіпотези полягає у порівнянні $\hat{S}_x(\omega)$, із значеннями $S_{кр}(\omega)$ заданої ймовірності, що приймаються в якості границь довірчого інтервалу $I_\alpha[S_x(0)]$ або $I_\alpha[S_q(\omega)]$. У цьому випадку:

$$S_{кр}(\omega) = \frac{\chi^2(\alpha, l)}{l}, \qquad (6.111)$$

де

$$l = \frac{2N - 0{,}5\tau_{\max}}{\tau_{\max}}, \qquad (6.112)$$

N - об'єм вибірки, $\tau_{\max}$ - число точок максимального зсуву корреляційної функції. Тоді для побудови довірчого інтервалу використовується рівність:

$$I_\alpha[S_x(0)] = \overline{S}_x(\omega)\frac{\chi^2(\alpha,l)}{l}, \qquad (6.113)$$

де $\overline{S}_x(\omega)$ - середній рівень спектральної щільності (відповідаючий "білому шуму"), що розраховується в інтервалі значень корреляційної функції. Вихід піків спектральної щільності за межі довірчого інтервалу буде свідчити про вірогідність частот коливань, що виявлені на спектрограмі.

При значному внеску у випадковий процес "червоного шуму" спектр випадкової функції порівнюється з довірчим інтервалом, який будується на основі рівності:

$$I_\alpha[S_q(\omega)] = S_q(\omega)\frac{\chi^2(\alpha,l)}{l}, \qquad (6.114)$$

де

$$S_q(\omega) = \overline{S}_x(\omega)\frac{1-r_x^2(\tau_1)}{1+r_x^2(\tau_1)-2r_x(\tau_1)\cos\dfrac{\pi k}{\tau_m}},$$

$$(k=\overline{1,m}) \qquad (6.115)$$

Вона утримує значення корреляційної функції $r_x(\tau_1)$ одиничного зсуву (τ_1) і максимального m - на спектрограмі ($m \sim 2\tau_{\max}$).

На рис.6.13 показуються спектрограми для випадкових приростів меридіональної компоненти швидкості вітру на висотах 15 (а) і 20 (б) км. в районі островів Воллоп. На висоті 15 км. є підстави побудувати довірчу границю для спектральної щільності, виходячи з припущення про "білий шум", а на висоті 20 км - з припущення про "червоний шум".

Приблизна оцінка довірчої границі для когерентності на виконується за формулою:

$$I_\alpha[\gamma(\omega)] = \frac{2}{\sqrt{l}}, \qquad (6.116)$$

де, як було показано вище:

$$l = \frac{2N - \frac{\tau_{\max}}{2}}{\tau_{\max}} \qquad (6.117)$$

число ступенів волі, N - довжина реалізації. Значення корегентності при 1% і 5% рівнях значущості α для різного числа ступенів волі по Гудману приводяться у табл.6.1

Таблиця 6.1 - Довірчі границі когерентності $\gamma(\omega)$

Рівень значущості α	Число ступенів волі			
	4	10	20	40
0,01	0,89	0,63	0,46	0,33
0,05	0,80	0,53	0,38	0,27

Із табл.6.1 випливає що, наприклад, при числі ступенів волі $l = 20$ корегентність, що дорівнює 0,38 і більше буде з імовірністю 0,95 вірогідною.

6.6 Особливості дослідження статистичної структури нестаціонарних часових рядів гідрометеорологічних характеристик

6.6.1 Виявлення періодичностей, які утримуються у випадкових часових рядах

Часові ряди метеорологічних величин, як показали численні дослідження, утримують періодичні компоненти, обумовлені хвильовою природою атмосферних процесів.

Існує ряд методів дослідження періодичностей, що містяться у часових рядах. Їх називають захованими періодичностями. Одним з найбільш зручних для реалізації на ЕОМ є метод, оснований на інтегральному перетворенню Фурьє. Він дає можливість без будь-яких додаткових досліджень отримати частоти, амплітуди та початкові фази періодичних компонент, захованих у часовій послідовності.

Часовий ряд $x(t)$, заданий на інтервалі $t \in [-\tau, \tau]$, можна розглядати як кусково-гладку функцію часу. Таку функцію у відповідності до теореми Діріхлє можна виразити суперпозицією простих гармонік

$$x(t) = \sum_{k=0}^{\infty} A_k \sin(\omega_k t + \varphi_k) , \qquad (6.118)$$

де A_k - амплітуда k - тої гармоніки, ω_k - її частота, φ_k - початкова фаза.

Рівність (6.118) може бути переписаною таким чином:

$$x(t) = \sum_{k=0}^{\infty} [a_k \cos \omega_k t + b_k \sin \omega_k t] , \qquad (6.119)$$

якщо позначити

$$a_k = A_k \sin \varphi_k , \qquad (6.120)$$

$$b_k = A_k \cos \varphi_k . \qquad (6.121)$$

Як свідчать формули (6.120) і (6.121),

$$\varphi_k = arctg \frac{a_k}{b_k} . \qquad (6.122)$$

Для кусково–гладкої функції $x(t)$, заданої на нескінченному інтервалі, справедливим є перетворення Фурьє:

$$F(i\omega) = \int_{-\infty}^{\infty} x(t) e^{-i\omega t} dt \ , \qquad (6.123)$$

де $i = \sqrt{-1}$ - уявна одиниця.

Часові ряди метеорологічних величин визначені на скінченому інтервалі. Вони можуть бути апроксимовані таким чином:

$$x(t) = \begin{cases} x(t), & коли \quad t \in [-\tau, \tau] \ ; \\ 0, & коли \quad |t| > \tau \ . \end{cases} \qquad (6.124)$$

Для такої функції перетворення Фурьє має вид:

$$F_l(i\omega) = \frac{1}{\tau} \int_{-\tau}^{\tau} x(t) e^{-i\omega t} dt \ . \qquad (6.125)$$

Якщо використати відому формулу Ейлера, то інтеграл (6.125) приймає таку форму:

$$F_\tau(i\omega) = u(\omega) - i\upsilon(\omega) \ , \qquad (6.126)$$

де

$$u(\omega) = \frac{1}{\tau} \int_{-\tau}^{\tau} x(t) \cos \omega t dt \ , \qquad (6.127)$$

$$\upsilon(\omega) = \frac{1}{\tau} \int_{-\tau}^{\tau} x(t) \sin \omega t dt \ . \qquad (6.128)$$

Рівності (6.127) і (6.128) є відповідно косинус - і синус – перетворення Фурьє функції $x(t)$, апроксимованої виразом (6.124). Коли частоти гармонічних компонент, що утримуються в $x(t)$, не є дуже близькими, то $u(\omega)$ і $\upsilon(\omega)$ мають вид кривих з різко означенними піками в точках $\omega = \omega_k$. Висота піків приблизно дорівнює амплітудам парної a_k і непарної b_k складових періодичного коливання з частотою ω_k, захованого в процесі $x(t)$. На тих самих частотах ω_k будуть спостерігатися піки амплітуд $A_k = A(\omega_k)$, оскільки $a_k \approx u(\omega_k)$ і $b_k \approx \upsilon(\omega_k)$ і

$$A(\omega_k) = [u^2(\omega_k) + \upsilon^2(\omega_k)]^{\frac{1}{2}} \ . \qquad (6.129)$$

З метою покращення селективних якостей перетворена Фурьє (6.127) і (6.128) в них вводять множники ("вікна"), які зменшують вплив значень $x(t)$, заданих поблизу границь

інтервалу визначення функції. Одним з таких "вікон" є множник Гіббса:

$$g(t) = \frac{\sin\frac{\pi t}{\tau}}{\frac{2t}{\tau}} \quad . \tag{6.130}$$

Очевидно, $g(\tau) = g(-\tau) = 0$. Максимум цієї функції спостерігається при $\tau = 0$ і дорівнює

$$g(0) = \frac{\pi}{2}\lim_{t\to 0}\frac{\sin\frac{\pi t}{\tau}}{\frac{\pi t}{\tau}} = \frac{\pi}{2} \quad .$$

Таким чином, графік функції $g(t)$ на інтервалі $[-\tau, \tau]$ має вид, зображений на рис.6.14.

Отже, перетворення, за допомогою яких може проводитися селекція періодичностей, мають вид:

$$u(\omega) = \frac{1}{\tau}\int_{-\tau}^{\tau}\frac{\sin\frac{\pi t}{\tau}}{\frac{2t}{\tau}}x(t)\cos\omega t dt \quad , \tag{6.131}$$

$$\upsilon(\omega) = \frac{1}{\tau}\int_{-\tau}^{\tau} \frac{\sin\frac{\pi t}{\tau}}{\frac{2t}{\tau}} x(t)\sin\omega t dt \ . \qquad (6.132)$$

Інтеграли (6.131) і (6.132) обчислюються одним із наближуючих методів.

Із-за обмеженості інтервалу й скінченної кількості точок завдання функції, інформації про функцію недостатньо для визначення параметрів гармонік з періодом $T > 2\tau$ і $T < \frac{\tau}{m}$. Отже, мінімально і максимально можливі гармоніки, що можуть бути виявлені, мають частоти, розташовані в інтервалі $(\omega_{\min}, \omega_{\max})$, де $\omega_{\min} = \frac{\pi}{\tau}$; $\omega_{\max} = \frac{m\pi}{\tau}$.

Інтервал дискретності $\Delta\omega$ при чисельному інтегруванні вибирається з урахуванням властивостей множника Гіббса. При його впровадженні у перетворення Фурьє можна гарантувати, що вплив амплітуд сусідніх за частотою гармонік не перебільшує 0,05 від амплітуди, якщо $\Delta\omega\tau \geq 4{,}5$. Звідси випливає, що крок при обчисленнях $u(\omega)$ і $\upsilon(\omega)$ визначається рівністю:

$$\Delta\omega = \frac{4{,}5}{\tau} \ . \qquad (6.133)$$

Періодичності, що отримуються у часових рядах $x(t)$, визначаються по піках амплітуд $A(\omega)$ на періодограмі (амплітудно-частотній характеристиці). На періодограмах існує ряд малозабезпечених піків, утруднюючих аналіз. Для їх ліквідації застосовують фільтр Тьюкі :

$$\widetilde{A}(\omega_i) = 0{,}25A(\omega_{i-1}) + 0{,}5A(\omega_i) + \\ + 0{,}25A(\omega_{i+1}) \quad . \qquad (6.134)$$

Визначення періодичностей, характерних для процесу $x(t)$, основується на побудові верхньої довірчої границі для амплітуд з заданою імовірністю при умові, що амплітуди підпорядковуються нормальному розподілу. Періоди T_k гармонік ω_k, які відповідають пікам амплітуд, що виходять за довірчу границю, ототожнюються з періодами гармонічних коливань, які утримуються у випадковому процесі $x(t)$. Для кожного з них знаходять початкову фазу

$$\varphi_k = arctg\frac{u(\omega_k)}{\upsilon(\omega_k)} \quad . \qquad (6.135)$$

Початкова фаза дає можливість знайти точку h_k на осі часу, яка є початком коливання.

На рис.6.15 приводиться амплітудно-частотна характеристика часового ряду меридіональної компоненти швидкості вітру в пункті Форт Шерман на висоті 45 км. Данні отримані в результаті ракетного зондування атмосфери в період

з 1967 по 1974 рік. З нього випливає, що з імовірністю 0,95 у процесі, що розглядається, утримуються коливання, які характеризуються статистичною значущістю, з періодами 6, 4 та 1-2 місяці (хвилі Маддена-Джуліана) й 15 днів. Останні, очевидно, треба інтерпретувати як хвилі Кельвіна. Якщо визначати періодичності з ймовірністю 0,68, то до перелічених періодичностей меридіонального вітру вийдуть ще декілька періодичностей, що відносяться до спектру хвиль Кельвіна.

Треба мати на увазі, що у деяких випадках на амплітудно-частотних характеристиках спостерігаються сплески амплітуд, найбільш часто на низьких частотах, які значно відрізняються від загального рівня коливань амплітуд гармонік (рис.6.16). Це означає, що ці гармоніки характеризуються найбільшою, порівняно до інших, енергією. У таких випадках насамперед треба перевірити гіпотезу про те, що такі сплески амплітуд належать до тієї ж генеральної сукупності, що й амплітуди інших гармонік. У разі неприйняття цієї гіпотези такі сплески при розрахунках середнього значення амплітуди $\overline{A}$ й середнього квадратичного відхилу амплітуд σ_A треба виключити із сукупності амплітуд і після цього розрахувати довірчу границю амплітуд з тією чи іншою ймовірністю для визначення статистично значущих періодичностей.

6.6.2 Згладжування часових рядів

Більшість метеорологічних величин являють собою нестаціонарні випадкові процеси. Основною причиною цього є те, що під впливом неоднаковості надходячої до земної поверхні кількості сонячної радіації протягом доби, сезону і року вони придбавають добовий, сезонний, річний хід. Багаторічні змінення характеру кліматоформуючих факторів приводять до виникнення трендів, тобто однонаправлених змінень метеорологічних величин протягом тривалого часу. Прикладом може бути відоме потепління клімату у першій половині двадцятого століття.

Дослідження статистичної структури метеорологічних величин основуються на послідовності їх значень у виді еквідистантних часових рядів. Останні можуть бути зображені як сума детермінованої $\hat{x}(t)$ і випадкової $x_3(t)$ компонент. У свою чергу, детермінована компонента складається з тренду $x_1(t)$ і періодичної компоненти $x_2(t)$, яка відбиває в залежності від інтервалу дискретності часового ряду віковий, річний або добовий хід процесу $x(t)$. Отже,

$$x(t) = x_1(t) + x_2(t) + x_3(t) \qquad (6.136)$$

При правильному вилученні детермінованої складової $\hat{x}(t)$, випадкова компонента може розглядатися як стаціонарні випадкові прирости.

Детермінована основа процесу вилучається шляхом фільтрації (або згладжування) вихідного часового ряду. Позначимо оператор згладжування через L і застосуємо його до рівності (6.136)

$$L[x(t)] = L[x_1(t)] + L[x_2(t)] + L[x_3(t)] \qquad (6.137)$$

Припустимо, що оператор L точно вилучає трендову компоненту, тобто

$$L[x_1(t)] = x_1(t) \qquad (6.138)$$

Якщо відняти від рівності (6.136) рівність (6.137), то з урахуванням (6.138) маємо

$$x(t) - L[x(t)] = x_2(t) - L[x_2(t)] + x_3(t) - \\ - L[x_3(t)] \ .$$

(6.139)

Важливим є питання про те, у якій мірі члени $L[x_2(t)]$ і $L[x_3(t)]$ можуть спотворювати справжні коливання залишкового ряду (6.139) та індукувати хибні коливання.

Одним з видів згладжування є ковзне осереднення, яке у загальному виді може бути зображене таким чином:

$$\hat{x}(t_k) = \frac{1}{n} \sum_{i=k-n/2}^{k+n/2} \alpha_i x(t_i) \ , \qquad (6.140)$$

де α_i - ваговий множник; n - кількість точок, по яких проводиться згладжування:

$$k = 1 + \frac{n}{2}; \ \ 2 + \frac{n}{2}; ...; N' + \frac{n}{2}; \ \ N' = N - n + 1$$

$N -$ число членів ряду.

Якщо в рівності (6.140) $\alpha_i = 1 \ \forall \ i = \overline{1, n}$, то оператор згладжування визначає просте ковзне осереднення, у якому вага всіх точок, котрі приймають участь при розрахуванні

середнього значення на інтервалі $\left[k-\frac{n}{2};k+\frac{n}{2}\right]$, однакова. Більш коректними є фільтри, що утримують тригонометричні

$$\alpha_i = 1+\cos\frac{2\pi(k-i)}{n} \qquad (6.141)$$

або експоненціальні

$$\alpha_i = \exp\left[-\frac{|k-i|}{n}\right] \qquad (6.142)$$

вагові множники. Вони зменшуються по відзначених формулами (6.141) і (6.142) законах від середини інтервалу згладжування до його кінців. Дійсно, нехай в рівності (6.141) величина i приймає такі значення: $i=k-\frac{n}{2}$; $i=k$; $i=k+\frac{n}{2}$. Тоді, очевидно, α_i мають значення $\alpha_{k-\frac{n}{2}}=0$; $\alpha_k=2$; $\alpha_{k+\frac{n}{2}}=0$. На рис.6.17 приводиться залежність вагового множника (6.141) від положення точок i на інтервалі $\left[k-\frac{n}{2};k+\frac{n}{2}\right]$.

Окрім виду вагового множника, результат фільтрації залежить й від кількості точок, по яких виконується згладжування. Вона, очевидно, визначається рівністю

$$n = \frac{\tau}{\Delta t}, \qquad (6.143)$$

де Δt - інтервал дискретності ряду.

Чим більше n, тим швидше реакція фільтру, але тим гірше його фільтруючі якості, і навпаки. Отже, задача полягає у правильному виборі періоду згладжування. Означені вище особливості ковзного осереднення приводять до того, що при надто великому періоді згладжування з детермінованої основи $\hat{x}(t)$ процесу відфільтровується визначна частина періодичної компоненти $x_2(t)$, яка переходить до випадкової компоненти $x_3(t)$. Навпаки, коли період осереднення малий, частина випадкової складової процесу переходить до детермінованої частини процесу, а випадкова компонента $x_3(t)$ придбаває властивості "білого шуму".

Для вибору періоду згладжування випадкової послідовності $x(t)$ при ковзному осередненні, треба дотримуватися таких рекомендацій. По-перше, необхідно, щоб період згладжування відповідав періодичності, яка існує у процесі $x(t)$. По-друге, значення періоду згладжування повинно відповідати періодичності, яку досліджувач хоче зберегти в детермінованій складовій випадкової послідовності в залежності від задачі, яку він намагається розв'язати. Наприклад, якщо треба у складовій $\hat{x}(t)$ зберегти річний хід метеорологічної величини $x(t)$, а коливання з меншими періодами необхідно отфільтрувати з вихіднго часовою ряду,

то ковзне осереднення проводять при такому числі значень випадкової величини n, яке відповідає річному інтервалу. Тоді на виході з фільтру ми отримаємо процес, який утримує, крім трендів, коливання з періодом один рік і більше, а коливання з меншими періодами перейдуть у випадкову складову. Остання має, як правило, властивості квазістаціонарного процесу. Тому для дослідження її статистичної структури використовують розглянуті вище, методи корреляційного і спектрального аналізу. На рис.6.18 приводиться графік детермінованої компоненти зональної складової швидкості вітру на висоті 35 км над районом Форт Шерман (приекваторіальна зона Північно-Американського континенту), яка отримана шляхом ковзного осереднення за допомогою косинус-фільтра (6.141). Періодом осереднення, дорівнює одному року. На цьому рисунку зображається також й вихідний випадковий процес - результати ракетного зондування атмосфери з дискретністю один тиждень. З рис. 6.18 випливає, по-перше, що вихідний випадковий процес, який досліджується, не є стаціонарним. По-друге, в результаті ковзного осереднення, крім слабко вираженої однорічної періодичності, чітко проявляється коливання з періодом два роки, яке називається квазідвохрічною періодичністю швидкості вітру. У багатьох роботах, які присвячені квазідвохрічних коливань швидкості вітру в стратосфері приекваторіальної зони, стверджується наявність західної й східної фаз коливань, що проявляються через один рік кожна. Рис. 6.18 ілюструє той факт, що західна фаза, яка відбивається максимуми функції, що утворюються в результаті ковзного осереднення і відповідають малим значенням східної (від'ємної) зональної складової швидкості вітру, не є таким періодом, протягом якого спостерігається виключно західна (додатня) компонента зонального вітру. Протягом західної фази квазідвохрічної періодичності, західні вітри часто, змінюються на східні. Більш того, східна компонента дає більший внесок при осередненні, результатом чого й є той факт, що практично, вся детермінована складова західної компоненти розташовується у від'ємній півплощині, тобто має східний напрямок. Навпаки, в

період східної фази квазідвохрічної періодичності спостерігається виключно східний зональний вітер.

У якості другого прикладу на рис.6.19 зображається детермінована складова зональної компоненти швидкості вітру на висоті 10 км над районом Уайт Сендз (субтропічна зона Північно-Американського континенту), що отримана також шляхом ковзного осереднення (з періодом осереднення один рік) результатів радіозондування атмосфери, що над цим районом проводилося в період з 1965 до 1974 рр. На графіку детермінованої основи виразно відбивається однорічна періодичність коливання швидкості зонального вітру з максимумами, що припадають на холодні місяці року. Коливання з меншими періодами, які для цього випадкового процесу виявляються з досить великою ймовірністю на амплітудно-частотній характеристиці, відфільтровуються й переходять, як зазначалося вище, у випадкову складову вихідного процесу. Цікавим є й те, що крім річного коливання , на детермінованій основі процесу (рис.6.19), проявляється також і двохрічне коливання. Воно визначається збільшенням через кожні два роки максимумів амплітуди річного коливання.

Детерміновані складові наведених прикладів характеризуються наявністю тільки періодичних коливань. Трендова компонента практично не проявляється. Різниця між вихідним процесом $x(t)$ і детермінованою його складовою $\hat{x}(t)$ дає випадкову компоненту $x_3(t)$, яка має властивості, близькі до стаціонарного процесу. Тому її називають квазістаціонарними приростами. Для дослідження їх статистичної структури використовують методи корреляційного і спектрального аналізу. У якості прикладу на рис.6.20 містяться спектральні щільності випадкових приростів зональної складової швидкості вітру на висотах 5 (а) і 10 (б) км над районом Уайт Сендз. Сплески спектральної щільності, що виходять за довірчу границю з ймовірністю 90 %, свідчать про те, що в цьому випадковому процесі значна енергія припадає на флуктуації, які мають часовий масштаб 2-4 місяці. Як вже

зазначалося, такі коливання швидкості вітру називають хвилями Маддена-Джуліана.

Наведені приклади свідчать про те, що за допомогою викладених вище методів статистичного аналізу нестаціонарних часових послідовностей, є можливість отримати важливі характеристики статистичної структури гідрометеорологічних процесів. Тому вони знаходять широке використання при проведенні відповідних наукових досліджень.

ЧАСТИНА II

БАГАТОВИМІРНИЙ СТАТИСТИЧНИЙ АНАЛІЗ МЕТЕОРОЛОГІЧНИХ ПРОЦЕСІВ І ПОЛІВ

1. Деякі загальні положення

Багатовимірний статистичний аналіз, як комплекс імовірно-статистичних методів, знаходить широке використання в задачах дослідження статистичної структури метеорологічних полів, побудові статистичних моделей метеорологічних прогнозів. Під метеорологічним полем ми будемо розуміти сукупність значень метеорологічної величини на упорядкованій множині точок трьохвимірного простору у фіксований момент часу. Якщо, при цьому, фіксується одна із координат, то поле називають двовимірним. Прикладом такого поля може бути поле геопотенціальних висот фіксованої ізобаричної поверхні, поле температури, поле зональної чи меридіональної компоненти швидкості вітру на фіксованій висоті тощо.

Однієї з головних задач метеорологічного обслуговування галузей народного господарства є розробка прогнозів метеорологічних величин. Прогноз - це імовірнісне висловлювання про стан процесу, що передбачається у визначеним наступний момент часу. Процес складання прогнозу називають прогнозуванням. Але для того, щоб здійснити прогнозування, необхідно розробити методи прогнозу. Основою для цього є аналіз процесів, які підлягають дослідженню.

Аналіз складається з трьох етапів: ретроспекції, діагнозу і проспекції. Розглянемо зміст кожного з цих етапів. Ретроспекцією називають вивчення інформації про стан процесу в минулому. Це дає можливість отримати систему факторів, що чинять вплив на розвиток процесу, провести вибір

оптимального їх складу і визначити вид моделі, адекватно відбиваючої динаміку процесу.

На стані діагнозу здійснюють формування з урахуванням ретроспективної інформації параметрів вибраної моделі.

Проспекцією називають проведення чисельних експериментів з розробленою моделлю з метою визначення вірогідності прогнозів, які розробляються на основі моделі, а також для удосконалення моделі. Якщо йдеться про статистичні моделі, то вхідною інформацією є статистичні сукупності метеорологічних величин, які відіграють роль впливаючих факторів. В деяких випадках у ролі впливаючих факторів виступають характеристики тих чи інших метеорологічних полів, які враховуються при побудові прогностичної моделі. Оцінки параметрів статистичної моделі метеорологічного прогнозу знаходяться за допомогою тих чи інших методів багатовимірного статистичного аналізу на основі цієї інформації.

Методи багатовимірного статистичного аналізу базуються на визначних аксіомах і теоремах лінійної алгебри і теорії імовірностей. Всі математичні об'єкти будемо розглядати та визначати у багатовимірному просторі. Арифметичним простором називають множину всіляких n - вимірних систем упорядкованих дійсних чисел $(a_1, a_2, ..., a_n)$. Сама ця система чисел називається точкою n - вимірного арифметичного простору $A(a_1, a_2, ..., a_n)$. Якщо визначити початок координат у точці $O(0,0,...,0)$, то точка A може ототожнюватися з n - вимірним вектором $\vec{A}(a_1, a_2, ..., a_n)$. Числа, що розташовуються в дужках у цьому випадку мають, сенс координат вектора $\vec{A}$.

Велике значення у практичних застосуваннях має метричний арифметичний простір. Метричним арифметичним простором називають такий простір, де будь-яким двом точкам цього простору A і B ставиться у відповідність дійсне число $\rho(A, B)$, яке задовольняє таким вимогам:

1) $\rho(A,B) \geq 0$; при $A = B$ $\rho(A,B) = 0$ (властивість невід'ємності);

2) $\rho(A,B) = \rho(B,A)$ (властивість симетричності);

3) Для будь-яких трьох точок цього простору A, B, C $\rho(A,B) \leq \rho(A,C) + \rho(B,C)$ (властивість трикутника).

Метричний арифметичний n - вимірний простір називають евклідовим R^n, якщо для точок $A(a_1, a_2, ..., a_n)$ і $B(b_1, b_2, ..., b_n)$ метрикою $\rho(A,B)$ є евклідова відстань, що визначається формулою

$$\rho(A,B) = \sqrt{\begin{matrix}(b_1 - a_1)^2 + (b_2 - a_2)^2 + \\ + ... + (b_n - a_n)^2\end{matrix}} = \left\| \overrightarrow{AB} \right\| \tag{1.1}$$

Цю метрику називають нормою вектора $\overrightarrow{AB}$. Властивості метрики $\rho(A,B)$, що перелічені вище, носять назву аксіоматики метричного простору. Для евклідового простору може бути визначеним скалярний добуток двох векторів

$$\vec{A} \times \vec{B} = \left\langle \vec{A}\vec{B} \right\rangle = \sum_{i=1}^{n} a_i b_i \tag{1.2}$$

Розглянемо вектор X_i з координатами

$$X_i = \begin{pmatrix} x_{1i} \\ x_{2i} \\ \dots \\ x_{ni} \end{pmatrix} \qquad (1.3)$$

і вектор X_k з координатами

$$X_k = \begin{pmatrix} x_{1k} \\ x_{2k} \\ \dots \\ x_{nk} \end{pmatrix} \qquad (1.4)$$

В метеорологічних дослідженнях є всі підстави розглядати вектори X_i і X_k як поля метеорологічних величин. Дійсно, оскільки поле метеорологічної величини визначається сукупністю її значень на мережі метеорологічних станцій (або на множині вузлів регулярної сітки точок), а положення кожної метеорологічної станції (вузла) добре відоме, то якщо розташувати значення метеорологічної величини відповідно визначеному порядку нумерації метеорологічних станцій (або вузлів), то отримаємо упорядковану систему дійсних чисел. Вони і складають вектор n - вимірного евклідового простору.

Як відомо, значення метеорологічної величини мають випадковий характер. Цей факт докладно обгрунтовувався вище. Тому поля метеорологічних величин будемо у подальшому розглядати як випадкові вектори n - вимірного евклідового простору. Доречі, випадковий вектор може відбивати не тільки

метеорологічне поле, а й вертикальний профіль метеорологічної величини, яка визначена на множині послідовних висот, або сукупність впливаючих факторів (предікторів), що відбивають стан того чи іншого метеорологічного процесу.

Метеорологічне поле відноситься до визначного моменту часу. Інформацію про множину метеорологічних полів (або вертикальних профілів чи векторів-предікторів) можна розглядати як відповідну множину із m випадкових n - вимірних векторів, або, якщо розташувати цю часову послідовність векторів один за одним, то як матрицю порядку $n \times m$ такого виду:

$$X = \begin{pmatrix} x_{11}x_{12}\ldots x_{1j}\ldots x_{1m} \\ x_{21}x_{22}\ldots x_{2j}\ldots x_{2m} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ x_{i1}x_{i2}\ldots x_{ij}\ldots x_{im} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ x_{n1}x_{n2}\ldots x_{nj}\ldots x_{nm} \end{pmatrix} \qquad (1.5)$$

Індекси елементів x_{ij} $(i = \overline{1,n};\ j = \overline{1,m})$ матриці X визначають, відповідно, номер метеорологічної станції (або вузли сітки) і час, до якого відноситься j - те поле. Отже, стовпці матриці X - це індивідуальні поля метеорологічної величини, а рядки - статистичні сукупності метеорологічної величини на кожній i - тій станції. Як вже відзначалося, елемент x_{ij} може мати смисл значення метеорологічної

величини на i - тій висоті або i - того впливаючого фактора $(i=\overline{1,n})$ у j - тий момент часу $(j=\overline{1,m})$.

Оскільки ми маємо статистичні сукупності метеорологічної величини для кожної метеорологічної станції, то можна розрахувати для кожної i - тої станції середнє значення цієї величини за формулою

$$\overline{x}_i = \frac{1}{m}\sum_{j=1}^{m} x_{ij} \qquad (i=\overline{1,n}) \tag{1.6}$$

Таким чином, ми отримаємо вектор середніх значень метеорологічної величини:

$$\overline{X} = \begin{pmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \dots \\ \overline{x}_i \\ \dots \\ \overline{x}_n \end{pmatrix}$$

який має смисл середнього поля чи середнього вертикального профілю метеорологічної величини в залежності від задачі, що розглядається.

Звичайно, кожний рядок матриці X також можна розглядати як вектор m - вимірного евклідового простору. Наприклад,

$$X_i = (x_{i1} x_{i2} \ldots x_{ij} \ldots x_{im}) \tag{1.7}$$

Віднімемо від кожної координати цього вектора середнє значення. Цю операцію називають операцією центрирування. Будемо мати вектор

$$\Delta X_i = (\Delta x_{i1} \Delta x_{i2} \ldots \Delta x_{ij} \ldots \Delta x_{im}) \tag{1.8}$$

де

$$(j = \overline{1, m}) \tag{1.9}$$

Очевидно, квадрат норми вектора (1.8) є

$$\|\Delta X_i\|^2 = \sum_{j=1}^{m} \Delta x_{ij}^2 = \sum_{j=1}^{m} (x_{ij} - \overline{x}_i)^2 \tag{1.10}$$

Звідси випливає, що

$$\frac{\|\Delta X_i\|^2}{m} = \frac{\sum_{j=1}^{m} (x_{ij} - \overline{x}_i)^2}{m} = \sigma_{x_i}^2 \tag{1.11}$$

Таким чином дисперсію метеорологічної величини в i - тій точці поля можна трактувати як зменшений в m - разів квадрат норми вектора (1.8). Тоді середній квадратичний відхил метеорологічної величини в i - тій точці (на i - тій метеорологічній станції) є

$$\sigma_{X_i} = \frac{\|\Delta X_i\|}{\sqrt{m}} \tag{1.12}$$

Знайдемо скалярний добуток векторів ΔX_i і ΔX_k. Будемо мати

$$\langle \Delta X_i \Delta X_k \rangle = \Delta X_i^{'} \Delta X_k = \sum_{j=1}^{m} (x_{ij} - \overline{x}_i)(x_{kj} - \overline{x}_k) \tag{1.13}$$

(індексом $'$ будемо позначати операцію транспонування матриць чи векторів).

Видно, що

$$\frac{\langle \Delta X_i^{'} \Delta X_k \rangle}{m} = \frac{1}{m} \sum_{j=1}^{m} (x_{ij} - \overline{x}_i)(x_{kj} - \overline{x}_k) = K_{ik} \tag{1.14}$$

є коваріацією між метеорологічною величиною в точках i та k метеорологічного поля.

Очевидно, з урахуванням рівностей (1.12) і (1.14) маємо:

$$\frac{\left\langle \Delta X_i' \Delta X_k \right\rangle}{\left\| \Delta X_i \right\| \left\| \Delta X_k \right\|} = \frac{K_{ik} m}{\sqrt{m}\sigma_{X_i} \sqrt{m}\sigma_{X_k}} = \frac{K_{ik}}{\sigma_{X_i} \sigma_{X_k}} = r_{ik} \tag{1.15}$$

Отже, корреляція, що визначає міру лінійного корреляційного зв'язку між метеорологічною величиною в точках i та k поля, є нормованим скалярним добутком векторів ΔX_i й ΔX_k . Тому корреляцію й називають інколи нормованим коваріаційним моментом. З іншого боку, за визначенням скалярного добутку

$$\left\langle \Delta X_i' \Delta X_k \right\rangle = \left\| \Delta X_i \right\| \left\| \Delta X_k \right\| \cos(\Delta X_i \overset{\wedge}{,} \Delta X_i) \tag{1.16}$$

Звідси

$$\frac{\left\langle \Delta X_i' \Delta X_k \right\rangle}{\left\| \Delta X_i \right\| \left\| \Delta X_k \right\|} = \cos(\Delta X_i \overset{\wedge}{,} \Delta X_i) = r_{ik} \ , \tag{1.17}$$

тобто корреляцію можна трактувати як косинус кута між векторами ΔX_i та ΔX_k.

Як відомо, коли величина X_i й X_k некоррельовані, то $r_{ik} = 0$. Але у цьому разі, як випливає з формул (1.15) і (1.17), вектори ΔX_i і ΔX_k є ортогональними. Отже поняття ортогональності та некоррельованості є однозначними, коли йдеться про випадкові величини.

Визначимо ще один цікавий факт. В лінійній алгебрі обгрунтовується так звана нерівність Коші-Буняковського

$$\left\langle \vec{A}\vec{B} \right\rangle \le \left\| \vec{A} \right\| \left\| \vec{B} \right\| \qquad (1.18)$$

Для векторів ΔX_i і ΔX_k , що розглядаються, вона має вигляд

$$\left\| \left\langle \Delta X_i \Delta X_k \right\rangle \right\| \le \left\| \Delta X_i \right\| \left\| \Delta X_k \right\| ,$$

або з урахуванням формул (1.12) і (1.14)

$$\left| K_{ij} \right| \le \sigma_{X_i} \sigma_{X_k} \qquad (1.19)$$

Розділивши обидві частини рівності (1.19) на її праву частину, прийдемо до висновку, що

$$\left| r_{ij} \right| \le 1, \qquad (1.20)$$

або

$$-1 \le r_{ij} \le 1 \qquad (1.21)$$

Ця властивість коефіцієнта корреляції, яка в розділі "корреляційний зв'язок гідрометеорологічних величин" декларувалась, тепер отримала обгрунтування.

Наведені вище співвідношення доконливо ілюструють те що, багатовимірний статистичний аналіз спирається, як визначилося вище, з одного боку на теореми і аксіоми лінійної алгебри, а з другого – на положення теорії імовірності та математичної статистики.

2. КОРРЕЛЯЦІЙНИЙ АНАЛІЗ МЕТЕОРОЛОГІЧНИХ ОБ'ЄКТІВ

2.1 Матриці коваріацій і корреляцій.

Розв'язок чисельних задач сучасної метеорології потребує знань про статистичну структуру метеорологічних полів, таких як, наприклад, поля опадів, температури, тиску та вологості повітря, швидкості вітру тощо. Сукупність m метеорологічних полів, що відносяться до визначених термінів спостереження, можна, як відзначалося вище, зображати матрицею порядку $n \times m$ вигляду

$$X = \begin{pmatrix} x_{11}x_{12}\ldots x_{1j}\ldots x_{1m} \\ x_{21}x_{22}\ldots x_{2j}\ldots x_{2m} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ x_{i1}x_{i2}\ldots x_{ij}\ldots x_{im} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ x_{n1}x_{n2}\ldots x_{nj}\ldots x_{nm} \end{pmatrix} \qquad (2.1.1)$$

Таким же чином можна скласти інформацію, наприклад, про m вертикальних профілів метеорологічної величини, що вимірюється на n рівнях атмосфери, або про множину m векторів-предікторів у випадку, коли вирішується проблема побудови статистичної моделі метеорологічного прогнозу. Можна навести приклади й інших задач. Щоб не перелічувати кожний раз ці задачі, будемо у подальшому інколи іменувати

поля, вертикальні профілі метеорологічних величин, вектори-предіктори й т.п. метеорологічними об'єктами.

Матриця (2.1.1) утримує великий об'єм інформації. Її стовпці є відповідними метеорологічними об'єктами. У матриці (2.1.1) концентрується інформація про m таких об'єктів. Рядки матриці являють собою, як вже зазначалося, часові ряди відповідної метеорологічної величини.

Таке матричне зображення метеорологічних об'єктів є дуже раціональним, оскільки дає можливість побудувати прості алгоритми дослідження їх статистичної структури.

Найбільш важлива інформація про статистичну структуру метеорологічних об'єктів міститься у матриці коваріацій. Алгоритм побудови цієї матриці складається з декількох етапів. Знайдемо, насамперед, вектор середніх значень метеорологічних величин

$$\overline{X} = \begin{pmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \dots \\ \overline{x}_i \\ \dots \\ \overline{x}_n \end{pmatrix} \qquad (2.1.2)$$

де

$$\overline{x}_i = \frac{1}{m} \sum_{j=1}^{m} x_{ij} \quad (i = \overline{1,n}) \qquad (2.1.3)$$

Підкреслимо ще раз, що коли йдеться про поля метеорологічних величин, то вектор (2.1.2) є середнє поле метеорологічної величини. У випадку дослідження вертикальної

статистичної структури метеорологічної величини, вектор (2.1.2) має смисл її середнього вертикального профілю.

На основі матриці (2.1.1) і вектора (2.1.2) визначимо відповідну матрицю центрованих елементів. Для цього від кожного елемента кожного рядка матриці (2.1.1) віднімемо відповідне середнє значення. Будемо мати

$$\Delta X = \begin{pmatrix} \Delta x_{11}\Delta x_{12}\ldots\Delta x_{1j}\ldots\Delta x_{1m} \\ \Delta x_{21}\Delta x_{22}\ldots\Delta x_{2j}\ldots\Delta x_{2m} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \Delta x_{i1}\Delta x_{i2}\ldots\Delta x_{ij}\ldots\Delta x_{im} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \Delta x_{n1}\Delta x_{n2}\ldots\Delta x_{nj}\ldots\Delta x_{nm} \end{pmatrix} \qquad (2.1.4)$$

де

$$\Delta x_{ij} = x_{ij} - \bar{x}_i \qquad (2.1.5)$$

Операція, що проведена над матрицею (2.1.1), називається, як зазначалось вище, операцією центрирування. Тоді матриця коваріацій K_X визначається таким матричним рівнянням:

$$K_X = \frac{1}{m-1}\Delta X'\Delta X \qquad (2.1.6)$$

Покажемо справедливість цього твердження. Для цього візьмемо більш просту матрицю порядку

$$\Delta X' = \begin{pmatrix} \Delta x_{11}\Delta x_{12}...\Delta x_{1j}...\Delta x_{1m} \\ \Delta x_{21}\Delta x_{22}...\Delta x_{2j}...\Delta x_{2m} \end{pmatrix}$$

Для такої матриці рівність (2.1.6) в координатній формі має вид

$$K_X \frac{1}{m-1}\begin{pmatrix} \Delta x_{11}\Delta x_{12}...\Delta x_{1j}...\Delta x_{1m} \\ \Delta x_{21}\Delta x_{22}...\Delta x_{2j}...\Delta x_{2m} \end{pmatrix} \times$$

$$\times \begin{pmatrix} \Delta x_{11}\Delta x_{21} \\ \Delta x_{12}\Delta x_{22} \\ ..... \\ \Delta x_{1j}\Delta x_{2j} \\ ..... \\ \Delta x_{1m}\Delta x_{2m} \end{pmatrix} =$$

$$= \begin{pmatrix} \frac{1}{m-1}\sum_{j=1}^{m}\Delta x^2_{1j} & \frac{1}{m-1}\sum_{j=1}^{m}\Delta x_{1j}\Delta x_{2j} \\ \frac{1}{m-1}\sum_{j=1}^{m}\Delta x_{2j}\Delta x_{1j} & \frac{1}{m-1}\sum_{j=1}^{m}\Delta x^2_{2j} \end{pmatrix} =$$

$$= \begin{pmatrix} \sigma_1^2 & K_{12} \\ K_{21} & \sigma_2^2 \end{pmatrix}$$

(2.1.7)

Узагальнюючи проведені обчислювання, будемо мати:

$$K_x = \begin{pmatrix} \sigma_1^2 & K_{12} & \dots & K_{1j} & \dots & K_{1n} \\ K_{21} & \sigma_2^2 & \dots & K_{2j} & \dots & K_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ K_{i1} & K_{i2} & \dots & K_{ij} & \dots & K_{in} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ K_{n1} & K_{n2} & \dots & K_{nj} & \dots & \sigma_n^2 \end{pmatrix} \qquad (2.1.8)$$

Елементи матриці (2.1.8) розраховуються по формулах:

$$\sigma_i^2 = \frac{1}{m-1} \sum_{S=1}^{m} \Delta x_{iS}^2 \qquad (2.1.9)$$

$$K_{ij} = \frac{1}{m-1} \sum_{S=1}^{m} \Delta x_{iS} \Delta x_{jS} \qquad (2.1.10)$$

Отже, як випливає з формули (2.1.9) і (2.1.10), на головній діагоналі матриці (2.1.8) розташовуються дисперсії метеорологічної величини. Порядковий номер дисперсії на діагоналі відповідає номеру метеорологічної станції, якщо йдеться про метеорологічні поля, номеру стандартної висоти, якщо йдеться про вертикальні профілі метеорологічних величин, або номеру предіктора, якщо досліджується статистичні

особливості системи предікторів при побудові моделі прогнозу. Інші елементи матриці (2.1.8) - відповідні коваріації.

Матриця коваріацій має такі властивості:

- її елементи є дійсними числами;
- вона є симетричною;
- матриця коваріацій є додатньо визначеною.

З останньої властивості випливає, що $\left|K_X\right| > 0$; (Прямими дужками позначається визначник). Доречі, матриця коваріацій, поряд з вектором математичних сподівань m_X, відіграє роль параметра щільності імовірностей багатовимірного нормального розподілу

$$f(X) = \frac{1}{(2\pi)^{\frac{n}{2}}\left|K_X\right|^{\frac{1}{2}}} \times \\ \times \exp\left\{-\frac{1}{2}(X - m_X)K_X^{-1}(X - m_X)\right\} \tag{2.1.11}$$

Маючи матрицю коваріацій, можна легко сформувати діагональну матрицю σ середніх квадратичних відхилів. Вона має вид:

$$\sigma = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_n \end{pmatrix} \tag{2.1.12}$$

Обернена матриця від діагональної матриці знаходиться дуже просто

$$\sigma^{-1} = \begin{pmatrix} \frac{1}{\sigma_1} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{1}{\sigma_n} \end{pmatrix} \qquad (2.1.13)$$

Якщо помножити ліворуч та праворуч матрицю коваріацій K_X на матрицю (2.1.13), то будемо мати матрицю корреляцій R_X

$$R_X = \sigma^{-1} K_X \sigma^{-1} \qquad (2.1.14)$$

Покажемо це на матрицях другого порядку:

$$\sigma^{-1}K_X\sigma^{-1} = \begin{pmatrix} \frac{1}{\sigma_1} & 0 \\ 0 & \frac{1}{\sigma_2} \end{pmatrix} \begin{pmatrix} \sigma_1^2 & K_{12} \\ K_{21} & \sigma_2^2 \end{pmatrix} \times$$

$$\times \begin{pmatrix} \frac{1}{\sigma_1} & 0 \\ 0 & \frac{1}{\sigma_2} \end{pmatrix} = \begin{pmatrix} \frac{\sigma_1^2}{\sigma_1^2} & \frac{K_{12}}{\sigma_1\sigma_2} \\ \frac{K_{21}}{\sigma_2\sigma_1} & \frac{\sigma_2^2}{\sigma_2^2} \end{pmatrix} =$$

$$= \begin{pmatrix} 1 & r_{12} \\ r_{21} & 1 \end{pmatrix} = R_X$$

Узагальнюючи цей результат на матриці порядку n, отримаємо

$$R_X = \begin{pmatrix} 1 & r_{12} & \dots & r_{1j} & \dots & r_{1n} \\ r_{21} & 1 & \dots & r_{2j} & \dots & r_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ r_{i1} & r_{i2} & \dots & r_{ij} & \dots & r_{in} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ r_{n1} & r_{n2} & \dots & r_{nj} & \dots & 1 \end{pmatrix} \qquad (2.1.15)$$

Матриця корреляцій має властивості, аналогічні властивостям матриці коваріацій, тобто вона дійсна, симетрична і додатньо визначена.

Матриці коваріацій і корреляцій утримують важливу інформацію про особливості статистичної структури метеорологічних об'єктів. Якщо, наприклад, йдеться про метеорологічні поля, то поряд з полем середніх значень, яке характеризує вектор (2.1.2), можна побудувати за допомогою матриці коваріацій поле дисперсій метеорологічної величини, або середніх квадратичних відхилів. У матриці корреляцій міститься інформація про структуру n полів корреляцій. Дійсно, елемент цієї матриці r_{ij} характеризує лінійний корреляційний зв'язок між метеорологічною величиною на i - тій та j - тій станціях. Інші елементи матриці корреляцій характеризують аналогічні корреляційні зв'язки з іншими метеорологічними станціями, тобто рядок чи відповідний стовпець матриці R_X складає поле корреляцій. Полюсом цього поля буде та i - та метеорологічна станція, елемент корреляційної матриці для якої розташовується на перетину i - того рядку і j - того стовпця матриці. Очевидно, він дорівнює одиниці. На карті відповідного масштабу проводять ізокорреляти - лінії, що з'єднують точки з однаковими значеннями корреляцій. Система ізокоррелят дає змогу проаналізувати характер поля корреляцій. У якості прикладу на рис.2.1 і 2.2. наводяться поля корреляцій місячних кількостей опадів у грудні для полюсів корреляцій у Дніпропетровську і Чернігові. До аналізу цих полів ми ще повернемося пізніше.

2.2. Однорідність та ізотропність метеорологічних полів

Властивості однорідності та ізотропності випадкових полів є дуже важливими. Якщо ці властивості притаманні метеорологічним полям, то дуже спрощуються процедури їх

статистичного аналізу. Насамперед, сформулюємо загальне означення однорідності та ізотропності випадкових полів.

Випадкове поле $X(\rho)$ визначене на множині $Д \subset R^n$, називається однорідним, якщо:

1) коли $\rho_1 \in Д$; $\rho_2 \in Д$, то $\rho_2 + \rho_1 \in Д$;

2) $M[X(\rho)] = const$;

3) $M[X(\rho_1)X(\rho_2)]$ залежить тільки від $l = \rho_1 - \rho_2$

(які раніше, літера M визначає операцію математичного сподівання). Частіше за все, використовуються випадкові поля, для яких $Д$ - група цілочислових точок у R^n і $Д = R^n$.

Однорідне випадкове поле $X(\rho)$, визначене на R^m, називається ізотропним, якщо його коваріаційна матрична функція $K(l)$ залежить тільки від норми $|l|$ вектора в R^m (m позначає кількість точок, у яких визначається поле).

З цих означень випливає, що у однорідних полів $\bar{x}_i = const$; $\sigma_i^2 = const$, $\forall i = \overline{1,n}$, а K_{ij} залежить тільки від відстані між i-того та j - тою точками поля. Властивість ізотропності потребує ще й незалежності просторової корреляційної функції від напрямків полів корреляції.

Метеорологічні поля, як правило, не мають властивостей однорідності й ізотропності. Завдяки тому, що їх структура залежить від циркуляційних факторів і місцевих умов, які суттєво розрізнюються в різних районах території, яку охоплює метеорологічне поле, середні значення й дисперсії метеорологічних величин в різних точках полів мають різні значення. До такого висновку приходять, навіть враховуючи точність статистичних оцінок зазначених вище параметрів. Але досить часто можна використати просте лінійне перетворення вихідних випадкових величин і перейти до інших, випадкових

величин φ_{ij}, для яких вимоги властивостей однорідності й ізотропності виконуються, точніше майже виконуються. Поля таких випадкових величин називають квазіоднорідними і квазіізотропними. Зазначене лінійне перетворення має такий вигляд:

$$x_{ij} = \overline{x}_i + \sigma_i \varphi_{ij} \tag{2.2.1}$$

Очевидно, величини φ_{ij} є центрованими і нормованими, тобто

$$\varphi_{ij} = \frac{x_{ij} - \overline{x}_i}{\sigma_i} \tag{2.2.2}$$

Легко бачити, що , по-перше,

$$\overline{\varphi}_{ij} = 0, \forall\ i = 1, n \tag{2.2.3}$$

і, по-друге, матриця $\varphi = \{\varphi_{ij}\}$ перетворюється з матриці ΔX, що визначається рівністю (2.1.4), за допомогою матричної рівності

$$\varphi = \sigma^{-1} \Delta X \tag{2.2.4}$$

Але тоді матрицею коваріацій випадкових величин φ_{ij} є матриця корреляцій вихідних випадкових величин. Дійсно, як зазначалося у попередньому розділі,

$$K_{\varphi} = \frac{1}{m-1}\varphi'\varphi \qquad (2.2.5)$$

Підставимо до матричної рівності вираз (2.2.4), будемо мати:

$$\begin{aligned} K_{\varphi} &= \frac{1}{m-1}\sigma^{-1}\Delta X \times (\sigma^{-1}\Delta X) = \\ &= \frac{1}{m-1}\sigma^{-1}\Delta X \times \Delta X\sigma^{-1} = \\ &= \sigma^{-1}\left(\frac{1}{m-1}\Delta X \times \Delta X\right)\sigma^{-1} = \sigma^{-1}K_X\sigma^{-1} = \\ &= R_X \end{aligned} \qquad (2.2.6)$$

Звідси виходить, що $\sigma_{\varphi}^2 = 1 \forall i = \overline{1,n}$, а полями коваріацій для випадкових величин φ є поля корреляцій вихідних випадкових величин X. Тепер, щоб зробити висновок про однорідність і ізотропність полів φ, досить впевнитись у тому, що поля корреляцій, які містяться у матриці (2.1.15), для всіх полюсів корреляції мають подібний вигляд і визначаються

системою майже концентричних ізокорелят. До таких висновків можливо в багатьох випадках прийти, ураховуючи наявність довірчих інтервалів для корреляцій. На рис.2.1, 2.2, де наводяться поля корреляцій місячних кількостей опадів на території України у грудні, видно, що поля корреляцій в обидвох випадках мають подібний вид. Отже, оскільки для центрованих і нормованих на середній квадратичний відхил місячних кількостей опадів, які визначаються рівністю (2.2.2), середні значення дорівнюють нулю, а дисперсії - одиниці для всіх точок поля, а поля коваріацій не змінюються при змінюванні полюса корреляцій, то поля цих метеорологічних величин можна вважати квазіодніродними.

Зробимо тепер перерізи полів корреляцій у різних напрямках (рис.2.1 і 2.2). Це приведе до отримання просторових корреляційних функцій. Для зазначених напрямків перерізу корреляційні функції містяться в табл.2.1.

Як видно, всі корреляційні функції розташовуються у достатньо вузькій смузі, що дає підставу вважати, що структура полів корреляцій не залежить від напрямку. Таким чином, нормовані й центровані поля місячних опадів у грудні є не тільки квазіоднорідними, але й квазіізотропними. Їх статистична структура буде характеризуватися просторівою корреляційною функцією, яка знаходиться шляхом осереднення корреляційних функцій для декількох перерізів. У випадку, що наводиться в табл.2.1, осереднена корреляційна функція міститься на рис.2.3. Із рис.2.3 випливає, що з достатньою точністю її можна апроксимувати співвідношенням

$$r(l) = 1 - \frac{l}{l_0}, \qquad (2.2.7)$$

де $l_0 = 16$ одиниць масштабу, а l - поточне значення одиниць масштабу.

Оскільки ця корреляційна функція визначається на інтервалі $0 < l < l_0$, то, як зазначалося в розділі 6.3.3.2, їй відповідає спектральна щільність

$$S(\Omega) = \frac{1}{\pi\Omega^2 l_0}(1 - \cos\Omega l_0) \qquad (2.2.8)$$

де $\Omega = \frac{2\pi}{l}$

Графік цієї спектральної щільності зображається на рис.2.4. Видно, що енергетичний спектр місячної кількості опадів в грудні на території України є вузькосмуговим. Це означає, що основні особливості полів опадів формуються під впливом процесів великого масштабу. Це відповідає дійсності, оскільки, як відомо, основний внесок в опади вносить циклонічна діяльність.

Таблиця 2.1 - Просторові корреляційні функції місячних кількостей опадів у грудні.

Полюси	Номери напрямків перетинів полів	Номери одиниць масштабу (l) просторових корреляційних функцій								
		1	2	3	4	5	6	7	8	9
Чернігів	1	0,95	0,92	0,82	0,68	0,65	0,62	0,59	0,54	0,48
	2	0,93	0,87	0,84	0,82	0,79	0,74	0,65	0,55	0,48
	3	0,94	0,89	0,81	0,77	0,73	0,65	0,64	0,51	0,43
Дніпропетровськ	4	0,89	0,78	0,69	0,64	0,58	0,55	0,52	0,43	0,32
	5	0,87	0,78	0,68	0,60	0,48	0,44	0,40	0,38	0,36
	6	0,93	0,89	0,87	0,85	0,83	0,81	0,75	069	0,61
	$\overline{r}(l)$	0,92	0,86	0,79	0,73	0,68	0,64	0,59	0,52	0,45

Продовження табл. 2.1

Полюси	Номери напрямків перетинів полів	Номери одиниць масштабу (l) просторових корреляційних функцій			
		10	11	12	13
Чернігів	1	0,44	0,40		
	2	0,40	0,36	0,33	0,28
	3	0,35	0,27	0,20	
Дніпропетровськ	4	0,23			
	5	0,34	0,33	0,31	0,27
	6	0,5			
	$\bar{r}(l)$	0,38	0,34	0,28	0,27

2.3 Статистична інтерполяція метеорологічних полів.

2.3.1 Поняття про інтерполяцію й екстраполяцію.

Інтерполяцією називається визначення функції $f(x_0)$ в точці x_0 при умові, що є відомими значення хоча б в одній з точок відрізка $x \le x_1$ і хоча б в одній з точок відрізка $x \ge x_2$.

Точка x_0 розташовується між точками x_1 і x_2, тобто $x_1 < x_0 < x_2$.

Екстраполяцією називається визначення функції в точці x_0, якщо є відомим хоча б одне із значень цієї функції в точці $x \le x_1$, або в точці $x \ge x_2$. У першому випадку $x_0 > x_1$, у другому - $x_0 < x_2$.

Визначення залишаються незмінними і у тому разі, коли йдеться про функцію $f(\rho)$, де $\rho = \rho(x, y, z, t)$.

Інтерполяція і екстрополяція використовується в багатьох метеорологічних задачах. Охарактеризуємо деякі з них.

1. Прогноз метеорологічної величини є не що інше, як екстраполяція цієї величини на визначний момент часу.

2. Для того, щоб реалізувати чисельні методи прогнозу метеорологічного поля, треба знати значення метеорологічної величини в вузлах деякої регулярної сітки точок. Відомими ж є значення цієї величини на мережі метеорологічних станцій, які розташовуються на різних відстанях одна від одної. Отже, треба вирішити задачу інтерполяції метеорологічної величини із мережі метеорологічних станцій на вузли регулярної сітки точок.

3. На основі дослідження точності інтерполяції в горизонтальному напрямку вирішується задача оптимізації мережі метеорологічних станцій.

4. За допомогою інтерполяції за часом вирішується задача визначення оптимальних термінів метеорологічних спостережень.

5. Бувають випадки, коли при критичному контролі результатів метеорологічних спостережень на деякій метеорологічній станції виникають сумніви відносно значення того чи іншого параметра атмосфери. За допомогою інтерполяції цієї метеорологічної величини з інших метеорологічних станцій в точку розташування цієї станції проводиться контроль сумнівного значення метеорологічної величини.

За способом реалізації методи інтерполяції і екстраполяції поділяють на три вида: динамічну, статистичну та формальну. Динамічна базується на рівняннях гідродинаміки атмосфери; статистична основується на закономірностях, які отримуються шляхом статистичного аналізу масового матеріалу; формальна - являє собою апроксимацію метеорологічного поля за допомогою тієї чи іншої системи функцій. Прикладом формальної інтерполяції є апроксимація метеорологічного поля $f(x,y)$ поліномом

$$f(x,y) = A_1x^3 + A_2y^3 + A_3x^2y + A_4xy^2 + \\ + A_5xy + A_6x^2 + A_7y^2 + A_8x + A_9y + A_{10} \tag{2.31}$$

Поліном (2.3.1) дає прийнятну точність апроксимації гладких полів, таких як, наприклад, поля осереднених за досить великий проміжок часу значень температури, атмосферного тиску, складових вектора вітру на визначеному рівні атмосфери. Досить часто для апроксимації метеорологічних полів використовується їх розклад по базісних поліномах

$$f(x,y) = \sum_{n=0}^{k}\sum_{m=0}^{k} a_{mn}P_m(x)P_n(y) \tag{2.3.2}$$

У якості поліномів $P_m(x)$ і $P_n(y)$ виступають ортогональні поліноми Чебишева, Ерміта, Лежандра, Лаггера, експоненціальні поліноми тощо.

Нехай маємо значення метеорологічної величини у точках, що визначаються радіусами – векторами ρ_i $f(\rho_1); f(\rho_2); f(\rho_3); ...; f(\rho_n)$. Ставиться задача

проінтерполювати значення функції в точку ρ_0. Позначимо його через $\hat{f}(\rho_0)$. Екстрапольоване значення функції, очевидно, можна записати таким чином:

$$\hat{f}(\rho_0) = F[f(\rho_1), f(\rho_2), ..., f(\rho_n)] \qquad (2.3.3)$$

Вигляд функціоналу F залежить від метода інтерполяції і положення точки ρ_0 відносно точок $\rho_1, \rho_2, ...\rho_n$. Позначивши точне значення функції через $f(\rho_0)$, отримаємо похибку інтерполяції

$$\hat{\delta}_f(\rho_0) = \hat{f}(\rho_0) - f(\rho_0) \qquad (2.3.4)$$

Виникає питання, від чого залежить ця похибка? Справа у тому, що в дійсності ми маємо діло не з функціями $f(\rho_i)$, а з функціями $\tilde{f}(\rho_i)$, які утримують похибку вимірювання $\delta_f(\rho_i)$, тобто

$$\tilde{f}(\rho_i) = f(\rho_i) + \delta_t(\rho_i) \qquad (2.3.5)$$

Таким чином,

$$\hat{f}(\rho_0) = F[\tilde{f}(\rho_1), \tilde{f}(\rho_2), ..., \tilde{f}(\rho_n)] \qquad (2.3.6)$$

Ясно, що від $\delta_f(\rho_i)$ будуть залежати й результати інтерполяції.

Часто в метеорологічних задачах використовують у якості функціоналу F лінійну форму, тобто

$$\hat{f}(\rho_0) = \sum_{i=1}^{n} a_i \tilde{f}(\rho_i) \qquad (2.3.7)$$

Інтерполяція, що проводиться за допомогою рівності (2.3.7), називається лінійною інтерполяцією. Коефіцієнти a_i є ваговими множниками (вагами).

2.3.2 Точність інтерполяції.

Як зазначалося у попередньому розділі, при інтерполяції будь-якої метеорологічної величини виникає похибка, яка визначається рівністю (2.3.4). Мірою цієї похибки будемо вважати її середній квадрат

$$E^2 = \overline{[\hat{\delta}_f(\rho_0)]^2}, \qquad (2.3.8)$$

де риска зверху позначає операцію осереднення.

Покажемо, що у випадку лінійної інтерполяції середній квадрат похибки інтерполяції визначається характеристиками статистичної структури поля метеорологічної величини, яке інтерполюється, а також похибки вимірювань. Для цього

підставимо в формулу (2.3.8) вирази (2.3.4) і (2.3.7). Будемо мати

$$E^2 = \overline{[\sum_{i=1}^{n} a_i \widetilde{f}(\rho_i) - f(\rho_0)]^2} \qquad (2.3.9)$$

Функції, що містяться в виразі (2.3.9), можна записати таким чином:

$$\widetilde{f}(\rho_i) = \bar{f}(\rho_i) + \widetilde{f}'(\rho_i) \qquad (2.3.10)$$

$$f(\rho_0) = \overline{f(\rho_0)} + f'(\rho_0) \qquad (2.3.11)$$

де штрихами позначаються пульсації функцій. З урахуванням цього, формула (2.3.9) перетворюється, тобто

$$E^2 = \overline{\left\{ \begin{array}{l} \left[\sum_{i=1}^{n} a_i \overline{f}(\rho_i) - \overline{f}(\rho_0) \right] + \\ + \left[\sum_{i=1}^{n} a_i f'(\rho_i) - f'(\rho_0) \right] \end{array} \right\}^2} \qquad (2.3.12)$$

Праву частину рівності (2.3.12) піднесемо до другого степеня. Отримаємо:

$$
\begin{aligned}
E^2 = & \overline{\left[\sum_{i=1}^{n} a_i \overline{f}(\rho_i) - \overline{f}(\rho_0)\right]^2} + \\
& + \overline{\left[\sum_{i=1}^{n} a_i \widetilde{f}'(\rho_i) - f'(\rho_0)\right]^2} + \\
& + 2\overline{\left[\sum_{i=1}^{n} a_i \overline{f}(\rho_i) - \overline{f}(\rho_0)\right]} \times \\
& \times \overline{\left[\sum_{i=1}^{n} a_i \widetilde{f}'(\rho_i) - f'(\rho_0)\right]}
\end{aligned}
\tag{2.3.13}
$$

Проаналізуємо праву частину виразу (2.3.13) використовуючи властивості оператора осереднення. Очевидно

$$
\begin{aligned}
\overline{\left[\sum_{i=1}^{n} a_i \overline{f}(\rho_i) - \overline{f}(\rho_0)\right]^2} &= \left[\begin{array}{l} \sum_{i=1}^{n} \overline{a_i \overline{f}(\rho_i)} - \\ - \overline{\overline{f}(\rho_0)} \end{array}\right]^2 = \\
&= \left[\sum_{i=1}^{n} a_i \overline{f}(\rho_i) - \overline{f}(\rho_0)\right]^2 = (\overline{\hat{\delta}})^2
\end{aligned}
\tag{2.3.14}
$$

є квадратом середньої похибки інтерполяції.

$$\overline{\left[\sum_{i=1}^{n} a_i \overline{f}(\rho_i) - \overline{f}(\rho_0)\right]\left[\sum_{i=1}^{n} a_i \widetilde{f}'(\rho_i) - f'(\rho_0)\right]}$$

$$= \left[\sum_{i=1}^{n} a_i \overline{f}(\rho_i) - \overline{f}(\rho_0)\right] \times$$

$$\times \left[\sum_{i=1}^{n} a_i \overline{\widetilde{f}'(\rho_i)} - \overline{f'(\rho_0)}\right] = 0,$$

оскільки

$$\overline{\widetilde{f}}'(\rho_i) = \overline{f}'(\rho_0) = 0$$

Позначимо

$$E'^2 = \overline{\left[\sum_{i=1}^{n} a_i \widetilde{f}'(\rho_i) - f'(\rho_0)\right]^2} \qquad (2.3.15)$$

Тоді маємо

$$E^2 = E'^2 + (\overline{\hat{\delta}})^2 \qquad (2.3.16)$$

Розглянемо смисл складової E'^2. Для цього у рівності (2.3.15) врахуємо, що

$$\widetilde{f}'(\rho_i) = \widetilde{f}(\rho_i) - \overline{f}(\rho_i) = f(\rho_i) - \overline{f}(\rho_i) + \\ + \delta_f(\rho_i) = f'(\rho_i) + \delta_f(\rho_i)$$

Отже

$$E'^2 = \overline{\left[\sum_{i=1}^{n} a_i[\widetilde{f}'(\rho_i) + \delta_f(\rho_i)] - f'(\rho_0)\right]^2} \tag{2.3.17}$$

Якщо піднести до другого степеня праву частину рівняння (2.3.17), то отримаємо:

$$E'^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \{\overline{f'(\rho_i) f'(\rho_j)} + \\ + \overline{\delta_f(\rho_i) f'(\rho_j)} + \overline{\delta_f(\rho_i) f'(\rho_j)} + \\ + \overline{\delta_f(\rho_i)\delta_f(\rho_j)}\} - 2\sum_{i=1}^{n} a_i[\overline{f'(\rho_i) f'(\rho_0)} - \\ - \overline{\delta_f(\rho_i) f'(\rho_0)}] + \overline{[f'(\rho_0)]^2} \tag{2.3.18}$$

або, оскільки

$$\begin{cases} \overline{f'(\rho_i)f'(\rho_j)} = K_f(\rho_i,\rho_j), \\ \overline{f'(\rho_i)f'(\rho_0)} = K_f(\rho_i,\rho_0), \\ \overline{f'(\rho_i)\delta_f(\rho_j)} = \overline{f'(\rho_j)\delta_f(\rho_i)} = \\ = K_{f\delta}(\rho_i,\rho_j), \\ \overline{\delta_f(\rho_i)f'(\rho_i)} == K_{f\delta}(\rho_i,\rho_0), \\ [\overline{f'(\rho_0)}]^2 = \sigma_f^2(\rho_0), \end{cases} \tag{2.3.19}$$

$$E'^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j [K_f(\rho_i,\rho_j) + 2K_{f\delta}(\rho_i,\rho_j) + + K_\delta(\rho_i,\rho_j)] - 2\sum_{j=1}^{n} a_i [K_f(\rho_i,\rho_0) - - K_{\delta f}(\rho_i,\rho_0)] + \sigma_f^2(\rho_0) \tag{2.3.20}$$

Як відомо, виконуються такі умови:

- похибки вимірювань не коррелюються з величиною, що вимірюється. Тому $K_{f\delta}(\rho_i,\rho_j) = 0; \forall i,j = \overline{0,n}$;

- похибки вимірювань в різних точках поля не коррелюють між собою, тобто

$$K_\delta(\rho_i,\rho_j) = \begin{cases} \sigma_\delta^2(\rho_i) & при \quad i = j \\ 0 & при \quad i \neq j \end{cases} \tag{2.3.20}$$

З урахуванням зазначених умов рівняння (2.3.20) спрощується і має вид

$$E'^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K_f(\rho_i,\rho_j) + \sum_{j=1}^{n} a_i^2 \sigma_\delta^2(\rho_i) -$$

$$- 2\sum_{i=1}^{n} a_i K_f(\rho_i,\rho_0) + \sigma_f^2(\rho_0) \qquad (2.3.21)$$

Отже, рівності (2.3.16) і (2.3.21) свідчать про те, що середній квадрат похибки інтерполяції E^2 складається з квадрата середньої похибки інтерполяції, а також величини E'^2, яка у повній мірі визначається характеристиками статистичної структури поля, яке інтерполюється, і дисперсією похибки вимірювань метеорологічної величини в точках поля.

2.3.3 Оптимальна інтерполяція.

Оперуючи з лінійною моделлю (2.3.7) ми прийняли припущення, що коефіцієнти a_i є знанними величинами. Але в дійсності ці коефіцієнти підлягають визначенню на основі інформації про статистичну структуру полів. Ясно, що метод визначення цих коефіцієнтів повинен бути таким, що б він давав змогу знайти найкращі у певному смислі значення коефіцієнтів. Таким чином, ця задача є задачою оптимізації моделі (2.3.7). Щоб вирішити таку оптимізаційну задачу, треба, по-перше, визначити критерій якості.

Природно вважати, що таким критерієм є мінімум середнього квадрату похибки інтерполяції E^2. Отже, задача полягає у тому, щоб знайти такі коефіцієнти $a_i (i = \overline{1,n})$, які б

приводили до мінімуму середнього квадрата інтерполяції. З врахуванням рівності (2.3.16) сформульована умова має такий вид

$$\frac{\partial E'^2}{\partial a_i} = 0, \ \forall \ i = \overline{1,n} \qquad (2.3.22)$$

Застосовуючи зазначену операцію диференціювання до рівняння (2.3.21), маємо

$$\frac{\partial E'^2}{\partial a_i} = 2\sum_{j=1}^{n} a_j [K_f(\rho_i, \rho_j) + \sigma_\delta^2(\rho_j)] - \\ - 2K_f(\rho_0, \rho_i) = 0 \\ (i = \overline{1,n}) \qquad (2.3.23)$$

або

$$\sum_{j=1}^{n} a_j [K_f(\rho_i, \rho_j) + \sigma_\delta^2(\rho_j)] = K_f(\rho_0, \rho_i) \\ (i = \overline{1,n}) \qquad (2.3.24)$$

Оскільки характеристики корреляційної структури метеорологічних полів і статистичні характеристики похибок вимірювань відомі (принципи досліджень корреляційного аналізу метеорологічних полів розглядалися в попередніх розділах) можна позначити

$$K_{ij} = K_f(\rho_i, \rho_j) + \sigma_\delta^2(\rho_j) \qquad (2.3.25)$$

і тоді маємо систему лінійних алгебраїчних рівнянь

$$\sum_{j=1}^{n} K_{ij} a_j = K(\rho_i, \rho_0) \qquad (2.3.26)$$

відносно шуканих коефіцієнті a_j лінійної інтерполяційної моделі (2.3.7). Систему рівнянь (3.2.26) можна записати в матричній формі

$$KA = K_0 \qquad (2.3.27)$$

якщо позначити

$$K = \begin{pmatrix} K_{11} & K_{12} & \dots & K_{1n} \\ K_{21} & K_{22} & \dots & K_{2n} \\ \dots & \dots & \dots & \dots \\ K_{n1} & K_{n2} & \dots & K_{nn} \end{pmatrix} \qquad (2.3.28)$$

$$K_0 = \begin{pmatrix} K_f(\rho_1, \rho_0) \\ K_f(\rho_2, \rho_0) \\ \dots \\ K_f(\rho_n, \rho_0) \end{pmatrix} \qquad (2.3.29)$$

$$A = \begin{pmatrix} a_1 \\ a_2 \\ \dots \\ a_n \end{pmatrix} \qquad (2.3.30)$$

Якщо $|K| \neq 0$, то розв'язок системи (3.2.27) буде мати вид

$$A = K^{-1} K_0 \qquad (2.3.31)$$

При вирішенні задачі оптимальної інтерполяції виникає ряд труднощів. Якщо дисперсії похибок вимірювання, взагалі кажучи, відомі, то коваріаційні моменти треба розраховувати на основі множини полів метеорологічної величини, що визначена в n точках поля. Для цього, по-перше, треба строго закріпити положення точок ρ_i і ρ_j. В дійсності при інтерполяції функції f в точку ρ_0 визначається деяка кількість впливаючих точок. Але заздалегідь невідомо, чи буде значення функції $f(\rho)$ у цих точках знаним на час виконання інтерполяції. Добре відомо, що в метеорологічній інформації бувають прогалини. Дуже рідкою є метеорологічна мережа на

океанічних просторах і дуже часто набуває потреба використовувати корабельну інформацію. Але положення корабля наперед визначити неможливо.

Задача значно спрощується для однорідних та ізотропних полів.

Для них, як відомо, $\overline{f} = const$; $\sigma_f^2 = const$, а $K_f(\rho_i, \rho_j) = K_f\left|\rho_i - \rho_j\right| = K_f(d_i)$, де d_i- відстані між точками поля. У більшості випадків метеорологічні поля не мають властивостей однорідності та ізотропності. Проте часто є можливість вирішити задачу оптимальної інтерполяції за допомогою модифікованих полів, які зазначеними властивостям підлягають.

Розглянемо функції:

$$\varphi_0 = \frac{f'(\rho_0)}{\sigma_f(\rho_0)} \qquad (2.3.32)$$

$$\varphi_i = \frac{\widetilde{f}'(\rho_i)}{\sigma_f(\rho_i)} \qquad (2.3.33)$$

Оскільки в чисельниках цих рівностей розташовуються пульсації функцій, тобто відхилення їх від середнього значення, то

$$\overline{\varphi}_i = 0, \ \forall \ i = \overline{0,n}, \ \sigma_{\varphi_i}^2 = 1, \ \forall \ i = \overline{0,n},$$

$K_{\varphi}(\rho_i, \rho_j) = r_f(\rho_i, \rho_j)$ - коваріаційна функція. З достатнього для практики точністю можна припустити, що $r_f(\rho_i, \rho_j) = r_f(d_i)$, тобто, що поля функції $\varphi(\rho_i)$- однорідні і ізотропні. Приклад такої корреляційної функції наводиться у попередньому розділі.
Якщо

$$\hat{f}'(\rho_0) = \sum_{i=1}^{n} a_i \tilde{f}'(\rho_i) \qquad (2.3.34)$$

то з урахуванням рівностей (2.3.32) і (2.3.33) маємо

$$\hat{\varphi}(\rho_0) = \sum_{i=1}^{n} b_i \varphi(\rho_i) \qquad (2.3.35)$$

де

$$b_i = \frac{\sigma_f(\rho_i)}{\sigma_f(\rho_0)} a_i \qquad (2.3.36)$$

Вирішуючи задачу оптимальної інтерполяції для функції $\varphi(\rho_i)$, що дає можливість без перелічених вище труднощів знайти коефіцієнти інтерполяції b_i, за допомогою рівності (2.3.36) можна, знаючи середні квадратичні відхили метеорологічної величини в точках поля, отримати коефіцієнти

a_i, тобто вид інтерполяційного поліному для вихідної метеорологічної величини $f(\rho)$.

Задача оптимальної інтерполяції - дуже важлива. На її основі здійснюється об'єктивний аналіз метеорологічних полів, який зводиться до знаходження значень метеорологічної величини у вузлах регулярної сітки точок, а також вирішується задача узгодження полів метеорологічних величин при реалізації чисельних (гідродинамічних) методів метеорологічних прогнозів.

2.4 Розклад вертикальних профілів метеорологічних величин у трикутному базису

2.4.1 Побудова розкладу випадкового вектора у трикутному базису

При розв'язках деяких практичних задач виникає необхідність мати діло з математичними моделями фізичних параметрів атмосфери (температури, атмосферного тиску, компонентів швидкості вітру тощо). Прикладом такої задачі є задача врахування впливів атмосферних збурень на динаміку літального апарату. В залежності від типу літального апарату, його призначення, конструктивних і аеродинамічних особливостей, використовується та чи інша система диференціальних рівнянь. У загальному виді її можна записати так:

$$\dot{Y} = F(Y, U, X, t), \quad Y(t_0) = Y_0 \qquad (2.4.1)$$

де $\dot{Y}$ - l - вимірний вектор похідних фазових координат літального апарата за часом, U - r - вимірний вектор керуючих сил і моментів; X - n - вимірний вектор атмосферних збурень,

Y_0 - l - вимірний вектор початкових умов для фазових координат; t - поточний час.

При визначних умовах, вектором атмосферних збурень може бути вектор, координати якого є значення метеорологічної величини на деяких стандартних висотах, тобто її вертикальний профіль. Математичні моделі вертикальних профілів можуть мати різний вид: розклад у базису власних векторів матриць коваріацій, ряди по ортогональних функціях (Чебишева, Ерміта, Лагера, експоненціальних функціях тощо). У деяких випадках більш корисним є застосування для моделювання вертикальних профілів або часових рядів метеорологічних величин трикутного базису. Розглянемо його структуру, властивості й алгоритм побудови.

Як вже неодноразово зазначалося, фізичні параметри атмосфери є нестаціонарними функціями координат трьохвимірного простору й часу. Розглянемо випадкову функцію $x(h)$, де h - висота над рівнем моря, що характеризує розподіл по висоті деякого параметра атмосфери в визначному пункті. Якщо є відомим математичне сподівання $m_X(h)$ і дисперсія $\sigma_X^2(h)$ цієї функції, то можна побудувати центровану й нормовану функцію

$$\varphi(h) = \frac{x(h) - m_x(h)}{\sigma_X(h)} \qquad (2.4.2)$$

Очевидно, для такої функції $m_\varphi == 0$; $K_\varphi(h,h') = R_X(h,h')$ (K_φ і R_X, як і раніше, визначають коваріацію і корреляцію відповідно). Отже, можна випадкову функцію $x(h)$ представити таким чином:

$$x(h) = m_X(h) + \sigma_X(h)\varphi(h) \qquad (2.4.3)$$

і створювати модель не вихідних параметрів атмосфери $x(h)$, а функцій $\varphi(h)$ на основі коваріаційної матриці

$$R_X = M(\varphi, \varphi') \qquad (2.4.4)$$

де φ - вертикальний профіль цієї модифікованої величини, тобто вектор визначений своїми координатами на тій же множині стандартних висот.

Розглянемо модель вертикального розподілу параметра атмосфери у виді розкладу його у трикутному базису

$$\varphi = P\xi \qquad (2.4.5)$$

де

$$P = \begin{pmatrix} P_{11} & 0 & 0 & \dots & 0 \\ P_{21} & P_{22} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ P_{n1} & P_{n1} & P_{n3} & \dots & P_{nn} \end{pmatrix} \qquad (2.4.6)$$

а ξ - вектор, що визначає координати випадкового вектора φ у базису P. Будемо вважати, що вектор ξ має такі властивості:

$$\begin{cases} M[\xi] = 0, \\ M[\xi\xi'] = E, \end{cases} \tag{2.4.7}$$

де E - одинична матриця.

Тоді

$$R_X = M[\varphi\varphi'] = M[P\xi\xi'P'] = PEP' = PP' \tag{2.4.8}$$

Отже задача полягає у тому, щоб знайти таку нижню трикутну матрицю P, добуток якої на відповідну верхню трикутну матрицю P' дає матрицю корреляцій.

Поставимо рівність (2.4.5) у праву частину рівняння (2.4.4). Виконавши операцію множення матриць і розгорнувши матрицю корреляцій (запишемо тільки верхню половину цих матриць які є симетричними, будемо мати:

$$\begin{pmatrix} r_{11} & r_{12} & r_{13} & \dots & r_{1n} \\ & r_{22} & r_{23} & \dots & r_{2n} \\ & & r_{33} & \dots & r_{3n} \\ & & & \dots & \dots \\ & & & & r_{nn} \end{pmatrix} =$$

$$= \begin{pmatrix} P_{11}^2 & P_{11}P_{21} & P_{11}P_{31} & \dots & P_{11}P_{n1} \\ & P_{21}^2 + P_{22}^2 & P_{21}P_{31} + P_{22}P_{32} & \dots & P_{21}P_{n1} + P_{22}P_{n2} \\ & & P_{11}^2 + P_{22}^2 + P_{33}^2 & \dots & P_{31}P_{n1} + P_{32}P_{n2} + P_{33}P_{n3} \\ & & & \dots & \dots \\ & & & & P_{n1}^2 + P_{n2}^2 + P_{n3}^2 + \dots + P_{nn}^2 \end{pmatrix} \quad (2.4.9)$$

Діагональні елементи матриці корреляцій, як відомо, дорівнюють $r_{11} = r_{22} = \dots = r_{nn} = 1$. Крім того, аналіз елементів матриці PP', котрі розташовуються на перетині її i - того рядку і j - того стовпця, показує, що вони є сумою виду:

$$\sum_{j=1}^{i} P_{ij} P_{kj} \quad (i,k = \overline{1,n})$$

Отже, з урахуванням рівності (2.4.9), для визначення елементів P_{ij} матриці (2.4.6) маємо $\frac{n(n+1)}{2}$ рівнянь виду:

$$r_{ik} = \sum_{j=1}^{i} P_{ij} P_{kj} \qquad (i,k = \overline{1,n}) \tag{2.4.10}$$

Оскільки матриця корреляцій відома, система цих рекурентних рівнянь легко розв'язується. Маючи на увазі рівняння (2.4.3), модель вертикальних профілів фізичних параметрів атмосфери можна представити у матричній формі таким чином:

$$X = m_X + \sigma_X P \xi, \tag{4.2.11}$$

де σ_X - діагональна матриця середніх квадратичних відхилів випадкової величини X. В таблиці 2.2 у якості прикладу міститься матриця корреляцій температури повітря влітку в пункті Уайт Сендз, а в табл.2.3 компоненти відповідної трикутної матриці P.

З урахуванням (2.4.7) можна вважати, що компоненти вектора ξ володіють властивостями нормально розподілених випадкових величин. Це дає можливість формувати такий вектор за допомогою датчика випадкових чисел, якщо, наприклад, розглядається задача визначення розсіювання траєкторій динамічної системи (2.4.1).

Можна змінити гіпотезу про випадкові координати вектора φ у трикутному базисі. Це приведе до зміни структури трикутної матриці. Будемо вважати, що координати $\theta_i (i = \overline{1,n})$ вектора φ в деякому трикутному базису мають такі властивості :

$$\begin{cases} M[\theta_i] = 0 \\ M[\theta_i \theta_j] = Д_i \delta_{ij} \end{cases} \qquad (2.4.12)$$

(δ_{ij} - символ Кронекера).

Тоді

$$M[\theta\theta'] = Д, \qquad (2.4.13)$$

де $Д$ - діагональна матриця, елементами якої є $Д_i$ - дисперсії випадкових компонент вектора θ.

Будемо вважати, що

$$\varphi = T\theta, \qquad (2.4.14)$$

де T - нижня трикутна матриця.

У цьому разі

$$R_x = M[\varphi\varphi'] = M[T\theta\theta'T'] = TДT' \qquad (2.4.15)$$

Оскільки є справедливою матрична рівність

$$Д = Д^{\frac{1}{2}} Д^{\frac{1}{2}}, \qquad (2.4.16)$$

і $Д^{\frac{1}{2}} = (Д^{\frac{1}{2}})'$,

то маємо

$$R_X = TД^{\frac{1}{2}}(Д^{\frac{1}{2}})'T' \qquad (2.4.17)$$

Порівнюючи рівності (2.4.8) і (2.4.17), маємо

$$P = TД^{\frac{1}{2}}, \qquad (2.4.18)$$

звідки

$$T = PД^{-\frac{1}{2}}, \qquad (2.4.19)$$

Остання рівність дає можливість просто знайти компоненти матриці T- трикутного базису, у якому проводиться розкладання випадкового вектора φ.

2.4.2 Канонічний розклад випадкового вектора

Для того, щоб впровадити розклад вертикального профілю метеорологічної величини у трикутному базису (2.4.14), треба знайти вектор θ проекцій вектора φ у цьому базису. З цією метою розглянемо канонічний розклад випадкового вектора φ за В.С.Пугачевим. Він має такий вид:

$$\varphi(h) = \sum_{v=1}^{n} \upsilon_v y_v(h) \qquad (2.4.20)$$

Функції $y_v(h)$ мають назву координатних функцій, а υ_v - випадкових коефіцієнтів розкладу. Ці елементи канонічного розкладу розраховуються по рекурентних формулах

$$\begin{cases} \upsilon_v = \varphi(h_v) - \sum_{i=1}^{v-1} \upsilon_i y_i(h_v); \\ Д_v^* = 1 - \sum_{i=1}^{v-1} Д_i^* [y_i(h_v)]^2; \\ y_v(h) = \dfrac{1}{Д_v^*} [r(h, h_v) - \sum_{i=1}^{v-1} Д_i^* y_i(h) y_i(h_v)] \end{cases} \qquad (2.4.21)$$

Із рівності (2.4.21) випливає, що

$$\begin{aligned} &Д_1^* = 1; y_1(h_1) = r(h_1, h_1) = 1; y_1(h_2) = \\ &= r(h_1, h_2); ...; y_1(h_n) = r(h_1, h_n)` \end{aligned} \qquad (2.4.22)$$

і, крім того,

$$y_\mu(h_v) = 0 \ при \ \ \mu > v \qquad (2.4.23)$$

Звідси виходить, що матриця координатних функцій є трикутною, а перший стовпець цієї матриці дорівнює першому стовпцю матриці корреляцій. Для другого стовпця маємо:

$$\begin{cases} y_2(h_2) = 1, Д_2^* = \dfrac{1}{1 - r^2(h_1, h_2)}; \\ y_2(h_3) = \dfrac{1}{1 - r^2(h_1, h_2)}[r(h_2, h_3) - \\ - r(h_1, h_2) r(h_1, h_3)]; \\ y_2(h_4) = \dfrac{1}{1 - r^2(h_1, h_2)}[r(h_2, h_4) - \\ - r(h_1, h_2) r(h_1, h_4)]; \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_2(h_n) = \dfrac{1}{1 - r^2(h_1, h_2)}[r(h_2, h_n) - \\ - r(h_1, h_2) r(h_1, h_n)]; \end{cases} \qquad (2.4.24)$$

і так далі для решти стовпців матриці координатних функцій.

Звернемося тепер до трикутної матриці T, що визначається рівністю (2.4.19).

Оскільки всі елементи матриці P відомі і, як очевидно, $Д_1 = M[\theta_1^2] = 1$, то елементи першого стовпця матриці T дорівнюють:

$$\begin{array}{l} t_{11} = 1; t_{21} = r(h_2, h_1); t_{31} = r(h_3, h_1); \ldots; \\ t_{n1} = r(h_n, h_1) \end{array} \qquad (2.4.25)$$

Як випливає з рівностей (2.4.10) і (2.4.13)

$$Д_2 = M[\theta_2^2] = 1 - r_{12}^2$$

(корреляції r_{ij} і $r(h_i, h_j)$ - рівнозначні). Тому елементи другого стовпця матриці T мають значення:

$$\begin{cases} t_{22} = 1; \\ t_{23} = \dfrac{1}{1 - r_{12}^2}(r_{23} - r_{12} r_{13}); \\ t_{24} = \dfrac{1}{1 - r_{12}^2}(r_{24} - r_{12} r_{14}); \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ t_{2n} = \dfrac{1}{1 - r_{12}^2}(r_{2n} - r_{12} r_{1n}); \end{cases} \qquad (2.4.26)$$

Таким же чином розраховуються елементи інших стовпців матриці T. Порівняння відповідних рівностей (2.4.24) і (2.4.26)

показує, що елементи відповідних стовпців матриці координатних функцій і матриці T тотожні.

Координати $\theta_i (i = \overline{1,n})$ випадкового вектора φ у трикутному базису T можна знайти із виразу (2.4.14), якщо визначити вектор φ. Оскільки матриця T - трикутна, матричне рівняння (2.4.14) утворює систему рекурентних формул

$$\begin{cases} \theta_1 = \varphi_1; \\ \theta_2 = \varphi_2 - t_{12}\theta_1; \\ \theta_3 = \varphi_3 - t_{32}\theta_2 - t_{31}\theta_1; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \theta_n = \varphi_n - t_{n,n-1}\theta_{n-1} - t_{n,n-2}\theta_{n-2} - \ldots - t_{n1}\theta_1, \end{cases} \tag{2.4.27}$$

або, якщо їх згорнути,

$$\theta_v = \varphi_v - \sum_{i=1}^{v-1} \theta_i t_{vi} \quad (v = \overline{1,n}), \tag{2.4.28}$$

що повністю співпадає з першим рівнянням системи (2.4.21). Отже, отримані вище рішення показують, що розклад випадкового вектора φ в трикутному базису T, що визначається матричним рівнянням (2.4.14), є тотожнім до канонічного розкладу (2.4.20) В.С. Пугачева. Однак алгоритм (2.4.14) - (2.4.19) і (2.4.28) має перевагу перед розв'язками, що отримані В.С.Пугачевим для побудови координатних функцій і визначення коефіцієнтів υ_v, оскільки вони дають змогу замінити дуже простими матричними рівняннями достатньо

громіздкі математичні викладки, які приводять до того ж результату. Отже, у якості вектора випадкових атмосферних збурень у правій частині моделі динамічної системи (2.4.1) (а також вертикального профілю фізичного параметра атмосфери взагалі) можна використовувати поряд з моделлю (2.4.11) модель.

$$X = m_X + \sigma T \theta \qquad (2.4.29)$$

Крім того, з рівнянь (2.4.5), (2.4.14) і (2.4.18) видно, що між векторами θ і ξ справедливим є взаємозв'язок

$$\theta = Д^{\frac{1}{2}} \xi \qquad (2.4.30)$$

який має смисл перетворення стиску. Треба зауважити, що розглянуті трикутні базиси можна використовувати для моделювання не тільки вертикальних профілів метеорологічних величин, але і для побудови моделей реалізацій будь-яких випадкових функцій.

3. КОМПОНЕНТНИЙ АНАЛІЗ МЕТЕОРОЛОГІЧНИХ ОБ'ЄКТІВ

3.1 Напрямки використання компонентного аналізу в метеорології

Застосування статистичних методів при розв'язках чисельних метеорологічних задач пов'язане з рядом труднощів. Це непідпорядкування в багатьох випадках метеорологічних величин нормальному закону розподілу, неоднорідність та неізотропність метеорологічних полів, нестаціонарність випадкових процесів, що характеризують змінення метеорологічних величин за часом і т.д. Існує ще одне суттєве обмеження, яке пов'язане з розробкою статистичних моделей метеорологічних прогнозів. Йдеться про дуже велику розосередженість вихідної інформації, яка необхідна для прийняття заключення про майбутній стан атмосферних процесів. Це утруднення, здається на перший погляд, можливо обійти шляхом збільшення числа впливаючих факторів (предікторів) і розширення об'ємів статистичних сукупностей. Але чим більше взяти предікторів, тим більшою стає розмірність матриці коваріацій і тим гіршою буде її обумовленість. Останній факт пояснюється тим, що збільшення в прогностичній моделі кількості предікторів приводить до збільшення тісноти корреляційних зв'язків між ними. Відомо, що при обчисленні коефіцієнтів рівняння моделі, якою є, наприклад, регресійне рівняння, або параметрів дискрімінантної функції, коли модель грунтується на теорії розпізнавання образів, необхідно виконувати процедуру обернення коваріаційної матриці. Якщо остання погано обумовлена (тобто визначник її має невелике значення), обернення матриці приводить до дуже ненадійних результатів. Таким чином, параметри статистичної моделі можуть містити в собі великі похибки, і модель не буде у необхідній мірі

адекватною процесу, що моделюється. Тому склад предікторів повинен бути у визначній мірі оптимальним. Це досягається за допомогою так званих методів "просіювання".

Часто ефективним виявляється й інший шлях. Він полягає у тому, що спочатку проводиться параметризація складу впливаючих факторів, тобто замість цих факторів у моделі використовуються нові змінні у вигляді лінійних комбінацій вихідних факторів. Нові змінні повинні бути, по-перше, взаємно некоррельованими (ортогональними). По-друге, щоб при можливо меншій їх кількості ураховувалася суттєва частина мінливості вихідних факторів (предиктантів). Переліченим вимогам відповідає компонентний аналіз, який часто в метеорологічній літературі називають методом "емпіричних ортогональних функцій (е.о.ф) або "природних ортогональних функцій".

Можна перелічити велику кількість ортогональних функцій, які можуть використовуватися, і інколи використовуються, для апроксимації двовимірних метеорологічних полів. Це, наприклад, поліноми Лежандра, Лаггера, Ерміта, Чебишева I і II роду і інші. Однак, розклади метеорологічних полів на їх основі є у великій мірі формальними у тому смислі, що коефіцієнти розкладів не несуть, як правило, смислової інформації. У назві "природні ортогональні функції" підкреслюється, наперед всього, той факт, що отримані на основі вихідних метеорологічних полів (або інших метеорологічних об'єктів) ці функції відбивають основні особливості статистичної структури полів. Завдяки некоррельованості нових змінних спрощується фізичний аналіз статистичних моделей, не виникає обчислювальних труднощів при розв'язаннях систем нормальних рівнянь, необхідних для визначення моделей. Не виникає великих проблем при необхідності повернутися до вихідних змінних.

Окрім перелічених переваг, нові змінні, яких називають головними компонентами, мають і самостійне значення. Вони часто виявляються відбиттям тих чи інших фізичних процесів, як і обумовлюють метеорологічну величину, що прогнозується, тобто несуть в собі смислове навантаження.

Компонентний аналіз може з успіхом застосовуватися й при розв'язках інших метеорологічних задач. Однією з них є стиск метеорологічної інформації: стиском інформації називають суттєве скорочення кількості інформації при збереженні основного її змісту. Це є надзвичайно важливою обставиною в задачі збереження інформації і її безпосереднього використання у практичних цілях.

Другою важливою задачею, котру дає можливість реалізувати компонентний аналіз, є задача фільтрації метеорологічної інформації. Суть її полягає у такому. Як відомо, поля метеорологічних величин, а також інші метеорологічні об'єкти, формуються під дією атмосферних процесів різних масштабів: процесів макромасштабу, синоптичного масштабу, мезомасштабу, процесів ще більш дрібного масштабу. Метеорологічна інформація утримує і шумову компоненту, яка обумовлена дрібномасштабними флуктуаціями, похибками вимірювань та первинної обробки результатів спостережень. Часто виникає необхідність, в залежності від характеру задачі при вивченні явищ погоди, зосередити увагу на процесах більших масштабів і відвернутися від розглядання складових, що обумовлені впливами процесів дрібних масштабів. Для цього також може застосовуватись компонентний аналіз.

3.2 Власні вектори і власні значення матриці коваріацій.

Нехай маємо деяке поле центрованих значень метеорологічної величини ΔX_j

$$\Delta X_j = \begin{pmatrix} \Delta x_{1j} \\ \Delta x_{2j} \\ .... \\ \Delta x_{nj} \end{pmatrix} \qquad (3.2.1)$$

Треба здійснити параметризацію цього поля, тобто виразити поле, яке визначене значеннями метеорологічної величини на множині точок простору, за допомогою декількох некоррельованих параметрів, які лінійно зв'язані з компонентами випадкового вектора (3.2.1) і які утримують основну інформацію про поле. В основі розв'язку цієї задачі лежить лінійне ортогональне перетворення вихідного поля (3.2.1) в базісу власних векторів матриці коваріацій (або корреляцій) полів цієї метеорологічної величини. Отже, першим етапом цієї задачі є визначення власних векторів. Для цього розглянемо матричне рівняння повної проблеми власних значень

$$K_X u_i = \lambda_i u_i \qquad (3.2.2)$$

В цьому рівнянні K_X - n - вимірна матриця коваріацій, u_i - i - тий власний вектор, λ_i - відповідне власне значення (власне число) матриці K_X. Рівняння (3.2.2) значить, що власний вектор так перетворюється матрицею коваріацій, як він трансформується шляхом множення його на власне число λ_i. Таке перетворення вектора, як відомо, є перетворенням розтягу, коли $\lambda_i > 1$, і перетворенням стиску, коли $\lambda_i < 1$.

Перетворимо матричне рівняння (3.2.2) таким чином:

$$K_X u_i - \lambda_i u_i = 0 \qquad (3.2.3)$$

Винесемо у рівнянні (3.2.3) вектор u_i за дужки. Отримаємо

$$(K_X - E\lambda_i)u_i = 0, \qquad (3.2.4)$$

де E - одинична матриця. Можна легко показати, що в координатній формі рівняння (3.2.4) виявляється системою однорідних лінійних алгебраїчних рівнянь

$$\begin{cases} (\sigma_1^2 - \lambda_i)u_{1i} + K_{12}u_{2i} + \ldots + K_{1n}u_{ni} = 0, \\ K_{21}u_{1i} + (\sigma_2^2 - \lambda_i)u_{2i} + \ldots + K_{2n}u_{ni} = 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ K_{n1}u_{1i} + K_{n2}u_{2i} + \ldots + \left(\sigma_n^2 - \lambda_n\right)u_{ni} = 0. \end{cases} \qquad (3.2.5)$$

у яких невідомими є компоненти i - того власного вектора

$$u_i = \begin{pmatrix} u_{1i} \\ u_{2i} \\ \ldots \\ u_{ni} \end{pmatrix}, \qquad (3.2.6)$$

а коефіцієнтами при невідомих - елементи матриці коваріацій. Але ця система, взагалі кажучи, розв'язаною бути не може, оскільки матриця $K_X - \lambda_i E$ утримує ще не визначені діагональні елементи, тому що невідомими є власні значення λ_i. Однак останні можуть бути визначеними, якщо використати відому теорему лінійної алгебри: система лінійних однорідних алгебраїчних рівнянь має нетривіальні розв'язки, якщо визначник цієї системи дорівнює нулю. Оскільки нас цікавлять саме такі розв'язки, зрівняємо до нуля визначник системи (3.2.5)

$$\begin{vmatrix} \sigma_1^2 - \lambda & K_{12} & \dots & K_{1n} \\ K_{21} & \sigma_2^2 - \lambda & \dots & K_{2n} \\ \dots & \dots & \dots & \dots \\ K_{n1} & K_{n2} & \dots & \sigma_n^2 - \lambda \end{vmatrix} = 0 \qquad (3.2.7)$$

Визначник (3.2.7) є алгебраїчне рівняння n - ого степеня відносно невідомого λ. Покажемо це на визначникові другого порядку

$$\begin{vmatrix} \sigma_1^2 - \lambda & K_{12} \\ K_{21} & \sigma_2^2 - \lambda \end{vmatrix} = 0$$

Очевидно ліва його частина має вид

$$\lambda^2 - (\sigma_1^2 + \sigma_2^2)\lambda + (\sigma_1^2\sigma_2^2 - K_{12}K_{21}) = 0$$

Поширюючи отриманий результат на визначник (3.2.7), маємо

$$\lambda^n + A_1\lambda^{n-1} + \ldots + A_{n-1}\lambda + A_n = 0 \qquad (3.2.8)$$

Коефіцієнти $A_1 \ldots A_n$ рівняння (3.2.8) є, як і у наведеному прикладі, визначними комбінаціями елементів матриці коваріацій. Рівняння (3.2.8) можна розв'язати за допомогою якого-небуть чисельного метода, наприклад, метода хорд чи метода ітерацій. Коли власні значення $\lambda_i (i = \overline{1,n})$ відомі, всі коефіцієнти системи рівнянь (2.3.6) стають повністю визначеними, і для кожного власного значення розв'язки системи (3.2.6) дають відповідний власний вектор. В застосовній математиці розроблені чисельні методи для розв'язання повної проблеми власних значень. Найбільш популярним є метод Якобі (метод обертань).

Існує така теорема: власні значення додатньо визначеної, симетричної і дійсної матриці є дійсними, додатніми і простими числами. Оскільки матриця коваріацій задовольняє умовам цієї теореми, власні значення мають зазначені в теоремі властивості.

Власні значення розташовують у порядку їх зменшення

$$\lambda_1 > \lambda_2 > \ldots > \lambda_n \qquad (3.2.9)$$

Означені властивості власних значень мають глибокий фізичний смисл, який буде розкритий декілька пізніше.

Першому власному значенню λ_1 відповідає перший власний вектор

$$u_1 = \begin{pmatrix} u_{11} \\ u_{21} \\ \dots \\ u_{n1} \end{pmatrix} \qquad (3.2.10)$$

другому власному значенню λ_2 - відповідає другий власний вектор

$$u_2 = \begin{pmatrix} u_{12} \\ u_{22} \\ \dots \\ u_{n2} \end{pmatrix} \qquad (3.2.11)$$

і так далі.

Якщо матриця коваріацій добре обумовлена, то перший власний вектор складається тільки з додатних компонент, у другого спостерігається одна зміна знака компонент, у третього - дві зміни і т.д. Власні вектори мають важливу властивість, яка визначається співвідношенням

$$\left\langle u_i u_j \right\rangle = u_i^{'} u_j = \delta_{ij} \left| u_i \right| \left| u_j \right| \qquad (3.2.12)$$

де δ_{ij} - символ Кронекера.

Вона значить, що скалярний добуток різних власних векторів дорівнює нулю, тобто що власні вектори ортогональні.

Найбільш часто власні вектори нормуються: замість векторів $u_i (i = \overline{1,n})$ використовуються власні вектори

$$W_i = \frac{u_i}{|u_i|} \tag{3.2.13}$$

Тоді, очевидно,

$$\langle W_i W_j \rangle = \frac{\langle u_i u_j \rangle}{|u_i||u_j|} = \delta_{ij} \frac{|u_i||u_j|}{|u_i||u_j|} = \delta_{ij} \tag{3.2.14}$$

Такі власні вектори називаються ортонормованими. Сукупність ортонормованих власних векторів складає ортогональну матрицю, що має такі властивості:

$$W'W = WW' = E \tag{3.2.15}$$

Для прикладу, в табл. 3.1 приводяться десять перших власних значень коваріаційної матриці місячних кількостей опадів у грудні на території України і відповідних власних векторів.

3.3 Ортогональні компоненти випадкових метеорологічних об'єктів

Зазначені властивості власних векторів дають можливість розглядати їх як базис n - вимірного евклідового простору R^n. У такому разі, можна провести розклад вектора ΔX_R, що являє собою той чи інший метеорологічний об'єкт (поле, вертикальний профіль, вектор предікторів і т.д.), у цьому базису. Відповідне лінійне перетворення має вид

$$W'\Delta X_j = z_j \qquad (3.3.1)$$

Оскільки базис власних векторів є ортогональним, то компоненти $z_{ij}\,(i = \overline{1,n})$ вектора z_j

$$z_k = \begin{pmatrix} z_{1j} \\ z_{2j} \\ \cdots \\ z_{ij} \\ \cdots \\ z_{nj} \end{pmatrix} \qquad (3.3.2)$$

є лінійно незалежними, а в статистичному смислі-некоррельованими

$$M[z_i z_j] = \begin{cases} \sigma^2_{z_i} & при \;\; i = j \\ 0 & при \;\; i \neq j \end{cases} \qquad (3.3.3)$$

Знайдемо тепер дисперсію $\sigma^2_{z_i}$ i - тої складової вектора ортогональних компонент, вважаючи, що вихідною інформацією є матриця $\Delta X = \{\Delta x_{ij}\}$ $(i = \overline{1,n}; j = \overline{1,m})$, яка утримує інформацію про m метеорологічних об'єктів. Відповідне ортогональне перетворення цієї матриці є

$$W^{'} \Delta X = z, \qquad (3.3.4)$$

де $z = \{z_{ij}\}$ $(i = 1,n; j = m)$ - матриця ортогональних компонент. Кожний рядок цієї матриці - це статистична сукупність i - тої ортогональної компоненти. Таким чином, для отримання оцінок перших двох моментів випадкових величин z_{ij} треба проводити осереднення по індексу j. Але перед тим, як виконувати відповідні розрахунки, запишемо рівняння (3.3.1) у скалярній формі

$$\begin{pmatrix} W_{11} & W_{21} & \dots & W_{S1} & \dots & W_{n1} \\ W_{12} & W_{22} & \dots & W_{S2} & \dots & W_{n2} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ W_{1i} & W_{21} & \dots & W_{Si} & \dots & W_{ni} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ W_{1n} & W_{2n} & \dots & W_{Sn} & \dots & W_{nn} \end{pmatrix} \begin{pmatrix} \Delta x_{1j} \\ \Delta x_{2j} \\ \dots \\ \Delta x_{Sj} \\ \dots \\ \Delta x_{nj} \end{pmatrix} =$$

$$= \begin{pmatrix} z_{1j} \\ z_{2j} \\ \dots \\ z_{ij} \\ \dots \\ z_{nj} \end{pmatrix} \qquad (3.3.5)$$

Після визначення матричного добутку в лівій його частині, отримаємо

$$z_{ij} = \sum_{S=1}^{n} W_{Si} \Delta x_{Sj} \quad (i = \overline{1,n}; j = 1,m) \qquad (3.3.6)$$

Знайдемо тепер середнє значення i- тої ортогональної компоненти $\bar{z}_i$

$$\bar{z}_i = \bar{z}_{ij}^{j} = \overline{\sum_{S=1}^{n} W_{Si}\Delta x_{Sj}} == \sum_{S=1}^{n} \overline{W_{Si}\Delta x_{Sj}} =$$
$$= \sum_{S=1}^{n} W_{Si}\overline{\Delta x_{Sj}} = 0,$$

(3.3.7)

оскільки Δx_{Sj} - центровані величини. Ураховуючи результат (3.3.7), а також рівність (3.3.6), знайдемо дисперсію i - тої ортогональної компоненти метеорологічного об'єкта

$$\sigma_{z_i}^2 = \frac{1}{m-1}\sum_{S=1}^{n} z_{ij}^2 = \frac{1}{m-1}\sum_{j=1}^{m}(\sum_{S=1}^{n} W_{Si}\Delta x_{Sj}) \times$$
$$\times (\sum_{p=1}^{n} W_{pi}\Delta x_{pj}) = \frac{1}{m-1}\sum_{j=1}^{n}(\sum_{S=1}^{n}\sum_{p=1}^{n} W_{Si}W_{pi}\Delta x_{Sj}\Delta x_{pj}) =$$
$$= \sum_{S=1}^{n}\sum_{p=1}^{n} W_{Si}W_{pi}\left(\frac{1}{m-1}\sum_{j=1}^{m}\Delta x_{Sj}\Delta x_{pj}\right)$$

Очевидно,

$$\frac{1}{m-1}\sum_{j=1}^{m}\Delta x_{Sj}\Delta x_{pj} = K_{Sp} \ ,$$

тобто коваріація метеорологічної величини, що досліджується (наприклад, коваріація між метеорологічною величиною в p - тій і S - тій точках метеорологічного поля). Отже,

$$\sigma_{z_i}^2 = \sum_{S=1}^{n}\sum_{p=1}^{n} W_{Si} W_{pi} K_{Sp} \qquad (3.3.8)$$

Можна легко показати, що

$$\sum_{S=1}^{n}\sum_{p=1}^{n} W_{Si} W_{pi} K_{Sp} = W_i' K_X W_i \qquad (3.3.9)$$

Маючи на увазі рівняння повної проблеми власних значень (3.2.2), знайдемо праву частину рівності (3.3.9). Для цього помножимо рівняння (3.2.2) ліворуч на вектор W_i' . Будемо мати:

$$W_i' K_X W_i = \lambda_i W_i' W_i = \lambda_i \qquad (3.3.10)$$

Тепер з рівностей (3.3.8), (3.3.9) і (3.3.10) випливає,

$$\sigma_{z_i}^2 = \lambda_i \qquad (3.3.11)$$

Рівність (3.3.11) визначає фізичний смисл власних значень матриці коваріацій: вони є дисперсіями відповідних

ортогональних компонент метеорологічних об'єктів. Це дуже важливий результат. Дійсно, згідно з співвідношенням (3.2.9), маємо

$$\sigma_{z_1}^2 > \sigma_{z_2}^2 > \ldots > \sigma_{z_n}^2 \qquad (3.3.12)$$

Крім того, власні значення мають ще одну важливу властивість:

$$\sum_{i=1}^{n} \lambda_i = t_r K_X \qquad (3.3.13)$$

Індексом t_r - позначається слід матриці (від англійського слова trace). Нагадаємо, що слідом матриці називають суму її елементів, що розташовані на головній діагоналі. Оскільки на головній діагоналі матриці коваріацій розташовуються дисперсії, то

$$\sum_{i=1}^{n} \lambda_i = \sum_{i=1}^{n} \sigma_{X_i}^2 \qquad (3.3.14)$$

Це означає, що сума всіх власних значень дорівнює сумарній дисперсії метеорологічного об'єкта.

Але як випливає з співвідношень (3.2.9) і (3.3.12) тепер відбувається розподіл дисперсій: дисперсія першої ортогональної компоненти z_1 є найбільшою. Дисперсії другої z_2 і слідуючих за нею ортогональних компонент, як правило, швидко зменшуються. Метеорологічні процеси мають одну

важливу властивість. Вона полягає у тому, що чим більший масштаб атмосферного процесу, тим більше значення дисперсії йому відповідає. Наприклад, дисперсія швидкості вітру у вільній атмосфері, яка характеризує мінливість крупномасштабних гілок загальної циркуляції атмосфери, має порядок 10^2 м^2/с^2. Швидкість вітру мезомасштабних утворень в атмосфері характеризується дисперсією, порядок якої 10^0-10^1 м^2/с^2, а мікромасштабних турбулентних вихорів - дисперсією порядку 10^{-2}-10^0 м^2/с^2 у залежності від масштабу турбулентності. Таким чином, співвідношення (3.3.12) характеризує, по-суті, розподіл дисперсій по масштабах атмосферних процесів, що утворюють ті чи інші метеорологічні об'єкти. Отже перші ортогональні компоненти $z_1, z_2, \ldots, z_K$ дисперсії яких вичерпують основну частину сумарної дисперсії, концентрують інформацію про найбільш крупномасштабні особливості метеорологічних об'єктів, що досліджуються. Тому їх називають головними компонентами метеорологічних полів, вертикальних профілів метеорологічних величин чи інших метеорологічних об'єктів.

Головні компоненти отримуються шляхом ортогонального перетворення метеорологічного об'єкта за допомогою відповідних власних векторів. Це випливає з формули (3.3.1). Тому відповідні власні вектори також утримують інформацію про особливості найбільш крупномасштабних атмосферних процесів, що відповідають за формування статистичної структури метеорологічних об'єктів.

3.4. Задача про стиск метеорологічної інформації.

При розв'язуванні великої кількості метеорологічних задач виникає необхідність мати діло із значними об'ємами інформації про значення метеорологічних величин, отриманих шляхом метеорологічних вимірювань та спостережень. Наприклад, побудова статистичних моделей метеорологічних прогнозів пов'язана з необхідністю використовувати численні поля

атмосферного тиску за минулі терміни. Але кожне поле визначається значеннями тиску на множині метеорологічних станцій, яких налічується у залежності від того, який регіон розглядається, декілька сотень. Виникає питання, як можливо використати таку інформацію у якості предіктора у статистичній моделі. Треба мати на увазі, що крім поля атмосферного тиску треба ураховувати ще й інші предіктори, які визначаються вибірками поля чи метеорологічних величин. Можливим є лише один шлях - виконати стиск інформації. Під стиском інформації розуміється значне скорочення її кількості при збереженні основної смислової інформації. Одним із методів, за допомогою яких можна досягти зазначеної мети, є компонентний аналіз.

Як вже відзначалося, за допомогою сукупності власних векторів. Матриці ковариацій, які являють собою ортогональний базіс евклідового простору, можна отримати замість випадкового вектора X вектор ортогональних компонент Z, дисперсіями котрих є власні значення. При цьому сумарна дисперсія поля розподіляється таким чином, що найбільша її частина припадає на декілька перших k ортогональних компонент $Z_i(i=\overline{1,k})$, які, як зазначалося вище називають головними компонентами. Вони й містять найбільш суттєву інформацію про структуру метеорологічного об'єкта, що підлягає дослідженню. Отже, ми можемо тепер замість n - вимірного вихідного вектора X, наприклад 200-вимірного вектора атмосферного тиску, яким визначається поле цієї метеорологічної величини, використати k - вимірний вектор головних компонент, яких у цьому векторі лише декілька $(k << n)$. Останні $n-k$ ортогональних компонент відносяться до дрібномасштабних взбурень та різних похибок, які утримуються у вихідній інформації. Стиснена таким чином інформація може зберігатися на магнітних носіях пам'яті.

Щоб визначити число k головних компонент, треба відповідно до смислу поставленої задачі, визначити частку η_k

сумарної дисперсії метеорологічного поля (або вектора предікторів), яка відповідає найбільш крупномасштабним особливостям поля. Оскільки справедливим є співвідношення

$$\sum_{i=1}^{n} \sigma_{X_i}^2 = t_r K_X = \sum_{i=1}^{n} \lambda_i , \qquad (3.4.1)$$

то, очевидно, число η можна визначити як

$$\eta_k = \frac{\sum_{i=1}^{k} \lambda_i}{t_r K_X} \qquad (3.4.2)$$

Число k, при якому $\eta \geq \eta_k$, де η_k установлюється дослідником (наприклад, треба зберегти не менше ніж 80% найбільш суттєвої інформації), і визначає ті головні компоненти вектора (3.3.4), котрі утримують найбільш важливу інформацію про статистичну структуру метеорологічних полів чи, наприклад, вектора предікторів.

3.5 Задача фільтрації інформації про метеорологічні поля.

Задача фільтрації метеорологічних полів у певному смислі, є зворотньою відносно задачі стиску інформації. Якщо

у першому випадку треба було сконцентрувати найбільш суттєву інформацію у декількох перших ортогональних компонентах, то в задачі фільтрації необхідно вихідну інформацію звільнити від інформації, що відноситься до дрібномасштабних збурень і похибок.

Очевидно, справедливим є рівняння

$$\frac{\sum_{i=1}^{k}\lambda_i}{t_r K_X} + \frac{\sum_{i=k+1}^{n}\lambda_i}{t_r K_X} = \eta + \delta = 1 \qquad (3.5.1)$$

Оскільки η_k, як було визначено у попередньому розділі, характеризує ту найбільш суттєву інформацію, яку треба залишити і яка міститься в перших k ортогональних компонентах метеорологічного поля (головних компонентах), то

$$\delta = \frac{\sum_{i=k+1}^{k}\lambda_i}{t_r K_X} \qquad (3.5.2)$$

відноситься до тієї інформації, від якої треба звільнитися.

Ми досягнемо цієї мети, якщо будемо вважати, що всі ортогональні компоненти z_j від $k+1$ до n-ої дорівнюють нулю. За таких умов вектор ортогональних компонент, наприклад, метеорологічного поля, має такий вид:

$$\widetilde{z}_j = \begin{pmatrix} z_{1j} \\ z_{2j} \\ \dots \\ z_{kj} \\ 0 \\ \dots \\ 0 \end{pmatrix} \qquad (3.5.3)$$

Тепер залишається виконати зворотнє перетворення за допомогою матриці власних векторів, тобто від вектора ортогональних компонент поля перейти до самого поля. Для цього залишимо рівняння (3.3.4) у виді:

$$W'\Delta\widetilde{X}_j = \widetilde{Z}_j \qquad (3.5.4)$$

і помножимо його ліворуч на матрицю W. Враховуючи властивість ортогональності матриці W (2.3.15) будемо мати

$$\Delta\widetilde{X}_j = W\widetilde{Z}_j \qquad (3.5.5.)$$

Таким чином, ми отримали поле метеорологічної величини з центрованими її значеннями в точках поля. Оскільки середні значення метеорологічної величини в точках поля відомі, легко можна отримати і саме фільтроване поле метеорологічної величини. Можна операцію фільтрації виконати іншим чином. Оскільки головні компоненти метеорологічного поля є результатом ортогонального перетворення поля першими k

власними векторами, а останні $n-k$ ортогональні компоненти характеризують інформацію, від якої треба звільнитися і вони отримані в результаті ортогонального перетворення поля останніми $n-k$ ортогональними векторами, то поставленої цілі можна досягти, якщо використовувати матрицю перетворення $\widetilde{W}$ розміру $n \times n$

$$\widetilde{W} = \begin{pmatrix} W_{11} & W_{12} & \dots & W_{1K} & 0 & \dots & 0 \\ W_{21} & W_{21} & \dots & W_{2K} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ W_{i1} & W_{i2} & \dots & W_{iK} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ W_{n1} & W_{n2} & \dots & W_{nK} & 0 & \dots & 0 \end{pmatrix} \qquad (3.5.6)$$

Тоді фільтроване метеорологічне поле отримаємо таким чином

$$\Delta\widetilde{X}_j = \widetilde{W} Z_J \qquad (3.5.7)$$

В рівності (3.5.7) вектор Z_j утримує всі ортогональні компоненти метеорологічного поля. Перетворення (3.5.5) і (3.5.7), очевидно є тотожними.

3.6 Приклади компонентного аналізу метеорологічних об'єктів.

Місячна кількість опадів q є однією із важливих кліматичних параметрів. Вона характеризує умови зволожування того чи іншого району, а її поля - умови зволоження територій, для яких ці поля побудовані. На рис.3.1 зображається поле місячних кількостей опадів у лютому 1980 року. Воно визначено значеннями цієї метеорологічної величини на мережі 21 метеорологічних станцій України. На карті методом лінійної екстраполяції проведені ізогієти - лінії однакових кількостей опадів. Таке поле, як вже зазначалося вище, може бути зображене двадцятиодновимірним вектором, а сукупність полів за n років – матрицею Q розмірності $21 \times n$

$$Q = \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1n} \\ q_{21} & q_{22} & \cdots & q_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ q_{21,1} & q_{21,2} & \cdots & q_{21,n} \end{pmatrix} \qquad (3.6.1)$$

В матриці (3.6.1) кожний стовпець визначає окреме поле місячних кількостей опадів, а кожний рядок - вибірку цієї метеорологічної величини на метеорологічній станції, умовний номер якої співпадає з номером рядка. В прикладі, що наводиться, здійснювався добір місячних кількостей опадів за період з 1951 до 1981 $(n = 31)$. На основі матриці (3.6.1) була розрахована матриця коваріацій K_Q порядка 21×21 і відповідна матриця корреляцій R_Q, для якої розв'язувалася повна проблема власних значень. В табл.3.1 наводяться першість власні значення, які вичерпують основну долю сумарної дисперсії змінних, і відповідні їм власні вектори.

Таблиця 3.1 - Власні значення λ_i і відповідні їм власні вектори місячних кількостей опадів у лютому.

Номер компо-нент	Номери власних векторів						
	i	1	2	3	4	5	6
	λ_i	19,25	1,27	0,27	0,17	0,02	0,01
1		0,22	-0,28	0,07	-0,20	-0,25	0,06
2		0,21	-0,29	0,17	-0,44	0,04	-0,43
3		0,22	-0,27	0,18	-0,06	0,19	0,19
4		0,21	-0,29	0,20	-0,31	0,17	-0,19
5		0,22	-0,12	-0,27	0,13	-0,40	-0,01
6		0,22	-0,05	-0,36	0,11	-0,27	-0,24
7		0,22	-0,22	-0,02	0,06	-0,20	0,29
8		0,22	-0,14	-0,15	0,14	-0,20	0,14
9		0,22	-0,22	0,07	0,10	0,04	0,32
10		0,23	-0,03	-0,05	0,20	0,21	-0,03
11		0,22	-0,12	0,08	0,26	0,24	0,23
12		0,22	0,03	0,29	0,43	0,28	-0,10
13		0,22	0,17	0,11	0,15	0,13	-0,10
14		0,22	0,266	0,26	0,09	-0,12	-0,13
15		0,23	0,01	-0,27	0,10	0,05	-0,15
16		0,22	0,12	-0,28	-0,02	0,18	-0,18
17		0,22	0,19	-0,36	-0,18	0,23	-0,18
18		0,20	0,35	-0,07	-0,46	0,06	0,54
19		0,22	0,22	-0,12	-0,09	0,18	-0,01
20		0,21	0,32	0,17	-0,14	-0,14	-0,11
21		0,21	0,32	0,40	0,03	-0,45	-0,18

Із табл. 3.1 випливає, що перші два власні значення вичерпують біля 97 сумарної дисперсії центрованих і нормованих на середній квадратичний відхил місячних кількостей опадів (слід матриці корреляцій). Тому найбільш крупномасштабні властивості статистичної структури полів

опадів утримуються в перших двох власних векторах. Перший власний вектор має однорідну структуру, що обумовлюється впливом на поля опадів найбільш крупномасштабного опадоутворюючого процесу - переміщення через Україну циклонів і атмосферних фронтів в загальному напрямку із заходу на схід. Другий власний вектор має області максимуму в південно-східній частині і мінімум в північно-західній частині України. Зазначений максимум обумовлюється впливом циклонів, що спостерігалися у цю пору і проходили з Чорного моря в північно-східному напрямку.

Відповідно до зазначеної структури власних значень і власних векторів, головна інформація про властивості множини полів опадів утримується в перших двох головних компонентах, рядки яких утворюються ортогональними перетвореннями матриці (3.6.1) першими двома власними векторами

$$\begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1n} \\ z_{21} & z_{22} & \cdots & z_{2n} \end{pmatrix} =$$

$$= \begin{pmatrix} W_{11} & W_{21} & \cdots & W_{21,1} \\ W_{12} & W_{22} & \cdots & W_{21,2} \end{pmatrix} \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1n} \\ q_{21} & q_{22} & \cdots & q_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ q_{21,1} & q_{21,2} & \cdots & q_{21,n} \end{pmatrix}$$

Таким чином, здійснюється операція стиску інформації про n двадцятиодновимірних полів місячних кількостей опадів.

Розглянемо тепер задачу, фільтрації поля опадів. Для цього необхідно втілити операцію, яка відповідає рівності (3.5.5). Приведемо її для поля місячних кількостей опадів, що міститься на рис.3.1. Відповідне фільтроване поле зображається на рис.3.2. Для освітлення результату фільтрації порівняємо

вихідне та фільтроване поля. На рис.3.1 виразно виділяються два максимуми опадів. Перший з них (вторинний) розташовується в районі Карпат, другий в Придніпровьї. Їх розділяє область мінімума, вісь якої приходить від Ізмаілу на Камянець-Подільський і далі у північному напрямку. Структура поля опадів яскраво відбиває вплив Карпат на опадоутворюючі процеси, а саме при перевалюванні циклонів і атмосферних фронтів, що надходять із південно-західного і західного напрямків, через Карпати, відбувається їх ослаблення, що обумовлює зону мінімуму опадів. При подальшому їх переміщені на схід над Придніпров'єм вони знову набувають активність, і кількості опадів зростає. На величину і конфігурацію максимуму над східною Україною чинять вплив і циклони, траєкторії яких приходили у цьому місяці з Чорного моря на північний схід через цю частину України. На відмінну від цього, фільтроване поле має лише один максимум місячних кількостей опадів, що розташовується над південно-східною Україною. Причина появи його вже обговорювалась. Необхідно підкреслити, що в результаті операції фільтрації в полі опадів зникає вторинний процес - коливання кількості опадів, яке проявляється шляхом утворення мінімуму над зоною, що межується з східними схилами Карпат.

Другий приклад використання компонентного аналізу відноситься до досліджень вертикальної структури полів швидкості вітру в тропосфері і стратосфері західної півкулі.

У якості вихідних даних виступають результати радіо- та ракетного зондування атмосфери над районом, який розташовується в помірній (острови Воллоп: 37^0 $50^{'}$ пн.ш., 75^0 $29^{'}$ з.д.). Північної Америки. Ряди зональної і меридіональної компонент швидкості вітру з тижневою дискретністю були сформовані для рівнів 5, 10, 15, 20, 25, 30, 35, 40, 45, 50 і 55км. Шар атмосфери 5-55 км за відомими властивостями поділявся на два шари: 5-25 км і 25-55 км. Значення компонент швидкості вітру на зазначених рівнях, що відносяться до першого чи другого шару, розглядалися як компоненти 5-вимірного у першому шарі і 7-вимірного у другому шарі вектора. Послідовності векторів складають матриці вихідних даних типа

(3.6.1). Для зазначеного пункту зондування атмосфери матриці зональної і меридіональної складових вітру мають порядок 5 х 528 - для першого, і порядок 7 х 528 - для другого шару атмосфери.

На основі цих матриць були розраховані матриці коваріацій, для яких розв'язувалася повна проблема власних значень. Розглянемо результати лише для зональної складової швидкості вітру.

В табл. 3.2 містяться значення перших двох власних значень матриці коваріацій, що вичерпують біля 90 сумарної дисперсії, і відповідні їм власні вектори для шару 5 - 25 км.

Таблиця 3.2 - Власні значення і власні вектори матриці коваріацій зональної складової швидкості вітру. Шар 5 - 25 км.

Висота, км	$\lambda_1 = 581.0$ $(\eta = 0{,}76)$	$\lambda_1 = 104.6$ $(\eta = 0{,}89)$
	W_1	W_2
5	0,3845	- 0,2241
10	0,6646	- 0,4549
15	0,4577	0,0339
20	0,2994	0,3892
25	0,3338	0,7683

В табл. 3.3 утримуються перші три власні значення матриці коваріацій для шару 25-55 км і відповідні їм власні вектори. Як видно, вони складають 85 сумарної дисперсії зональної складової швидкості вітру в цьому шарі атмосфери.

Таблиця 3.3 - Власні значення і власні вектори матриці коваріацій зональної складової швидкості вітру. Шар 25-55 км.

Висота км	$\lambda_1 = 419.6$ $(\eta = 0.60)$	$\lambda_2 = 121.8$ $(\eta = 0.77)$	$\lambda_1 = 59.4$ $(\eta = 0.85)$
	W_1	W_2	W_3
25	0,0208	- 0,0763	0,0085
30	0,0922	- 0,1735	- 0,0511
35	0,1716	- 0,2637	- 0,0826
40	0,4162	- 0,7887	0,2881
45	0,4893	0,0391	- 0,5666
50	0,5263	0,2597	- 0,3437
55	0,5216	0,4511	0,6843

Табл. 3.2 і 3.3 свідчать про те, що основні риси вертикальної структури зональної складової швидкості вітру утримуються в перших двох головних компонентах в шарі 5 - 25 км і в перших трьох головних компонентах в шарі 25 - 55 км. Вони були отримані шляхом ортогонального перетворення матриць вихідних даних за допомогою приведених в табл.3.2 і 3.3 власних векторів. Часові послідовності головних компонент підлягали операції фільтрації за допомогою перетворення Фурьє (дивись розділ 6.6) першої частини, в результаті чого виявлені періодичності, що сховані в часових рядах головних компонент.

Таблиця 3.4 - Періодичності, що утримуються в часових рядах головних компонент зональної складової швидкості вітру (острови Воллоп).

Номер головної компонен-ти	Частота (ω,1/тижд)	Період (T,тижд.)	Амплітуда, (A, м*с$^{-1}$)	Фазовий зсув (φ,рад)
шар 5 - 25 км				
1	0,132	47,7	12,8	1,09
2	0,132	47,7	3,0	1,53
	0,251	25,0	1,8	0,67
3	0,284	22,1	1,2	0,11
	0,540	11,6	1,1	-1,07
	1,231	5,1	1,1	-1,22
	2,093	3,0	1,2	0,04
	2,166	2,9	1,1	0,91
шар 25 - 55 км				
1	0,132	47,7	62,0	1,42
	0,251	25,0	12,8	0,41
2	0,132	47,7	4,7	-0,35
	0,183	34,3	2,9	-1,11
	0,388	16,2	3,4	-0,88
	0,523	12,0	2,8	-1,50
	0,785	8,0	2,6	-1,29
	2,326	2,7	2,3	0,77
3	0,063	99,3	3,3	-0,67
	0,217	28,9	2,7	-0,76
	0,338	18,6	3,3	0,02
	0,388	16,2	2,9	-0,01
	0,491	12,8	2,9	0,23
	0,675	9,3	2,3	-1,29
	1,282	4,9	2,4	-0,55

Із таблиці 3.4 випливає, що зональній складовій швидкості вітру в шарі 5 - 25 км (тропосфера і нижня стратосфера) притаманні 3 періодичності: річна - в першій, річна і піврічна - в другій головній компоненті. Амплітуда річної періодичності у першій головній компоненті є досить великою (12,8 м/с). Річна

періодичність, що проявляється в другій головній компоненті має у 4 рази меншу амплітуду ніж у першій компоненті. Третя ж головна компонента включає широкий спектр періодичних коливань.

В середній і верхній стратосфері (шар 25-55 км) перша головна компонента утримує річну і піврічну періодичності з дуже великими амплітудами. Друга і третя головні компоненти характеризуються наявністю широких спектрів періодичних коливань. У другій головній компоненті крім річної проявляються восьмимісячна, чотирьохмісячна періодичності, а також 2-3 місячні і хвилі з періодом 2-3 неділі (хвилі Кельвіна). Останні можуть бути віднесеними до хвиль Кельвіна. Цікавим є той факт, що у третій головній компоненті крім періодичностей, що утримуються в другій головній компоненті зональної складової швидкості вітру, проявляється квазідвохрічна періодичність. Це свідчить про те, що квазідвохрічна періодичність швидкості вітру є притаманною стратосфері не тільки приекваторіальної зони, а і стратосфері середніх широт.

Наведений приклад показує, що компонентний аналіз дає можливість провести ретельне дослідження властивостей крупномасштабних складових загальної циркуляції атмосфери, виявити шляхом статистичного аналізу головних компонент швидкості вітру періодичності, які притаманні швидкості вітру в тропосфері і стратосфері.

4. РЕГРЕСІЙНІ МОДЕЛІ МЕТЕОРОЛОГІЧНОГО ПРОГНОЗУ

4.1 Лінійна множинна регресія.

4.1.1 Рівняння множинної лінійної регресії.

Раніше вже розглядалося рівняння лінійної регресії вида $\bar{y}(x) = ax + b$, яке відбиває корреляційний зв'язок між двома випадковими величинами x і y. Але у метеорології, дуже рідко реалізується зв'язок такої форми, оскільки стан того чи іншого процесу, а також тієї чи іншої метеорологічної величини, що його відбиває, обумовлюється великою кількістю впливаючих факторів. Наприклад, при побудові прогностичної моделі для прогнозу місячної кількості опадів у деякому пункті на визначний термін треба враховувати, що опади обумовлюються, по-перше, полем атмосферного тиску біля земної поверхні над досить великим районом, по-друге, вони залежать від структури термобаричних полів на ряді рівнів вільної атмосфери. В число впливаючих факторів (предікторів) треба включити й деякі диференціальні характеристики метеорологічних полів, наприклад, адвекцію тепла, адвекцію вихору швидкості вітру і інші. Характер циклонічної діяльності, яка й обумовлює, головним чином, інтенсивність опадів, залежить у великій мірі від тепловмісту поверхневих вод Північної Атлантики, особливо в їх частині, що називається енергоактивними зонами океану. Великий вплив здійснює також й положення границі морського льоду. Можна перелічувати й інші гідрометеорологічні величини, які можуть грати роль предікторів.

Отже, якщо раніше йшлося про вплив однієї випадкової величини
на іншу, то тепер необхідно будувати математичну модель, яка б відбивала впливи множини предікторів на величину, що

прогнозується (предіктант). У якості такої статистичної моделі може бути лінійна модель множинної регресії.

Нехай предіктори й предіктант - центровані випадкові величини

$$\begin{cases} x_i = X_i - \overline{X}_i, (i = \overline{1,n}) \\ y = Y - \overline{Y} \end{cases} \quad (4.1.1)$$

Тоді рівняння лінійної регресії має вид

$$y = b_1 x_1 + b_2 x_2 + ... + b_n x_n, \quad (4.1.2)$$

де $b_1, b_2, ..., b_n$ - коефіцієнти (параметри) регресії.

Коефіцієнти регресії повинні бути визначеними на основі статистичних сукупностей предіктанта і предікторів. Розглянемо метод визначення коефіцієнтів регресії спочатку для рівняння

$$y = b_1 x_1 + b_2 x_2, \quad (4.1.3)$$

а потім отриманий результат розповсюдимо на рівняння (4.1.2).

Для визначення коефіцієнтів регресії b_1, b_2 використаємо метод найменших квадратів. Його метрика має вид

$$\sigma^2 = \sum_{i=1}^{m} \left[y_i - \left(b_1 x_{1i} + b_2 x_{2i} \right) \right]^2 \quad (4.1.4)$$

де m - об'єм статистичних сукупностей.
Необхідною й достатньою умовою мінімуму цієї метрики є умова

$$\begin{cases} \dfrac{\partial \delta^2}{\partial b_1} = 0 \\ \dfrac{\partial \delta^2}{\partial b_2} = 0 \end{cases} \tag{4.1.5}$$

Підставимо до рівнянь (4.1.5) метрику (4.1.4), отримаємо

$$\begin{cases} \dfrac{\partial \delta^2}{\partial b_1} = -2\sum\limits_{i=1}^{m}\left(y_i x_{1i} - b_1 x_{1i}^2 - b_2 x_{1i} x_{2i}\right) = 0 \\ \dfrac{\partial \delta^2}{\partial b_2} = -2\sum\limits_{i=1}^{m}\left(y_i x_{2i} - b_1 x_{1i} x_{2i} - b_2 x_{2i}^2\right) = 0 \end{cases} \tag{4.1.6}$$

або

$$\begin{cases} b_1 \sum\limits_{i=1}^{m} x_{1i}^2 + b_2 \sum\limits_{i=1}^{m} x_{1i} x_{2i} = \sum\limits_{i=1}^{m} y_i x_{1i} \\ b_2 \sum\limits_{i=1}^{m} x_{1i} x_{2i} + b_2 \sum\limits_{i=1}^{m} x_{2i}^2 = \sum\limits_{i=1}^{m} y_i x_{2i} \end{cases} \tag{4.1.7}$$

Система рівнянь (4.1.7) називається системою нормальних рівнянь. Вона є системою неоднорідних лінійних алгебраїчних

рівнянь. Оскільки величини x_i і y_i - центровані, то вірними є співвідношення

$$\begin{cases} \sum_{i=1}^{m} x_{1i}^2 = m\sigma_{x_1}^2, \\ \sum_{i=1}^{m} x_{2i}^2 = m\sigma_{x_2}^2, \\ \sum_{i=1}^{m} x_{1i}x_{2i} = m\sigma_{x_1}\sigma_{x_2}r_{x_1x_2}, \\ \sum_{i=1}^{m} y_i\, x_{1i} = m\sigma_y\sigma_{x_1}r_{yx_1}, \\ \sum_{i=1}^{m} y_i\, x_{2i} = m\sigma_y\sigma_{x_2}r_{yx_2}, \end{cases} \qquad (4.1.8)$$

Підставимо співвідношення (4.1.8) в рівняння (4.1.7). Після очевидних спрощень будемо мати

$$\begin{cases} b_1\sigma_{X_1} + b_2\sigma_{X_2}r_{X_1X_2} = \sigma_y r_{yX_1} \\ b_1\sigma_{X1}r_{X_1X_2} + b_2\sigma_{X2} = \sigma_y r_{yX_2} \end{cases} \qquad (4.1.9)$$

Розв'язки цієї системи алгебраїчних рівнянь мають такий вид

$$b_1 = \frac{r_{yX_1} - r_{yX_2} r_{X_1X_2}}{1 - r^2_{X_1X_2}} \frac{\sigma_y}{\sigma_{X_1}} \qquad (4.1.10)$$

$$b_1 = \frac{r_{yX_2} - r_{yX_1} r_{X_1X_2}}{1 - r^2_{X_1X_2}} \frac{\sigma_y}{\sigma_{X_2}}$$

(4.1.11)

Систему рівнянь (4.1.9) можна записати у матричній формі

$$R_X \sigma B = \sigma_y R_{yX} \qquad (4.1.12)$$

де R_X - корреляційна матриця, σ - діагональна матриця середніх квадратичних відхилів, B і R_{yX} стовпці шуканих коефіцієнтів та корреляцій між предіктантом і предікторами відповідно. Покажемо це на матрицях і векторах другого порядку. В координатній формі рівняння (4.1.12) має вид

$$\begin{pmatrix} 1 & r_{X_1X_2} \\ r_{X_1X_2} & 1 \end{pmatrix} \begin{pmatrix} \sigma_{X_1} & 0 \\ 0 & \sigma_{X_2} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} \sigma_y & r_{yX_1} \\ \sigma_y & r_{yX_2} \end{pmatrix},$$

або після виконання операцій в лівій частині останнього рівняння,

$$\begin{pmatrix} b_1\sigma_{X_1} + b_2\sigma_{X_2}r_{X_1X_2} \\ b_1\sigma_{X_1}r_{X_1X_2} + b_2\sigma_{X_2} \end{pmatrix} = \begin{pmatrix} \sigma_y & r_{yX_1} \\ \sigma_y & r_{yX_2} \end{pmatrix} \qquad (4.1.13)$$

Ураховуючи означення рівності двох векторів, приходимо до системи рівнянь (4.1.9). Ясно, що якщо матричне рівняння (4.1.12) є тотожнім системі (4.1.9), то воно відбиває і систему нормальних рівнянь будь-якого порядку

$$\begin{cases} b_1\sigma_{X_1} + b_2\sigma_{X_2}r_{X_1X_2} + ... + b_n\sigma_{X_n}r_{X_1X_n} = \\ = \sigma_y r_{yX_1} \\ b_1\sigma_{X_1}r_{X_2X_1} + b_2\sigma_{X_2} + ... + b_n\sigma_{X_n}r_{X_1X_n} = \\ = \sigma_y r_{yX_2} \\ ..... \\ b_1\sigma_{X_1}r_{X_2X_1} + b_2\sigma_{X_2}r_{X_nX_2} + ... + b_n\sigma_{X_n} = \\ = \sigma_y r_{yX_n} \end{cases}$$

(4.1.14)

В матричній формі розв'язок цієї системи рівнянь має вид

$$B = \sigma_y \sigma^{-1} R_X^{-1} R_{yX}, \qquad (4.1.15)$$

де $B = \{b_i\}_{n\times 1}; R_X = \{r_{X_iX_j}\}_{n\times n}; R_{yX} = \{r_{yX_i}\}_{n\times 1}$

У практичній реалізації система рівнянь (4.1.14) розв'язується одним з чисельних методів, наприклад, методом Гаусса.

В рівняння (4.1.2), як зазначалося, змінні є центрованими. Підставляючи до нього величини (4.1.1), після простих перетворень отримаємо

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_n X_n , \qquad (4.1.16)$$

де

$$b_0 = \overline{Y} - (b_1 \overline{X}_1 + b_2 \overline{X}_2 + \ldots + b_n \overline{X}_n) \qquad (4.1.17)$$

Вільний член b_0, як випливає з рівності (4.1.17), характеризує ту частину середнього значення предіктанта, котра вичерпується предікторами. Вона у визначній мірі характеризує адекватність регресійної прогностичної моделі.

В деяких випадках виникає можливість будувати регресійну модель на статистично незалежних предікторах, коли $r_{X_i X_j} = 0, \quad \forall \quad i, j = \overline{1, n}$. Тоді в системі (4.1.17) залишаються діагональні члени, і формули, що визначають коефіцієнти регресії, значно спрощуються і мають вид:

$$\begin{cases} b_1 = r_{yX_1} \dfrac{\sigma_y}{\sigma_{X_1}}, \\ b_2 = r_{yX_2} \dfrac{\sigma_y}{\sigma_{X_2}}, \\ \ldots\ldots\ldots\ldots\ldots \\ b_n = r_{yX_n} \dfrac{\sigma_y}{\sigma_{X_n}}. \end{cases} \qquad (4.1.18)$$

Прикладом такого випадку є побудова регресії між деякою метеорологічною величиною y та головними компонентами вектора предікторів

$$y = a_1 z_1 + a_2 z_2 + \ldots + a_k z_k \qquad (4.1.19)$$

дисперсіями яких, як зазначалося вище, є власні значення матриці коваріації λ_i. Тоді коефіцієнти α_i визначаються співвідношеннями

$$\begin{cases} a_1 = r_{yz_1} \dfrac{\sigma_y}{\sqrt{\lambda_1}} \\ \ldots\ldots\ldots\ldots\ldots \\ a_k = r_{yz_k} \dfrac{\sigma_y}{\sqrt{\lambda_k}} \end{cases} \qquad (4.1.20)$$

Головні компоненти вектора предікторів розраховуються шляхом ортогонального перетворення вектора предікторів матрицею власних векторів.

Лінійні регресійні моделі можна побудувати не тільки для прогнозу метеорологічних величин, а й для прогнозу полів метеорологічної характеристики y. Для цього формується матриця

$$Y = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1m} \\ y_{21} & y_{22} & \cdots & y_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ y_{n1} & y_{n2} & \cdots & y_{nm} \end{pmatrix} \qquad (4.1.21)$$

Котра містить в собі m полів цієї величини, що визначаються в n точках поля. На основі матриці коваріацій

$$K_y = \frac{1}{m-1} \Delta Y \Delta Y' \qquad (4.1.22)$$

знаходиться матриця власних векторів і відповідні власні значення й знаходяться головні компоненти поля, що прогнозується. Прогностична модель утримує k лінійних рівнянь множинної регресії, за допомогою яких розраховуються k головних компонент поля величини y. Обернене ортогональне перетворення спрогнозованих головних компонент в базису власних векторів дає прогностичне поле.

4.1.2 Множинний коефіцієнт корреляції.

У регресійному аналізі важливу роль відіграє множинний коефіцієнт корреляції. У відмінності від парного коефіцієнта корреляції він характеризує тісноту лінійного корреляційного зв'язку не з одним а з цілою системою предікторів $x_1, x_2, \ldots, x_n$, тобто є мірою адекватності регресійної прогностичної моделі. Для того, щоб обгрунтувати це, а також отримати алгоритм розрахунку множинного коефіцієнта корреляції, розглянемо рівняння регресії виду

$$y = b_1 x_1 + b_2 x_2 \qquad (4.1.23)$$

Запишемо метрику найменших квадратів

$$\delta^2 = \sum_{i=1}^{m} S^2_{y \bullet x_1 x_2} = \sum_{i=1}^{m} [y_i - (b_i x_{1i} + b_2 x_{2i})]^2 \qquad (4.1.24)$$

і розкриємо праву його частину.

$$\sum_{i=1}^{m} S^2_{y \bullet x_1 x_2} = \sum_{i=1}^{m} (y_i^2 - b_1 y_i x_{1i} - b_1 y_i x_{2i} -$$

$$- b_2 y_i x_{1i} - b_2 y_i x_{2i} + b_1^2 x_{1i}^2 + b_1 b_2 x_{1i} x_{2i} +$$

$$+ b_1 b_2 x_{1i} x_{2i} + b_2^2 x_{2i}^2) \qquad (4.1.25)$$

Якщо провести групування членів правої частини таким чином:

$$\sum_{i=1}^{m} S^2_{y\bullet x_1 x_2} = \sum_{i=1}^{m} y_i^2 - b_1 \sum_{i=1}^{m} y_i x_{1i} - b_2 \sum_{i=1}^{m} y_i x_{2i} +$$

$$+ b_1 (b_1 \sum_{i=1}^{m} x_{1i}^2 + b_2 \sum_{i=1}^{m} x_{1i} x_{2i} - \sum_{i=1}^{m} y_i x_{1i}) + b_2 \times$$

$$\times (b_1 \sum_{i=1}^{m} x_{1i} x_{2i} + b_2 \sum_{i=1}^{m} x_{2i}^2 - \sum_{i=1}^{m} y_i x_{2i})$$

то прийдемо до формули

$$\sum_{i=1}^{m} S^2_{y\bullet x_1 x_2} = \sum_{i=1}^{m} y_i^2 - b_1 \sum_{i=1}^{m} y_i x_{1i} - b_2 \sum_{i=1}^{m} y_i x_{2i}, \tag{4.1.26}$$

оскільки останні два члени, як випливає з формул (4.1.17), дорівнюють нулю.

У рівнянні (4.1.26) запровадимо формули (4.1.8). Будемо мати

$$\sum_{i=1}^{m} S^2_{y\bullet x_1 x_2} = m\sigma_y^2 - b_1 m \sigma_{x_1} \sigma_y r_{yx_1} - \\ - b_2 m \sigma_{x_2} \sigma_y r_{yx_2} \tag{4.1.27}$$

Розділимо обидві частини рівності (4.1.27) на загальний об'єм статистичної сукупності m і введемо позначення

$$\frac{\sum_{i=1}^{m} S^2_{y\bullet x_1 x_2}}{m} = \sigma^2_{y\bullet x_1 x_2} \qquad (4.1.28)$$

Очевидно рівність (4.1.28) має смисл дисперсії нев'язки рівняння (4.1.29). Якщо його поділити на дисперсію предіктанта, то будемо мати відносну дисперсію нев'язки прогностичного рівняння

$$S^2_{y\bullet x_1 x_2} = \frac{\sigma^2_{y\bullet x_1 x_2}}{\sigma^2_y} \qquad (4.1.29)$$

Отже, з урахуванням формул (4.1.28), (4.1.29), (4.1.10) і (4.1.11), маємо

$$S^2_{y\bullet x_1 x_2} = 1 - \frac{r^2_{yx_1} + r^2_{yx_2} - 2r_{yx_1} r_{yx_2} r_{x_1 x_2}}{1 - r^2_{x_1 x_2}} \qquad (4.1.30)$$

Величина

$$R^2_{y\bullet x_1 x_2} = \frac{r^2_{yx_1} + r^2_{yx_2} - 2r_{yx_1} r_{yx_2} r_{x_1 x_2}}{1 - r^2_{x_1 x_2}} \qquad (4.1.31)$$

і характеризує тісноту лінійного корреляційного зв'язку між предіктором y і предіктантами x_1 і x_2. Її називають коефіцієнтом множинної детермінації. Корінь квадратний з цієї величини

$$R_{y\bullet x_1 x_2} = \sqrt{R^2_{y\bullet x_1 x_2}} \qquad (4.1.32)$$

є коефіцієнт множинної корреляції. З формул (4.1.30)-(4.1.32) випливає, що

$$R_{y\bullet x_1 x_2} = \sqrt{1 - S^2_{y\bullet x_1 x_2}} \qquad (4.1.33)$$

Формула (4.1.33) є мірою адекватності прогностичної моделі. Дійсно, $S^2_{y\bullet x_1 x_2}$ характеризує ту частину предіктанта, котра не вичерпується моделлю. Чим вона менша, тим більш адекватною є модель. Якщо припустити, що $S^2_{y\bullet x_1 x_2} = 0$ (цього на практиці, звичайно, не буває), то модель тотожно відбиває процес, що моделюється. Але у цьому випадку $R_{y\bullet x_1 x_2} = 1$. Навпаки, коли $\sigma^2_{y\bullet x_1 x_2} = \sigma^2_y$, $S^2_{y\bullet x_1 x_2} = 1$ (такий прогноз називають випадковим), то $R_{y\bullet x_1 x_2} = 0$. Отже область зміни множинного коефіцієнта корреляції є

$$0 < R_{y\bullet x_1 x_2} < 1 \qquad (4.1.34)$$

Таким чином, чим більше значення множинного коефіцієнта корреляції, тим більш обумовленою є ця статистична модель. Узагальнимо тепер отримані результати на рівняння, що утримує n предікторів. Для цього до корреляційної матриці другого порядку, що відповідає рівнянню (4.1.23) додамо рядок й стовпець коефіцієнтів корреляції між предіктантом і предікторами

$$\widetilde{R} = \begin{pmatrix} 1 & r_{yx_1} & r_{yx_2} \\ r_{yx_1} & 1 & r_{x_1x_2} \\ r_{yx_2} & r_{x_1x_2} & 1 \end{pmatrix} \qquad (4.1.35)$$

Таку матрицю називають розширеною. Її визначник дорівнює

$$\left|\widetilde{R}\right| = \left(1 - r_{x_1x_2}^2\right)\left(1 - \frac{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{yx_1}r_{yx_2}r_{x_1x_2}}{1 - r_{x_1x_2}^2}\right) \qquad (4.1.36)$$

Перший співмножник правої частини рівності (4.1.36) є, очевидно, визначник матриці корреляції $\left|R_x\right|$. Якщо на нього поділити обидві частини співвідношення (4.1.36), то будемо мати

$$\frac{\left|\widetilde{R}\right|}{\left|R_x\right|} = 1 - \frac{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{yx_1} r_{yx_2} r_{x_1x_2}}{1 - r_{x_1x_2}^2} \qquad (4.1.37)$$

Другий член правої частини рівності (4.1.37) є, очевидно, множинний коефіцієнт детермінації $R_{y\bullet x_1x_2}^2$. Тоді, як свідчить рівність (4.1.33), ліва частина (4.1.37) дорівнює відносній залишковій дисперсії $S_{y\bullet x_1x_2}^2$. Отже з урахуванням цього маємо

$$R_{y\bullet x_1x_2} = \sqrt{1 - \frac{\left|\widetilde{R}\right|}{\left|R_x\right|}} \qquad (4.1.38)$$

Поширюючи отриманий результат на n предікторів, запишемо формулу для визначення множинного коефіцієнта корреляції

$$R_{y\bullet x_1x_2\ldots x_n} = \sqrt{1 - \frac{\left|\widetilde{R}\right|}{\left|R_x\right|}} \qquad (4.1.39)$$

Але тепер

$$\widetilde{R} = \begin{pmatrix} 1 & r_{yx_1} & r_{yx_1} & \dots & r_{yx_n} \\ r_{yx_1} & 1 & r_{x_1x_2} & \dots & r_{x_1x_n} \\ r_{yx_2} & r_{x_1x_2} & 1 & \dots & r_{x_2x_n} \\ \dots & \dots & \dots & \dots & \dots \\ r_{yx_n} & r_{x_nx_1} & r_{x_nx_2} & \dots & 1 \end{pmatrix} \qquad (4.1.40)$$

розширена матриця для корреляційної матриці R_x n-ого порядку. У цьому випадку

$$\frac{\left|\widetilde{R}\right|}{\left|R_x\right|} = S^2_{y\bullet x_1x_2\dots x_n} \qquad (4.1.41)$$

і рівність (4.1.39) можна записати так:

$$R_{y\bullet x_1x_2\dots x_n} = \sqrt{1 - S^2_{y\bullet x_1x_2\dots x_n}} \qquad (4.1.42)$$

Множний коефіцієнт корреляції $R_{y\bullet x_1x_2\dots x_n}$ характеризує тісноту корреляційного зв'язку між предіктантом і системою предікторів $x_1, x_2, \dots, x_n$, тобто є мірою адекватності моделі метеорологічного прогнозу $y = f(x_1, x_2, \dots, x_n)$, яка має вид лінійного рівняння множинної регресії. Чим він більший, тим більшою мірою адекватності характеризується прогностична модель.

Оцінити міру адекватності моделі можна й іншим шляхом, а саме шляхом перевірки статистичної гіпотези H_0 про те, що залишкова дисперсія $\sigma^2_{y \bullet x_1 x_2 \ldots x_n}$ значуще відрізняється від дисперсії предіктанта. Якщо гіпотеза H_0 відхиляється, то це означає, що прогноз по моделі не відрізняється від випадкового. Перевірка гіпотези H_0 відбувається за критерієм Фішера, який формується таким чином:

$$F = \frac{\sum_{i=1}^{m} (y_{\Phi_i} - \bar{y}_{\Phi})^2 \Big/ m-1}{\sum_{i=1}^{m-n} (y_{\Phi_i} - y_{p_i})^2 \Big/ m-n-1} \qquad (4.1.43)$$

де m - об'єм вибірок, n - число предікторів, що включені в модель, y_{ϕ} і y_p - фактичні й розрахункові значення предіктанта.

Треба мати на увазі, що об'єм вибірки повинен у 10 і більше разів перевищувати число предікторів.

Із рівності (4.1.43) випливає, що залишкова дисперсія розташовується в знаменнику. Гіпотеза H_0 не відхиляється, коли $F < F_{кр}(\alpha, \nu_1, \nu_2)$, де $\nu_1 = m-1$; $\nu_2 = m-n-1$.

Вірогідність прогнозу на основі регресійної моделі перевіряють й за допомогою коефіцієнта корреляції між фактичними та розрахунковими значеннями предіктанта. Для цього використовують як навчаючу, так і перевірочну сукупності предіктанта й предікторів. Ясно, що модель тим

краще відповідає величині, що прогнозується, чим ближчим до одиниці є коефіцієнт корреляції.

4.2 Методи добору оптимального складу предікторів

Створення самої регресійної моделі метеорологічного прогнозу, тобто оцінка коефіцієнтів $b_1, b_2, ..., b_n$ рівняння регресії

$$y = b_1 x_1 + b_2 x_2 + ... + b_n x_n \qquad (4.2.1)$$

є суто статистичною задачею, якщо визначені предіктори й для них сформовані статистичні сукупності. Але для визначення складу предікторів необхідно, наперед всього, розв'язати фізичну задачу, тобто на основі наших знань про фізику процесу формування метеорологічної величини чи поля фізичного параметру атмосфери що прогнозується, визначити склад впливаючих факторів. Всі вони складають множину потенційних предікторів.

Може скластися враження, що чим більше предікторів буде враховано при побудові прогностичної моделі, тим більш обумовленою вона буде й тим кращими будуть результати прогнозування. Але це далеко не так. Включення до моделі великої кількості предікторів збільшує порядок матриці корреляцій. У зв'язку з тим, що серед потенційних предікторів є визначна кількість пов'язаних між собою високими корреляційними зв'язками, матриця корреляцій може стати погано обумовленою. Оскільки її елементами є статистичні оцінки корреляцій, котрі, як відомо, отримують похибки, то така властивість матриці корреляцій може привести до значних похибок при оцінках коефіцієнтів регресії і, як наслідок, до погіршення якості прогностичної моделі. Щоб уникнути

переліченних проблем, із множини потенційних предікторів добирають ті, які виявляються статистично значущими. Цю операцію називають операцією "просіювання" предікторів. Розглянемо деякі алгоритми "просіювання" предікторів.

4.2.1 Просіювання предікторів за допомогою частинного коефіцієнта корреляції

4.2.1.1 Поняття про частинний коефіцієнт корреляції.

Раніше ми розглядали парні коефіцієнти корреляції, які визначали лінійний корреляційний зв'язок однієї випадкової величини y з якою-небудь другою. При цьому ми вважали, що на випадкову величину y діють й інші випадкові величини. Але нас не цікавило питання саме які це величини. В дійсності, предіктори, як правило, статистично зв'язані між собою і при побудові статистичних моделей ці зв'язки треба враховувати. Це можна здійснити за допомогою частинних коефіцієнтів корреляції. Приведемо визначення частинного коефіцієнта корреляції й розглянемо алгоритм його оцінки. Припустимо, що на випадкову величину y діють дві випадкові величини x_1 і x_2. Частинним коефіцієнтом корреляції між випадковими величинами y і x_1 $r_{yx_1 \bullet x_2}$ називають коефіцієнт корреляції між ними при умові, що вплив другої випадкової величини x_2 на y вже є врахованим. Таким же чином визначається частинний коефіцієнт корреляції $r_{yx_2 \bullet x_1}$

Будемо вважати, що нам відома матриця корреляцій

$$R_x = \begin{pmatrix} 1 & r_{x_2 x_1} \\ r_{x_2 x_1} & \end{pmatrix} \qquad (4.2.2)$$

і вектор парних корреляцій між y і x_1 , та x_2

$$R_{yx} = \begin{pmatrix} r_{yx_1} \\ r_{yx_2} \end{pmatrix} \qquad (4.2.3)$$

На їх основі сформуємо розширену матрицю корреляцій

$$R = \begin{pmatrix} 1 & r_{yx_1} & r_{yx_2} \\ r_{yx_1} & 1 & r_{x_1 x_2} \\ r_{yx_2} & r_{x_1 x_2} & 1 \end{pmatrix} \qquad (4.2.4)$$

Як очевидно, вона утворюється з матриці R_x шляхом добавлення до неї рядка та стовпця, що складаються з координат вектора R_y. На основі матриці (4.2.4) розрахуємо мінори $|R_x|, Д_{yx_i}, Д_{\overline{yx_i}} \quad (i = 1,2)$. Мінори $Д_{yx_i}$ складаються таким чином: стовпець, на першому місці котрого розташовується парна корреляція r_{yx_i}, переставляється на перше місце, а на його місці ставиться перший стовпець і, після цього, викреслюють перші рядок і стовпець. Очевидно

$$Д_{yx_1} = r_{yx_1} - r_{yx_2} r_{x_1x_2} \tag{4.2.5}$$

$$Д_{yx_2} = r_{yx_2} - r_{yx_1} r_{x_1x_2} \tag{4.2.6}$$

Означення мінора $Д_{\overline{yx_1}}$ має такий смисл: це мінор визначника $|R|$, який не утримує парної корреляції r_{yx_1}. Очевидно ми його будемо мати, якщо викреслимо із мінора $|R|$ рядок і стовпець, що утримують цю парну корреляцію. Отже

$$Д_{\overline{yx_1}} = \begin{vmatrix} 1 & r_{yx_2} \\ r_{yx_2} & 1 \end{vmatrix} = 1 - r_{yx_2}^2 \tag{4.2.7}$$

$$Д_{\overline{yx_2}} = \begin{vmatrix} 1 & r_{yx_1} \\ r_{yx_1} & 1 \end{vmatrix} = 1 - r_{yx_1}^2 \tag{4.2.8}$$

Частинні коефіцієнти корреляції визначаються таким чином:

$$r_{yx_1 \bullet x_2} = \frac{Д_{yx_1}}{\sqrt{|R_x| Д_{\overline{yx_1}}}} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{\left(1 - r_{x_1x_2}^2\right)\left(1 - r_{yx_2}^2\right)}} \tag{4.2.9}$$

$$r_{yx_2 \bullet x_1} = \frac{Д_{yx_2}}{\sqrt{|R_x| Д_{\overline{yx_2}}}} = \frac{r_{yx_2} - r_{yx_1} r_{x_1 x_2}}{\sqrt{\left(1 - r_{x_1 x_2}^2\right)\left(1 - r_{yx_1}^2\right)}} \tag{4.2.10}$$

Виникає питання, яка суттєва інформація утримується в частинних коефіцієнтах корреляції? Щоб відповісти на нього, розглянемо такий приклад. Нехай парні коефіцієнти корреляції між випадковими величинами мають значення: $r_{yx_1} = 0.70; r_{yx_2} = 0.90;\quad r_{x_1 x_2} = 0.80.$ Що можна сказати про ці випадкові величини? Звісно те, що випадкова величина y характеризується дуже тісними корреляційними зв'язками і з величиною x_1, і з величиною x_2. Але треба звернути увагу на те, що дві останні випадкові величини теж зв'язані дуже тісним корреляційним зв'язком між собою. Отже, щоб визначити яка з величин x дійсно чинить вплив на величину y, треба розрахувати частинні коефіцієнти корреляції за допомогою формул (4.52) і (4.53). Розрахунки дають такі їх значення: $r_{yx_1 \bullet x_2} = -0.08;\ r_{yx_2 \bullet x_1} = 0.79.$ Таким чином ясно, що в дійсності на випадкову величину y чинить вплив випадкова величина x_2. Корреляційний зв'язок y з величиною x_1, якщо урахувати її зв'язок з величиною x_2, є не тільки незначним, але навіть має зворотній характер.

Поширюючи отриманий алгоритм розрахунків частинних коефіцієнтів корреляції на n змінних, треба побудувати розширену матрицю

$$R = \begin{pmatrix} 1 & r_{yx_1} & r_{yx_2} & r_{yx_3} & \dots & r_{yx_k} & \dots & r_{yx_n} \\ r_{yx_1} & 1 & r_{x_1x_2} & r_{x_1x_3} & \dots & r_{x_1x_k} & \dots & r_{x_1x_n} \\ r_{yx_2} & r_{x_2x_1} & 1 & r_{x_2x_3} & \dots & r_{x_2x_k} & \dots & r_{x_2x_n} \\ r_{yx_3} & r_{x_3x_1} & r_{x_3x_2} & 1 & \dots & r_{x_3x_k} & \dots & r_{x_3x_n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ r_{yx_k} & r_{x_kx_1} & r_{x_kx_2} & r_{x_kx_3} & \dots & 1 & \dots & r_{x_kx_n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ r_{yx_n} & r_{x_nx_1} & r_{x_nx_2} & r_{x_nx_3} & \dots & r_{x_nx_k} & \dots & 1 \end{pmatrix}$$

(4.2.11)

і на її основі визначити мінори $\left|R_x\right|, Д_{yx_k}, Д_{\overline{yx}_k}$ $\left(k = 1,2,...,n\right)$ за тими ж правилами, як це було зроблено у попередньому випадку для матриці третього порядку.

Тоді

$$r_{yx_k \bullet x_1, x_2, \dots x_{k-1}, x_{k+1}, \dots, x_n} = \frac{Д_{yx_k}}{\sqrt{\left|R_x\right| Д_{\overline{yx}_k}}} \qquad (4.2.12)$$

мінори $\left|R_x\right|, Д_{yx_k}, Д_{\overline{yx}_k}$ мають порядок n, матриця R - порядок $n+1$.

Частинні коефіцієнти корреляції використовуються у ряду методів об'єктивного добору предікторів. Розглянемо деякі з них.

4.2.1.2 Просіювання предікторів за методом включення.

Метод включення основується на проведенні ряду послідовних операцій, які полягають у розрахунках частинних коефіцієнтів корреляцій, що враховують на кожному кроці вплив нового предіктора. Для спрощення, запровадимо позначення $r_{yx_j} = r_{0j}$, $r_{x_i x_j} = r_{ij}$ і розширену матрицю корреляцій позначимо таким чином:

$$R = \begin{pmatrix} 1 & r_{12} & r_{13} & \dots & r_{1n} & r_{01} \\ r_{21} & 1 & r_{23} & \dots & r_{2n} & r_{02} \\ r_{31} & r_{32} & 1 & \dots & r_{3n} & r_{03} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ r_{n1} & r_{n2} & r_{n3} & \dots & 1 & r_{0n} \\ r_{01} & r_{02} & r_{03} & \dots & r_{0n} & 1 \end{pmatrix} \qquad (4.2.13)$$

На відмінну від матриці (4.2.11) розширення матриці корреляцій між потенціальними предікторами $R_x = \left(r_{ij}\right)_{n\times n}$ відбулось шляхом добавлення рядка і стовпця корреляцій між предіктантом і предікторами не на перших, як у матриці (4.2.11), а на останніх місцях.

У якості першого предіктора приймається той k - тий предіктор із множини x_j $\left(j = \overline{1,n}\right)$всіх потенційних предікторів, якому відповідає найбільше значення парних коефіцієнтів корреляції за модулем: $\left|r_{0k}\right| = \max_j \left|r_{0j}\right|$. Після цього на основі матриці (4.2.13) розраховується матриця

частинних коефіцієнтів корреляції між предіктантом і предікторами при умові, що вплив k - того предіктора, якому приписується номер перший, вже врахований. Ця матриця має такий вид:

$$R^{(1)} = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & r_{23.1} & r_{24.1} & \dots & r_{2n.1} & r_{02.1} \\ 0 & r_{32.1} & 1 & r_{34.1} & \dots & r_{3n.1} & r_{03.1} \\ 0 & r_{42.1} & r_{43.1} & 1 & \dots & r_{4n.1} & r_{04.1} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & r_{n2.1} & r_{n3.1} & r_{n4.1} & \dots & 1 & r_{on.1} \\ 0 & r_{02.1} & r_{03.1} & r_{04.1} & \dots & r_{0n.1} & 1 \end{pmatrix} \tag{4.2.14}$$

Нема ніяких труднощів зрозуміти, чому всі елементи першого рядка і першого стовпця, крім першого елемента, матриці $R^{(1)}$ дорівнюють нулю. Дійсно, наприклад, другий елемент першого рядка, як очевидно, є частинним коефіцієнтом корреляції $r_{12.1}$. Якщо його розрахувати за формулою (4.2.10), в котрій $y = x_1$, то отримуємо такий результат:

$$r_{12.1} = \frac{r_{12} - r_{11}r_{12}}{\sqrt{\left(1 - r_{12}^2\right)\left(1 - r_{11}^2\right)}} \tag{4.2.15}$$

Оскільки $r_{11}=1$, то приходимо до невизначеності типу $\frac{0}{0}$. Якщо здійснити граничний перехід в рівності (4.2.15), застосувавши правило Лопіталя, то будемо мати:

$$\lim_{r_{11}\to 1} r_{12.1} = \lim_{r_{11}\to 1} \frac{\frac{\partial}{\partial r_{11}}\left(r_{12}-r_{11}r_{12}\right)}{\frac{\partial}{\partial r_{11}}\left[\sqrt{\left(1-r_{12}^2\right)\left(1-r_{11}^2\right)}\right]} =$$

$$= \lim_{r_{11}\to 1} \frac{-r_{12}\sqrt{\left(1-r_{12}^2\right)\left(1-r_{11}\right)}}{-2r_{11}\left(1-r_{12}^2\right)} = 0$$

Таким же чином пояснюється рівність нулю й інших членів перших рядка і стовпця матриці $R^{(1)}$.

Після цього розглядаються частинні коефіцієнти корреляції між предіктантом і предікторами, що містяться в останньому стовпці (або рядку) матриці $R^{(1)}$ й вибирається з них той, який за абсолютною величиною є найбільшим. Відповідаючий йому предіктор включається до складу предікторів, що мають статистичну значущість, і йому приписується номер другий. Після цього розраховується матриця $R^{(2)}$, елементами якої є частинні коефіцієнти корреляції при умові, що вплив першого і другого предікторів вже врахований. Якщо відкинути перший рядок і стовпець матриці (4.2.14), то матриця $R^{(2)}$ має такий вид:

$$R^{(2)} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & r_{34.12} & \dots & r_{3n.12} & r_{03.12} \\ 0 & r_{43.12} & 1 & \dots & r_{4n.12} & r_{04.12} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & r_{n3.12} & r_{n4.12} & \dots & 1 & r_{0n.12} \\ 0 & r_{03.12} & r_{04.12} & \dots & r_{0n.12} & 1 \end{pmatrix} \tag{4.2.16}$$

Третій предіктор визначається шляхом порівняння частинних корреляцій між предіктантом і предікторами, що утримуються в останньому стовпці матриці (4.2.16), за зазначеним вище правилом. Процедура продовжується до тих пір, доки на деякому $S+1$ - ому етапі всі частинні коефіцієнти між предіктантом і предікторами, що залишилися, втрачають статистичну значущість. Гіпотеза H_0 про статистичну незначущість цих частинних коефіцієнтів корреляції на рівні значущості α перевіряється за допомогою критерію Стьюдента

$$t = \frac{\left| r_{0(S+1).12\dots S} \right|}{\sigma_r} \tag{4.2.17}$$

Вона не відхиляється, якщо $t < t(\alpha, \nu)$, де $\nu = m - S$. Отже, ці S предікторів, що вийшли до складу статистично значущих, й являють собою основу при побудові статистичних моделей метеорологічного прогнозу.

4.2.13 Просіювання предікторів за методом покрокової

регресії

Цей метод отримав таку назву тому, що прогностичне рівняння лінійної регресії формується за допомогою деякої кількості послідовних кроків. Основою його є також використання частинних коефіцієнтів корреляції.

Перш за все, на основі статистичних сукупностей всіх n потенційних предікторів й предіктанта об'ємом m розраховується вектор парних коефіцієнтів корреляції між предіктантом і предікторами

$$R_{yx} = \begin{pmatrix} r_{yx_1} \\ r_{yx_2} \\ \cdots \\ r_{yx_k} \\ \cdots \\ r_{yx_n} \end{pmatrix} \qquad (4.2.18)$$

Проводиться аналіз координат цього вектора й визначається найбільший з них за абсолютною величиною. Нехай, наприклад, це буде r_{yk}. Зрозуміло, що є всі підстави вважати, що предіктор x_k значущим чином впливає на формування предіктанта, тобто його треба включити в склад оптимальних предікторів. Йому приписують номер перший $(x_k = x_1)$, а інші предіктори перенумеровують й переходять до виконання покрокової процедури.

1. За допомогою методу найменших квадратів (МНК) знаходять оцінку рівняння регресії предіктанта y на перший предіктор x_1, тобто

$$y^{(1)} = a_1^{(1)} x_1 \qquad (4.2.19)$$

(одиниця в дужках визначає номер кроку).

Знаходять перші нев'язки

$$\varepsilon_{1i} = y_i - y_i^{(1)}, \quad \forall i = \overline{1,m} \qquad (4.2.20)$$

і далі розраховують коефіцієнти корреляції $r_{\varepsilon_1 x_j} \left(j = \overline{2,n}\right)$.

Оскільки величини ε_{1i} характеризують значення предіктанта, з яких вилучається вплив першого предіктора, то коефіцієнти корреляції $r_{\varepsilon_1 x_j}$ мають сенс частинних коефіцієнтів корреляції. Далі проводиться аналіз всіх $n-1$ отриманих коефіцієнтів корреляції і з них визначається найбільший за абсолютною величиною. Нехай це буде x_s Тоді йому визначається номер другий $\left(x_s = x_2\right)$, і він включається до складу оптимальних предікторів. Інші предіктори знову перенумеровують й переходять до другого кроку.

2. За допомогою МНК будують рівняння регресії предіктанта y на предіктори x_1 і x_2

$$y^{(2)} = a_1^{(2)} x_1 + a_2^{(2)} x_2 \qquad (4.2.21)$$

й визначають другі нев'язки

$$\varepsilon_{2i} = y_i - y_i^{(2)}, \qquad (4.2.22)$$

тобто вилучають зі значень предіктанта впливи предікторів x_1 і x_2. Далі розраховують коефіцієнти корреляції $r_{\varepsilon_2 x_j}\left(j = \overline{3,n}\right)$, аналіз котрих дає підставу для вибіру третього предіктора x_3 і т.д. Процедура триває до тих пір, доки отримані на деякому l - тому кроку всі частинні коефіцієнти корреляції $r_{\varepsilon_2 x_j}\left(j = \overline{3,n}\right)$ втрачають статистичну значущість. Гіпотеза H_0 про це перевіряється за допомогою відомого критерію Стьюдента, який визначається таким же чином, як і в розділі 4.2.1.2. Треба мати на увазі, що регресійна модель створюється при умові, що об'єми статистичних сукупностей є досить великими, що дає підставу вважати, що розподіл коефіцієнтів корреляції близький до нормального.

Отже, за методом покрокової регресії ми відбираємо з n потенційних предікторів l статистично значущих й отримуємо лінійне рівняння регресії, яке при умові, що всі змінні спочатку центруються й нормуються на середній квадратичний відхил, має вид

$$y = a_1 x_1 + a_2 x_2 + \ldots + a_l x_l \qquad (4.2.23)$$

Разом з цим, при виконанні покрокової процедури на кожному кроку як допоміжний контролюючий параметр, розраховується відповідний множинний коефіцієнт корреляції. До речі, множинний коефіцієнт корреляції може використовуватися й для відбору значущих предікторів. Розглянемо цей метод у наступному розділі.

4.2.2 Відбір статистично значущих предікторів на основі множинного коефіцієнта корреляції

Як і у попередньому методі, основою являється вектор парних коефіцієнтів корреляції (4.2.18) між предіктантом і предікторами. Аналіз складових цього вектора дає можливість визначити той предіктор, парний коефіцієнт корреляції якого з предіктантом є найбільшим за абсолютною величиною. Він зараховується до складу оптимальних предікторів і йому приїдається номер перший. Останні предіктори перенумеровуються. Після цього розраховуються множинні коефіцієнти кореляції $R_{y.x_1x_j}\ (j=2,3,...,n)$. Зрозуміло, що у якості другого оптимального предіктора повинен виступати той s-тий предіктор, для якого $R_{y.x_1x_j}=\max_j R_{y.x_1x_j}$ $(j=2,3,...,n)$ є найбільшим. Йому приписується номер другий, а інші предіктори знову перенумеровуються. Далі проводять розрахунки множинних коефіцієнтів корреляції $R_{y.x_1x_2x_j}\left(j=\overline{3,n}\right)$ і визначається той l- тий предіктор x_l, для якого $R_{y.x_1x_2x_l}=\max_j R_{y.x_1x_2x_j}\left(j=\overline{3,n}\right)$. Він включається до складу оптимальних предікторів за номером третім, а інші предіктори перенумеровуються знову. Виникає питання, до яких же пір треба продовжувати цю процедуру? Справа у тому, що з включенням нового предіктора до складу статистично значущих предікторів значення відповідного множинного коефіцієнта корреляції збільшується, і при включенні деякого k - того предіктора він переходить у стан насичення. Характер зміни множинного коефіцієнта корреляції на кожному кроку розрахунків схематично зображується на рис. 4.1. Стан насичення виникає тоді, коли гіпотеза H_0 про

значущість розбіжностей між $R_{y.x_1x_2...x_k}$ і $R_{y.x_1x_2...x_kx_{k+1}}$ відхиляється. Перевірка цієї статистичної гіпотези проводиться за допомогою F критерія, який знаходиться за формулою

$$F = \frac{R^2_{y.x_1x_2...x_kx_{k+1}} - R^2_{y.x_1x_2...x_k}}{1 - R^2_{y.x_1x_2...x_k}} \times \frac{m-k+1}{k-1} \tag{4.2.24}$$

Зазначена гіпотеза відхиляється, коли виконується умова $F < F_{кр}(\alpha, \nu_1, \nu_2)$, де $\nu_1 = k-1$, а $\nu_2 = m-k+1$, α - рівень значущості. Гіпотеза H_0 перевіряється після кожного кроку розрахунків множинних коефіцієнтів корреляції. Незначущість розбіжностей між $R_{y.x_1x_2...x_k}$ і $R_{y.x_1x_2...x_kx_{k+1}}$ позначає що включення до складу факторів прогностичної моделі предіктора x_{k+1} не приводить до суттєвого зменшення остаточної дисперсії порівняно з випадком, коли модель будується на k попередньо добраних предікторів.

4.3 Система нелінійних рівнянь регресії із зворотними зв'язками.

4.3.1 Загальна постановка задачі.

Застосування у всіх випадках лінійної прогностичної моделі навряд чи можна вважати обгрунтованим. Звичайно, використання у прогностичній моделі полінома першого степеня у певній мірі виправдане тим, що обчислення коефіцієнтів апроксимуючого полінома за допомогою методу найменших квадратів більш високих степенів поєднано з великими обчислювальними труднощами. Але ці труднощі можна подолати, якщо для побудови поліномів більш високих степенів використати метод імовірносної апроксимації.

Статистичні моделі метеорологічних прогнозів розробляються для передобчислювання ряду елементів погоди. Вони, по-перше, мають, взагалі кажучи, загальні предіктори і, по-друге, статистично пов'язані між собою. Якщо необхідно за допомогою статистичних моделей скласти прогноз зв'язаних одне з одним комплексу метеорологічних величин або вивчити особливості взаємодії елементів певного метеорологічного явища, то ці цілі можуть бути досягнутими на основі системи апроксимуючих поліномів зі зворотніми зв'язками.

Для побудови такої математичної моделі треба:

а) визначити вид апроксимуючих поліномів для системи предікторів, тобто структуру моделі;

б) знайти коефіцієнти апроксимуючих поліномів для системи предіктантів, тобто характеристики моделі;

в) оцінити ступінь адекватності моделі.

Перелічені задачі є задачами проблеми ідентифікації у широкому смислі оператора прогностичної або еволюційної системи.

У якості вихідної статистичної моделі розглянемо систему із k рівнянь регресії третього порядку

$$\hat{y}_l = a_0^{(l)} + \sum_{i=1}^{m} a_i^{(l)} x_i + \sum_{\substack{i,j=1 \\ (i \le j)}}^{m} a_{ij}^{(l)} x_i x_j +$$

$$+ \sum_{\substack{i,j,s=1 \\ (i \le j \le s)}}^{m} a_{ijs}^{(l)} x_i x_j x_s + \sum_{\substack{\nu=1 \\ (\nu \ne l)}}^{k} \alpha_\nu^{(l)} y_\nu \qquad (4.3.1)$$

$$\left(l = \overline{1,k}\right)$$

Останній член рівняння (4.3.1) характеризує внесок зворотніх зв'язків між предіктантами (вихідними параметрами моделі). Поліноми вище третього степеня як правило не вживаються, оскільки вони не приводять до суттєвого уточнення, але значно збільшують об'єми розрахунків.

Будемо вважати, що предіктори є ортогональними, мають нульове математичне сподівання, одиничну дисперсію, тобто є центрованими і нормованими, і мають нормальний розподіл, а параметри на виході моделі - центровані і нормовані. Ці припущення не являють собою суттєві обмеження, оскільки існує ряд процедур ортогоналізації випадкових величин. Однією з них є процедура компонентного аналізу, яка докладно розглядалася у розділі 3. Вимога про нормальний розподіл випадкових величин є звичайною у регресійному аналізі, а також у багатьох інших методах багатовимірного статистичного аналізу. Система рівнянь (4.1.1) є первісною у тому смислі, що у подальшому структура її рівнянь буде уточнюватися. Це відноситься як до степенів предікторів, так і до їх переліку. Спочатку ж будемо вважати, що кожний предіктант системи (4.1.1) обумовлюється однаковим складом предікторів.

4.3.2 Система твірних функцій.

Нехай у наявності є статистичні сукупності предікторів $x_i\left(i=\overline{1,m}\right)$ і предіктантів $y_l\left(l=\overline{1,k}\right)$, котрі мають властивості, що визначені вище. Задача полягає у тому, що треба визначити коефіцієнт $a_0, a_i, a_{ij}, a_{ijs}$ і $k-1$ коефіцієнтів α_ν, для кожного l того полінома $\left(l=\overline{1,k}\right)$ системи рівнянь (4.1.1). Для вирішення цієї задачі будемо використовувати метод імовірносної апроксимації. Суть цього методу полягає у такому.

Нехай на множенні X визначені випадкові функції $y\left(x\right)$, які апроксимуються поліномами (4.1.1). Похибка апроксимації функції $y_l\left(x\right)$ визначається рівністю:

$$\varepsilon_l\left(x\right)=y_l\left(x\right)-\hat{y}_l\left(x\right) \qquad (4.3.2)$$

Дисперсія похибки ε_l дорівнює

$$Д_\varepsilon=\overline{\varepsilon_l^2}-\left(\overline{\varepsilon_l}\right)^2, \qquad (4.3.3)$$

де

$$\overline{\varepsilon_l}=M[\varepsilon_l\left(x\right)]$$

- визначає операцію математичного сподівання.

Визначимо коефіцієнти полінома (4.1.1) при умові мінімуму дисперсії похибки апроксимації. Її очевидно, можна досягнути за допомогою системи рівнянь:

$$\begin{cases} \dfrac{\partial Д_\varepsilon}{\partial a_0} = 0 \\ \dfrac{\partial Д_\varepsilon}{\partial a_p} = 0 \qquad \left(p = \overline{1,m}\right), \\ \dfrac{\partial Д_\varepsilon}{\partial a_{ps}} = 0 \qquad \left(p,s = \overline{1,m}\right), \\ \dfrac{\partial Д_\varepsilon}{\partial a_{psr}} = 0 \qquad \left(p,s,r = \overline{1,m}\right), \\ \dfrac{\partial Д_\varepsilon}{\partial \alpha_\mu} = 0 \qquad \left(\mu = \overline{1,k}; \mu \neq l\right) \end{cases} \qquad (4.3.4)$$

В рівняннях (4.3.4) і далі індекс l тимчасово пропускається. Якщо врахувати співвідношення (4.3.3), то систему (4.3.4) можна переписати таким чином:

$$\begin{cases} \dfrac{\partial \overline{\varepsilon^2}}{\partial a_0} - 2\overline{\varepsilon}\,\dfrac{\partial \overline{\varepsilon}}{\partial a_0} = 0, \\ \dfrac{\partial \overline{\varepsilon^2}}{\partial a_p} - 2\overline{\varepsilon}\,\dfrac{\partial \overline{\varepsilon}}{\partial a_p} = 0 \qquad \left(p = \overline{1,m}\right), \\ \dfrac{\partial \overline{\varepsilon^2}}{\partial a_{ps}} - 2\overline{\varepsilon}\,\dfrac{\partial \overline{\varepsilon}}{\partial a_{ps}} = 0 \qquad \left(p,s = \overline{1,m}\right), \\ \dfrac{\partial \overline{\varepsilon^2}}{\partial a_{psr}} - 2\overline{\varepsilon}\,\dfrac{\partial \overline{\varepsilon}}{\partial a_{psr}} = 0 \qquad \left(p,s,r = \overline{1,m}\right), \\ \dfrac{\partial \overline{\varepsilon^2}}{\partial \alpha_\mu} - 2\overline{\varepsilon}\,\dfrac{\partial \overline{\varepsilon}}{\partial \alpha_\mu} = 0 \qquad \left(\alpha_\mu = \overline{1,k}; \mu \neq l\right) \end{cases} \tag{4.3.5}$$

Беручи до уваги рівність (4.3.5), знайдемо

$$\frac{\partial \overline{\varepsilon}}{\partial a_0} = -\frac{\partial \hat{y}}{\partial a_0} = -1$$

$$\frac{\partial \overline{\varepsilon^2}}{\partial a_0} = \frac{\overline{\partial \varepsilon^2}}{\partial a_0} = 2\overline{\varepsilon}\,\frac{\partial \overline{\varepsilon}}{\partial a_0} = -2\overline{\varepsilon}$$

Отже, перше рівняння системи (4.3.5) перетворюється у тотожність

$$-2\overline{\varepsilon} + 2\overline{\varepsilon} = 0$$

Це свідчить про те, що умова для визначення коефіцієнта a_0 є відсутньою. Тому використаємо ще одну умову:

$$\overline{\varepsilon} = 0 \qquad (4.3.6)$$

Рівність (4.3.6) разом з системою (4.3.5) визначає, що ми бажаємо мати такі коефіцієнти апроксимуючого поліному (4.3.1), які дають нульову середню похибку апроксимації при мінімальній дисперсії похибки. У цьому й полягає метод імовірносної апроксимації.

За отриманими результатами, сформульовані вище умови приводять до такої системи рівнянь:

$$\begin{cases} \overline{\varepsilon} = 0 \\ \dfrac{\partial \overline{\varepsilon^2}}{\partial a_p} = 0 \qquad \left(p = \overline{1,m}\right), \\ \dfrac{\partial \overline{\varepsilon^2}}{\partial a_{ps}} = 0 \qquad \left(p,s = \overline{1,m}\right), \\ \dfrac{\partial \overline{\varepsilon^2}}{\partial a_{psr}} = 0 \qquad \left(p,s,r = \overline{1,m}\right), \\ \dfrac{\partial \overline{\varepsilon^2}}{\partial \alpha_\mu} = 0 \qquad \left(\mu = \overline{1,k}; \mu \neq l\right) \end{cases} \qquad (4.3.7)$$

Звернемося, спочатку, до похідної $\frac{\partial \overline{\varepsilon^2}}{\partial a_p}$. Очевидно,

$$\frac{\partial \overline{\varepsilon^2}}{\partial a_p} = M\left[\frac{\partial \varepsilon^2}{\partial a_p}\right] = 2M\left[\frac{\partial \varepsilon}{\partial a_p}\right] =$$
$$= 2M\left[(y-\hat{y})\frac{\partial(y-\hat{y})}{\partial a_p}\right] = 2M\left[\hat{y}\frac{\partial \hat{y}}{\partial a_p}\right] -$$
$$- 2M\left[y\frac{\partial \hat{y}}{\partial a_p}\right]$$

(4.3.8)

Інші похідні в рівняннях (4.3.8) розраховуються аналогічно. Відновимо індекс l і застосуємо такі позначення:

$$\begin{cases} z_0^{(l)} = M[y_l] & \left(l = \overline{1,k}\right), \\ z_p^{(l)} = M[y_l x_p] & \left(p = \overline{1,m}\right), \\ z_{ps}^{(l)} = M[y_l x_p x_s] & \left(p,s = \overline{1,m}\right), \\ z_{psr}^{(l)} = M[y_l x_p x_s x_r] & \left(p,s,r = \overline{1,m}\right) \\ r_{l\mu} = M[y_l y_\mu] & \left(\mu = \overline{1,k}; \mu \neq l\right) \end{cases}$$

(4.3.9)

Диференціювання рівнянь (4.3.1) по їх коефіцієнтах дає:

$$\frac{\partial \hat{y}_l}{\partial a_p^{(l)}} = x_p; \frac{\partial \hat{y}_l}{\partial a_{ps}^{(l)}} = x_p x_s; \frac{\partial \hat{y}_l}{\partial a_{psr}^{(l)}} = x_p x_s x_r;$$

$$\frac{\partial \hat{y}_l}{\partial \alpha_\mu} = y_\mu$$

(4.3.10)

Отже, враховуючи (4.3.10), а також (4.3.9), умови (4.3.7) можна переписати таким чином:

$$\begin{cases} z_0^{(l)} = M\left[\hat{y}_l\right] & \left(l = \overline{1,k}\right), \\ z_p^{(l)} = M\left[\hat{y}_l x_p\right] & \left(p = \overline{1,m}\right), \\ z_{ps}^{(l)} = M\left[\hat{y}_l x_p x_s\right] & \left(p,s = \overline{1,m}\right), \\ z_{psr}^{(l)} = M\left[\hat{y}_l x_p x_s x_r\right] & \left(p,s,r = \overline{1,m}\right) \\ r_{l\mu} = M\left[\hat{y}_l y_\mu\right] & \left(\mu = \overline{1,k}; \mu \neq l\right) \end{cases}$$

(4.3.11)

Ліві частини рівностей (4.3.11), як свідчить система рівностей (4.3.9), не залежать від виду апроксимуючого полінома, а визначаються значеннями змінної y, що складають статистичну сукупність цієї випадкової величини, і законом розподілу випадкового вектора X. Як було зазначено вище, будемо вважати, що всі компоненти вектора предікторів X

підпорядковуються нормальному закону з нульовими математичними сподіваннями та одиничними дисперсіями.

Підставимо у праві частини системи (4.3.11) рівняння (4.3.1). Тоді, враховуючи властивості компонент вектора X і випадкових величин y_l, отримаємо систему твірних функцій.

$$z_0^{(l)} = a_0^{(l)} + \sum_{\substack{i,j=1 \\ (i \le j)}}^{m} a_{ij}^{(l)} M[x_i x_j] = 0 \qquad (4.3.12)$$

$$z_p^{(l)} = \sum_{i=1}^{m} a_i^{(l)} M[x_i x_p] +$$

$$+ \sum_{\substack{i,j,t=1 \\ (i \le j \le t)}}^{m} a_{ijt}^{(l)} M[x_i x_j x_t x_p] + \sum_{\substack{\nu=1 \\ (\nu \ne l)}}^{k} \alpha_\nu^{(l)} M[y_\nu x_p]$$

$$\left(p = \overline{1,m}\right) \qquad (4.3.13)$$

$$z_{ps}^{(l)} = a_0^{(l)} M[x_p x_s] +$$

$$+ \sum_{\substack{i,j=1 \\ (i \le j)}}^{m} a_{ij}^{(l)} M[x_p x_s x_i x_j] + \sum_{\substack{\nu=1 \\ (\nu \ne l)}}^{k} \alpha_\nu^{(l)} M[y_\nu x_p x_s]$$

$$\left(p \le s = \overline{1,m}\right) \qquad (4.3.14)$$

$$z_{psr}^{(l)} = \sum_{i=1}^{m} a_i^{(l)} M[x_i x_p x_s x_r] +$$

$$+ \sum_{\substack{i,j,t=1 \\ (i \le j \le t)}}^{m} a_{ijt}^{(l)} M[x_i x_j x_t x_p x_s x_r] +$$

$$+ \sum_{\substack{\nu=1 \\ (\nu \ne l)}}^{k} \alpha_\nu^{(l)} M[y_\nu x_p x_s x_r] \qquad (4.3.15)$$

$$\left(p \le s \le r = \overline{1,m}\right)$$

$$z_{l\mu} = \sum_{i=1}^{m} a_i^{(l)} M[y_\mu x_i] + \sum_{\substack{i,j=1 \\ (i \le j)}}^{m} a_{ij}^{(l)} M[y_\mu x_i x_j] +$$

$$+ \sum_{\substack{i,j,t=1 \\ (i \le j \le t)}}^{m} a_{ijt}^{(l)} M[y_\mu x_i x_j x_t] + \sum_{\substack{\nu=1 \\ (\nu \ne l)}}^{k} \alpha_\nu^{(l)} M[y_\nu y_\mu]$$

$$\left(\mu = \overline{1,k}; \mu \ne l\right) \qquad (4.3.16)$$

Функції (4.3.12) - (4.3.16) дають можливість отримати коефіцієнти рівнянь (4.3.1).

4.3.3 Коефіцієнти системи поліномів першого степеня.

Система рівнянь першого степеня, очевидно, має такий вид:

$$\hat{y}_l = a_0^{(l)} + \sum_{i=1}^{m} a_i^{(l)} x_i + \sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} \alpha_\nu^{(l)} x_\nu \qquad (4.3.17)$$

$$\left(l = \overline{1,k}\right)$$

З урахуванням зазначених вище властивостей предіктантів і предікторів для системи рівнянь регресії (4.3.17) твірні функції стають такими:

$$\begin{cases} a_0^l = 0, \\ z_p^{(l)} = \sum_{i=1}^{m} a_i^{(l)} M[x_i x_p] + \sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} \alpha_\nu^{(l)} M[y_\nu x_p], \\ z_{l\mu} = \sum_{i=1}^{m} a_i^{(l)} M[y_\mu x_i] + \sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} \alpha_\nu^{(l)} M[y_\nu y_\mu] \end{cases} \qquad (4.3.18)$$

Оскільки $M[x_i, x_p] = \delta_{ip}$ - символ Кронекера (нагадаємо, що розглядається ортонормована система предікторів), то з

урахуванням рівностей (4.3.8) друге рівняння системи (4.3.18) дає співвідношення, що визначає коефіцієнти

$$a_l^{(p)} = z_p^{(l)} - \sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} \alpha_\nu^{(l)} z_p^{(\nu)} \qquad (4.3.19)$$

$$\left(p = \overline{1,m}\right)$$

Коефіцієнти зворотнього зв'язку $\alpha_\nu^{(l)}$, що входять як в рівність (4.3.19), так і в систему рівнянь (4.3.17) безпосередньо, визначаються третьою твірною функцією системи (4.3.18). Підставляючи до неї значення $\alpha_\nu^{(l)}$ із (4.3.19), будемо мати:

$$z_{\mu l} = \sum_{p=1}^{m} [z_p^{(l)} - \sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} \alpha_\nu^{(l)} z_p^{(\nu)}] z_p^{(\mu)} + \sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} \alpha_\nu^{(l)} r_{\mu\nu}$$

або після елементарних перетворень

$$\sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} (r_{\mu\nu} - \sum_{p=1}^{m} z_p^{(\mu)} z_p^{(\nu)}) \alpha_\nu^{(l)} = r_{\mu l} - \sum_{p=1}^{m} z_p^{(l)} z_p^{(\mu)} \qquad (4.3.20)$$

Позначимо

$$A_{\mu\nu}^{(l)} = r_{\mu l} - \sum_{p=1}^{m} z_p^{(\mu)} z_p^{(\nu)} \quad \left(\mu,\nu = \overline{1,k}; \mu,\nu \neq l\right) \tag{4.3.21}$$

$$B_{\mu}^{(l)} = r_{\mu l} - \sum_{p=1}^{m} z_p^{(\mu)} z_p^{(\nu)} \quad \left(\mu = \overline{1,k}; \mu \neq l\right) \tag{4.3.22}$$

Тоді рівність (4.3.20) приймає вид:

$$\sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} A_{\mu\nu}^{(l)} \alpha_{\nu}^{(l)} = B_{\mu}^{(l)} \quad \left(\mu = \overline{1,k}; \mu \neq l\right), \tag{4.3.23}$$

що являє собою систему $\left(k-1\right)$ лінійних рівнянь відносно невідомих $\alpha_{\nu}^{(l)}$. Запишемо її в матричній формі

$$A^{(l)} \alpha^{(l)} = B^{(l)}, \tag{4.3.24}$$

де

$$A^{(l)} = \{A_{\mu\nu}^{(l)}\}_{(k-1),(k-1)} \tag{4.3.25}$$

квадратна матриця порядку $(k-1)$, що утворюється із k - вимірної матриці шляхом вилучення рядка і стовпця з номером l ; $\alpha^{(l)}$ і $B^{(l)}$ - вектори

$$\alpha^{(l)} = \begin{pmatrix} \alpha_1^{(l)} \\ \alpha_2^{(l)} \\ \dots \\ \alpha_{l-1}^{(l)} \\ \alpha_{l+1}^{(l)} \\ \dots \\ \alpha_k^{(l)} \end{pmatrix} \qquad (4.3.26)$$

$$B^{(l)} = \begin{pmatrix} B_1^{(l)} \\ B_2^{(l)} \\ \dots \\ B_{l-1}^{(l)} \\ B_{l+1}^{(l)} \\ \dots \\ B_k^{(l)} \end{pmatrix} \qquad (4.3.27)$$

З рівняння (4.3.21) випливає, що матриці $A^{(l)}$ є симетричними.

Якщо $\left|A^{(l)}\right| \neq 0$, то вектор шуканих коефіцієнтів (4.3.26) визначається однозначно

$$\alpha^{(l)} = A^{(l)^{-1}} B^{(l)} \tag{4.3.28}$$

Після цього стає можливим визначення коефіцієнтів a_p^l за формулою (4.3.19) і побудувати поліноми (4.3.17) для всіх значень l.

4.3.4 Коефіцієнти системи поліномів другого степеня

Звернемося тепер до системи поліномів другого степеня

$$\hat{y}_l = a_0^{(l)} + \sum_{i=1}^{m} a_i^{(l)} x_i + \sum_{\substack{i,j=1 \\ (i \le j)}}^{m} a_{ij}^{(l)} x_i x_j + \sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} \alpha_\nu^{(l)} y_\nu \tag{4.3.29}$$

Твірні функції дають для системи (4.3.29) такі рівняння, за допомогою котрих можна розрахувати всі коефіцієнти $a_0^{(l)}, a_i^{(l)}, a_{ij}^{(l)}, \alpha_\nu^{(l)}$:

$$a_0^{(l)} + \sum_{\substack{i,j=1 \\ (i \le j)}}^{m} a_{ij}^{(l)} M[x_i x_j] = 0, \tag{4.3.30}$$

$$z_p^{(l)} = \sum_{i=1}^{m} a_i^{(l)} M[x_i x_p] + \sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} \alpha_\nu^{(l)} M[y_\nu x_p] ,$$

$$\left(p = \overline{1,m}\right)$$

(4.3.31)

$$z_{ps}^{(l)} = a_0^{(l)} M[x_p x_s] +$$

$$+ \sum_{\substack{i,j=1 \\ (i \leq j)}}^{m} a_{ij}^{(l)} M[x_p x_s x_i x_j] + \sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} \alpha_\nu^{(l)} M[y_\nu x_p x_s]$$

$$\left(p \leq s = \overline{1,m}\right)$$

(4.3.32)

$$z_{l\mu} = \sum_{i=1}^{m} a_i^{(l)} M[y_\mu x_i] + \sum_{\substack{i,j=1 \\ (i \leq j)}}^{m} a_{ij}^{(l)} M[y_\mu x_i x_j] +$$

$$+ \sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} \alpha_\nu^{(l)} M[y_\nu y_\mu]$$

$$\left(\mu = \overline{1,k}; \mu \neq l\right)$$

(4.3.33)

З властивостей величин x_i виходить, що

$$M[x_p x_s x_i x_j] =$$
$$= \begin{cases} 3, & коли \quad i = j = p = s \\ 1, & коли \quad i = j; p = s; i \neq p \ \ i \ \ т.д. \\ 0, & в \ інших \ випадках \end{cases}$$

(4.3.34)

Тоді маючи на увазі (4.3.9) і (4.3.34), із рівнянь (4.30)-(4.3.34) отримаємо:

$$a_0^{(l)} = -\sum_{p=1}^{m} a_{pp} \quad \left(l = \overline{1,k}\right) \tag{4.3.35}$$

$$a_p^{(l)} = z_p^{(l)} - \sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} \alpha_\nu^{(l)} z_p^{(\nu)} \tag{4.3.36}$$
$$\left(p = \overline{1,m}\right)$$

$$a_{pp}^{(l)} = \frac{z_{pp}^{(\nu)}}{2} - \sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} \alpha_\nu^{(l)} \frac{z_{pp}^{(\nu)}}{2} \quad \left(p = \overline{1,m}\right) \tag{4.3.37}$$

$$a_{ps}^{(l)} = z_{ps}^{(l)} - \sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} \alpha_\nu^{(l)} z_{ps}^{(\nu)} \quad \left(p < s = \overline{1,m}\right) \tag{4.3.38}$$

Як видно, коефіцієнти (4.3.35)-(4.3.38) залежать від коефіцієнтів $\alpha_{\nu}^{(l)}$. Отже, задача побудови поліномів (4.3.29) буде остаточно розв'язаною, якщо будуть відомі коефіцієнти зворотнього зв'язку $\alpha_{\nu}^{(l)}$. Для їх визначення, як і у випадку поліномів першого степеня, підставимо у праву частину рівняння (4.3.33) рівняння (4.3.36)-(4.3.38). З урахуванням (4.3.9) після деяких перетворень будемо мати:

$$\sum_{\substack{\nu=1 \\ (\nu\neq l)}}^{k}\Big[r_{\mu\nu}-\sum_{p=1}^{m}\Big(z_p^{(\mu)}z_p^{(\nu)}+\frac{z_{pp}^{(\mu)}z_{pp}^{(\nu)}}{2}\Big)-$$

$$-\sum_{\substack{p,s=1 \\ (p<s)}}^{m}z_{ps}^{(\mu)}z_{ps}^{(\nu)}\Big]\alpha_{\nu}^{(l)}=r_{\mu l}-$$

$$-\sum_{p=1}^{m}\Big(z_p^{(\mu)}z_p^{(l)}+\frac{z_{pp}^{(\mu)}z_{pp}^{(\nu)}}{2}\Big)-\sum_{\substack{p,s \\ (p<s)}}^{m}z_{ps}^{(\nu)}z_{ps}^{(\mu)}$$

$$\left(\mu=\overline{1,k};\mu\neq l\right) \qquad (4.3.39)$$

Позначимо

$$A_{\mu\nu}^{(l)} = r_{\mu\nu} - \sum_{p=1}^{m}\left(z_p^{(\mu)} z_p^{(\nu)} + \frac{z_{pp}^{(\mu)} z_{pp}^{(\nu)}}{2}\right) - \\ - \sum_{\substack{p,s=1 \\ (p<s)}}^{m} z_{ps}^{(\nu)} z_{ps}^{(\mu)} \tag{4.3.40}$$

$$B_{\mu}^{(l)} = r_{\mu l} - \sum_{p=1}^{m}\left(z_p^{(\mu)} z_p^{(l)} + \frac{z_{pp}^{(\mu)} z_{pp}^{(\nu)}}{2}\right) - \\ - \sum_{\substack{p,s \\ (p<s)}}^{m} z_{ps}^{(\nu)} z_{ps}^{(\mu)} \tag{4.3.41}$$

$$\left(l = \overline{1,k}; \mu = \overline{1,k}; \mu \neq l\right)$$

Це дає можливість рівність (4.3.39) переписати таким чином:

$$\sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} A_{\mu\nu}^{(l)} \alpha_{\nu}^{(l)} = B_{\mu}^{(l)} \quad \left(l = \overline{1,k}; \mu = \overline{1,k}; \mu \neq l\right) \tag{4.3.42}$$

Ми знову прийшли для кожного l до системи лінійних алгебраїчних рівнянь типу (4.3.24), розв'язком якої є вектор (4.3.26) шуканих коефіцієнтів зворотнього зв'язку $\alpha_{\nu}^{(l)}$. Структура матриць коефіцієнтів цих систем аналогічна структурі матриць (4.25), але елементи матриць тепер розраховуються за формулою (4.3.40). Після визначення

коефіцієнтів $\alpha_{\nu}^{(l)}$ формули (4.3.35)-(4.3.38) дають можливість отримати всі коефіцієнти системи поліномів (4.3.29).

4.3.5 Коефіцієнти системи поліномів третього степеня

Нехай маємо, нарешті, систему поліномів третього степеня зі зворотніми зв'язками

$$\hat{y}_l = a_0^{(l)} + \sum_{i=1}^{m} a_i^{(l)} x_i + \sum_{\substack{i,j=1 \\ (i\le j)}}^{m} a_{ij}^{(l)} x_i x_j + \\ + \sum_{\substack{i,j,t=1 \\ (i\le j\le t)}}^{m} a_{ijt}^{(l)} x_i x_j x_t + \sum_{\substack{\nu=1 \\ (\nu\ne l)}}^{k} \alpha_{\nu}^{(l)} y_{\nu} \qquad (4.3.43)$$

Для визначення коефіцієнтів системи (4.3.43) необхідно використовувати всі твірні функції (4.3.12)-(4.3.16). Маючи на увазі систему (4.3.9), із рівняння (4.3.12) і (4.3.14) відповідно отримуємо

$$a_0^{(l)} = -\sum_{p=1}^{m} a_{pp}^{(l)} \quad \left(l = \overline{1,k}\right) \qquad (4.3.44)$$

$$a_{pp}^{(l)} = \frac{z_{pp}^{(l)}}{2} - \sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} a_{\nu}^{(l)} \frac{z_{pp}^{(\nu)}}{2} \quad \left(p = \overline{1,m}\right) \ , \qquad (4.3.45)$$

$$a_{ps}^{(l)} = z_{ps}^{(l)} - \sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} \alpha_{\nu}^{(l)} z_{ps}^{(\nu)} \ ,$$
$$\left(p = \overline{1,m}\right) \ . \qquad (4.3.46)$$

Ці формули повністю співпадають з відповідними формулами для поліномів другого порядку. З перелічених вище властивостей предікторів випливає, що

$$M[x_i x_j x_t x_p x_s x_r] =$$
$$i = j = t = p = s = r \ ;$$
$$= \begin{cases} 15, & \text{якщо} \\ 3, & \text{якщо } i = j = t = p; s = r; s \neq p \\ 1, & \text{якщо } i = j; t = p; s = r; j \neq p; p \neq s \\ 0, & \text{в інших випадках} \end{cases} \quad \begin{matrix} \\ i \ т.д. \\ i \ т.д. \\ \\ . \end{matrix}$$

(4.3.47)

Ураховуючи (4.3.47), знайдемо інші коефіцієнти системи (4.3.43). Щоб визначити коефіцієнти $a_p^{(l)}$, використуємо функції (4.3.13). Будемо мати:

$$a_p^{(l)} = z_p^{(l)} - 3a_{ppp}^{(l)} - \sum_{j=1}^{p-1} a_{jjp}^{(l)} - \sum_{j=p+1}^{m} a_{pjj}^{(l)} -$$

$$- \sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} \alpha_\nu^{(l)} z_p^{(\nu)} \qquad \left(p = \overline{1,m}\right)$$

(4.3.48)

Очевидно, коефіцієнти $a_p^{(l)}$ залежать від коефіцієнтів при третьому степені змінних у рівняннях регресії. Із функцій (4.3.15) ці коефіцієнти визначаються. Маємо

$$z_{ppp}^{(l)} = 3a_p^{(l)} + 15a_{ppp}^{(l)} + 3\sum_{j=1}^{p-1} a_{jjp}^{(l)} + 3\sum_{j=p+1}^{m} a_{pjj}^{(l)} +$$

$$+ \sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} \alpha_\nu^{(l)} z_p^{(\nu)} \qquad \left(p = \overline{1,m}\right)$$

(4.3.49)

$$z_{pss}^{(l)} = a_p^{(l)} + 2a_{pss}^{(l)} + \sum_{j=1}^{p-1} a_{jjp}^{(l)} + \sum_{j=p+1}^{m} a_{pjj}^{(l)} +$$

$$+ 3a_{ppp}^{(l)} + \sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} \alpha_\nu^{(l)} z_{pss}^{(\nu)}$$

$$\left(p < s = \overline{1,m}\right)$$

(4.3.50)

$$z_{pps}^{(l)} = a_s^{(l)} + 2a_{pps}^{(l)} + \sum_{j=1}^{p-1} a_{jjs}^{(l)} + \sum_{j=s+1}^{m} a_{sjj}^{(l)} +$$

$$+ 3a_{sss}^{(l)} + \sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} \alpha_\nu^{(l)} z_{pps}^{(\nu)} \qquad (4.3.51)$$

$$\left(p < s = \overline{1,m}\right)$$

$$z_{psr}^{(l)} = a_{psr}^{(l)} + \sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} \alpha_\nu^{(l)} z_{psr}^{(\nu)} \qquad (4.3.52)$$

$$\left(p < s < r = \overline{1,m}\right)$$

Якщо, розв'язати рівняння (4.3.48)-(4.3.51) разом і знайти коефіцієнти $a_{psr}^{(l)}$ з рівняння (4.3.52), то будемо мати:

$$a_{ppp}^{(l)} = \frac{1}{6}[z_{ppp}^{(l)} - 3z_p^{(l)} - \sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} \alpha_\nu^{(l)} (z_{ppp}^{(\nu)} - 3z_p^{(\nu)})]$$

$$\left(p = \overline{1,m}\right) \qquad (4.3.53)$$

$$a_{pss} = \frac{1}{2}[z_{pss}^{(l)} - z_p^{(l)} - \sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} \alpha_\nu^{(l)}(z_{pss}^{(\nu)} - z_p^{(\nu)})]$$

$$\left(p < s = \overline{1,m}\right) \tag{4.3.54}$$

$$a_{pps} = \frac{1}{2}[z_{pps}^{(l)} - z_s^{(l)} - \sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} \alpha_\nu^{(l)}(z_{pps}^{(\nu)} - z_s^{(\nu)})]$$

$$\left(p < s = \overline{1,m}\right) \tag{4.3.55}$$

$$a_{psr} = z_{psr} - \sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} \alpha_\nu^{(l)} z_{psr}^{(\nu)}$$

$$\left(p < s < r = \overline{1,m}\right) \tag{4.3.56}$$

Тепер з'являється можливість отримати остаточну формулу для коефіцієнтів $a_p^{(l)}$. Після деяких перетворень отримаємо:

$$a_p^{(l)} = \frac{1}{2}\{(m+4)z_p^{(l)} - \sum_{j=1}^{p-1} z_{jjp}^{(l)} - \sum_{j=p}^{m} z_{pjj}^{(l)} -$$

$$- \sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} \alpha_\nu^{(l)} [(m+4)z_p^{(\nu)} - \sum_{j=1}^{p-1} z_{jjp}^{(\nu)} - \sum_{j=p}^{m} z_{pjj}^{(\nu)}]\}$$

$$\left(p = \overline{1,m}\right) \qquad (4.3.57)$$

Як і для двох випадків, що попередньо розглядалися, всі коефіцієнти рівнянь (4.3.43) окрім вільного члена, залежать від коефіцієнтів зворотних зв'язків. Визначення останніх завершує рішення задачі побудови системи поліномів третього степеня. Щоб визначити коефіцієнти $\alpha_\nu^{(l)}$ необхідно, як і в попередніх випадках, використати твірну функцію (4.3.16). Підставляючи до правої частини значення коефіцієнтів $a_p^{(l)}, a_{ps}^{(l)}, a_{psr}^{(l)}$ із формул (4.3.44)-(4.3.46), (4.3.53)-(4.3.57), з урахуванням рівнянь (4.3.9) після деяких перетворень знову приходимо до системи рівнянь виду:

$$\sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} A_{\mu\nu}^{(l)} \alpha_\nu^{(l)} = B_\mu^{(l)} \quad \left(\mu = \overline{1,k}; \mu \neq l\right), \qquad (4.3.58)$$

або в матричній формі

$$A^{(l)} \alpha^{(l)} = B^{(l)}, \qquad (4.3.59)$$

яка вже розглядалася вище. Вектор $\alpha^{(l)}$ визначається рівністю (4.3.26), а елементи матриці $A^{(l)}$ і вектора вільних членів $B^{(l)}$ такими формулами:

$$A_{\mu\nu}^{(l)} = r_{\mu\nu} - \frac{1}{2}\sum_{p=1}^{m}\{[(m+4)z_p^{(\nu)} - \sum_{j=1}^{p-1} z_{jjp}^{(\nu)} -$$

$$- \sum_{j=p}^{m} z_{pjj}^{(\nu)}]z_p^{(\mu)} + z_{pp}^{(\nu)}z_{pp}^{(\mu)} + \frac{z_{ppp}^{(\nu)}z_{ppp}^{(\mu)}}{3} -$$

$$- z_p^{(\nu)}z_{ppp}^{(\mu)}\} - \frac{1}{2}\sum_{p,s=1}^{m}[(z_{pss}^{(\nu)} - z_p^{(\nu)})z_{pss}^{(\mu)} + (z_{pps}^{(\nu)} -$$

$$- z_s^{(\nu)})z_{pps}^{(\mu)} + 2z_{ps}^{(\nu)}z_{ps}^{(\mu)}] - \sum_{p,s,r=1}^{m} z_{psr}^{(\nu)}z_{psr}^{(\mu)}$$

$$(\mu,\nu = \overline{1,k}; \nu,\mu \neq l)$$

(4.3.60)

$$B_{\mu}^{(l)} = r_{\mu l} - \frac{1}{2}\sum_{p=1}^{m}\{[(m+4)z_p^{(l)} - \sum_{j=1}^{p-1} z_{jjp}^{(l)} -$$

$$- \sum_{j=p}^{m} z_{pjj}]z_p^{(\mu)} + z_{pp}^{(l)}z_{pp}^{(\mu)} + \frac{z_{ppp}^{(l)}z_{ppp}^{(\mu)}}{3} -$$

$$- z_p^{(l)}z_{ppp}^{(\mu)}\} - \frac{1}{2}\sum_{\substack{p,s=1 \\ (p<s)}}^{m}[(z_{pss}^{(l)} - z_p^{(l)})z_{pss}^{(\mu)} + (z_{pps}^{(l)} -$$

$$- z_s^{(l)})z_{pps}^{(\mu)} + 2z_{ps}^{(\nu)}z_{ps}^{(\mu)}] - \sum_{\substack{p,s,r=1 \\ (p<s<r)}}^{m} z_{psr}^{(l)}z_{psr}^{(\mu)}$$

$$(\mu = \overline{1,k}; \nu,\mu \neq l)$$

(4.3.61)

Таким чином, і для поліномів третього степеня визначення $(k-1)$ коефіцієнтів $a_{\nu}^{(l)}$ зводиться до розв'язків систем лінійних неоднорідних алгебраїчних рівнянь, коефіцієнти при невідомих в яких і вільні члени розраховуються по отриманих формулах.

4.4 Статистичний аналіз регресійних моделей

Слідуючим кроком, який треба зробити після отримання параметрів регресійної моделі, це статистичний аналіз рівнянь регресії. Він складається з таких етапів:

а) перевірка гіпотези про статистичну значущість оцінок параметрів моделі;

б) перевірка гіпотези про адекватність і інформативність моделі.

Зупинимося, наперед всього, на перший з перелічених задач.

4.4.1 Перевірка гіпотези про статистичну значущість параметрів лінійної регресійної моделі

Як зазначалося вище, на основі МНК вектор оцінок коефіцієнтів лінійної множинної регресії визначається матричним рівнянням

$$B = \sigma_y \sigma^{-1} R_x^{-1} R_{xy} \qquad (4.4.1)$$

Але можна показати, що для цього можна використати й еквівалентне рівняння

$$B = F^{-1}XY \qquad (4.4.2)$$

де $F = XX^{'}$ - інформаційна матриця Фішера, X - матриця порядку $n \times m$ центрованих значень предікторів, Y - стовпець центрованих значень предіктанта. З іншого боку

$$F^{-1} = \frac{1}{\sigma_y^2}\begin{pmatrix} \sigma^2(b_1) & K_{b_1b_2} & \dots & K_{b_1b_n} \\ K_{b_2b_1} & \sigma^2(b_2) & \dots & K_{b_2b_n} \\ \dots & \dots & \dots & \dots \\ K_{b_nb_1} & K_{b_nb_2} & \dots & \sigma^2(b_n) \end{pmatrix} =$$

$$= \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1n} \\ c_{21} & c_{22} & \dots & c_{2n} \\ \dots & \dots & \dots & \dots \\ c_{n1} & c_{n2} & \dots & c_{nn} \end{pmatrix} = C \qquad (4.4.3)$$

В рівності (4.4.3) $\sigma^2(b_i)$ - дисперсія i-того коефіцієнта регресії, а $K_{b_ib_j}$ - коваріація відповідних коефіцієнтів. Отже, як очевидно, діагональні елементи матриці дорівнюють C

$$C_{ii} = \frac{\sigma^2(b_i)}{\sigma_y^2} \quad (i = 1,2,...,n) \tag{4.4.4}$$

Звідки

$$\sigma(b_i) = \sqrt{C_{ii}}\sigma_y \tag{4.4.5}$$

Можна показати, крім того, що

$$XY = \begin{pmatrix} \sum_{i=1}^{m} x_{1i} y_i \\ \sum_{i=1}^{m} x_{2i} y_i \\ \dots \\ \sum_{i=1}^{m} x_{ni} y_i \end{pmatrix} = m\sigma_y \sigma R_{xy} \tag{4.4.6}$$

В рівності (4.4.6) m - об'єм вибірок, σ_y - середній квадратичний відхил предіктанта, σ - діагональна матриця середніх квадратичних відхилів предікторів. Компануючи рівності (4.4.2), (4.4.3) і (4.4.6), маємо

$$B = m\sigma_y C\sigma R_{xy} \qquad (4.4.7)$$

Якщо звернутися тепер до рівняння (4.4.1), то видно, що

$$m\sigma_y C\sigma = \sigma_y \sigma^{-1} R_{xy} \ , \qquad (4.4.8)$$

звідки

$$C = \frac{1}{m}\sigma^{-1} R_x^{-1} \sigma^{-1} \qquad (4.4.9)$$

Розраховуючи обернену матрицю R_x^{-1}, можна легко знайти C_{ii} $\left(i = \overline{1,n}\right)$діагональні елементи матриці (4.4.9), а за допомогою формули (4.4.5), змінюючи σ_y на середній квадратичний відхил відтворюванності $S_{\hat{y}}$ (про нього докладно буде йтись нижче), і оцінку середнього квадратичного відхилу коефіцієнтів регресії

$$S\left(b_i\right) = \sqrt{C_{ii}} S_{\hat{y}} \qquad (4.4.10)$$

Тепер можна побудувати довірчий інтервал для коефіцієнтів регресії

$$\Delta b_i = S(b_i) t_{кр}\left(\frac{\alpha}{2}, f_l\right), \qquad (4.4.11)$$

де $t_{кр}$ - критерій Стьюдента для рівня значущості $\frac{\alpha}{2}$ і числа степенів волі $f_l = m - n$, де n - число коефіцієнтів регресії.

Гіпотеза H_0 про статистичну значущість коефіцієнта регресії не відхиляється, якщо

$$|b_i| > \Delta b_i \qquad (4.4.12)$$

При цьому треба переконатися, що коефіцієнти корреляції

$$r_{b_i bj} = \frac{K_{b_i b_j}}{\sigma(b_i)\sigma(b_j)}, \qquad (4.4.13)$$

які не складно отримати на основі недіагональних елементів матриці C, близькі до нуля, тобто що параметри моделі некоррельовані.

Оцінка статистичної значущості коефіцієнтів регресії може проводитися за допомогою критерію Стьюдента

$$t = \frac{|b_i|}{S_{b_i}}$$

Середній квадратичні відхили коефіцієнтів регресії визначаються формулою (4.4.10). Гіпотеза про статистичну значущість коефіцієнтів регресії не відхиляється, коли $t > t_{кр}(\alpha, \nu)$. Але вилучати ті коефіцієнти регресії, для яких ця нерівність не виконується, або не виконується нерівність

(4.4.12), треба дуже обережно. Може статися, що вилучення їх приведе до зниження інформативності моделі. Щоб уникнути цієї ситуації, треба включити в модель й ті коефіцієнти регресії, для яких критерій Стьюдента не дуже відрізняється від критичного значення.

Члени рівняння регресії, для коефіцієнтів яких нерівність (4.4.12) не виконується, вилучаються з рівняння регресії.

4.4.2 Аналіз статистичної значущості коефіцієнтів системи рівнянь множинної нелінійної регресії.

У всіх наведених вище викладках, які відносяться до побудови рівнянь нелінійної регресії зі зворотніми зв'язками, вважалося, що вектори предікторів X в k рівняннях регресії, що складають систему (4.3.1), мають одне й те ж число компонент $x_i \ \left(i=\overline{1,n}\right)$. У дійсності кожний предіктант, що описується l - тим рівнянням системи (4.3.1), може мати різну кількість предікторів, які взагалі кажучи, мають і різну фізичну природу. Однак ці обставини не накладають яких-небудь обмежень на загальність отриманих результатів. Дійсно, всі коефіцієнти рівнянь системи залежать тільки від оцінок моментів, які можна розрахувати на основі статистичних сукупностей предікторів і предіктанта, тобто вони знаходяться незалежно. Таким чином результати, що отримані для одного із рівнянь системи (4.3.1), не чинять ніякого впливу на коефіцієнти інших рівнянь. Більш того, зовсім не обов'язково, щоб рівняння системи мали однакову структуру. Навпаки, задача полягає в тому, щоб шляхом статистичного аналізу рівнянь системи (4.3.1) здійснити вибір їх найбільш прийнятної структури.

Першим етапом аналізу структури системи прогностичних рівнянь є оцінка значущості предікторів з точки зору внеску їх в дисперсію предіктанта, що обумовлюється дією всієї сукупності предікторів. Щоб отримати потрібні для цього

рівняння, знайдемо дисперсію предіктанта $\hat{y}_l$. Це можна зробити, застосовуючи відому рівність

$$Д[\hat{y}_l] = M[\hat{y}_l^2] - \left(M[\hat{y}_l]\right)^2 \qquad (4.4.14)$$

Для спрощення викладок, проведемо їх для полінома першого степеня. Для поліномів більш високих степенів запишемо кінцеві результати. Піднесемо до квадрату рівність (4.4.17) і застосуємо операцію математичного сподівання. Будемо мати:

$$M[\hat{y}_l^2] = a_0^{(l)2} + \sum_{i=1}^{m} a_i^{(l)2} M[x_i^2] +$$

$$+ 2a_0^{(l)} \sum_{i=1}^{m} a_i^{(l)} M[x_i] + 2 \sum_{\substack{i,j=1 \\ (i<j)}}^{m} a_i^{(l)} a_j^{(l)} M[x_i x_j] +$$

$$+ 2a_0^{(l)} \sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} \alpha_\nu^{(l)} M[y_\nu] + 2 \sum_{i=1}^{m} a_i^{(l)} \sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} \alpha_\nu^{(l)} M[x_i y_\nu] +$$

$$+ \sum_{\substack{\nu,\mu=1 \\ (\nu,\mu \neq l)}}^{k} \alpha_\nu^{(l)} \alpha_\mu^{(l)} M[y_\nu y_\mu]$$

$$\left(l = \overline{1,k}\right)$$

(4.4.15)

Рівність (4.4.15), якщо врахувати властивості предіктанта і предікторів, що розглядалися вище, а також систему (4.4.9), буде мати такий вид:

$$M[\hat{y}_l^2] = a_i^{(l)2} + \sum_{i=1}^{m} a_i^{(l)2} + 2\sum_{i=1}^{m} a_i^{(l)} \sum_{\substack{\nu=1 \\ (\nu \neq l)}}^{k} \alpha_\nu^{(l)} z_i^{(\nu)} +$$

$$+ \sum_{\substack{\nu,\mu=1 \\ (\nu,\mu \neq l)}}^{k} \alpha_\nu^{(l)} \alpha_\mu^{(l)} r_{\nu\mu}]$$

$$\left(l = \overline{1,k}\right)$$

(4.4.16)

Підставимо тепер замість суми по індексу ν у передостанньому члені правої частини (4.4.16) її значення із рівняння (4.3.19). Після деяких перетворень отримаємо:

$$M[\hat{y}_l^2] = a_i^{(l)2} + 2\sum_{i=1}^{m} a_i^{(l)} z_i^{(l)} - \sum_{i=1}^{m} a_i^{(l)2} +$$

$$+ \sum_{\substack{\nu,\mu=1 \\ (\nu,\mu \neq l)}}^{k} \alpha_\nu^{(l)} \alpha_\mu^{(l)} r_{\nu\mu}]$$

$$\left(l = \overline{1,k}\right)$$

(4.3.17)

Другий член рівності (4.4.14) знаходиться просто. З урахуванням властивостей предікторів і предіктанта він дорівнює

$$M[\hat{y}_l] = a_l^{(0)} = 0\,, \qquad (4.4.18)$$

оскільки для поліномів першого степеня $a_0^{(l)} \equiv 0$. Отже дисперсія предіктанта дорівнює:

$$Д^{(1)}[\hat{y}_l] = \sum_{i=1}^{m}[2z_i^{(l)}a_i^{(l)} - a_i^{(l)2}] + \\ + \sum_{\substack{\nu,\mu=1 \\ (\nu,\mu\neq l)}}^{k} \alpha_\nu^{(l)}\alpha_\mu^{(l)}r_{\mu\nu} \tag{4.4.19}$$

Із рівняння (4.4.19) впливає, що вона складається з двох частин: $Д_x^{(1)}[\hat{y}_l]$ - частини, що обумовлюється впливами предікторів, і $Д_y^{(1)}[\hat{y}_l]$ - частини, що обумовлюється дією зворотніх зв'язків. Очевидно

$$Д_x^{(1)}[\hat{y}_l] = \sum_{i=1}^{m}[2z_i^{(l)}a_i^{(l)} - a_i^{(l)2}], \tag{4.4.20}$$

$$Д_y^{(1)}[\hat{y}_l] = \sum_{\substack{\nu,\mu=1 \\ (\nu,\mu\neq l)}}^{k} \alpha_\nu^{(l)}\alpha_\mu^{(l)}r_{\mu\nu} \tag{4.4.21}$$

З формули (4.4.20) видно, що її структура дає змогу визначити внесок кожного предіктора в дисперсію, що обумовлюється впливами всіх предікторів. Він очевидно дорівнює:

$$\chi_p^{(1)}(\hat{y}_l) = 2z_p^{(l)}a_p^{(l)} - a_p^{(l)2} \tag{4.4.22}$$

Таким чином,

$$Д_x^{(1)}[\hat{y}_l] = \sum_{p=1}^{n} \chi_p^{(1)}(\hat{y}_l) \qquad (4.4.23)$$

Аналогічно

$$Д_y^{(1)}[\hat{y}_l] = \sum_{s=1}^{k} \theta_s^{(1)}(\hat{y}_l)\ , \qquad (4.4.24)$$

де

$$\theta_s^{(1)}[\hat{y}_l\ ] = \alpha_s^{(l)} \sum_{\substack{\mu=1 \\ (\mu \neq l)}}^{k} \alpha_\mu^{(l)} r_{s\mu}\ ] \qquad (4.4.25)$$

$$\left(l = \overline{1,k}\right)$$

характеризує внесок s - того зворотнього зв'язку в дисперсію, що обумовлюється впливами всіх зворотніх зв'язків. Рівняння (4.4.24) і (4.4.25) залишаються незмінними для поліномів другого і третього степенів, а величини $\chi_p\left(y_l\right)$ визначаються рівняннями:

$$\chi_p^{(2)}(\hat{y}_l) = 2\left(z_p^{(l)} a_p^{(l)} + z_{pp}^{(l)} a_{pp}\right) -$$
$$-\left(a_p^{(l)2} + 2a_{pp}^{(l)2}\right) + \sum_{j=p+1}^{m}\left(2z_{pj}^{(l)} a_{pj}^{(l)} - a_{pj}^{(l)2}\right),$$
$$\left(p = \overline{1,m}; l = \overline{1,k}\right)$$

(4.4.26)

для поліномів другого степеня;

$$\chi_p^{(3)}(\hat{y}_l) = 2\left(z_p^{(l)} a_p^{(l)} + z_{pp}^{(l)} a_{pp}^{(l)} + z_{ppp}^{(l)} a_{ppp}^{(l)}\right) -$$
$$-[a_p^{(l)2} + 6a_p^{(l)} a_{ppp}^{(l)} + (2a_p^{(l)} + 6a_{ppp}^{(l)})(\sum_{j=1}^{p-1} a_{jjp}^{(l)} +$$
$$+ \sum_{j=p+1}^{m} a_{pjj}^{(l)}) + 2a_{pp}^{(l)2} + 15a_{ppp}^{(l)\,2}] +$$
$$+ \sum_{j=p+1}^{m} \{2(z_{pj}^{(l)} a_{pj}^{(l)} + z_{pjj}^{(l)} a_{pjj}^{(l)} + z_{ppj}^{(l)} a_{ppj}^{(l)}) - [a_{pj}^{(l)2} +$$
$$+ 2a_{pjj}^{(l)\,2} + 2a_{ppj}^{(l)\,2} + a_{ppj}^{(l)}(\sum_{\eta=1}^{j-1} a_{\eta\eta j}^{(l)} + \sum_{\eta=j+1}^{m} a_{j\eta\eta}^{(l)}) +$$
$$+ a_{pjj}^{(l)}(\sum_{\eta=1}^{p-1} a_{\eta\eta p}^{(l)} + \sum_{\eta=p+1}^{m} a_{p\eta\eta}^{(l)})] + \sum_{s=j+1}^{m} 2z_{pjs}^{(l)} a_{pjs}^{(l)} -$$
$$a_{pjs}^{(l)\,2})\},$$
$$\left(p = \overline{1,m}; l = \overline{1,k}\right)$$

(4.4.27)

для поліномів третього степеня.

Тепер можна перевірити гіпотезу про незначущість внеску перших s предікторів, якщо розрахувати $\chi_p(\hat{y}_l)$ і розташувати їх у порядку збільшення. Гіпотеза H_0 формулюється так: дисперсія, що обумовлюється внесками останніх $m-t$ - предікторів на рівні α значуще розрізняється від дисперсії, що обумовлюється дією всіх предікторів. Для цього застосовується критерію Фішера

$$F = \frac{Д_x[\hat{y}_l]}{\sum_{i=1}^{m-t} \chi_i^t[\hat{y}_l]} \qquad (4.4.28)$$

Гіпотеза не відхиляється, якщо $F > F_{кр}[\alpha; N-m; N-(m-t)]$. Таким же чином перевіряється гіпотеза про значущість внеску кожного зворотнього зв'язку.

Отримані результати дають можливість вирішити задачу і про статистичну значущість членів нелінійних рівнянь регресії, які мають другу і третю степені. Це можна зробити шляхом порівняння Фішера

$$F = \frac{\chi_p^t[\hat{y}_l]}{\chi_i^{(t-1)}[\hat{y}_l]}, \quad t = 2,3 \qquad (4.4.29)$$

З критичним значенням $F_{кр}[\alpha; N-m; N-(m-k)]$, де k - кількість членів другого чи третього порядків, які не роблять

суттєвого внеску в дисперсію, що обумовлюється впливом p - того предіктора.

Дослідження впливу нелінійностей і предікторів, звичайно, треба починати з рівнянь, що отримують більш високий степінь змінних.

4.4.3 Перевірка гіпотез про адекватність та інформативність регресійних моделей

Перевірка гіпотези H_0 про адекватність моделей виконується шляхом порівняння дисперсій неадекватності $S^2_{над}$ і відтворюванності $S^2_{\hat{y}}$. Перша з них розраховується за допомогою рівняння:

$$S^2_{над} = \frac{\sum_{j=1}^{N_1}\left(y_j - \hat{y}_j\right)^2}{N_1 - m} \qquad (4.4.30)$$

де y_j - значення предіктанта із вихідної сукупності, $\hat{y}_j$ - значення предіктанта, що відтворюється за рівнянням регресії, N_1 - число чисельних експериментів з моделлю, m - число параметрів у моделі.

Дисперсія відтворюванності $S^2_{\hat{y}}$ - це є дисперсія нев'язки рівняння, яка вже розглядалася у розділі 4.1.2. Вона визначається рівнянням:

$$S_{\hat{y}}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{N-1} \qquad (4.4.31)$$

де N - об'єм вихідної сукупності предіктанта. Гіпотеза про адекватність моделі перевіряється за допомогою критерію Фішера, що розраховується за формулою:

$$F_p = \frac{S_{над}^2}{S_{\hat{y}}^2} \qquad (4.4.32)$$

Модель вважається адекватною, якщо

$$F_p \le F_{кр}[\alpha; N-k; N-1]$$

Гіпотеза про інформативність моделі перевіряється шляхом порівняння дисперсії предіктора

$$S_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}{N-1} \qquad (4.4.33)$$

із дисперсією відтворюванності також за допомогою критерію Фішера

$$F_R = \frac{S_y^2}{S_{\hat{y}}^2} \qquad (4.4.34)$$

Якщо $F_R > F_{кр}[\alpha; N-1; N-1]$, то модель, що досліджується, описує результати прогнозування краще, ніж просте середнє значення предіктанта, тобто модель інформативна. На практиці висувається більш жорстка вимога. Вважають наприклад, що модель має передбачальну властивість, тобто є інформативною, якщо $F_R \geq 1{,}5F_{кр}$. Мірою корисності є також множинний коефіцієнт корреляції R або множинний коефіцієнт детермінації R^2, оскільки

$$R^2 = 1 - \frac{S_{\hat{y}}^2}{S_y^2} \qquad (4.4.35)$$

Наближення коефіцієнта множинної корреляції до одиниці свідчить про задовільну інформативність моделі. Навпаки, якщо він наближається до нуля, це свідчить про відсутність корреляції між предіктантом і всіма предікторами, тобто про неінформативність побудованої моделі.

4.5 Приклади застосування регресійних моделей.

Спочатку розглянемо приклад, що ілюструє процес побудови лінійного рівняння регресії як моделі прогнозу концентрації SO_2 для одного з районів Одеси. У якості предікторів були розглянуті 30 характеристик стану пограничного шару атмосфери, які впливають на формування полів концентрації інгредієнта. Методом "включення" (розділ 4.2.1.2) було визначено дев’ять таких найбільш значущих предікторів: середня концентрація SO_2 за попередню добу (предіктор №1), температура повітря (предіктор №2), напрямок

вітру (предіктор №3), швидкість вітру (предіктор №4), відносна вологість повітря (предіктор №5), верхня границя приземної інверсії (предіктор №6), нижня границя припіднятої інверсії (предіктор №7), верхня її границя (предіктор №8), вертикальний градієнт температури повітря в пограничному шарі (предіктор №9). Перелік цих метеорологічних величин характеризує перенос і турбулентну дифузію домішок в приземному шарі атмосфери.

Оскільки термін прогнозу дорівнював 12 годин, то всі предіктори, окрім першого, добиралися за попередній дванадцяти годинний термін. Вектор корреляцій між предіктантом y і переліченими предікторами X має такий вид:

$$R_{xy} = \begin{pmatrix} 0.837 \\ -0.214 \\ 0.310 \\ 0.196 \\ 0.507 \\ 0.210 \\ 0.414 \\ 0.374 \\ -0.196 \end{pmatrix}$$

Лінійне рівняння регресії будувалося за допомогою метода покрокової регресії (розділ 4.2.1.3). Нижче наводяться результати відповідного ітераційного процесу

1) Крок 1-й

Вільний член $b_0 = 5{,}527 * 10^{-12}$

Коефіцієнти регресії

$b_1 = 0{,}837$, коефіцієнт Стьюдента t = 8,79. Предіктор 1

Множинний коефіцієнт корреляції R = 0,837.

2) Крок 2-й

Вільний член $b_0 = 5{,}380 * 10^{-12}$

Коефіцієнт регресії

$b_1 = 0{,}847$, коефіцієнт Стьюдента t = 9,19 Предіктор 1

$b_2 = 0{,}148$, коефіцієнт Стьюдента t = 1,50 Предіктор 8

Множинний коефіцієнт корреляції R = 0,848

3) Крок 3

Вільний член $b_0 = 5{,}32 * 10^{-12}$

Коефіцієнти регресії

$b_1 = 0{,}872$, коефіцієнт Стьюдента t = 9,81 Предіктор 1

$b_2 = 0{,}173$, коефіцієнт Стьюдента t = 1,80 Предіктор 8

$b_3 = 0{,}145$, коефіцієнт Стьюдента t = 1,64 Предіктор 2

Множинний коефіцієнт корреляції R = 0,860.

4) Крок 4

Вільний член $b_0 = 5{,}33 * 10^{-12}$

Коефіцієнт регресії

$b_1 = 0{,}867$, коефіцієнт Стьюдента t = 9,92 Предіктор 1

$b_2 = 0{,}199$, коефіцієнт Стьюдента t = 2,11 Предіктор 8

$b_3 = 0{,}116$, коефіцієнт Стьюдента t = 1,33 Предіктор 2

$b_4 = -0{,}099$, коефіцієнт Стьюдента t = 1,29 Предіктор 9

Множинний коефіцієнт корреляції R = 0,865

5) Крок 5

Вільний член $b_0 = 5{,}74 * 10^{-12}$

Коефіцієнт регресії

$b_1 = 0{,}826$, коефіцієнт Стьюдента t = 9,53 Предіктор 1

$b_2 = 0{,}197$, коефіцієнт Стьюдента t = 2,10 Предіктор 8

$b_3 = 0{,}103$, коефіцієнт Стьюдента t = 1,19 Предіктор 2

$b_4 = - 0{,}086$, коефіцієнт Стьюдента t = 0,99 Предіктор 9

$b_5 = 0{,}079$, коефіцієнт Стьюдента t =0,91 Предіктор 5

Множинний коефіцієнт корреляції R = 0,867

6) Крок 6

Вільний член $b_0 = - 6{,}075 * 10^{-12}$

Коефіцієнти регресії

b_1 = 0,826, коефіцієнт Стьюдента t = 9,49. Предіктор 1
b_2 = 0,215, коефіцієнт Стьюдента t = 2,28. Предіктор 8
b_3 = 0,095, коефіцієнт Стьюдента t = 1,09. Предіктор 2
b_4 = 0,094, коефіцієнт Стьюдента t = 1,08. Предіктор 9
b_5 = 0,078, коефіцієнт Стьюдента t = 0,90. Предіктор 5
b_6 = 0,078, коефіцієнт Стьюдента t = 0,55. Предіктор 3
Множинний коефіцієнт корреляції R = 0,866

7) Крок 7

Вільний член $b_0 = - 7{,}673 * 10^{-12}$

Коефіцієнти регресії

b_1 = 0,815, коефіцієнт Стьюдента t = 9,45. Предіктор 1
b_2 = 0,263, коефіцієнт Стьюдента t = 3,89. Предіктор 8
b_3 = 0,074, коефіцієнт Стьюдента t = 0,86. Предіктор 2
b_4 = - 0,087, коефіцієнт Стьюдента t = 1,01. Предіктор 9
b_5 = 0,086, коефіцієнт Стьюдента t = 1,00. Предіктор 5
b_6 = 0,098, коефіцієнт Стьюдента t = 0,68. Предіктор 3
b_7 = 0,176, коефіцієнт Стьюдента t = 1,89. Предіктор 7
Множинний коефіцієнт корреляції R = 0,869.

8) Крок 8

Вільний член $b_0 = - 9{,}180 * 10^{-12}$

Коефіцієнт регресії

b_1 = 0,814, коефіцієнт Стьюдента t = 9,41. Предіктор 1
b_2 = 0,363, коефіцієнт Стьюдента t = 3,87. предіктор 8
b_3 = 0,071, коефіцієнт Стьюдента t = 0,82. Предіктор 2
b_4 = - 0,98, коефіцієнт Стьюдента t = 1,13. Предіктор 9
b_5 = 0,080, коефіцієнт Стьюдента t = 0,93. Предіктор 5
b_6 = 0,118, коефіцієнт Стьюдента t = 0,83. Предіктор 3
b_7 = - 0,179, коефіцієнт Стьюдента t = 1,91. Предіктор 7
b_8 = 0,033, коефіцієнт Стьюдента t = 0,38. Предіктор 4
Множинний коефіцієнт корреляції R = 0,868

9) Крок 9

Вільний член $b_0 = - 9{,}40 * 10^{-12}$

Коефіцієнт регресії

b_1 = 0,815, коефіцієнт Стьюдента t = 9,41. Предіктор 1
b_2 = 0,362, коефіцієнт Стьюдента t = 3,86. Предіктор 8

b_3 = 0,073, коефіцієнт Стьюдента t = 0,84. Предіктор 2
b_4 = - 0,097, коефіцієнт Стьюдента t = 1,13. Предіктор 9
b_5 = 0,080, коефіцієнт Стьюдента t = 0,92. Предіктор 5
b_6 = 0,120, коефіцієнт Стьюдента t = 0,85. Предіктор 3
b_7 = - 0,177, коефіцієнт Стьюдента t = 1,89. Предіктор 7
b_8 = 0,035, коефіцієнт Стьюдента t = 0,40. Предіктор 4
b_9 = 0,057, коефіцієнт Стьюдента t = 0,16. Предіктор 6.
Множинний коефіцієнт корреляції R = 0,868.

Порівнюючи 8-й і 9-й кроки, приходимо до висновку, що ітераційний процес усталюється. Дійсно, невеликі різниці між коефіцієнтами регресії восьмого і дев'ятого кроків виявляються тільки у третьому знаку, коефіцієнти Стьюдента теж змінюються незначно, а множинний коефіцієнт регресії зовсім не змінюється.

У попередньому розділі зазначалось, що формальне відкидання коефіцієнтів регресії у лінійній моделі по значеннях коефіцієнта Стьюдента може привести до втрати адекватності й інформативності моделі. Це особливо відноситься до випадку, коли вибірки предіктанта і предікторів відносно невеликі, внаслідок чого елементи матриці корреляцій предікторів можуть утримувати значні похибки. Тому остаточний висновок відносно виду моделі у такому випадку треба робити на основі чисельних експериментів з моделлю. У прикладі, що розглядається, чисельні експерименти показали, що задовільну адекватність показує прогностична модель для прогнозу концентрації $SO_2 - \hat{q}$ з предікторами, для яких коефіцієнт Стьюдента $t \geq 1$. Як видно це такі предіктори: середня концентрація SO_2 за попередню добу (x_1), верхня границя припіднятої інверсії (x_8), нижня її границя (x_7) і вертикальний градієнт температури повітря у пограничному шарі атмосфери (x_9).

Отже прогностична модель має вид:

$$\hat{q} = 0.8x_1 - 0.177x_7 + 0.362x_8 - 0.097x_9$$

В рівнянні регресії предіктори і предіктанти є центрованими і нормованими на середній квадратичний відхил. Значення множинного коефіцієнта корреляції свідчить про те, що адекватність моделі є задовільною. Легко можна побачити, що дисперсія похибки прогнозу складає біля 24% від дисперсії інгредієнта.

Розглянемо тепер приклад, який відноситься до системи нелінійних рівнянь регресії зі зворотніми зв'язками (4.3.1). На основі цієї системи була побудована статистична модель теплих туманів, що дуже часто утворюються в холодному півріччі над районами північно-західного Причорномор'я.

Багаточисельні експериментальні вимірювання показали, що спектри крапель таких туманів описуються гама-розподілом, який докладно розглядався в розділі (2.4).

$$f(r) = \frac{\alpha^{\lambda} r^{\lambda-1} e^{-\alpha r}}{\Gamma(\lambda)}, \qquad (4.5.1)$$

де r - радіус краплі; α - параметр масштабу; λ - параметр форми, $\Gamma(\lambda)$ - гама-функція. Крім того, важливими характеристиками спектра є r_{mod} - модальний радіус крапель N_{r_m} - концентрація крапель у модальному частковому інтервалі спектра.

Перелічені параметри спектра крапель розглядалися у якості предіктантів. Для них шляхом "просіювання" (розділ 4.2.1.2) був визначений оптимальний склад впливаючих факторів: радіаційний баланс B, який характеризує суму всіх радіаційних потоків тепла на рівні добору проб мікроструктури,

швидкість вітру υ_2 на цьому рівні (рівень 2 м), градієнт швидкості вітру в шарі 2-15 м $\Delta\upsilon_{2-16}$, температура повітря T_2 - на рівні 2 м, температура підстилаючої поверхні T_0, градієнт температури ΔT_{0-2} в шарі 0-2 м і градієнт температури ΔT_{2-12} в шарі 2-12 м.

На основі рядів значень перелічених параметрів внутрішньої (параметри спектра крапель туманів) і зовнішньої (параметри приземного шару атмосфери, що зазначені вище) структури туманів за тривалий термін дали змогу знайти коефіцієнти системи поліномів (4.3.42), яка є статистичною моделлю теплих туманів. Очевидно, модель складається з чотирьох нелінійних рівнянь регресії зі зворотніми зв'язками.

Аналіз коефіцієнтів регресії мікрофізичних характеристик туманів по параметрах приземного шару є дуже складним і більш доцільно визначити внесок кожного з впливаючих факторів у дисперсію характеристик мікроструктури. Як зазначалося вище, це можна зробити за допомогою формул, що розглядаються в попередньому розділі. В табл. 4.1 містяться частки в процентах дисперсії параметрів мікроструктури, що утворюється впливом кожного впливаючого фактора, відносно загальної дисперсії, яка обумовлюється дією всієї системи факторів.

Таблиця 4.1- Внесок (%) впливаючих факторів в дисперсію мікрофізичних параметрів туманів.

Пара-метри	Ф а к т о р и

мікро-струк-тури	B	υ_2	$\Delta\upsilon_{2-16}$	T_0	T_2	ΔT_{0-2}	ΔT_{2-12}
α	53,6	19,6	1,8	11,4	5,3	4,0	4,3
λ	5,4	26,8	0,9	27,6	25,0	0,2	14,1
r_{mod}	7,1	69,6	18,7	2,1	0,1	1,5	0,9
N_{r_m}	3,3	3,5	10,2	24,3	52,1	5,1	1,5

Із табл.4.1 випливає, що ширина спектру розмірів крапель туманів, яка визначається параметром α, головним чином змінюється під впливом коливань радіаційного балансу B, швидкості вітру υ_2, а також температури підстилаючої поверхні. Форма спектра λ залежить від швидкості вітру, температури підстилаючої поверхні і температури повітря на рівні 2м, а також вертикального градієнту температури. Ці залежності знаходять добру фізичну обгрунтованість. Притоки і стоки короткохвильової і довгохвильової радіації обумовлюють температурний режим підстилаючої поверхні і нижнього шару повітря, а останній визначає умови конденсації водяної пари, а швидкість вітру поряд з вертикальним градієнтом температури - умови розвитку турбулентності, яка, з свого боку, впливає на характеристики спектру за рахунок перемішування об'ємів повітря в тумані і механізму турбулентної коагуляції. На змінювання модального радіуса впливають головним чином швидкість вітру, градієнт швидкості вітру, а на концентрацію крапель модального часткового інтервалу - температура на рівні 2м і температура підстилаючої поверхні, а також градієнт швидкості вітру в приземному шарі атмосфери.

Одначе при розгляданні туману як деякої складної системи впливи характеристик стану приземного шару повітря виявляється значно складнішими, оскільки параметри мікроструктури туману зв'язані між собою. В моделі ці

взаємозв'язки віддзеркалюють коефіцієнти зворотніх зв'язків α_v. Значення цих коефіцієнтів утримуються в табл. 4.2.

Таблиця 4.2 - Коефіцієнти зворотніх зв'язків.

Параметри мікро-структури	α	λ	r_{mod}	N_{r_m}
α	-	1,00	- 0,19	0,23
λ	0,61	-	0,39	-0,18
r_{mod}	-0,28	0,96	-	0,02
N_{r_m}	0,62	-0,79	0,03	-

Як випливає з табл. 4.2, поширення спектра крапель (збільшення α) супроводиться збільшенням параметра форми λ, тобто зменшенням асиметрії розподілу крапель по розмірах. З іншого боку, поширення спектра при незмінному параметру форми зв'язане зі зменшенням модального радіусу крапель. При незмінному параметрі масштабу α зменшення асиметрії розподілу (збільшення λ) зв'язане зі збільшенням модального радіусу, але при цьому кількість крапель модального часткового інтервалу радіусів крапель зменшується, тобто спектр становиться більше розтягнутим по розмірах крапель.

Отже, статистична модель туману у виді системи нелінійних рівнянь регресії зі зворотніми зв'язками дає досить задовільні з фізичної точки зору результати, які співпадають з висновками теорії полідисперсних хмар і туманів.

5. ФАКТОРНИЙ АНАЛІЗ

5.1 Основні ідеї факторного аналізу

Розглянемо систему метеорологічних величин x_i $\left(i=\overline{1,n}\right)$, які є центрованими, тобто $M[x_i]=0$, $\forall i=\overline{1,n}$. Задача полягає у тому, щоб виразити кожну з них за допомогою k деяких випадкових величин $f_j\left(j=\overline{1,k}\right)$, які є однаковими для всіх метеорологічних величин, але входять у кожну з них із своєю вагою p_{ij}. Фактори f_j не повністю вичерпують вихідні змінні x_i. Для цього треба додати ще деякі випадкові величини υ_i, які називають залишками. Отже

$$x_i=\sum_{j=1}^{k}P_{ij}f_j+\upsilon_i\ \left(i=\overline{1,n}\right) \qquad (5.1.1)$$

При цьому $k<<n$. Якщо у рівності (5.1.1) розгорнути суму і придати всі значення індексу i, то прийдемо до системи рівнянь:

$$\begin{cases} x_1=p_{11}f_1+p_{12}f_2+...+p_{1k}f_k+\upsilon_1, \\ x_2=p_{21}f_1+p_{22}f_2+...+p_{2k}f_k+\upsilon_2, \\ ..... \\ x_1=p_{n1}f_1+p_{n2}f_2+...+p_{nk}f_k+\upsilon_n, \end{cases} \qquad (5.1.2)$$

Видно, що всі вихідні величини x_i, як зазначалося вище, виражаються через однакові випадкові величини f_j $(j=1,k)$. Проте вони входять в різні величини з різними ваговими коефіцієнтами p_{ij}.

Введемо позначення

$$X=\begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} \tag{5.1.3}$$

$$F=\begin{pmatrix} f_1 \\ f_2 \\ \dots \\ f_k \end{pmatrix} \tag{5.1.4}$$

$$V=\begin{pmatrix} \upsilon_1 \\ \upsilon_2 \\ \dots \\ \upsilon_n \end{pmatrix} \tag{5.1.5}$$

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ p_{n1} & p_{n2} & \cdots & p_{nk} \end{pmatrix} \qquad (5.1.6)$$

Тоді систему рівняння (5.1.2) можна записати в матричній формі

$$X = PF + V \qquad (5.1.7)$$

Фактори f_j будемо називати узагальненими факторами. Будемо вважати їх такими, що

$$M[FF^{'}] = E \qquad (5.1.8)$$

(E - одинична матриця). Умова (5.1.8) означає, що узагальнені фактори некоррельовані й мають одиничну дисперсію. Будемо вважати, крім того, що залишки υ_i є незалежними, тобто

$$M[V \cdot V^{'}] = Д, \qquad (5.1.9)$$

де $Д$ - діагональна матриця, що складається з дисперсій d_i залишків υ_i

$$Д = \begin{pmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & d_n \end{pmatrix} \qquad (5.1.10)$$

і, що залишки некорреляються з факторами. На цій підставі знайдемо матрицю коваріацій векторів X

$$K = M[X \cdot X^{'}] = M\{(PF + V)(F^{'}P^{'} + V^{'})\} =$$
$$= PM[F \cdot F^{'}]P^{'} + M[VF^{'}] \cdot P^{'} + PM[F \cdot V] +$$
$$+ M[V \cdot V^{'}]$$

Ураховуючи перелічені властивості узагальнених факторів і залишків, отримаємо

$$K = PP^{'} + Д \qquad (5.1.11)$$

Отже, коваріаційну матрицю системи метеорологічних величин X можна виразити через матрицю вагів узагальнених факторів і діагональну матрицю дисперсій залишків.

5.2 Оцінки вагів узагальнених факторів і

дисперсій залишків

Задача полягає у тому, щоб на основі вибіркової матриці коваріацій $\hat{K}$ знайти незміщені, ефективні та умотивовані оцінки факторних вагів p_{ij} і дисперсій залишків d_i $\left(i=\overline{1,n};\ j=\overline{1,k}\right)$. Будемо шукати розв'язок цієї задачі на основі методу максимальної правдоподібності. Він полягає у тому, що за визначним правилом формується функція правдоподібності і знаходять такі шукані параметри, котрі придбавають максимум функцій правдоподібності. Будемо вважати, що система випадкових величин x_i $\left(i=\overline{1,n}\right)$ має багатовимірний нормальний закон розподілу, Тоді функція правдоподібності має такий вид

$$L=-\frac{1}{2}n\ln\left|K\right|-\frac{n}{2}tr\left(K^{-1}\cdot\hat{K}\right), \qquad (5.2.1)$$

де tr - слід матриці, що розташовується в дужках.

Функція правдоподібності, як свідчить рівность (5.1.11), є функцією двох шуканих змінних p_{ij} і d_i. Отже її максимум визначається за допомогою рівнянь

$$\frac{\partial L}{\partial p_{ij}}=0\ ; \qquad \frac{\partial L}{\partial d_i}=0 \qquad (5.2.2)$$

Результат диференціювання приводить до матричних рівнянь

$$P'K' - P'K^{-1}\hat{K}K^{-1} = 0, \qquad (5.2.3)$$

$$diag\left(K^{-1} - K^{-1}\hat{K}K^{-1}\right) = 0 \qquad (5.2.4)$$

Введемо матрицю

$$N = P'Д^{-1}P \qquad (5.2.5)$$

і використаємо тотожність

$$P'K^{-1} = \left(E + N\right)^{-1}P'Д^{-1} \qquad (5.2.6)$$

Щоб побачити, що матрична рівність (5.2.6) це тотожність, помножимо вираз (5.2.6) спочатку праворуч на K, а потім ліворуч на $\left(E + N\right)$. Отримаємо

$$P'Д^{-1}K = \left(E + N\right)P' = P' + NP' \qquad (5.2.7)$$

Замість матриці N поставимо її значення із рівності (5.2.5) і врахуємо рівняння (5.1.11). Будемо мати:

$$P'Д^{-1}K = P' + P'Д^{-1}P \cdot P' = P' +$$
$$+ P'Д^{-1}\left(K - Д\right) = P' + P'Д^{-1}K - P'Д^{-1}Д =$$
$$= P'Д^{-1}K$$

Підставимо тепер рівність (5.2.5) і тотожність (5.2.6) в рівняння (5.2.3). Після нескладних перетворень будемо мати

$$P' = N^{-1}P'Д^{-1}\left(\hat{K} - Д\right) \qquad (5.2.8)$$

Рівняння (5.2.4) помножимо ліворуч на матрицю $Д = K - PP'$ (5.2.9). Отримаємо:

$$diag\left(K - \hat{K}\right) = 0 \qquad (5.2.10)$$

Це рівняння означає, що при $i = j \;\; K_{ii} = \hat{K}_{ii}$, тобто діагональні елементи генеральної і вибіркової матриць коваріацій дорівнюють одне одному. Враховуючи це, за допомогою рівності (5.2.9) отримаємо

$$d_i = \hat{K}_{ii} - \sum_{\nu=1}^{k} P_{i\nu}^2 \qquad (5.2.11)$$

Отже, оскільки вибіркова коваріаційна матриця відома, рівняння (5.2.11) дає можливість знайти дисперсії залишків після розрахування вагових коефіцієнтів узагальнених факторів. Для цього помножимо рівняння (5.2.8) ліворуч на матрицю N, а праворуч на матрицю $Д^{-1}P$. Будемо мати

$$N^2 = P'Д^{-1}\left(\hat{K} \; - Д\right)Д^{-1}P \qquad (5.2.12)$$

Матриця N^2 є діагональною. Її елементи являють собою перші k власні значення матриці $Д^{-1}\left(\hat{K} - Д\right)$, а рядки матриці P - її відповідні власні вектори. Оскільки вибіркова коваріаційна матриця $\hat{K}$ може бути розрахованою, зазначена особливість дає можливість побудувати достатньо простий аргумент ітераційного процесу розв'язку рівнянь (5.2.8) і (5.2.11), за допомогою якого можна знайти оцінки вагів p_{ij} і дисперсій d_i. Ітераційний процес описується такою системою рівностей:

$$\begin{cases} h_{\mu(\nu)} = P'_{\mu(\nu)} Д^{-1}_{(\nu)}; \\ \alpha_{\mu j(\nu)} = h'_{\mu(\nu)} P_{j(\nu)}; \\ g_{\mu(\nu)} = h_{\mu(\nu)} \hat{K} - P'_{\mu(\nu)} - \sum_{j=1}^{\mu-1} \alpha_{\mu j(\nu)} P'_{j(\nu+1)}; \\ \beta_{\mu(\nu)} = g'_{\mu(\nu)} h_{\mu(\nu)}; \\ P_{\mu(\nu+1)} = \dfrac{1}{\sqrt{\beta_{\mu(\nu)}}} g'_{\mu(\nu)}; \end{cases} \tag{5.2.13}$$

де μ - номер узагальненого фактора; $\nu = \overline{0, r}$ - номер ітерації; P'_{μ} - μ - тий рядок матриці P' ($'$ - означає, як і раніше, операцію транспонування).

Для організації ітераційної процедури необхідно визначити число k узагальних факторів, які разом з залишком υ_i описують множину вихідних змінних x_i. Ця задача вирішується шляхом перевірки гіпотези H_0 про наявність рівно k узагальнених факторів. Вона проводиться за допомогою критерія χ^2, що знаходиться за формулою

$$\chi^2 = n\left(\frac{1}{2}\sum_{i=1}^{n} y_{ii}^2 + \sum_{i,j=1}^{n} y_{ij}y_{ji}\right), \qquad (5.2.14)$$

де y_{ij} $(i,j=\overline{1,n})$ - елементи матриці $Y = Д^{-1}\left(\hat{K}_0 - PM^{'}\right)$, а матриця $M^{'}$ визначається так: $M^{'} = N^{-1}P^{'}Д^{-1}\hat{K}_0$. Матриця $\hat{K}_0$ формується шляхом заміни елементів головної діагоналі найбільшими елементами відповідних стовпців виборочної матриці коваріацій $\hat{K}$ (при визначені вагів першого узагальненого фактора) або залишкових матриць (останні утворюються в результаті виключення факторів, для яких оцінки факторів вже проведені). Критерій χ^2 порівнюється з табличним для числа ступенів волі

$$\nu = \frac{1}{2}[(n-k)^2 + (n+k)] \qquad (5.2.15)$$

і вітряного рівня значущості. Число ітерацій визначається співвідношенням:

$$\left|P_{\mu j(m+1)} - P_{\mu j(m)}\right| < \varepsilon, \;\; j = \overline{1,n} \qquad (5.2.16)$$

де ε - визначена точність розрахунків, n - порядок матриці коваріацій.

Після отримання за процедурою (5.2.13) матриць P і $Д$ можна розрахувати вектор F значень узагальнених факторів для вектора X за формулою

$$F = \left(E + N\right)^{-1} P' Д^{-1} X, \qquad (5.2.17)$$

Це дає можливість виразити сукупність n вихідних характеристик через невелику кількість k узагальнених факторів.

5.3 Приклади факторного аналізу метеорологічної інформації.

Одним з прикладів, що можна навести для ілюстрації можливостей факторного аналізу, є дослідження на його основі особливостей температурно-вологісного режиму території України.

Температурно-вологісний режим можна охарактеризувати за допомогою трьох метеорологічних величин: середніми за місяць температурою і вологістю повітря та місячною кількостю опадів. Ці характеристики дають уявлення про середні за місяць притік та стік вологи. У довідниках по клімату наводяться значення перелічених вище параметрів атмосфери. Однак комплексний аналіз їх структури над великими регіонами значно утруднюється, особливо коли йдеться про прогноз

температурно-вологісного режиму. Використання факторного аналізу значно спрощує цю задачу.

Для розрахування вагів узагальненого фактора, дисперсій залишків і дослідження статистичної структури узагальненого фактора температурно-вологісного режиму України використовувалась мережа з 14 довгорядних метеорологічних станцій з об'ємами вибірок середніх місячних температур x_1, відносних вологостей x_2 і сім опадів x_3, що дорівнюють $n = 75$ (з 1990 по 1974 рр.). На основі цих даних були розраховані матриці взаємних коваріацій для кожної метеорологічної станції. Ці матриці, як зрозуміло, мають третій порядок. Перевірка гіпотези H_0 про кількість узагальнених факторів показали, що зазначені три метеорологічні величини визначаються одним узагальненим фактором f з вектором вагів $P = (p_1 p_2 p_3)$, тобто

$$x_1 = p_1 f + \upsilon_1,$$

$$x_2 = p_2 f + \upsilon_2,$$

$$x_3 = p_3 f + \upsilon_3,$$

де υ_1, υ_2, υ_3 - відповідні залишки. У якості прикладу, в табл.5.1 - приводяться ваги узагальненого фактору і дисперсії залишків для березня і липня.

Таблиця 5.1 - Ваги узагальненого фактору температурно-вологісного режиму і дисперсії залишків.

Пункт	березень				липень			
	Ваги			d_i^2	Ваги			d_i^2
	p_1	p_2	p_3		p_1	p_2	p_3	
Львів	0,49	-0,31	0,71	0,45	0,47	-0,62	0,50	0,48
Чернівці	0,94	-0,28	0,38	0,39	0,80	-0,65	0,61	0,33
Житомир	0,79	-0,26	0,39	0,41	0,96	-0,57	0,53	0,22
Київ	0,70	-0,43	0,67	0,38	0,94	-0,69	0,70	0,16
Умань	0,58	-0,42	0,66	0,45	0,78	-0,56	0,64	0,26
Кіровоград	0,98	-0,40	0,47	0,28	0,98	-0,62	0,58	0,07
Миколаїв	0,79	-0,37	0,50	0,39	0,91	-0,67	0,54	0,12
Херсон	0,01	-0,73	0,06	0,49	0,84	-0,64	0,69	0,08
Одеса	0,91	-0,79	0,45	0,34	0,94	-0,54	0,64	0,18
Симферо-піль	0,96	-0,46	0,38	0,33	0,87	-0,55	0,48	0,30
Полтава	-0,15	-0,93	-0,09	0,20	0,98	-0,66	0,64	0,05
Харків	-0,92	-0,10	-0,30	0,24	0,99	-0,66	0,60	0,05
Дніпропет-ровськ	-0,96	-0,38	-0,25	0,14	0,91	-0,65	0,66	0,11
Луганськ	-0,19	-0,96	-0,22	0,14	0,99	-0,63	0,62	0,05

З Табл. 5.1 випливає, що в березні при одному і тому ж узагальненому факторі (наприклад $f > 0$) спостерігаються додатні аномалії (відхилення від середнього значення) температури ($x_1 > 0$) в правобережній і від'ємні аномалії температури ($x_1 < 0$) в лівобережній Україні, які супроводжуються від'ємними аномаліями відносної вологості повітря ($x_2 < 0$) і додатніми сумами опадів в правобережній ($x_3 > 0$) і від'ємними аномаліями в лівобережній Україні.

Отже, в березні температурно-вологісний режим лівобережної і правобережної частин України суттєво розрізняються. Це можна пояснити тим, що в цей час східна частина України більш часто знаходиться під дією антициклонічного баричного поля, що приводить до більш

низького температурного фону і меншої кількості опадів, ніж в західній частині країни.

В липні циркуляційний режим території України однорідний. Наслідком цього і є однорідний температурно-вологісний режим території, що відбивається на однорідному розподілу вагів узагального фактору в температурі, вологості і опадів.

Звертає на себе увагу той факт, що крім західних областей (Львів, Чернівці) дисперсії залишків в липні суттєво менші, ніж в березні, що свідчить про те, що мінливість метеорологічних величин, що розглядаються, влітку на більшості території України значно менша, ніж в березні.

Наведемо ще один приклад використання факторного аналізу в метеорологічних дослідженнях. Йдеться про дослідження вертикальної статистичної структури вітру в тропосфері та стратосфері (шар 5 – 55 км). Вихідними даними є вертикальні профілі зональної складової швидкості вітру, які задавалися у виді одинадцятивимірного вектора зональної швидкості вітру. Значення швидкості вітру на кожному рівні через 5 км висоти отримані за допомогою радіо- і ракетного зондування атмосфери в трьох пунктах західної півкулі: Антігуа ($\varphi = 17.2^0 N; \lambda = 61.8^0 W$), Канаверал ($\varphi = 28.5^0 N; \lambda = 80.5^0 W$) і о. Валлоп ($\varphi = 37.8^0 N; \lambda = 75.5^0 W$). Кількість векторів складала: в першому пункті - 539, в другому і третьому пунктах - 528. На основі цих даних для зазначених районів були розраховані одинадцятивимірні вибіркові матриці коваріацій $\hat{K}$, на основі яких розраховувалися за ітераційною процедурою ваги узагальнених факторів зональних компонент швидкості вітру, які, як було показано в попередніх розділах, є однаковими для зонального вітру на всіх одинадцяти рівнях шару 5 - 55 км. В пунктах Канаверал і о. Уоллоп інформацію про вертикальну структуру зонального вітру вичерпують два узагальнені фактори, а в пункті Антігуа - три. Ваги узагальнених факторів містяться в табл. 5.2.

Таблиця 5.2 - Ваги узагальненого фактора зональної компоненти швидкості вітру в тропосфері і стратосфері західної півкулі.

№ п\п	Z, км	Антігуа			Канаверал		о.Уоллоп	
		P_1	P_2	P_3	P_1	P_2	P_1	P_2
1	5	0,26	-0,78	0,14	0,17	- 0,87	0,25	- 0,90
2	10	0,06	-0,91	0,17	0,28	- 0,92	0,20	- 0,91
3	15	0,01	-0,91	0,20	0,38	- 0,87	0,37	- 0,85
4	20	0,23	-0,62	0,51	0,60	- 0,67	0,61	- 0,69
5	25	0,21	-0,21	0,86	0,80	- 0,38	0,87	- 0,31
6	30	0,23	-0,20	0,89	0,83	- 0,44	0,82	- 0,39
7	35	0,49	-0,30	0,63	0,88	- 0,38	0,88	- 0,35
8	40	0,80	-0,16	0,28	0,94	- 0,25	0,92	- 0,30
9	45	0,85	-0,22	0,26	0,90	- 0,30	0,93	- 0,26
10	50	0,88	-0,27	0,17	0,90	- 0,25	0,92	- 0,25
11	55	0,77	-0,02	0,11	0,88	- 0,19	0,88	- 0,24

Аналіз даних табл. 5.2 показує, що, по-перше, загальна структура вагів узагальнених факторів в зональній складовій швидкості вітру залежить від широти. В тропічній зоні вона суттєво відрізняється від того, як розподіляються ваги за висотами в помірних широтах. Це визначається тим, що значно розрізнюються циркуляційні механізми в тропічній зоні й помірних широтах. По-друге, незважаючи на те, що пункти Канаверал і о.Уоллоп розташовуються на значній відстані один від одного, однорідність циркуляційних механізмів, а саме переважаючий західно-східний перенос повітряних мас в тропосфері і в стратосфері взимку і східно-західний перенос в стратосфері влітку, приводить до однорідної структури вагів першого і другого узагальнених факторів в цих пунктах. В тропосфері (5 - 15 км), великі від'ємні ваги приходяться на другий фактор, в стратосфері (25 - 55 км), великі додатні ваги -

на перший фактор. Між зазначеними шарами атмосфери розташовується шар, який є перехідним між тропосферою і шаром 25-55 км. Як відомо, цей шар нижньої стратосфери від 15 до 25 км характеризуються послабленою інтенсивністю як зонального переносу, так і внутрішньорічних і довгоперіодних періодичностей швидкості вітру.

В тропічній зоні шар атмосфери 5 - 55 км по однорідних умовах циркуляції треба розділити на три шари: шар 5 - 15 км, де спостерігаються великі від'ємні ваги другого узагальненого фактору (доречі, як і в середніх широтах), шар 25 - 35 км, що характеризується великими вагами третього узагальненого фактору і шар 40 - 55 км, у котрого великі додатні ваги припадають на перший узагальнений фактор. Шар атмосфери 15 - 25 км, як і в помірній зоні, є перехідним.

Наведені приклади показують, що факторний аналіз є достатньо продуктивним методом статистичного аналізу метеорологічних об'єктів (метеорологічних полів і вертикальних профілів метеорологічних величин), який дозволяє отримати важливу інформацію про фізичні особливості атмосферних явищ і процесів.

6. КЛАСИФІКАЦІЯ І КЛАСТЕР

6.1 Класифікація на основі мінімуму функції відстані

Розглянемо вектори, компоненти яких характеризують ті чи інші параметри фізичного стану атмосфери або ту чи іншу ситуацію, що склалася в атмосфері. Будемо називати ці вектори векторами ситуації або образами. Припустимо також, що по деяких властивостях вектори ситуацій можна об'єднати у класи. Класи векторів, що мають деяку схожість називають кластерами. Серед векторів ситуацій кластера є такий образ, котрий є репрезентативним для усього кластера, тобто має деякі риси, притаманні всім образам цього кластера. Цей вектор (образ) називають еталоном або центром кластера. В деяких випадках кластер може мати декілька еталонів. Задача класифікації складається, по-суті, з двох задач: а) побудови правила, за яким можна визначити еталон і, таким чином, сформувати кластер; б) визначити міру близькості некластерізованого вектора до того чи іншого кластера.

Природною мірою близькості є відстань між некластерізованим вектором і кластером. За цією мірою, очевидно, вектор X_1 (рис.6.1) треба віднести до кластера V_i ($X_1 \in V_i$). Видимих підстав для віднесення вектора X_2 до кластерів V_i чи V_j нема, але це можна зробити, якщо розглянути відстані між вектором X_2 і центрами кластерів, які означаються жирною крапкою на рис. 6.1. Будемо відносити некластерізований вектор до того кластера, деяка функція відстані від котрого до центра цього кластера є найменшою порівняно з аналогічною функцією відстані до центрів інших кластерів. Це означає, що кластерізація проводиться на основі мінімума функції відстані. Розглянемо два випадки:

а) Випадок єдиного еталону в кластері.

Будемо вважати, що в кластері тільки один еталон, або центр кластера. Знайдемо відстань між вектором X і центром Z_i i-того кластера U_i. У якості відстані будемо розглядати евклідову відстань $Д_i$

$$Д_i = |X - Z_i| = \sqrt{(X - Z_i)'(X - Z_i)}, \qquad (6.1.1)$$

де вертикальними дужками позначається норма вектора.

Очевидно, коли

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_m \end{pmatrix} \qquad (6.1.2)$$

і

$$Z_i = \begin{pmatrix} z_{1i} \\ z_{2i} \\ \dots \\ z_{mi} \end{pmatrix}, \qquad (6.1.3)$$

то

$$Д_i = \sqrt{\begin{matrix}(x_1 - z_{1i})^2 + (x_2 - z_{2i})^2 + ... + \\ + (x_m - z_{mi})^2\end{matrix}}. \qquad (6.1.4)$$

Припустимо, що ми маємо M кластерів $U_i (i = 1,2,...,M)$, кожний з яких визначається єдиним еталоном $Z_1, Z_2, ..., Z_i, ..., Z_j, ..., Z_M$. Тоді, очевидно, $X \in U_i$, якщо $Д_i < Д_j$, $\forall j = \overline{1, M}$, $j \neq i$.

Можна розглянути й іншу функцію відстані, а саме:

$$Д_i^2 = |X - Z_i|^2 = (X - Z_i)'(X - Z_i). \qquad (6.1.5)$$

В матричному співвідношенні (6.1.5) виконаємо операцію множення.

$$Д_i^2 = X'X - Z_i'X - X'Z_i + Z_i'Z_i. \qquad (6.1.6)$$

Оскільки $Z_i'X = X'Z_i$, маємо

$$Д_i^2 = X'X + Z_i'Z_i - X'Z_i, \qquad (6.1.7)$$

або

$$Д_i^2 = X^{'}X - 2(X^{'}Z_i - \frac{1}{2}Z_i^{'}Z_i). \qquad (6.1.8)$$

Впровадимо позначення

$$d_i\left(X\right) = X^{'}Z_i - \frac{1}{2}Z_i^{'}Z_i. \qquad (6.1.9)$$

Із формул (6.1.8) і (6.1.9) випливає, що функція $Д_i^2$ буде мінімальною, якщо функція $d_i\left(X\right)$ є максимальною. Такий же висновок відноситься і до евклідової відстані $Д_i$. Отже, розв'язувальне правило має такий вид:

$$X \in U_i, \text{ якщо } d_i\left(X\right) > d_j\left(X\right), \forall j = \overline{1,M} \;\; i \neq j. \qquad (6.1.10)$$

Легко бачити, що функція відстані $d_{i(X)}$ є лінійною формою відносно координат вектора X з коефіцієнтами - координатами відомого вектора Z_i тобто

$$d_i\left(X\right) = x_1 z_{1i} + x_2 z_{2i} + ... + x_m z_{mi} - \frac{1}{2}\sum_{s=1}^{m} z_{si}^2 \qquad (6.1.11)$$

б) Випадок множини еталонів у кластері.

Нехай маємо M кластерів, але кожний з них характеризується деяким числом еталонів, тобто для i - того кластера U_i маємо еталони $z_i^{(1)}, z_i^{(2)},..., z_i^{(p)}$. Тоді знайдемо евклідову відстань між вектором X і кожним з центрів i - того кластера. Виберемо з них ту, яка є найменшою. Нехай це буде відстань до l - того центру $z_i^{(l)}$, тобто

$$Д_i = \min_l \left| X - z_i^{(K)} \right| \ \left(K = 1,2,...,l...,p\right). \tag{6.1.12}$$

Те ж саме зробимо для останніх кластерів. Тоді розв'язувальне правило є

$$X \in U_i \text{, якщо } Д_i < Д_j \ \ \forall j = \overline{1,M}\left(j \neq i\right) \tag{6.1.13}$$

Це розв'язувальне правило можна записати й для функції відстані

$$d_i\left(X\right) = \max_l \{X^{'} z_i^{(K)} - \frac{1}{2}\left(z_i^K\right)^{'} z_i^{(K)}\},$$
$$\left(K = 1,2,...,l,..., p\right). \tag{6.1.14}$$

Воно має вид:

$$X \in U_i, \text{якщо} d_i\left(X\right) > d_j\left(X\right) \ \ \forall j = \overline{1,M}\left(j \neq i\right) \tag{6.1.15}$$

Критерій мінімуму евклідової відстанні (максимуму функції відстанні) знаходить найбільш часте застосування в алгоритмах класифікації. Але треба мати на увазі, що установлення критерію близькості, за допомогою якого ми можемо віднести некластерізований вектор X до одного з уже відомих кластерів, ще не вирішує задачу кластерізації. Необхідно, крім того, установити правило або систему правил, які у системі даних дозволять ідентифікувати кожний з центрів кластерів і, таким чином, сформувати кластери. Такі правила складають міри схожості образів і критерії кластерізації. Розрізняють метричні і неметричні міри схожості. Однією з метричних мір схожості може розглядатися вже знайома нам метрика - евклідова відстань $Д = |X - Z|$. Іншою мірою схожості є число Махаланобіса

$$\Delta^2 = (X - \mu)' K^{-1} (X - \mu), \qquad (6.1.16)$$

де μ і K - відповідно вектор математичних сподівань і матриця коваріацій.

Звичайно, останню міру схожості можна використовувати тільки тоді, коли нам відомі ймовірності характеристики образів (вектори математичних сподівань і матриця коваріацій) для кожного з кластерів. Вона є дуже зручною для вирішення задачі віднесення некластерізованного вектора до того чи іншого кластера, але не може використовуватися для формування кластерів, якщо апріорі невідомі зазначені ймовірностні характеристики.

Окрім метричних використовуються й неметричні міри схожості. Однією з них може бути міра, що визначається таким співвідношенням:

$$S(X,Z)=\frac{X'Z}{|X|\cdot|Z|} \qquad (6.1.17)$$

У скалярній формі рівність (6.1.17) має вид:

$$S(X,Z)= =\frac{x_1z_1+x_2z_2+\ldots+x_mz_m}{\sqrt{x_1^2+x_2^2+\ldots+x_m^2}\sqrt{z_1^2+z_2^2+\ldots+z_m^2}} \qquad (6.1.18)$$

Вираз (6.1.18), очевидно, має смисл косинуса кута між векторами X і Z, або, як було показано в розділі 1, коефіцієнта кореляції між випадковими величинами X і Z, що визначені відповідними векторами

$$S(X,Z)=\cos\left(\hat{X,Z}\right)=r_{xz} \qquad (6.1.19)$$

Звичайно, ця міра ніякого відношення до метричних критеріїв не має. Існують й інші непараметричні міри схожості.

Вибір міри схожості ще не вирішує проблему кластерізації, оскільки треба ще визначити критерій, за яким відбувається розбиття простору даних на кластери, тобто необхідно ще визначити критерій кластерізації. Тут можливо використовувати два підходи. Перший з них є евристичним. В основу його покладається попит й інтуїція. Інтуїтивно зрозуміло, що критерієм близькості образів у кластері може бути евклідова відстань. Але цей критерій є відносним. Його треба

доповнювати порогом - деякою величиною евклідової відстанні, починаючи з якої треба образ відносити до того чи іншого кластера. Іншим підходом є підхід, що має в основі деякий критерій якості, який в залежності від характеру задачі підлягає мінімізації або максимізації. Прикладом такого критерія є сума квадратів відхилів векторів від деякого середнього вектора кластера

$$\Delta = \sum_{j=1}^{N_c} \sum_{X \in S_i} \left| X - m_i \right|^2 \qquad (6.1.20)$$

де

$$m_i = \frac{1}{N_i} \sum_{X \in S_i} X_i \qquad (6.1.21)$$

середній вектор i - того кластера, N_c - кількість кластерів, S_i - простір векторів, що відносяться до i - того кластера, N_i - кількість векторів i - того кластера.

6.2 Алгоритми кластерізації

6.2.1 Алгоритм максимальної відстані

Як випливає з назви алгоритма, він полягає у визначенні максимальної відстані із всіх мінімальних. Він складається з системи послідовних кроків. Для ілюстрації складемо дві

таблиці - таблицю описань образів X і таблицю центрів кластерів Z (рис.6.2)

1-й крок. Один з векторів довільно визначають центром першого кластера Z_1 (нехай це буде вектор $X_1 : X_1 \to Z_1$).

2-й крок. Визначаються відстані між всіма векторами і центром першого кластера

$$Д(X_i, Z_1),\ i = 2,3,... \tag{6.2.1}$$

3-й крок. Знаходимо максимальну із цих відстаней

$$\max Д(X_i, Z_1),\ i = 2,3,... \tag{6.2.2}$$

нехай це буде, наприклад,

$$\max Д(X_i, Z_1) = Д(X_6, Z_1) \tag{6.2.3}$$

Це дає підставу визначити у якості центра другого кластера вектор $X_6 : (X_6 \to Z_2)$.

4-й крок. Визначаються відстані всіх векторів до центрів двох кластерів Z_1 і Z_2

$$Д(X_i, Z_1),\ (i = 1,2,..., N) \tag{6.2.4}$$

$$Д(X_j, Z_1),\ (j = 1,2,..., N) \tag{6.2.5}$$

5-й крок. Знаходять для всіх відстаней мінімальні для кожної групи

$$\min Д(X_i, Z_1), \tag{6.2.7}$$

$$\min Д(X_j, Z_1), \tag{6.2.8}$$

6-й крок. Із всіх цих мінімальних відстаней визначають максимальну

$$\max \min Д(X_l, Z_m), \quad l = 1,2,..., N; m = 1,2 \tag{6.2.9}$$

Нехай це буде відстань, що відноситься до вектора X_l.

7-й крок. Ця максимальна відстань порівнюється з відстанню між центрами кластерів $Д(Z_1, Z_2)$ і визначається поріг для кластерізації. Якщо

$$\max \min Д(X_l, Z_m) \geq \frac{1}{2} Д(Z_1, Z_2), \tag{6.2.10}$$

то вектор X_l визначається у якості центра третього кластера $Z_3 : (X_l \to Z_3)$

8-й крок. Знаходяться відстані

$$Д(X_i, Z_1); Д(X_j, Z_2); Д(X_k, Z_3)$$
$$i, j, k = \overline{1, N}. \qquad (6.2.11)$$

9-й крок. Із кожної групи цих відстаней знаходяться мінімальні

$$\min Д(X_i, Z_1); \qquad (6.2.12)$$

$$\min Д(X_j, Z_2); \qquad (6.2.13)$$

$$\min Д(X_k, Z_3). \qquad (6.2.14)$$

10-й крок. Із цих мінімальних відстаней знаходять максимальну

$$\max \min Д(X_s, Z_m) \left(s = \overline{1, N}; m = 1,2,3\right) \qquad (6.2.15)$$

11-й крок. Цю відстань $Д(X_s, Z_m)$ порівнюють з відстанями між центрами кластерів

$$Д(Z_i, Z_j) \ (i, j = 1,2,3; i < j), \qquad (6.2.16)$$

що дає можливість визначити пороги. Якщо

$$Д(X_s, Z_m) \geq \frac{1}{2} Д(Z_i, Z_j), \qquad (6.2.17)$$

то вектор X_s визначається центром четвертого кластера Z_4. Якщо пороги не задовольняються, то робота алгоритму припиняється і процес кластерізації переходить до передостаннього кроку.

12-й крок. Вектори, що залишилися після визначення центрів кластерів, розподіляються по кластерах, які вже мають свої центри за розв'язувальним правилом:

$$X \in S_j, \text{ якщо } Д(X_i, Z_j) \leq Д(X_i, Z_k) \qquad (6.2.18)$$

13 - й крок. Уточнюються центри кластерів. У якості центрів кожного кластера $\widetilde{Z}_j$ встановлюється середній по кластеру вектор

$$\widetilde{Z}_j = \frac{1}{N_j} \sum_{X \in S_j} X; \qquad (6.2.19)$$

На цьому процес кластерізації завершується.

6.2.2 Алгоритм ИСОМАД (ітераційний самоорганізуючий метод аналізу даних)

Алгоритм ИСОМАД - є алгоритмом евристичного характеру, в основу якого покладені евклідові відстанні, які використовуються з визначеними порогами.

У якості вихідної інформації виступає матриця X, що утримує вектори (образи), які підлягають кластерізації. Нехай ці образи мають мірність M, а кількість їх N. Тоді матриця має такий порядок:

$$X = \left(x_{ij}\right)_{N\times M} \tag{6.2.20}$$

Крім того, формується матриця гіпотетичних центрів кластерів

$$Z = \left(z_{ij}\right)_{N_c\times M} \tag{6.2.21}$$

де N_c - кількість гіпотетичних кластерів. Кількість кластерів N_c визначається дослідником довільно, використовуючи попередні знання про суть процесів, що досліджуються. Тому ці кластери й називають гіпотетичними. На основі такої інформації довільно визначаються й гіпотетичні центри цих кластерів векторів, що складають матрицю X.

Крім того, визначаються: IT - кількість ітерацій, яку при роботі алгоритму не хотілось би перебільшувати, параметр Q - мінімальне число векторів, які можуть складати кластер. Часто вважають, що $Q = 2$.

До вихідної інформації відноситься ще і параметр Q_s, котрий характеризує загальну компактність всіх ознак x_i. Він

розраховується таким чином. По-перше, знаходяться по сукупностях i - тих компонент всіх образів їх середнє значення

$$\overline{x}_i = \frac{1}{N}\sum_{j=1}^{N} x_{ij}\,, \qquad (6.2.22)$$

По-друге, розраховують дисперсію цих компонент

$$\sigma_i^2 = \frac{1}{N-1}\sum_{j=1}^{N}\left(x_{ij} - \overline{x}_i\right)^2, \left(i = \overline{1, M}\right) \qquad (6.2.23)$$

Нарешті, знаходять середню дисперсію для всіх ознак

$$\overline{\sigma}^2 = \frac{1}{M}\sum_{i=1}^{M}\sigma_i^2\,. \qquad (6.2.24)$$

Параметр $Q_s = \sqrt{\overline{\sigma^2}}$. (6.2.25)

Друга частина алгоритму складається з ітераційного процесу.

1-й крок. Розрахунок евклідових відстаней між кожним j -тим вектором матриці X і центром i - того кластера

$$Д_{ij} = \sqrt{\sum_{k=1}^{M}\left(x_{jk} - z_{ik}\right)^2} \quad \left(j = \overline{1, N}; i = \overline{1, N_c}\right)$$

(6.2.26)

(N_c кількість гіпотетичних кластерів). В результаті утворюється матриця

$$Д = \left(Д_{ij}\right)_{N_c \times N} \qquad (6.2.27)$$

(N - кількість векторів, що утримуються в матриці X).

2-й крок. Проводиться рознесення векторів матриці X по кластерах. Для цього використовується вирішальне правило

$$X_j \in S_i, \text{ якщо } Д_{ij} \le Д_{lj} \quad i, l = \overline{1, N_c}, j = \overline{1, N}$$

(6.2.28)

де S_i - i - тий кластер, центр якого Z_i.

3-й крок. Підраховується кількість векторів N_i, що потрапили в кожний з кластерів ($i = \overline{1, N_c}$). Отримане число векторів порівнюється з параметром Q. Якщо $N_i < Q$, то i - тий кластер ліквідується і число N_c зменьшується на одиницю. Якщо $N_c = 1$, то процес кластерізації завершується. Це означає, що всі вектори потрапили до одного кластеру.

4-й крок. Проводиться корекція центрів кластерів. У якості нових центрів кластерів беруться вектори, координати

яких є середні значення відповідних координат всіх векторів, що складають кожний кластер, тобто

$$\widetilde{z}_{ik} = \frac{1}{N_i}\sum_{j=1}^{N_i} x_{ik}\,,\ \left(k=\overline{1,M};i=\overline{1,N_c}\right). \qquad (6.2.29)$$

5-й крок. Знаходяться евклідові відстані від кожного вектора, що відносяться до кожного i - того кластера, до центра свого кластера

$$\widetilde{Д}_{ij} = \sqrt{\sum_{j=1}^{M_i}(x_{jk}-\widetilde{z}_{ik})^2}\,,\ \left(j=\overline{1,N_i};i=\overline{1,N_c}\right) \qquad (6.2.30)$$

де $\widetilde{z}_{ik}$ - відповідна координата відкоректованого центра i - того кластера.

6-й крок. Розраховується середня з цих відстаней

$$\overline{Д}_i = \frac{1}{N_c}\sum_{j=1}^{N_i}\widetilde{Д}_{ij}\,, \qquad (6.2.31)$$

7-й крок. Визначається середня зважена відстань по всіх кластерах

$$\overline{Д} = \frac{1}{N}\sum_{i=1}^{N_c} N_i\overline{Д}_i \qquad (6.2.32)$$

8-й крок. Розраховується вектор середніх квадратичних відхилів, координати якого є середні квадратичні відхили відповідних координат векторів визначного кластера від відповідних координат його центра, тобто

$$\sigma_{ik} = \sqrt{\frac{\sum_{j=1}^{N_i}(x_{jk} - \bar{z}_{ik})^2}{N_i}}, \ \left(i = \overline{1, N_c}\right) \qquad (6.2.33)$$

В результаті отримаємо для кожного кластера вектор-рядок

$$\sigma_i = (\sigma_{ik})_{1\times M} \qquad (6.2.34)$$

9-й крок. Для кожного вектора $\sigma_i \ \left(i = \overline{1, N_c}\right)$ визначається $\max \sigma_{ik} \ \left(i = \overline{1, N_c}\right)$, які порівнюються з Q_s. Якщо $\max \sigma_{ik} \geq Q_s$ і $\overline{Д}_i \geq \overline{Д}$ і $N_i \geq 2(Q+1)$, то i -тий кластер розбивається на два нові кластера, а i -тий кластер ліквідується. Параметр N_c збільшується на одиницю. Центри нових кластерів визначаються у виді векторів, усі координати яких дорівнюють координатам i - того ліквідованого кластера, окрім k - тої координати, якій відповідає $\max \sigma_{ik}$. Для цих k - тих координат центрів нових кластерів, які позначимо $Z_i^{(+)}$ і $Z_i^{(-)}$, впроваджуються формули

$$z_{ik}^{(+)} = \bar{z}_{ik} + 0.5 \max \sigma_{ik} \qquad (6.2.35)$$

$$z_{ik}^{(-)} = \bar{z}_{ik} - 0.5 \max \sigma_{ik} \qquad (6.2.36)$$

Після цього процес кластерізації починається спочатку, тобто повертаємося до 1-го кроку.

10-й крок. Якщо $\overline{Д}_i < \overline{Д}$, то це означає, що кластери розташовуються недалеко один від одного. Тоді знаходять евклідові відстані

$$\tilde{\tilde{Д}}_{ij} = \sqrt{\sum_{k=1}^{M} \left(\bar{z}_{ik} - \bar{z}_{jk} \right)^2} \quad (i, j = \overline{1, N_c}) \qquad (6.2.37)$$

$\tilde{\tilde{Д}}_{ij}$ порівнюється з порогом Q_s. Якщо для деяких кластерів $\tilde{\tilde{Д}}_{ij} < Q_s$, то це означає, що ці кластери розташовуються близько один до одного, і є підстави для їх злиття.

Після злиття цих кластерів розраховуються координати нового центра кластера як середнє зважене із координат центрів кластерів, що зливаються.

Якщо $\tilde{\tilde{Д}}_{ij} > Q_s$, то процедура завершується.

6.3 Приклади кластерного аналізу метеорологічної інформації

Алгоритм, що розглядався у попередньому розділі, був застосований при розв'язуванні задачі виділення на території Українських Карпат районів з аналогічною структурою місячних кількостей опадів. У якості вихідної інформації виступали дані про значення цієї кліматичної характеристики на 60 метеорологічних станціях і постах за період з 1881 до 1980 рр. Для кожної метеорологічної станції розраховувалися середні значення місячних кількостей опадів, і об'єктами кластерізації, таким чином, становили 60 дванадцативимірних векторів.

Згідно з алгоритмом ІСОМАД, окрім матриці вихідних даних для реалізації ітераційної процедури треба визначити апріорну інформацію: гіпотетичне число кластерів N_s і найменше число векторів Q, що складають кластер. Приймалось $N_s = 2$ і $Q = 5$. Крім того визначалися гіпотетичні центри кластерів. У їх якості приймалися вектори середніх значень місячних кількостей опадів на метеорологічних станціях Івано-Франківськ і Міжгор'є.

В результаті реалізації алгоритма множина з 60 векторів середньомісячних кількостей опадів розділилася на 4 кластери, що об'єднують райони Українських Карпат з подібним режимом зволоження. Ці райони зображені на рис.6.3.

Як видно, цими районами є: Предкарпаття (кластер 1), центральна частина Карпат, що включає найбільш високі хребти (кластер 2), південно-західні схили Карпат (кластер 3), та Закарпаття (кластер 4). Перший кластер складають 17, другий 23, третій 11 і четвертий 8 векторів, координати яких є середньомісячними кількостями опадів на відповідних метеорологічних станціях і постах.

Формування кластерів за подібною структурою середньомісячних кількостей опадів, тобто за подібним режимом зволоження, привело одночасно до виділення однотипних ландшафтних зон цієї гірської системи. У цьому можна легко удосконалитися, розглядаючи рис. 6.4. На ньому зображені полігони повторюваностей абсолютних висот метеорологічних станцій і постів, що увійшли в кожний із районів, відповідаючих визначеному кластеру.

З рис. 6.4 випливає, що в першому районі (Предкарпаття) найбільша повторюваність висот станцій 67% припадає на діапазон висот 200 - 400 м, у четвертому районі (Закарпаття) – 88% станцій мають абсолютні висоти від 100 до 200 м, у той час коли у другому районі (центральна частина Карпат) більш 60% станцій розташовуються на висотах від 500 до 1000 м і більше. У третьому районі (південно-західні схили Карпат) найбільше число станцій розташовується на висоті 200 – 400 м, тобто як і в першому районі. Таке співпадіння районів з подібним режимом опадів, з одного боку, і подібним ландшафтом, з іншого, є відбиттям впливу рельєфу на опадоутворюючі процеси і підтверджує об'єктивність районування території методами кластерного аналізу.

В табл.6.1 наводяться координати векторів-центрів відповідних кластерів, що мають сенс осереднених по кожному району середіх значень місячних кількостей опадів. В їх величинах виявляються розбіжності між районами у режимах зволоження. Із табл. 6.1 випливає, що центральні Карпати (кластер 2) і південно-західні схили Карпат (кластер 3) характеризуються порівняно з іншими районами більшими кількостями опадів у всіх місяцях року. Якщо порівняти їх між собою, то видно, що у більшості випадків в центральних Карпатах спостерігається більша кількість опадів, ніж на південно-західних схилах. Це особисто притаманне у літні місяці з травня до серпня, тобто коли суттєву роль у формуванні опадів відіграє теплова конвекція.

Таблиця 6.1 - Координати векторів - центрів кластерів (мм).

№ клас-терів	М і с я ц і р о к у							
	1	2	3	4	5	6	7	8

1	40,1	39,4	40,6	54,1	76,6	104,0	104,0	89,8
2	67,3	69,5	73,8	81,2	110,7	143,9	143,7	120,4
3	68,1	67,9	64,3	68,9	84,9	109,9	109,9	98,4
4	57,0	52,7	49,1	53,6	60,1	96,1	86,0	75,7

Продовження табл.6.1

Місяці року			
9	10	11	12
65,2	56,9	51,3	48,8
99,1	99,3	96,8	89,9
84,1	84,1	87,6	87,4
63,6	68,5	72,7	

У період з грудня до лютого, коли основна маса опадів в Карпатах обумовлюється перевалюванням циклонів з південного заходу тобто з Угорської низини, кількість опадів на південно-західних схилах Карпат і в центральних Карпатах є приблизно однаковою.

Виявляє інтерес той факт, що в холодну пору року (з жовтня до березня) в Закарпатті випадає більше опадів, ніж в Предкарпатті, у той час коли з квітня до вересня середні по району місячні кількості опадів в Предкарпатті виявляються суттєво більшими. Це пояснюється як більш розвинутою конвекцією над більш перерізаною місцевістю, так і проходженням холодних фронтів з північного заходу уздовж східних схилів Карпат, які при перевалюванні через центральні Карпати розмиваються і над Закарпаттям стають слабко вираженими.

Розглянемо другий приклад, який відноситься до задачі класифікації екологічної інформації.

Для контролю якості атмосферного повітря у великих промислових містах установлюється мережа контрольно-вимірювальних пунктів, на яких виконуються вимірювання концентрацій шкідливих домішок у повітрі (SO_2, CO, NO_2 і т. д.). Розроблені також і автоматизовані системи контролю. Але після вимірювань виникає потреба проводити оцінку

вірогідності вимірюваних концентрацій інгредієнтів, оскільки під впливом випадкових джерел (наприклад, автомобіль, що припаркувався біля пункту вимірювань) концентрація того чи іншого інгредієнта може не відповідати промислово-транспортній ситуації і стану пограничного шару атмосфери.

Стан пограничного шару атмосфери, який чинить великий вплив на розсіювання шкідливих домішок, можна охарактеризувати сукупність метеорологічних величин, які складають вектори ситуацій. До нього треба включити й концентрації, що склалися у попередній термін, оскільки вони складають визначний фон забруднення. Перелік метеорологічних характеристик, які є компонентами вектора ситуацій, визначаться за допомогою метода "просіювання" предікторів. Ці методи докладно розглядаються в розділі 4.

Застосування цих методів на вибірках даних метеорологічних і аерологічних вимірювань, а також концентрації сірчаного ангідриду SO_2 в Москві дало можливість встановити таку систему статистично значущих предікторів:

1. Значення концентрації SO_2 за попередню добу, що отримані шляхом ковзного осереднення часових послідовностей концентрації (методи ковзного осереднення часових рядів розглядаються в першій частині підручника).
2. Значення концентрації SO_2 за дві години до терміну контролю.
3. Вертикальний градієнт температури повітря у нижньому кілометровому шарі атмосфери.
4. Зональна складова швидкості вітру на рівні 100 м.
5. Модуль швидкості вітру на висоті 100 м.
6. Меридіональна складова швидкості вітру біля земної поверхні.
7. Меридіональна складова швидкості вітру на висоті 200 м.
8. Висота нижньої границі припіднятої інверсії температури повітря.

Таким чином, вектори ситуацій складалися із 8 компонент. При проведенні чисельних експериментів було використано 325 таких векторів. Кластерізація векторів ситуацій виконувалося за допомогою алгоритму ICOMAД.

Для реалізації алгоритму, як зазначалося вище, необхідні такі вихідні дані - матриця векторів, що підлягають кластерізації $X = \left(x_{ij}\right)_{MN}$, M - кількість компонентів у векторі ситуацій (кількість рядків матриці); N - кількість векторів (кількість стовпців матриці); IT - кількість ітерацій; $Z = \left(z_{ij}\right)_{N_c M}$ - матриця гіпотетичних центрів кластерів; N_c - кількість кластерів; Q - мінімальне число вибіркових векторів, які можуть складати кластер. В задачі, що розглядається, було встановлено два гіпотетичних кластерів $\left(N_c = 2\right)$, центри яких визначені довільним чином з вихідної сукупності векторів ситуацій. Вони знаходяться в табл. 6.2.

Таблиця 6.2 - Гіпотетичні центри кластерів.

Кластер	Предіктори							
	1	2	3	4	5	6	7	8
z_1	0,55	0,25	0,4	10,4	9,0	0,8	6,0	0,0
z_2	1,38	1,25	-0,6	7,7	6,0	0,9	6,4	111,0

(значення концентрації SO_2 приводяться у відносних одиницях).

Окрім перелічених, необхідно визначити параметр Q_s, який є мірою середньоквадратичної відстані від центра кластера. Для його визначення, як було зазначено у попередньому розділі, спочатку розраховується дисперсія кожного з визначених предікторів, потім загальна дисперсія всіх компонент для всієї множини векторів у припущенні, що компоненти векторів є

незалежними. Корінь квадратний з цієї дисперсій й приймається за параметр Q_s. Після цього виконується ітераційна процедура, яка для алгоритму ІСОМАД описується в розділі 6.2.

В результаті реалізації процедури сукупність векторів ситуацій була розділена на 3 кластери. В перший кластер увійшло 222, у другий - 44, а в третій - 59 векторів ситуацій. Були визначені координати центрів цих кластерів. Вони приводяться в табл. 6.3.

Таблиця 6.3 - Координати центрів виділених кластерів.

Клас-тери	П р е д і к т о р и							
	1	2	3	4	5	6	7	8
1	1,14	1,16	- 0,02	2,7	0,2	1,5	6,2	26,0
2	1,04	1,02	0,24	1,1	0,1	2,4	5,4	95,0
3	1,01	1,00	0,28	0,3	0,7	3,0	5,7	170,0

Із табл. 6.3 випливає, що перший кластер виділяється порівняно з іншими підвищеним вмістом домішок в попередню добу і за 2 години до строку вимірювань, інверсійною стратифікацією в нижньому кілометровому шарі атмосфери, підвищеним значенням зональної компоненти швидкості вітру на висоті 100 м, підвищеним значенням вертикального градієнта меридіональної компоненти швидкості вітру в нижньому 200 метровому шарі повітря, більш низьким розташуванням нижньої границі висотної інверсії температури.

Третій кластер відрізняється від другого, головним чином, суттєво меншим значенням швидкості зонального вітру на висоті 200 м й більш високим розташуванням нижньої границі висотної інверсії температури.

Отже, кластерний аналіз дав можливість із вихідної множини векторів ситуацій сформувати три підмножини, які відповідають суттєво різним станам пограничного шару

атмосфери, з одного боку, і відзначаються внутрішньокластерною схожістю, з другого.

Відповідно до виділених кластерів були сформовані вибірки виміряних концентрацій SO_2 і на їх основі розраховані середні значення і дисперсії концентрацій. Вони розташовуються в табл. 6.4.

Таблиця 6.4 - Середні значення і дисперсії концентрацій.

Параметр	К л а с т е р		
	1	2	3
$\overline{q}_2$	1,24	0,96	0,81
σ_q	0,102	0,062	0,044

Як видно з табл. 6.4, середні значення і дисперсії концентрацій SO_2 при трьох різних типах станів пограничного шару атмосфери розрізняються один від одного. Однак, виникає питання чи суттєвими є ці розбіжності. На це питання дає відповідь перевірка статистичних гіпотез про значущість цих розбіжностей середніх і дисперсій між кластерами. Перевірка гіпотез проводилася на основі критеріїв Фішера F і Стюдента t. Результатами перевірки цих гіпотез містяться в табл. 6.5.

Таблиця 6.5 - Значення критеріїв F і t при = 0,05.

Кластери	К р и т е р і ї			
	F	$F_{кр}$	t	$t_{кр}$
1 - 2	1,64	1,48	6,45	1,96
1 - 3	2,32	1,34	12,32	1,96
2 - 3	1,42	1,60	3,22	1,96

Як виходить з табл. 6.5, середні значення і дисперсії концентрацій SO_2 для сукупностей, що створюються в результаті кластерізації векторів стану пограничного шару атмосфери, розрізняються значуще, за винятком дисперсії для другого й третього кластерів. Але ці кластери значуще розрізняються по середніх значеннях. Отже, ці три сукупності концентрацій SO_2 є вибірками із різних генеральних сукупностей.

Таким чином, кластерний аналіз дав можливість не тільки розділити вихідну множину векторів ситуацій на підмножини, кожна з яких об'єднує схожі у визначній мірі метеорологічні умови в пограничному шарі атмосфери, але й отримати відповідні їм різні по статистичних властивостях сукупності концентрацій інгредієнтів.

Розбіжності в статистичних характеристиках сукупностей концентрацій SO_2 з фізичної точки зору добре узгоджуються із середнім станом пограничного шару атмосфери, що характеризується центрами кластерів.

Підвищення значення середньої концентрації SO_2 при ситуаціях, що належать до першого кластеру, пояснюється найбільш стійкою стратифікацією повітря й низьким положенням висотної інверсії температури, що сприяє накопиченню домішок біля земної поверхні, а підвищене значення зональної компоненти швидкості вітру й вертикального градієнта його меридіональної компоненти обумовлюють найбільшу дисперсію концентрації інгредієнта. Зменшення стійкості повітря, що характерне для кластерів 2 і 3, з одного боку, і суттєве збільшення висоти нижньої границі висотної інверсії температури, з другого боку, сприяє розсіюванню домішок у більшому шарі атмосфери і, таким чином, зменшенню концентрації SO_2 біля земної поверхні. Зменшення швидкості вітру й невеликі її градієнти

обумовлюють наявність порівняно невеликих дисперсій концентрації інгредієнта.

7. МЕТОДИ ПРИЙНЯТТЯ АЛЬТЕРНАТИВНИХ РІШЕНЬ ВІДНОСНО МЕТЕОРОЛОГІЧНИХ ОБ'ЄКТІВ

7.1 Принципи побудови схеми альтернативного прогнозу

Регресійний аналіз дає можливість побудувати прогностичну модель, на основі якої розраховують значення метеорологічної величини, для котрої розроблялася регресійна модель. Інакше кажучи, на виході прогностичної моделі ми маємо значення предіктанта у виді числа. Але в практиці прогнозування часто виникає потреба передбачення метеорологічного явища. Наприклад, гроза може виникати чи не виникати, туман може утворитися чи ні, обледеніння літака може виникнути, а може не виникнути і т.д. Відповідно до цього, необхідно складати прогноз здійснення або нездійснення того чи іншого метеорологічного явища. Такий прогноз називають альтернативним.

Як і у попередніх прогностичних моделях, альтернативний прогноз має у своїй основі систему предікторів - метеорологічних величин, що описують стан атмосфери, сприятливий або несприятливий для розвитку атмосферного явища, що прогнозується.

Позначимо цю систему предікторів за допомогою вектора

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_i \\ \dots \\ x_n \end{pmatrix} \qquad (7.1.11)$$

Сформульована задача є типічною задачею розпізнавання образів. Суть її полягає у тому, що, по-перше, необхідно розділити весь простір образів на два підпростори, у першому з яких явище відбувається, а в другому - ні. По-друге, треба побудувати правило, за допомогою котрого можна віднести образ, який підлягає розпізнаванню, до того чи іншого підпросторів.

Нехай ми маємо множину V векторів-предікторів (образів), що складають простір зображень R_V. Припустимо, що цей простір розділяється на два підпростори R_{V_1} і R_{V_2}. У першому з них розташовується множина V_1 образів X, при яких явище відбувається, а у другому - множина V_2 образів X, коли явище не відбувається. Ясно, що

$$V_1 \cup V_2 = V,\ V_1 \cap V = \emptyset$$

Наперед всього, як зазначалося вище, треба побудувати поверхню, яка б розділяла підпростори R_{V_1} і R_{V_2}. Наведемо для пояснення прості приклади.

Нехай простір буде двовимірним $R_V = R_V(x_1, x_2)$ (рис.7.1). Тоді ми маємо на площині (x_1, x_2) лінію $x_1 = f(x_2)$, що розділяє підпростір R_{V_1} від підпростору R_{V_2}. Досить простим буде і випадок трьохвимірного простору $R_V = R_V(x_1, x_2, x_3)$ (рис.7.2). У цьому випадку підпростори R_{V_1} і R_{V_2} розділяє деяка поверхня у трьохвимірному просторі, рівняння якої має вид $x_3 = \varphi(x_1, x_2)$.

Всі такі випадки можна легко собі уявити. Більш складні умови виникають, коли розглядаються образи із багатовимірного простору $R_V = R_V(x_1, x_2, ..., x_n)$. Поверхня, що розділяє цей простір на підпростори R_{V_1} і R_{V_2} називається розділяючою гіперповерхнею, а її рівняння має вид:

$$F(x_1, x_2, .., x_n) = 0 \qquad (7.1.2)$$

Після цього, як зазначалося, треба отримати правило, за допомогою якого є підстава віднести вектор X, що підлягає розпізнаванню, до підпростору R_{V1} або підпростору R_{V_2}. Це правило називають розв'язувальним правилом. Якщо відповідно до нього приймається рішення, що $X \in R_{V_1}$, то явище прогнозується, якщо приймається рішення, що $X \in R_{V_2}$, то явище не прогнозується. Етап, що складається з побудови розділяючої гіперповерхні та розв'язувального правила, носить назву етапа навчання. Прийняття рішення про

належність вектора X до підпросторів R_{V_1} чи R_{V_2} називають етапом розпізнавання. Множина V векторів-предікторів, на основі якої реалізується перелічені етапи, називається навчаючою сукупністю. Крім неї, створюється ще й перевірочна сукупність, яка використовується для перевірки адекватності моделі альтернативного прогнозу.

Наведене вище приводить до висновку, що задача альтернативного прогнозу є по-суті задачою розпізнавання образів. Розглянемо основні ідеї теорії розпізнавання.

Позначимо через H_1 гіпотезу, що образ $X \in V_1$. Альтернативною буде гіпотеза H_2 про те, що $X \in V_2$. Задача розпізнавання полягає у тому, що треба знайти правило, яке дозволяє обгрунтовано прийняти гіпотезу H_1 або H_2. Всіляка процедура перевірки гіпотез передбачає, що приймаючи те чи інше рішення, ми можемо припустити помилку 1-го чи 2-го роду. Нагадаємо, що помилку 1-го роду ми припускаємо, коли відкидаємо правильну гіпотезу. Помилка 2-го роду пов'язана з прийняттям невірної гіпотези.

Будемо вважати, що відомими є умовні ймовірності класів V_1 і V_2

$$P\left(x_1, x_2, ..., x_n / V_1\right) \tag{7.1.3}$$

і

$$P\left(x_1, x_2, ..., x_n / V_2\right). \tag{7.1.4}$$

Позначимо ймовірність помилки 1-го роду через P_a , а 2-го роду через P_b. Знаючи ймовірності (7.1.3) і (7.1.4), а також

апріорні ймовірності $P(V_1)$ і $P(V_2)$ класів V_1 і V_2, можна розрахувати ймовірності помилок 1-го і 2-го роду. Вони визначаються формулами

$$P_a = P(V_1)\int_{R_{V_2}}\ldots\int P(x_1, x_2, \ldots, x_n / V_1)dx_1 dx_2 \ldots dx_n \tag{7.1.5}$$

$$P_b = P(V_2)\int_{R_{V_1}}\ldots\int P(x_1, x_2, \ldots, x_n / V_2)dx_1 dx_2 \ldots dx_n \tag{7.1.6}$$

Приймаючи ту чи іншу гіпотезу, або те чи інше класифікуюче рішення, ми рискуємо зробити одну із зазначених помилок. Із теорії статистичних рішень випливає, що класифікуюче рішення повинно бути таким, щоб середній риск (середня ціна) прийняття рішення був мінімальним. Середній риск визначається формулою

$$W = \delta_a P_a + \delta_b P_b \tag{7.1.7}$$

де δ_a і δ_b - ціна помилки 1-го і 2-го роду відповідно.

Ціни помилок 1-го і 2-го роду можуть дуже розрізнятися. Для ілюстрації цього наведемо такий приклад. Синоптик, складаючий прогноз погоди, на авіаційній метеорологічній станції (АМСТ) чекає проходження холодного фронту другого роду. Аналіз синоптичних матеріалів дає підставу передбачати, що при його проходженні будуть спостерігатися шквали. Як відомо, шквали можуть мати таку

силу, що приводять до руйнування легкомоторних літаків на стоянках, якщо їх не закріпити. Але міркування синоптика не привели до правильного висновку і він не попередив керівництво про небезпеку, тобто відкинув правільну гіпотезу, що є помилкою 1 роду. Це привело до великих матеріальних втрат. Розглянемо тепер іншу ситуацію. Синоптик безпідставно спрогнозував шквал при проходженні холодного фронту, який не відбувся, тобто він прийняв невірну гіпотезу і здійснив помилку 2-го роду. Наслідком цього була хибна тривога. Доречі, помилку 2-го роду й називають "помилкою хибної тривоги". Ціни цих двох помилок у нашому випадку непорівнянні. Але у деяких задачах є можливість вважати, що помилки 1-го та 2-го роду мають однакові ціни ($\delta_a = \delta_b$).

Маючи на увазі рівності (7.1.7), (7.1.6) і (7.1.5) запишемо рівняння для середнього риску у формі

$$W = \delta_a P(V_1) \cdot$$
$$\cdot \int_{R_{V2}} \ldots \int P(x_1, x_2, \ldots, x_n / V_1) dx_1 dx_2 \ldots dx_n +$$
$$+ \delta_b P(V_2) \int_{R_{V_1}} \ldots \int P(x_1, x_2, \ldots, x_n / V_2) dx_1 dx_2 \ldots dx_n$$

(7.1.8)

Ясно, що

$$\int\ldots\int_{R_{V_1}} P(x_1, x_2, \ldots, x_n / V_1) dx_1 dx_2 \ldots dx_n +$$
$$+ \int\ldots\int_{R_{V_2}} P(x_1, x_2, \ldots, x_n / V_1) dx_1 dx_2 \ldots dx_n = 1$$

(7.1.9)

Звідси

$$\int\ldots\int_{R_{V_2}} P(x_1, x_2, \ldots, x_n / V_1) dx_1 dx_2 \ldots dx_n =$$
$$= 1 - \int\ldots\int_{R_{V_1}} P(x_1, x_2, \ldots, x_n / V_1) dx_1 dx_2 \ldots dx_n$$

(7.1.10)

Підставляючи рівність (7.1.10) у формулу (7.1.8), отримаємо

$$W = \int\ldots\int_{R_{V_1}} [\delta_b P(V_2) P(x_1, x_2, \ldots, x_n / V_2) -$$
$$- \delta_a P(V_1) P(x_1, x_2, \ldots, x_n / V_1)] dx_1 dx_2 \ldots dx_n +$$
$$+ \delta_a P(V_1)$$

(7.1.11)

Із рівняння (7.1.11) випливає, що середній риск може досягати мінімуму, коли члени, що розташовані у квадратних дужках, менші нуля, тобто

$$\delta_b P(V_2) P(x_1, x_2, ..., x_n / V_2) - \\ - \delta_a P(V_1) P(x_1, x_2, ..., x_n / V_1) < 0 \qquad (7.1.12)$$

Цій нерівності є еквівалентною така нерівність:

$$\delta_a P(V_1) P(x_1, x_2, ..., x_n / V_1) > \\ > \delta_b P(V_2) P(x_1, x_2, ..., x_n / V_2) \qquad (7.1.13)$$

Оскільки всі величини, що складають цю нерівність, додатні числа, нерівність (7.1.13) можна переписати таким чином:

$$\frac{P(x_1, x_2, ..., x_n / V_1)}{P(x_1, x_2, ..., x_n / V_2)} > \frac{\delta_b P(V_2)}{\delta_a P(V_1)} \qquad (7.1.14)$$

Ліва частина нерівності (7.1.14) є деякою функцією вектора-предіктора

$$\lambda(x_1, x_2, .., x_n) = \frac{P(x_1, x_2, ..., x_n / V_1)}{P(x_1, x_2, ..., x_n / V_2)} \qquad (7.1.15)$$

Її називають функцією подібності.
Величину

$$\frac{\delta_b P(V_2)}{\delta_a P(V_1)} = \theta \qquad (7.1.16)$$

називають порогом. Таким чином, ми прийшли до такого розв'язувального правила:

$$\begin{aligned} &\text{вектор } X \in V_1 \text{ , якщо } \lambda(x_{1,} x_2, ..., x_n) > \theta \\ &\text{вектор } X \in V_2 \text{ , якщо } \lambda(x_{1,} x_2, ..., x_n) < \theta \end{aligned} \qquad (7.1.17)$$

Якщо є підстави вважати, що $\delta_a = \delta_b$ і $P(V_1) = P(V_2)$, то $\theta = 1$ і розв'язувальне правило має вид:

$$\begin{aligned} &\text{вектор } X \in V_1 \text{ , якщо } \lambda(x_{1,} x_2, ..., x_n) > 1 \\ &\text{вектор } X \in V_1 \text{ , якщо } \lambda(x_{1,} x_2, ..., x_n) < 1 \end{aligned} \qquad (7.1.18)$$

Розв'язувальне правило може використовуватись і в іншому виді. Дійсно, якщо нерівність (7.1.14) є справедливою, то справедливою є й нерівність

$$\ln \frac{P(x_1, x_2, ..., x_n / V_1)}{P(x_1, x_2, ..., x_n / V_2)} > \ln \frac{\delta_b P(V_2)}{\delta_a P(V_1)} \qquad (7.1.19)$$

або

$$\ln P(x_1, x_2, ..., x_n / V_1) - \ln P(x_1, x_2, ..., x_n / V_2) + + \frac{\delta_a P(V_1)}{\delta_b P(V_2)} > 0$$

(7.1.20)

Функцію

$$F(x_1, x_2, ..., x_n) = \ln P(x_1, x_2, ..., x_n / V_1) - - \ln P(x_1, x_2, ..., x_n / V_2) + \frac{\delta_a P(V_1)}{\delta_b P(V_2)}$$

(7.1.21)

називають дискримінантною функцією якщо використовується дискримінантна функція, то розв'язувальне правило придбаває вид:

$$X \in V_1 \text{ , якщо } F(x_{1,} x_2, ..., x_n) > 0$$
$$X \in V_2 \text{ , якщо } F(x_{1,} x_2, ..., x_n) < 0 \qquad (7.1.22)$$

Ясно, що рівняння $F(x_1, x_2, ..., x_n) = 0$ є рівнянням розділяючої поверхні для підпросторів R_{V_1} і R_{V_2}.

Методи, що основані на теорії статистичних рішень, мають такі обмеження: для їх реалізації необхідно знати

щільності умовних розподілів образів у класах V_1 і V_2. Ці закони розподілів є багатовимірними, і на основі множин векторів-предікторів класів V_1 і V_2 отримання їх аналітичного виду - це дуже складна задача. Тому вважають, що вид законів розподілу є відомим. У такому разі задача зводиться до необхідності на основі вибірок векторів-предікторів отримати оцінки параметрів цих законів. Ця процедура носить назву востановлення закону розподілу. При практичних реалізаціях цих методів вважають найбільш часто, що класи векторів-предікторів підпорядковуються умовним нормальним розподілам. У дійсності ці припущення строго не виконуються. Дуже добре відомо, що для багатьох метеорологічних величин, які виступають у ролі предікторів, нормальний закон розподілу не виконується. Але, як показує попит, це не вносить суттєвих похибок, якщо більшість предікторів має одномодальний розподіл. Ця умова у більшості випадків виконується.

7.2 Побудова розв'язувального правила для альтернативного прогнозу на основі багатовимірного нормального розподілу

Будемо вважати, що вектори-предіктори

$$X_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \dots \\ x_{nj} \end{pmatrix} \qquad (7.2.1)$$

підпорядковуються багатовимірному нормальному розподілу. Параметрами його, як відомо, є вектори математичних сподівань

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_n \end{pmatrix} \qquad (7.2.2)$$

і матриці коваріацій $K = \{K_{ij}\}_{n\times n}$. Тому щільності умовних розподілів для класів V_1 і V_2 мають вид:

$$P(x_1, x_2, \dots, x_n / V_1) = \frac{1}{(2\pi)^{n/2} |K_1|^{1/2}} \cdot \cdot \exp[-\frac{1}{2}(X - \mu_1)' K_1^{-1}(X - \mu_1)], \qquad (7.2.3)$$

$$P(x_1, x_2, \dots, x_n / V_2) = \frac{1}{(2\pi)^{n/2} |K_2|^{1/2}} \cdot \cdot \exp[-\frac{1}{2}(X - \mu_2)' K_2^{-1}(X - \mu_2)], \qquad (7.2.4)$$

де μ_1, μ_2, K_1, K_2 - вектори математичних сподівань і матриці коваріацій першого та другого класів.

Підставимо рівності (7.2.3) і (7.2.4) до дискримінантної функції (7.1.21). Отримаємо

$$F(x) = -\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|K_1| -$$
$$-\frac{1}{2}(X-\mu_1)' K_1^{-1}(X-\mu_1) + \frac{n}{2}\ln 2\pi +$$
$$\frac{1}{2}\ln|K_2| + \frac{1}{2}(X-\mu_2)' K_2^{-1}(X-\mu_2) +$$
$$+ \ln\frac{P(V_1)}{P(V_2)}$$

або після скорочень

$$F(x) = \frac{1}{2}[(X-\mu_2)' K_2^{-1}(X-\mu_2) -$$
$$-(X-\mu_1)' K_1^{-1}(X-\mu_1) + \ln\frac{|K_2|}{|K_1|}] + \ln\frac{P(V_1)}{P(V_2)} \quad (7.2.5)$$

Вважалося, що ціни помилок першого і другого роду однакові $\delta_a = \delta_b$. Дискримінантна функція $F(x)$, яка визначається формулою (7.2.5), носить назву квадратичної дискримінантної функції. Така назва пов'язується з тим, що перші два члени у квадратних дужках являють собою квадратичні форми, тобто многочлени степеня не більше другого.

Щоб використовувати дискримінантну функцію (7.2.5) при складанні альтернативного прогнозу гідрометеорологічного явища, треба здійснити "етап навчання". Він полягає у такому.

Нехай ми маємо m векторів-предікторів $X \in V_1$, тобто коли явище, що прогнозується здійснювалося, а також k векторів-предікторів $X \in V_2$, коли воно не спостерігалося. Множина векторів X об'ємом $V_l = m + k$ - навчаюча сукупність використовується для того,
щоб знайти статистичні оцінки векторів математичних сподівань класів μ_1 і μ_2, якими є вектори середніх значень $\overline{X}_1$ і $\overline{X}_2$ предікторів, а також коваріаційних матриць $\hat{K}_1$ і $\hat{K}_2$. На її основі знаходять також і апріорні ймовірності класів $P(V_1)$ і $P(V_2)$. Проведенням цих обчислювань і завершується етап навчання розпізнаючої системи (7.2.5). Після нього можна переходити до етапу розпізнавання, тобто прогнозу. Він полягає у тому, що вектор предікторів X підставляється у рівність (7.2.5) і знаходиться значення дискримінантної функції, що дає можливість за допомогою розв'язувального правила (7.1.22) віднести цей вектор до класу V_1, тобто прогнозувати це атмосферне явище, або до класу V_2, тобто явище не прогнозувати.

Дискримінантна функція (7.2.5) утримує операції обернення коваріаційних матриць K_1 і K_2. Ця операція може привести до негативних наслідків коли матриці коваріацій є

погано визначними. У такому разі похибки, що містяться в коваріаціях, можуть привести до великих похибок коефіцієнтів квадратичних форм і, таким чином, до помилок на стані розпізнавання. У ряді випадків ці похибки можуть бути більшими ніж ті похибки, які ми робимо, приймаючи умову

$$K_1 = K_2 = K \tag{7.2.6}$$

тобто вважаючи, що матриці коваріацій класів однакові. Частіше за всього приймається умова:

$$K = \frac{K_1 + K_2}{2} \tag{7.2.7}$$

Якщо прийняти умову (7.2.6) у дискримінантній функції (7.2.5) і вважати що $P(V_1) = P(V_2)$ (останню умову можна виконати, формуючи належним чином навчаючу сукупність), то будемо мати

$$F(x) = \frac{1}{2}[(X - \mu_2)' K_2^{-1}(X - \mu_2) -$$
$$-(X - \mu_1)' K_1^{-1}(X - \mu_1)] = (\mu_1 - \mu_2)' K^{-1} X +$$
$$+ \frac{1}{2}[\mu_2' K^{-1} \mu_2 - \mu_1' K^{-1} \mu_1] \tag{7.2.8}$$

Дискримінантна функція (7.2.8) є лінійною дискримінантною функцією. Дійсно, у першому члені правої частини рівності (7.2.8) добуток $\left(\mu_1 - \mu_2\right)' K^{-1}$ дає вектор-рядок, а добуток його з вектором-стовпцем X - це скалярний добуток цих векторів, тобто лінійна форма відносно координат вектора X. Другий член рівності (7.2.8) є скаляр, котрий можна розрахувати, як і вектор коефіцієнтів лінійної форми $\left(\mu_1 - \mu_2\right)' K^{-1}$, заздалегідь ще на етапі навчання.

Використання дискримінантного аналізу значно спрощується, якщо є підстави вважати коваріації рівними нулю. Це можна зробити, якщо всі недіагональні елементи матриць коваріацій класів V_1 і V_2 значно менші від діагональних, тобто дисперсій, і ними можна знехтувати.

Тоді

$$K = \begin{pmatrix} \sigma_1^2 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \sigma_n^2 \end{pmatrix} \qquad (7.2.9)$$

а її обернена матриця

$$K^{-1} = \begin{pmatrix} \dfrac{1}{\sigma_1^2} & 0 & \ldots & 0 \\ 0 & \dfrac{1}{\sigma_2^2} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \dfrac{1}{\sigma_n^2} \end{pmatrix}, \qquad (7.2.10)$$

тобто операція обернення матриць коваріацій значно спрощується. При цьому спрощується й процедура розрахування коефіцієнтів квадратичних форм у квадратичній дискримінантній функції (7.2.5). Зауважимо, що саме такий випадок ми будемо мати, якщо у якості предікторів будемо використовувати головні компоненти векторів-предікторів.

При умові, коли виконується умова (7.2.6) і, крім того, можна вважати коваріаційну матрицю діагональною, значно спрощується вид і лінійної дискримінантної функції (7.2.8). У цьому випадку вона дорівнює

$$F(x) = \sum_{i=1}^{n} \frac{\mu_{1i} - \mu_{2i}}{\sigma_i^2} x_i + \sum_{i=1}^{n} \frac{\mu_{2i}^2 - \mu_{1i}^2}{\sigma_i^2} \qquad (7.2.11)$$

Звичайно, при практичному використанні розглянутих дискримінантних функцій замість складових векторів математичних сподівань і дисперсій предікторів використовують їх статистичні оцінки, тобто середні значення і вибіркові дисперсії предікторів.

7.3 Приклад застосування дискримінантного аналізу в задачі прогнозу інтенсивності забруднення повітря шкідливими домішками

В розділі 4 розглядалася задача прогнозу концентрації шкідливих домішків (на прикладі концентрації SO_2) на основі лінійної регресійної моделі. Але бувають випадки, коли стоїть питання не про конкретне значення концентрації, а треба визначити чи буде концентрація інгредієнта перебільшувати деяке граничне значення, чи не буде. Така задача повинна розв'язуватися з метою контролю вірогідності вимірюванних концентрацій інгредієнтів автоматизованою системою

контролю якості атмосферного повітря. Зазначені задачі розв'язуються за допомогою алгоритмів теорії розпізнування образів. Розглянемо результати розв'язку однієї з таких задач.

У якості вихідної інформації виступали данні про концентрації сірчаного газу (предіктант), а також результати метеорологічних та аерологічних вимірювань (предіктори). До складу потенціальних предікторів були віднесені характеристики атмосфери, які чинять безпосередній вплив на поширення шкідливих домішків в пограничному шарі атмосфери. Це температура і відносна вологість повітря, вектор швидкості вітру, його зональна та меридіональна складові на рівнях 0,02; 0,4; 0,2; 0,5; 1,0 і 2.0 км, вертикальний градієнт швидкості вітру в шарах 0 - 0,5 і 0 - 1,0 км, вертикальний градієнт температури повітря в шарах 0 - 0,5; 0 - 1,0; 0 - 2,0; 0,5 - 1,0; 0,5 - 2,0 км, наявність приземної і припіднятої інверсій температури, положення їх границь, характеристики турбулентності тощо. Всього до складу потенціальних предікторів увійшло 48 метеорологічних характеристик, у тому числі середньодобові концентрації SO_2 у попередній добі, а також значення концентрації інгредієнта за дві і шість годин до терміну, що контролюється. За допомогою методу "включення" (розділ 4) були визначені статистично значущі предіктори. До їх складу увійшли: середньодобова концентрація SO_2 у попередню добу (x_1), концентрація SO_2 за дві години до терміну контролю (x_2), меридіональна складова швидкості вітру на рівні 0,2 км (x_6), модуль швидкості вітру на рівні 0,1 (x_5) км, відносна вологість повітря (x_3), напрямок вітру на рівні 0,1 км (x_4), градієнт температури в шарі 0 - 0,5 км (x_7). Навчальна сукупність складалася з семивимірних векторів ситуацій, координатами яких є перелічені статистично значущі характеристики.

Зазначена множина V векторів ситуацій поділялася на два класи V_1 і V_2. При цьому пороговим критерієм було

середнє значення концентрації SO_2, що розраховувалося на основі реалізації предіктанта. Клас V_1 складався з 349, а клас V_2 - з 211 векторів.

На основі цих підмножин векторів предікторів були побудовані нелінійна і лінійна дискримінантні функції. Чисельні експерименти, що проводилися на цих дискримінантних моделях дали таку ймовірність p правильної класифікації: нелінійна модель p = 0,74; лінійна модель p = 0,76.

Викликає інтерес питання про те, які з предікторів дають найбільший внесок у зазначений результат класифікації. Для відповіді на це питання були організовані чисельні експерименти з лінійною моделлю. Вони полягали у тому, що з векторів предікторів послідовно виключалися ті чи інші предіктори, після чого визначалося значення дискримінантної функції, проводилася класифікація й розраховувалася ймовірність правильної класифікації. Результати цих чисельних експериментів містяться в табл. 7.1, де хрестиком зазначається предіктор, що зберігається у векторі ситуацій, а рискою той, що вилучається.

Таблиця 7.1 - Імовірність правильної класифікації при різному складі векторів - ситуацій.

Номер предік-тора	І м о в і р н і с т ь								
	76	70	75	75	75	76	75	75	65
1	+	+	+	+	+	+	+	+	-
2	+	+	+	+	+	+	+	+	-
3	+	-	+	+	+	-	-	-	+
4	+	-	-	+	+	+	+	-	+
5	+	-	-	-	+	+	+	-	+
6	+	-	-	-	-	+	-	+	+
7	+	-	-	-	-	+	-	-	+

З табл. 7.1 випливає, що результати розпізнавання при різних наборах предікторів змінюється незначно, якщо у векторі ситуації зберігаються концентрації за зазначені попередні терміни, тобто перший і другий предіктори, імовірність правильної класифікації різко падає, якщо їх вилучити.

Отже, моделі розпізнавання, що організовані таким чином, є у значній мірі інерційними, слабко реагуючими на змінювання фізичних умов, що сприяють накопиченню чи розсіюванню домішку в атмосфері. Цей факт пояснюється тим, що при організованому за такими принципами навчання в систему ознак включаються різнорідні стани атмосфери, при яких рівні забруднення повітря можуть коливатися у значних межах. Тому статистичні зв'язки між координатами векторів предікторів, що характеризують стан пограничного шару атмосфери, і рівнями забруднення виявляються менше значущими, ніж залежності між концентраціями SO_2, які контролюються, і концентраціями в попередні терміни.

Більш ефективною виявляється двохступінчаста система розпізнавання. Вона полягає у тому, що спочатку виконується розбиття векторів ситуацій на кластери, що характеризують визначні стани пограничного шару атмосфери, а потім в кожному кластері проводиться розпізнавання стану забруднення атмосферного повітря інгредієнтом. Принципи кластерного аналізу, розглядалися у попередньому розділі. Тому стосовно до задачі, що розглядається, наведемо лише результати кластерізації.

Процес кластерізації векторів предікторів втілювався за допомогою алгоритму ІСОМАД. Вся множина із 560 векторів, характеристика яких розглядалася вище, була розділена на три кластери. До першого з них увійшло 450 , до другого - 75, до третього - 95 векторів предікторів, координати центрів цих кластерів містяться в табл. 7.2.

Таблиця 7.2 - Координати центрів кластерів.

Клас-тер	Номер предіктора						
	1	2	3	4	5	6	7
I	0,53	0,32	83,0	220,0	6,8	1,9	0,51
II	2,23	2,90	77,0	200,0	7,9	7,5	0,41
III	2,24	2,74	81,0	30,0	5,9	-1,7	-0,13

Аналізуючи результати, що утримуються в табл. 7.2, можна побачити, що перший кластер характеризується меншим порівняно з іншими кластерами вмістом домішку у попередні терміни, порівняно високим значенням відносної вологості повітря і швидкості вітру і відносно великим додатнім вертикальним градієнтом температури. Порівняно з першим, другий кластер утримує приблизно на порядок більші величини попередніх концентрацій інгредієнта, знижені значення відносної вологості повітря й більші швидкості вітру. Йому відповідає також більш стійкий стан пограничного шару повітря. Якщо середні значення попередніх концентрацій SO_2 у третьому кластері мало відрізняються від цих значень у другому кластері, то координати центрів другого і третього кластерів, що відносяться до параметрів стану пограничного шару атмосфери, розрізняються значно. В третьому кластері більш високою є відносна вологість повітря, змінюється на протилежний напрямок переносу, меншою є його швидкість, переважує інверсійна стратифікація пограничного шару атмосфери.

Оскільки кожному вектору предікторів ставиться у відповідність визначне значення предіктанта – концентрації SO_2, розбиттю векторів ситуацій на кластери відповідає розбиття множини значень предіктанта на підмножини, які відповідають виділеним кластерам. Емпіричні розподіли їх частостей міститься на рис. 7.3.

Як видно, ці розподіли для різних кластерів значно розрізняються. Першому кластеру (I) відповідають великі частості незначних концентрацій SO_2, а концентрації, що перевищують дві відносні одиниці практично відсутні. До другого кластеру (II) належать концентрації SO_2 від 1 до 5 відносних одиниць. При третьому кластері (III) ситуацій спостерігається відносні концентрації SO_2 від 0,5 до 7 одиниць. Крім областей визначення, розподіли розрізняються і формою. Якщо для концентрацій SO_2 першого кластеру векторів предікторів розподіл є близьким до експоненціального, то для другої і третьої підмножини концентрацій SO_2 криві розподілу мають моду. При цьому імовірність модального значення предіктанта другого кластеру перебільшує імовірність модального значення у третьому кластері майже у два рази. Розподіл імовірностей концентрацій SO_2, що відповідають векторам предікторів третього кластеру, характеризується, окрім того, значною правосторонньою асиметрією.

Двохступінчаста система класифікації полягає у такому: спочатку визначається належність вектора предікторів, що пред'являється для розпізнавання, до одного із трьох кластерів шляхом порівняння евклідових відстаней між цим вектором і центрами кластерів. Оскільки, як зазначалося вище, першому кластеру відповідають малі величини концентрацій SO_2, то попадання вектора ситуацій до першого кластеру свідчить про те, що відповідне йому значення концентрацій домішку повинно бути невеликим.

Вектори ситуацій, що відносяться до другого і третього кластерів, можуть супроводитись як великими, так й малими концентраціями інгредієнта. Тому віднесення вектора ситуацій до одного з них ще не дає можливості прийняти рішення відносно вірогідності концентрації домішку, що вимірюється автоматизованою системою. Необхідно побудувати розв'язувальне правило, яке дало б змогу виконати подальшу

класифікацію вже всередині кожного з кластерів. Для цього були сформовані навчальні сукупності шляхом розбиття кожного з останніх кластерів на два класи. Порогом при цьому встановлювалось середнє значення предіктанта, розрахованого для кожного кластера. В результаті із загальної кількості векторів предікторів другого кластеру до першого класу віднесено 31, для другого - 44 векторів. Із третього кластеру були сформовані підмножини, які складалися відповідно з 53 і 42 векторів.

Результати чисельних експериментів для двохступінчастої системи містяться в табл. 7.3. Розглядаються ймовірності правильного розпізнавання, що отримані за допомогою побудованих на основі перелічених вище навчальних сукупностей нелінійних і лінійних дискримінантних функцій, для ситуацій, котрі відносяться до другого й третього кластерів. Як зазначалося вище, першому кластеру векторів ситуацій відповідають тільки невеликі концентрації інгредієнта і тільки такі концентрації треба вважати вірогідними. Імовірності правильного розпізнавання, що містяться в табл. 7.3, отримані, крім того, при різних наборах предікторів.

Таблиця 7.3 - Імовірності правильного розпізнавання при різних наборах предікторів.

Набір предікторів моделі	Вид	І м о в і р н і с т ь	
		третій кластер	другий кластер
1,2.3,4, 5,6,7	лінійна нелінійна	0,66 0,80	0,64 0,79
1,2,3	лінійна нелінійна	0,59 0,42	0,63 0,48

Як випливає з табл. 7.3, у другому і третьому кластері при повному наборі предікторів нелінійні дискримінантні функції дають більшу імовірність правильного розпізнавання, ніж лінійні. Порівняно з одноступінчастою двохступінчаста модель є більш чутливою до предікторів, що характеризують стан пограничного шару атмосфери: виключення чотирьох останніх предікторів значно зменшує імовірність правильної класифікації, причому найбільш чутливими виявляються нелінійні дискримінантні моделі. Цей факт показує, що двоступінчасті моделі класифікації є більш фізично обгрунтованими, ніж одноступінчасті.

Треба мати на увазі, що потреби побудови двохступінчастих моделей розпізнавання образів визначаються дослідником у залежності від характеру задачі, що розв'язується. У багатьох інших випадках і одноступінчаста модель дає цілком прийнятні результати.

Зміст

процесів і полів

Додаток 1

Значення функції $f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$

t	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0,0	3989	3989	3989	3988	3986	3984	3982	3980	3977	3973
0,1	3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0,2	3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0,3	3814	3802	3700	3778	3765	3752	3739	3725	3712	3697
0,4	3683	3668	3653	3637	3621	3605	3589	3572	3555	3538
0,5	3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0,6	3332	3312	3292	3271	3251	3230	3209	3187	3166	3144
0,7	3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
0,8	2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0,9	2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1,0	2420	2396	2371	2347	2323	2299	2275	2251	2227	2203
1,1	2179	2155	2131	2107	2083	2059	2036	2012	1989	1965
1,2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1,3	1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1,4	1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1,5	1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1,6	1109	1092	1074	1057	1040	1023	1006	0989	0973	0957
1,7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1,8	0790	0775	0761	0748	0734	0721	0707	0694	0681	0669
1,9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2,0	0540	0529	0519	0508	0498	0488	0478	0468	0459	0449
2,1	0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2,2	0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2,3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2,4	0224	0219	0213	0203	0203	0198	0194	0189	0184	0180
2,5	0175	0171	0167	0163	0158	0154	0151	0147	0143	0139

Продовження Додатку 1

t	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
2,6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2,7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2,8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2,9	0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
3,0	0044	0043	0042	0040	0039	0038	0037	0036	0035	0034

Додаток 2

Значення інтегралу ймовірності $\Phi(t) = \frac{2}{\sqrt{2\pi}} \int_0^t e^{-\frac{t^2}{2}} dt$

t	$\Phi(t)$	t	$\Phi(t)$	t	$\Phi(t)$	t	$\Phi(t)$
0,00	0,00000	0,30	0,23582	0,60	0,45149	0,90	0,63188
01	00798	31	24344	61	45814	91	63718
02	01596	32	25103	62	46474	92	64243
03	02393	33	25860	63	47131	93	64763
04	03191	34	26614	64	47783	94	65278
0,05	0,03988	0,35	0,27366	0,65	0,48431	0,95	0,65789
06	04784	36	28115	66	49075	96	66294
07	05581	37	28862	67	49714	97	66795
08	06376	38	29605	68	50350	98	67291
09	07171	39	30346	69	50981	99	67783
0,10	0,07966	0,40	0,31084	0,70	0,51607	1,00	0,68269
11	08759	41	31819	71	52230	01	68750
12	09552	42	32552	72	52848	02	69227
13	10348	43	33280	73	53461	03	69699
14	11134	44	34006	74	54070	04	70166
15	11924	45	34729	75	54675	1,05	70628
16	12712	46	35448	76	55275	06	71086
17	13499	47	36164	77	55870	07	71538
18	14285	48	36877	78	56461	08	71986
19	15069	49	37587	79	57047	09	72429
0,20	0,15852	0,50	0,38292	0,80	0,57629	1,10	0,72867
21	16633	51	38995	81	58206	11	73300
22	17413	52	39694	82	58778	12	73729
23	18191	53	40389	83	59346	13	74152
24	18967	54	41080	84	59909	14	74571
25	19741	55	41768	85	60468	15	74986
26	20514	56	42452	86	61021	16	75395
27	21284	57	43132	87	61570	17	75800
28	22052	58	43809	88	62114	18	76200
29	22818	59	44481	89	62653	19	76595

Продовження Додатку 2

t	$\Phi(t)$	t	$\Phi(t)$	t	$\Phi(t)$	t	$\Phi(t)$
1,20	0,76986	1,55	0,87886	1,90	0,94257	2,25	0,97555
21	77372	56	88124	91	94387	26	97618
22	77754	57	88358	92	94514	27	97679
23	78130	58	88589	93	94639	28	97739
24	78502	59	88817	94	94762	29	97798
25	78870	1,60	0,89040	95	94882	2,30	0,97855
26	79233	61	89260	96	95000	31	97911
27	79592	62	89477	97	95116	32	97966
28	79945	63	89690	98	95230	33	98019
29	80295	64	89899	99	95341	34	98072
1,30	0,80640	65	90106	2,00	0,95450	35	98123
31	80980	66	90309	01	95557	36	98172
32	81316	67	90508	02	95662	37	98221
33	81648	68	90704	03	95764	38	98269
34	81975	69	90897	04	95865	39	98315
35	82298	1,70	0,91087	05	95964	2,40	0,98360
36	82617	71	91273	06	96060	41	98405
37	82931	72	91457	07	96155	42	98448
38	83241	73	91637	08	96247	43	98490
39	83547	74	91814	09	96338	44	98531
1,40	0,83849	75	91988	2,10	0,96427	45	98571
41	84146	76	92159	11	96514	46	98611
42	84439	77	92327	12	96599	47	98649
43	84728	78	92492	13	96683	48	98686
44	85013	79	92655	14	96765	49	98723
45	85294	1,80	0,92814	15	96844	2,50	0,98758
46	85571	81	92970	16	96923	51	98793
47	85844	82	93124	17	96999	52	98826
48	86113	83	93275	18	97074	53	98859
49	86378	84	93423	19	97148	54	98891
1,50	0,86639	1,85	0,93569	2,20	0,97219	2,55	0,98923
51	86696	86	93711	21	97289	56	98953
52	87149	87	93852	22	97358	57	98983
53	87398	88	93989	23	97425	58	99012
54	87644	89	94124	24	97491	59	99040

Продовження Додатку 2

t	$\Phi(t)$	t	$\Phi(t)$	t	$\Phi(t)$	t	$\Phi(t)$
2,60	0,99068	2,95	0,99682	3,30	0,99903	3,65	0,99974
61	99095	96	99692	31	99907	66	99975
62	99121	97	99702	32	99910	67	99976
63	99146	98	99712	33	99913	68	99977
64	99171	99	99721	34	99916	69	99978
65	99195	3,00	0,99730	35	99919	3,70	0,99978
66	99219	01	99739	36	99922	71	99979
67	0,99241	02	99747	37	99925	72	99980
68	99263	03	99755	38	99928	73	99981
69	99285	04	99763	39	99930	74	99982
2,70	0,99307	05	99771	3,40	0,99933	75	99982
71	99327	06	99779	41	99935	76	99983
72	99347	07	99786	42	99937	77	99984
73	99367	08	99793	43	99940	78	99984
74	99386	09	99800	44	99942	79	99985
75	99404	3,10	0,99806	45	99944	3,80	0,99986
76	99422	11	99813	46	99946	81	99986
77	99439	12	99819	47	99948	82	99987
78	99456	13	99825	48	99950	83	99987
79	99473	14	99831	49	99952	84	99988
2,80	0,99489	15	99837	3,50	99953	85	99988
81	99505	16	99842	51	99955	86	99989
82	99520	17	99848	52	99957	87	99989
83	99535	18	99853	53	99958	88	99990
84	99549	19	99858	54	99960	89	99990
85	99563	3,20	0,99863	55	99961	3,90	0,99990
86	99576	21	99867	56	99963	91	99991
87	99590	22	99872	57	99964	92	99991
88	99602	23	99876	58	99966	93	99992
89	99615	24	99880	59	99967	94	99992
2,90	0,99627	25	99855	3,60	0,99968	95	99992
91	99639	26	99889	61	99969	96	99992
92	99650	27	99892	62	99971	97	99993
93	99661	28	99896	63	99972	98	99993
94	99672	29	99900	64	99973	99	99993

Додаток 3

Значення функції Пуассона $P(X = m) = \frac{\lambda^m e^{-\lambda}}{m\ !}$

X	λ									
	0,1	**0,2**	**0,3**	**0,4**	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**	**1,0**
0	0,9048	8187	7408	6703	6065	5488	4966	4493	4066	3679
1	0905	1637	2222	2681	3033	3293	3476	3595	3659	3679
2	0045	0164	0333	0536	0758	0988	1217	1438	1647	1839
3	0002	0011	0033	0072	0126	0198	0284	0383	0494	0613
4	0000	0001	0003	0007	0016	0030	0050	0077	0111	0253
5	-	-	-	0001	0002	0004	0007	0012	0020	0031
6	-	-	-	-	-	-	0001	0002	0003	0005
7	-	-	-	-	-	-	-	-	-	0001

Продовження Додатку 3

X	λ									
	1,1	**1,2**	**1,3**	**1,4**	**1,5**	**1,6**	**1,7**	**1,8**	**1,9**	**2,0**
0	0,3329	3012	2725	2466	2231	2019	1827	1653	1496	1353
1	3662	3614	3543	3452	3347	3230	3106	2975	2842	2707
2	2014	2169	2303	2417	2510	2584	2640	2678	2700	2707
3	0738	0867	0998	1128	1255	1378	1496	1607	1710	1805
4	0203	0260	0324	0395	0471	0551	0636	0723	0812	0902
5	0045	0063	0084	0111	0141	0176	0216	0260	0309	0361
6	0008	0013	0018	0026	0035	0047	0061	0078	0098	0120
7	0001	0002	0003	0005	0008	0011	0015	0020	0027	0034
8	-	-	0001	0001	0001	0002	0003	0005	0006	0009
9	-	-	-	-	-	-	0001	0001	0001	0002

Продовження Додатку 3

X	λ									
	2,1	**2,2**	**2,3**	**2,4**	**2,5**	**2,6**	**2,7**	**2,8**	**2,9**	**3,0**
0	0,1225	1108	1003	0903	0821	0743	0672	0608	0550	0498
1	2572	2438	2306	2177	2052	1931	1815	1703	1596	1494
2	2700	2681	2652	2613	2565	2510	2450	2384	2314	2240
3	1890	1964	2083	2090	2136	2176	2205	2225	2234	2240
4	0992	1087	1169	1254	1336	1414	1488	1557	1622	1680
5	0417	0476	0538	0602	0668	0735	0804	0872	0941	1008
6	0146	0175	0206	0241	0278	0319	0362	0407	0455	0504
7	0004	0055	0068	0083	0099	0118	0140	0163	0188	0216
8	0012	0015	0020	0025	0031	0039	0047	0057	0068	0081
9	0003	0004	0005	0007	0009	0011	0014	018	0022	0027
10	0001	0001	0001	0002	0002	0003	0004	0005	0006	0008

Продовження Додатку 3

X	λ									
	3,5	**4,0**	**4,5**	**5,0**	**6,0**	**7,0**	**8,0**	**9,0**	**10,0**	**11,0**
0	0,0302	0183	0111	0067	0025	0009	0003	0001	-	-
1	1057	0733	0500	0333	0149	0064	0027	0011	0005	0002
2	1850	1465	1125	0842	0446	0223	0107	0050	0023	0010
3	2158	1954	1687	1404	0892	0521	0286	0150	0076	0037
4	1888	1954	1898	1755	1339	0912	0573	0338	0189	0102
5	1327	1563	1708	1755	1606	1277	0916	1277	0916	0224
6	0771	1042	1281	1462	1606	1490	1221	0911	0631	0411
7	0386	0595	0824	1044	1377	1490	1396	1171	0901	0646
8	0169	0298	0463	0653	1033	1304	1396	1318	1126	0888
9	0066	0132	0232	0363	0638	1014	1241	1318	1251	1058
10	0023	0053	0104	0181	0413	0710	0993	1186	1251	1194
11	0007	0019	0043	0082	0225	0452	0722	0970	1137	1194
12	0002	0006	0016	0034	0113	0264	0481	0728	0948	1094

Додаток 4

Значення критерію Стьюдента при рівні значущості a і числа ступенів волі ν

ν	**Двостороння критична область, α**							
	0,2	**0,1**	**0,05**	**0,02**	**0,01**	**0,005**	**0,002**	**0,001**
1	3,08	6,31	12,71	31,82	63,66	127,32	318,30	636,61
2	1,89	2,92	4,30	6,96	9,92	14,09	22,33	31,60
3	1,64	2,35	3,18	4,54	5,84	7,45	10,21	12,92
4	1,53	2,13	2,78	3,75	4,60	5,60	7,17	8,61
5	1,48	2,02	2,57	3,36	4,03	4,77	5,89	6,87
6	1,44	1,94	2,45	3,14	3,71	4,32	5,21	5,96
7	1,41	1,89	2,36	3,00	3,50	4,03	4,79	5,41
8	1,40	1,86	2,31	2,90	3,36	3,83	4,50	5,04
9	1,38	1,83	2,26	2,82	3,25	3,69	4,30	4,78
10	1,37	1,81	2,23	2,76	3,17	3,58	4,14	4,59
11	1,36	1,80	2,20	2,72	3,11	3,50	4,02	4,44
12	1,36	1,78	2,20	2,68	3,05	3,43	3,93	4,32
13	1,35	1,77	2,16	2,65	3,01	3,37	3,85	4,22
14	1,34	1,76	2,14	2,62	2,98	3,33	3,79	4,14
15	1,34	1,75	2,13	2,60	2,95	3,29	3,73	4,07
16	1,34	1,75	2,12	2,58	2,92	3,25	3,69	4,02
17	1,33	1,74	2,11	2,57	2,90	3,22	3,65	3,97
18	1,33	1,73	2,10	2,55	2,88	3,20	3,61	3,92
19	1,33	1,73	2,09	2,54	2,86	3,17	3,58	3,88
20	1,33	1,72	2,09	2,53	2,85	3,15	3,55	3,85
21	1,32	1,72	2,08	2,52	2,83	3,14	3,53	3,82
22	1,32	1,72	2,07	2,51	2,82	3,12	3,51	3,79
23	1,32	1,71	2,07	2,50	2,81	3,10	3,48	3,77
24	1,32	1,71	2,06	2,49	2,90	3,09	3,47	3,75
25	1,32	1,71	2,06	2,49	2,79	3,08	3,45	3,73
26	1,32	1,71	2,06	2,48	2,78	3,07	3,44	3,71
27	1,31	1,70	2,05	2,47	2,77	3,06	3,42	3,69
28	1,31	1,70	2,05	2,47	2,76	3,05	3,41	3,67
29	1,31	1,70	2,05	2,46	2,76	3,04	3,40	3,66
30	1,31	1,70	2,04	2,46	2,75	3,03	3,39	3,65

Продовження Додатку 4

ν	**Двостороння критична область, α**							
	0,2	**0,1**	**0,05**	**0,02**	**0,01**	**0,005**	**0,002**	**0,001**
40	1,30	1,68	2,02	2,42	2,70	2,97	3,31	3,55
60	1,30	1,67	2,00	2,39	2,66	2,91	3,23	3,46
120	1,29	1,66	1,98	2,36	2,62	2,85	3,16	3,37
	1,28	1,64	1,96	2,33	2,58	2,81	3,09	3,29

Додаток 5

Значення критерію Фішера F
для рівня значущості 0,05

ν_1 - число ступенів волі для більшої дисперсії

ν_2 - число ступенів волі для меншої дисперсії

ν_1	ν_2								
	12	**14**	**16**	**20**	**24**	**30**	**40**	**50**	**75**
10	2,9	2,9	2,8	2,8	2,7	2,7	2,7	2,6	2,6
11	2,8	2,7	2,7	2,7	2,6	2,6	2,5	2,5	2,5
12	2,7	2,6	2,6	2,5	2,5	2,5	2,4	2,4	2,4
13	2,6	2,6	2,5	2,5	2,4	2,4	2,3	2,3	2,3
14	2,5	2,5	2,4	2,4	2,4	2,3	2,3	2,2	2,2
15	2,5	2,4	2,3	2,3	2,3	2,3	2,2	2,2	2,2
16	2,4	2,4	2,3	2,3	2,2	2,2	2,2	2,1	2,1
17	2,4	2,3	2,3	2,3	2,2	2,2	2,1	2,1	2,0
18	2,3	2,3	2,3	2,3	2,2	2,1	2,1	2,0	2,0
19	2,3	2,3	2,2	2,2	2,1	2,1	2,0	2,0	2,0
20	2,3	2,2	2,2	2,1	2,1	2,0	2,0	2,0	1,9
21	2,3	2,2	2,2	2,1	2,1	2,0	2,0	1,9	1,9
22	2,2	2,2	2,1	2,1	2,0	2,0	1,9	1,9	1,9
23	2,2	2,1	2,1	2,0	2,0	2,0	1,9	1,9	1,8
24	2,2	2,1	2,1	2,0	2,0	1,9	1,9	1,9	1,8
25	2,2	2,1	2,1	2,0	2,0	1,9	1,9	1,8	1,8
26	2,2	2,1	2,1	2,0	2,0	1,9	1,9	1,8	1,8
27	2,1	2,1	2,0	2,0	1,9	1,9	1,8	1,8	1,8
28	2,1	2,1	2,0	2,0	1,9	1,8	1,8	1,8	1,8
29	2,1	2,1	2,0	1,9	1,9	1,9	1,8	1,8	1,7
30	2,1	2,0	2,0	1,9	1,9	1,8	1,8	1,8	1,7
32	2,1	2,0	2,0	1,9	1,9	1,8	1,8	1,7	1,7
34	2,1	2,0	2,0	1,9	1,8	1,8	1,7	1,7	1,7
36	2,0	2,0	1,9	1,9	1,8	1,8	1,7	1,7	1,7
38	2,0	2,0	1,9	1,9	1,8	1,8	1,7	1,7	1,6
40	2,0	2,0	1,9	1,8	1,8	1,7	1,7	1,7	1,6
42	2,0	1,9	1,9	1,8	1,8	1,7	1,7	1,6	1,6
44	2,0	1,9	1,9	1,8	1,8	1,7	1,7	1,6	1,6
46	2,0	1,9	1,9	1,8	1,8	1,7	1,7	1,6	1,6

Продовження Додатку 5

48	2,0	1,9	1,9	1,8	1,7	1,7	1,6	1,6	1,6
50	2,0	1,9	1,9	1,8	1,7	1,7	1,6	1,6	1,6
100	1,9	1,8	1,8	1,7	1,6	1,6	1,5	1,5	1,4
200	1,8	1,7	1,7	1,6	1,6	1,5	1,7	1,4	1,4
∞	1,8	1,7	1,6	1,6	1,5	1,5	1,4	1,4	1,3

Додаток 6

Значення $\chi^2(\alpha,\nu)$ для різних значень чисел ступенів волі та рівня значущості

ν \ α	0,995	0,990	0,975	0,95	0,90	0,80	0,70	0,30
1	0,0393	0,0157	0,0982	0,0393	0,0158	0,0642	0,148	1,07
2	0,0100	0,0201	0,0506	0,103	0,211	0,446	0,713	2,41
3	0,0717	0,115	0,216	0,352	0,584	1,00	1,42	3,67
4	0,207	0,297	0,484	0,711	1,06	1,65	2,19	4,88
5	0,412	0,554	0,831	1,15	1,61	2,34	3,00	6,06
6	0,676	0,872	1,24	1,64	2,20	3,07	3,83	7,23
7	0,989	1,24	1,69	2,17	2,83	3,82	4,67	8,38
8	1,34	1,65	2,18	2,73	3,49	4,59	5,53	9,52
9	1,73	2,09	2,70	3,33	4,17	5,38	6,39	10,7
10	2,16	2,56	3,25	3,94	4,87	6,18	7,27	11,8
11	2,60	3,05	3,82	4,57	5,58	6,99	8,15	12,9
12	3,07	3,57	4,40	5,23	6,30	7,81	9,03	14,0
13	3,57	4,11	5,01	5,89	7,04	8,63	9,93	15,1
14	4,07	4,66	5,63	6,57	7,79	9,47	10,8	16,2
15	4,60	5,23	6,26	7,26	8,55	10,3	11,7	17,3

Продовження Додатку 6

ν \ α	0,995	0,990	0,975	0,95	0,90	0,80	0,70	0,30
16	5,14	5,81	6,91	7,96	9,31	11,2	12,6	18,4
17	5,70	6,41	7,56	8,67	10,1	12,0	13,5	19,5
18	6,26	7,01	8,23	9,39	10,9	12,9	14,4	20,6
19	6,84	7,63	8,91	10,1	11,7	13,7	15,4	21,7
20	7,43	8,26	9,59	10,9	12,4	14,6	16,3	22,8
21	8,03	8,90	10,3	11,6	13,2	15,4	17,2	23,9
22	8,64	9,54	11,0	12,3	14,0	16,3	18,1	24,9
23	9,26	10,2	11,7	13,1	14,8	17,2	19,0	26,0
24	9,89	10,9	12,4	13,8	15,7	18,1	19,9	27,1
25	10,5	10,5	13,1	14,6	16,5	18,9	20,9	28,2
26	11,2	12,2	13,8	15,4	17,3	19,8	21,8	29,2
27	11,8	12,9	14,6	16,2	18,1	20,7	22,7	30,3
28	12,5	13,6	15,3	16,9	18,9	21,6	23,6	31,4
29	13,1	14,3	16,0	17,7	19,8	22,5	24,6	32,5
30	13,8	15,0	16,8	18,5	20,6	23,4	25,5	33,5
35	17,2	18,5	20,6	22,5	24,8	27,8	30,2	38,9
40	20,7	22,2	24,4	26,5	29,1	32,3	34,9	44,2
45	24,3	25,9	28,4	30,6	33,4	36,9	39,6	49,5
50	28,0	29,7	32,4	34,8	37,7	41,1	44,3	54,7
75	47,2	49,5	52,9	56,1	59,8	64,5	68,1	80,9
100	67,3	70,1	74,2	77,9	82,4	87,9	92,1	106,9

Продовження Додатку 6

ν \ α	0,20	0,10	0,05	0,025	0,010	0,005	0,001
1	1,64	2,71	3,84	5,02	6,63	7,88	10,8
2	3,22	4,61	5,99	7,38	9,21	10,6	13,8
3	4,64	6,25	7,81	9,35	11,3	12,8	16,3
4	5,99	7,78	9,49	11,1	13,3	14,9	18,5
5	7,29	9,24	11,1	12,8	15,1	16,7	20,5
6	8,56	10,6	12,6	14,4	16,8	18,5	22,5
7	9,80	12,0	14,1	16,0	18,5	20,3	24,3
8	11,0	13,4	15,5	17,5	20,1	22,0	26,1
9	12,2	14,7	16,9	19,0	21,7	23,6	27,9
10	13,4	16,0	18,3	20,5	23,2	25,2	29,6
11	14,6	17,3	19,7	21,9	24,7	26,8	31,3
12	15,8	18,5	21,0	23,3	26,2	28,3	32,9
13	17,0	19,8	22,4	24,7	27,7	29,8	34,5
14	18,2	21,1	23,7	26,1	29,1	31,3	36,1
15	19,3	22,3	25,0	27,6	30,9	32,8	37,7

Продовження Додатку 6

ν \ α	0,20	0,10	0,05	0,025	0,010	0,005	0,001
16	20,5	23,5	26,3	28,8	32,0	34,3	39,3
17	21,6	24,8	27,6	30,2	33,4	35,7	40,8
18	22,8	26,0	28,9	31,5	34,8	37,2	42,3
19	23,9	27,0	30,1	32,9	36,2	38,6	43,8
20	25,0	28,4	31,4	34,2	37,6	40,0	45,3
21	26,9	29,6	32,7	35,5	38,9	41,4	46,8
22	27,3	30,8	33,9	36,0	40,3	42,8	48,3
23	28,4	32,0	35,2	38,1	41,6	44,2	49,7
24	29,6	33,0	36,4	39,4	43,0	45,6	51,2
25	30,7	34,4	37,7	40,6	44,3	46,9	52,6
26	31,8	35,6	38,9	41,0	45,6	48,3	54,1
27	32,9	36,7	40,1	43,2	47,0	49,6	55,5
28	34,0	37,9	41,3	44,5	48,3	51,0	56,9
29	35,1	39,1	42,6	45,7	49,6	52,3	58,3
30	36,3	40,3	43,8	47,0	50,9	53,7	59,7
35	41,8	46,1	49,8	53,2	57,3	60,3	66,6
40	47,3	51,8	55,8	59,3	63,7	66,8	73,4
45	52,7	57,5	61,7	65,4	70,0	73,2	80,1
50	58,2	63,2	67,5	71,4	76,2	79,5	86,7
75	85,1	91,1	96,2	100,8	106,4	110,3	118,6
100	111,7	118,5	124,3	129,6	135,6	140,2	149,4