Міністерство освіти і науки України

Малинський фаховий коледж

Біометрія

Навчальний посібник для самостійної роботи студентів

спеціальності 205 Лісове господарство
ОПП «Лісове господарство»
для студентів ОС «Бакалавр»

2021

Укладач: Якименко О.Г., викладач Малинського фахового коледжу, кандидат педагогічних наук

Викладено лекційний матеріал дисципліни «Біометрія» для самостійної роботи студентів.

Для студентів освітнього ступеню бакалавр, галузі знань 20 Аграрні науки та продовольство , спеціальності 205 Лісове господарство (ОПП Лісове господарство)

Розглянуто і схвалено схвалено на засіданні кафедри лісівництва та захисту лісу
Протокол від 02.09. 2021 року № 2

Завідувач кафедри _____ / Я.Д.Фучило /

МОДУЛЬ І.
СТАТИСТИЧНЕ СПОСТЕРЕЖЕННЯ В БІОМЕТРІЇ. СТАТИСТИЧНІ ПОКАЗНИКИ ТА РОЗПОДІЛ БІОМЕТРИЧНИХ ОЗНАК

1. ПРЕДМЕТ І МЕТОДИ БІОМЕТРІЇ ЯК НАУКИ. ОСНОВНІ ІСТОРИЧНІ ЕТАПИ ФОРМУВАННЯ БІОМЕТРІЇ ЯК НАУКИ. ВИЗНАЧЕННЯ ОСНОВНИХ ПОНЯТЬ

1.1. Предмет і зміст біометрії як науки

З формальної точки зору біометрія являє собою сукупність математичних методів, що застосовують в біологічних дослідженнях. Найбільш тісно біометрія зв’язана з статистикою і теорією ймовірностей.

Сучасна біометрія - це розділ біології, змістом якої є планування спостережень та статистична обробка їх результатів.

Характерною особливістю біометрії є також те, що її методи застосовуються при аналізі не окремих фактів, а їх сукупностей, тобто явищ масового характеру, в колі яких виявляються закономірності, що не характерні окремим спостереженням.

Звичайно, не всі дослідження в біології опираються на біометрію. В біології з успіхом застосовуються і чисто описові методи, що не потребують кількісної оцінки отриманих результатів. Але там, де дослідження проводяться з використанням обліку або міри, застосування біометрії стає безумовно необхідним. В таких випадках зневажання методами біометрії або неправильне їх застосування призводить до невиправданих затрат праці і часу, а головне - є досить малодостовірним, а іноді і помилковим результатом.

Біометрія – формальна наука. Застосування її до аналізу досліджуваних явищ потребує обережності. Гекслі порівнював біометрію з млином, який меле все, що засипають, але цінність змеленого цілком визначається цінністю засипаного.

Біометрія призвана озброїти дослідників методами статистичного аналізу, виховувати в них статистичне мислення, розкриваючи перед ними діалектику взаємозв’язку між часткою і цілим, причиною і наслідком, випадковим і необхідним у явищах живої природи.

Отже, об'єктами біометричного аналізу можуть бути найрізноманітніші явища й процеси, що відбуваються як в живій, так і неживій природі. ***Предметом*** біометрії є розміри і кількісні співвідношення між масовими явищами в природі, закономірності їх формування, розвитку, взаємозв'язку.

У наведеному визначенні предмета біометрії підкреслюються дві принципові його особливості. По-перше, біометрія вивчає кількісний бік явищ в природі, а по-друге, вона вивчає не поодинокі, а масові явища.

Вивчаючи кількісний бік явищ, біометрія відбиває його у числах-показниках, характеризуючи цим конкретну міру явищ. Водночас вона встановлює загальні властивості, виявляє схожість і різницю окремих властивостей досліджуваних об'єктів, групує їх, виявляючи певні типи процесів і явищ, які вивчаються.

Явища в природі динамічні, вони безперервно змінюються й розвиваються, що неодмінно позначається на їх розмірах, співвідношеннях і пропорціях. Значення розглядуваних кількісних характеристик залежать від конкретних умов простору і часу.

Інша особливість предмета біометрії зумовлюється масовістю явищ в природі, їх повторюваністю у просторі або з плином часу.

Для масового явища характерна участь у ньому багатьох елементів, істотні властивості яких однакові або схожі між собою. Наявність будь-яких властивостей у окремого, поодинокого елемента – випадковість. Проте тільки-но численні елементи об'єднуються в одне ціле, сукупна дія випадковостей дає результат, практично незалежний від випадку.

Розглядаючи явища як масові й спираючись на облік усієї сукупності фактів, що їх стосуються, біометрія мовою чисел характеризує ступінь розвитку таких явищ, напрям і швидкість їх змін, щільність взаємозв'язків і взаємозалежностей. Усе це дає підстави стверджувати, що біометрія – могутній засіб пізнання життя.

1.2. Історичні етапи формування біометрії як науки

Біометрія, як відносно самостійна наукова дисципліна сформувалася у другій половині XIX ст. Але її витоки йдуть із більш раннього періоду в розвитку природничих наук: до того часу, коли виміри біологічних об’єктів стали розглядатися як метод наукового пізнання. Це відбувалося в ті роки, коли почало формуватися капіталістичне суспільство, що потребувало більш точних знань про природу. Актуальним для того часу став афоризм Г. Галілея (1564–1642) “вимірюй все, що можна виміряти і роби придатним для вимірювання, все що не вимірюється”.

В 1614 з’явилася книга Сантаріо (1561–1636) “Про статистику в медицині”. В 1680 вийшла книга Бореллі (1608–1679) “Про рух тварин” У 1768 р. французький гіпполог Бурнеля видає працю “Екстер’єр коня”, в якій наводиться набір вимірів, що необхідні для визначення придатності коней до тої чи іншої служби. Характерно, що в той же час, тобто в XVII ст., розвивається

військова антропологія, яка опирається на результати вимірів тіла чоловіків призовного віку з метою відбору осіб придатних до військової служби.

У 1718 р. в Лондоні вийшла в сів книга французького математика А. де Муавра (1667–1754) “Вчення про випадок”. Він виміряв ріст у 1375 дорослих жінок і розташував результати вимірів в ряд, він знайшов закономірність, що відповідає відомому в теорії ймовірностей закону нормального розподілу. Виникла необхідність інтеграції методів біології з методами теорії ймовірностей і математичною статистикою.

Теорія ймовірностей і математична статистика виникли в середині XVII ст. незалежно одна від одної. Стимулом до розвитку теорії ймовірностей став розвиток грошових відносин в буржуазному суспільстві. Відому роль в цьому відіграли азартні ігри, що стали простими моделями, які дали змогу помітити закономірності в поведінці випадкових явищ масового характеру.

У витоків теорії ймовірностей стояли вчені П. Ферма (1601–1665) та Б. Паскаль (1623–1662), а також голландський математик та природничник Х. Гюйгенс (1629–1696).

Вагомий внесок в становлення теорії ймовірностей внесли Я. Бернуллі (1654–1705) та А. де Муавр. Однак, найбільший розвиток отримала теорія в працях таких видатних математиків, як П. Лаплас (1749–1827), К. Гадес (1777–1555), С. Пуассон (1781–1840), а також П.Л. Чебишев та його Петербурзька школа.

Розвиток математичної статистики тісно пов’язаний з проблемами розвитку управління державою. До середини XVII ст. в економічно розвинутих країнах Європи назріла гостра необхідність пошуку нових методів аналізу статистичних даних (по демографії, торгівлі, страхуванні, тощо) та їх теоретичного обґрунтування. Задача зводилася до того, щоб із частини робити висновок про стан цілого. Розробка теорії вибіркового методу зближувала математичну статистику з висновками теорії ймовірностей, що стало важливим етапом на шляху зародження біометрії.

Першим, хто вдало поєднав методи антропології і соціальної статистики з висновками теорії ймовірностей та математичної статистики, був бельгійський антрополог і статистик А. Кетле (1796–1874). У 1835 р. у Брюсселі вийшла в світ книга Кетле “Про людину і розвиток її здібностей або досвід соціальної фізики”. На великому фактичному матеріалі Кетле вперше показав, що самі різні фізичні особливості людини і навіть поведінка підпорядкована загальному закону розподілу ймовірностей, що описується формулою Гауса–Лапласа. В праці “Антропологія” (1871) Кетле показав, що відкриті ним статистичні закономірності розповсюджуються не тільки на людське суспільство, але й на всі живі істоти.

Досліди Кетле стали визначною подією в історії статистичної науки Він одним з перших переконливо довів, що випадковості, які спостерігаються в живій природі, внаслідок їх повторюваності виявляють внутрішню тенденцію, яку можна досліджувати та описувати точними математичними методами.

Створення математичного апарату біометрії належить англійській школі біометриків XIX ст. на чолі якої стояли Ф. Гальтон (1822–1911) та К. Пірсон (1857–1936). Ця школа виникла під впливом праць Ч. Дарвіна (1809–1882). Одним з перших, хто відчув вплив праць Ч.Дарвіна, був його двоюрідний брат Ф.Гальтон. Сильний вплив на Гальтона мали також праці Кетле. Починаючи з 1865 р. Гальтон опублікував ряд праць по антропології та генетиці. На великому фактичному матеріалі він підтвердив висновок Кетле щодо розподілу ознак усіх біологічних об'єктів за формулою Гаусса-Лапласа.

К. Пірсон - професор Лондонського університету, учень Гальтона, створив математичний апарат біометрії, розвинув вчення про різні типи кривих розподілу, розробив метод моментів (1894) та критерій згідності "хі-квадрат" (1890). Пірсон ввів у біометрію такі показники, як середнє квадратичне відхилення (1896). Йому належить вдосконалення методів кореляції і регресії Гальтона (1896, 1898). Разом з Гальтоном і Уельдоном, Пірсон організував випуск журналу "Біометрика" (1901) редактором якого був до кінця свого життя. Цей журнал зіграв велику роль в пропаганді біометрії й розвитку англійської школи.

Але не дивлячись на великі результати Гальтона й Пірсона, їхні спроби застосувати ці методи для вирішення проблеми спадковості організмів виявилися невдалими. Датський вчений В. Йогансен (1857–1927), вказуючи на ці недоліки в своїх дослідах з квасолею, прийшов до важливого висновку, що біологічні проблеми мають вирішуватися за допомогою математики, але не як математичні задачі. "Статистиці – писав Йогансен – завжди має передувати біологічний аналіз, бо інакше результати можуть бути "статистичною брехнею". Математика має надавати допомогу, а не використовуватися в якості керівної ідеї".

Значення біометрії для біології було визнане не відразу: по-перше тому, що статистичні методи базувалися на великих вибірках, а по-друге – вимагали великої кількості обчислень.

Стан став мінятися після того, як була сформована теорія малої вибірки. Піонером у цій галузі став учень К. Пірсона – В. Госсет (1876–1937), що друкував статті в журналі "Біометрика" під псевдонімом "Стьюдент". Подальший розвиток теорія малої вибірки отримала в статтях Пірсона, а також Р. Фішера (1890–1962), який вніс вагомий внесок у розвиток біометрії.

Р. Фішер створив основи планування експериментів - теорії, що у наш час отримала подальший розвиток і стала відносно самостійним розділом біометрії.

Великий вклад в розвиток біометрії внесли і російські вчені: С.Н. Бернштейн (1880–1968), А.Я. Хінчин (1894–1958), Е.Е. Слуцький (1880–1848), А.І. Хотимський (1892–1979), Б.С. Ястремський (1877–1962), В.І. Романовський (1879–1954) і особливо А.Н. Комаров та його школа.

В історії біометрії можна виділити кілька періодів, або етапів розвитку:

**Перший період*, описовий, бере початок у XVII ст. В цей час відбувається перехід від словесного опису і елементарних обліків біологічних об'єктів до їх чисельних характеристик. Вимірювання розглядається як метод наукового пізнання живої природи.

**Другий період*, що почався в першій половині XIX ст., знаменується працями А. Кетле. В цей час закладаються основи біометрії як науки, метою якої є не опис явищ, а їх аналіз, що спрямований на відкриття статистичних закономірностей, що діють у сфері масових явищ. Біометрію розглядають одночасно і як науку, і як метод наукового пізнання.

**Третій період* - формалістський, що характеризується виникненням і розвитком англійської біометричної школи на чолі з Ф.Гальтоном та К. Пірсоном. В той час створюють математичний апарат біометрії і роблять спроби застосувати його для вивчення проблем спадковості і мінливості організмів.

**Четвертий період*, раціоналістичний, що почався у 1902 р. з класичних досліджень В. Йогансена , який довів, що в галузі біологічних досліджень перше місце має посідати біологічний експеримент, а не математика. Математичні методи мають застосовуватися при обробці експериментальний даних.

*П'ятий період у розвитку біометрії відкривають класичні роботи Стьюдента і Р. Фішера. В цей час створюються основи теорії малої вибірки, теорії планування експериментів, вводяться у зміст біометрії нові терміни і поняття. Все це пов'язане з революційними змінами у біології, з руйнуванням застарілих принципів і понять в галузі дослідницьких робіт з посиленням математизації біології. Відбувається все більша спеціалізація біометрії, застосування її методів в різних галузях біології, медицини, антропології тощо.

**Шостий період* почався в другій половині XX ст., він тісно пов'язаний з розвитком кібернетики і обчислювальної техніки і характеризується розширенням як кола завдань, які вирішує біометрія, так і широким застосуванням математичного апарату, що потребує великих обчислень.

1.3. Основні категорії статистики

З питанням про предмет біометрії пов'язані поняття ***статистичної закономірності*** та ***статистичної сукупності.***

Закономірність – це повторюваність, послідовність і порядок у масових процесах. Виявити і виміряти статистичну закономірність можна лише з урахуванням дії закону великих чисел, основними принципами якого є масовість і причинна зумовленість явищ. Згідно з цими принципами закони розвитку природи виразно виявляються лише в досить численній сукупності подій. Об'єктивною основою існування статистичних закономірностей є складне переплетіння причин, які формують масовий процес, – основних, спільних для всіх подій масового процесу, та індивідуальних для кожної з них окремо, але випадкових для маси. У разі великої кількості подій вплив випадкових причин взаємно врівноважується, завдяки чому закон стає видимим.

Отже, статистичні закономірності притаманні лише сукупностям. Саме сукупність, а не окремий елемент є тією базою реального світу, відносно якої можна встановлювати конкретні закони.

Статистична сукупність – це певна множина елементів, поєднаних умовами існування й розвитку.

У реальному житті існує складне поєднання різних сукупностей та їх елементів. Базою вивчення конкретної статистичної закономірності є та сукупність, елементи якої – носії підпорядкованих цій закономірності характеристик.

Внаслідок об'єднання елементів у сукупність виникають якісно нові системні властивості. Вони відбивають спільність і відмінність, сталість і мінливість, повторюваність і неповторність властивостей, зв'язків і співвідношень елементів. Системні властивості становлять сутність статистичної закономірності. Відбиваючи характер дії об'єктивних законів розвитку природи в конкретних умовах простору і часу, статистичні закономірності виявляються по-різному.

Специфічна риса біометрії – узагальнення даних. Передумовою та початком такого узагальнення має бути вимірювання, тобто приписування явищу числових значень. Біометричним еквівалентом властивостей, притаманних елементам сукупності, є ***ознака.*** Кожний елемент сукупності характеризується багатьма ознаками, значення яких змінюються від елемента до елемента або від одного періоду до іншого. Ознака, яка набуває в межах сукупності різних значень, називається такою, що варіює, а відмінність, коливання значень ознаки – ***варіацією.*** Наприклад, ознаки людини: вік, стать, сімейний стан, освіта тощо.

Одні ознаки виражаються числами, інші – словесно. Їх називають відповідно ***кількісними*** **і** ***атрибутивними*** (описовими або якісними). Серед атрибутивних ознак одні чітко окреслені (стать, колір), інші невизначені (суб'єктивні оцінки, твердження, думки).

Ознаки мають різний рівень вимірювання, що відображується у відповідних типах шкал. Тип шкали можна визначити допустимими перетвореннями її чисел або допустимими арифметичними діями з цими числами. Згідно з класифікацією шкал за рівнем вимірювання – від «слабкої» до «сильної» – вирізняють три їх типи: номінальну, порядкову, метричну. Чим вищий рівень шкали, тим ширше коло відповідних допустимих перетворень чисел, тим більше арифметичних дій реалізується.

Номінальна шкала – шкала найменувань. «Оцифрування» ознак цієї шкали виконується так, щоб подібним елементам відповідало одне й те саме число, а неподібним – різні числа. Очевидно, число відіграє роль символу. Для ідентифікації найменувань шкали використовуються натуральні числа 1, 2, 3, ... або певні числові коди.

Номінальні ознаки, які мають лише два протилежні значення (наприклад, чоловіча / жіноча стать), називають альтернативними. Їх ідентифікують числами «1» або «0» залежно від наявності чи відсутності властивості.

Порядкова (рангова) шкала встановлює не лише відношення подібності елементів, а й відношення послідовності – порядку.

Це відношення типу «більше ніж», «краще ніж» і т. ін. Кожній позначці шкали приписується число – ранг. Такими числами можуть бути: 1, 2, 3 ... *n; 0,* 25, 50, 75, 100; -2, –1,0, 1,2, тобто значення будь-якої монотонно зростаючої функції, що відповідають послідовності значень ознаки, не враховуючи відстань між ними.

Метрична шкала – це звичайна шкала дійсних чисел. За допомогою метричної шкали вимірюються натурально-речові явища, ресурси та результати експерименту. Вибір одиниці такої шкали залежить від природи, матеріального змісту явища, конкретних завдань дослідження та практичної доцільності. За характером варіації ознаки метричної шкали поділяються на дискретні та неперервні.

Дискретні ознаки мають лише окремі цілочислові значення: кількість пелюсток у квітці, кількість дітей у сім'ї тощо. ***Неперервні ознаки*** мають будь-які значення в певних межах варіації. Наприклад, вік людини в межах від 0 до 100 і більше років. Таке визначення неперервної ознаки дещо умовне, її можна подати квазідискретною величиною (вік – числом виповнених років). До неперервних належать також розрахункові ознаки, а саме: народжуваність, урожайність, тощо.

Окремо взяті елементи будь-якої сукупності характеризуються практично необмеженим числом різних ознак. Які саме з цих ознак підлягають вимірюванню в конкретному дослідженні, залежить від його мети.

Оскільки біометрія вивчає масові процеси, індивідуальні значення ознак систематизуються, зводяться в єдине ціле. Узагальнюючою характеристикою явищ є ***біометричний показник.*** На відміну від ознак, які реєструються, показники розраховуються. Це може бути простий підсумок елементів сукупності або значень ознаки, результат порівняння двох величин або складніших розрахунків.

1.4. Біометрична методологія

Біометрична методологія – це комплекс спеціальних, притаманних лише біометрії методів і прийомів дослідження. Вона ґрунтується на загальнофілософських (діалектична логіка) і загальнонаукових (порівняння, аналіз, синтез) принципах.

Згідно з принципами діалектичної логіки біометрія будь-яке явище розглядає не ізольовано, а у взаємозв'язку з іншими, виявляє фактори, які спричинюють варіацію значень ознак у межах сукупності, оцінює ефекти впливу факторів і щільність причинно-наслідкових зв'язків.

Біологічні явища динамічні, тому біометрія вивчає їх у розвитку, оцінюючи тенденції та циклічні коливання, інтенсивність динаміки та структурних зрушень.

Розглядаючи сукупності елементів, біометрія, з одного боку, визначає в них схожі риси і відмінності, об'єднує елементи в групи, вирізняючи окремі типи й форми явищ, а з іншого – узагальнює інформацію як за окремими групами (типами), так і за сукупністю в цілому.

Особливості методології пов'язані, по-перше, з точним вимірюванням і кількісним описуванням масових явищ; по-друге, з використанням узагальнюючих показників для характеристики об'єктивних закономірностей.

Будь-яке біометричне дослідження послідовно проходить три етапи. Перший етап – збирання первинного матеріалу, реєстрацією фактів чи опитуванням респондентів. На другому етапі зібрані дані підлягають систематизації та групуванню – від характеристики окремих елементів переходять до узагальнюючих показників у формі абсолютних, відносних чи середніх величин. Третій етап передбачає аналіз варіації, динаміки, взаємозв'язків.

Етапи об'єднуються метою дослідження. На кожному з них застосовуються ті методи, які можуть дати глибоку й всебічну характеристику явищ, що вивчаються.

На другому етапі – етапі узагальнення даних – елементи сукупності класифікують за певними ознаками, наприклад, народжених можна класифікувати за статтю та місцем народження. Впорядковану таким чином сукупність називають ***варіаційним рядом.*** Залежно від способу класифікації розрізняють ***ряди розподілу*** **та** ***ряди динаміки. Ряд розподілу*** – це результат класифікації, ***групування*** елементів сукупності (станом на певний момент чи за певний інтервал часу). За допомогою групувань виокремлюються характерні властивості та різноякісні типи явищ. ***Ряд динаміки*** класифікує значення показників у часі (за періодами чи моментами часу), описує динаміку розвитку масового процесу.

В арсеналі біометричних методів аналізу – методи вивчення варіації, диференціації та сталості, швидкості та інтенсивності розвитку, узагальнюючі індекси, регресійні моделі тощо. Вивчаючи різноманітні явища та процеси, біометричний метод пристосовується до їх особливостей. Одна річ, скажімо, збирання даних про демографічні процеси (народжуваність, смертність, міграцію), інша – про екологічний стан довкілля. Але в будь-якому дослідженні виявляються притаманні цьому методу особливості – масовість даних, кількісне вимірювання, узагальнення.

Аналітичні можливості біометричних методів поглиблюються завдяки використанню компактної та раціональної форми подання результатів узагальнення інформації, а також аналізу виявлених закономірностей. Такими формами є ***таблиці та графіки.***

Передумовою використання біометричних методів у конкретному дослідженні має бути визначення суті явища, що вивчається, його істотних властивостей. ***Теоретичний аналіз*** дає всебічне уявлення про природу й логіку предмета пізнання. Це – об'єктивна основа методологічних рішень.

2. ЕЛЕМЕНТИ ТЕОРІЇ ЙМОВІРНОСТЕЙ

1. Подія. Всі можливі результати випробовувань називаються елементарними подіями ω. Сукупність всіх елементарних подій $\Omega=\{\omega\}$ називається множиною (або простором) елементарних подій.

Для спрощення будемо вважати, що простір Ω скінчений або облікований. В цьому випадку подією називається люба підмножина множина елементарних подій Ω . Кажуть, що в результаті дослідження здійснилася подія A, якщо здійснилася елементарна подія $\omega \in A$. Нехай множина $\varnothing$ є неможливою подією Φ, яка при дослідженні здійснитися ніколи не може. Сама множина

елементарних подій є достовірною подією **Ω** , яка здійсниться внаслідок кожного випробовування, тому, що результат кожного дослідження є елементарною подією. Інші підмножини простору елементарних подій **Ω** називаються випадковими подіями **A, B, C, ...** В результаті дослідження випадкова подія може або здійснюватися, або не здійснюватися, при цьому перед дослідженням результат не визначений.

2. Відношення між подіями. Відношення між подіями можна розглядати як відношення між відповідними підмножинами множини елементарних подій **Ω**.

1) **Наступність**: **A⊂B**, якщо із здійсненням події **A** випливає здійснення події **B**, або подія **A** викликає за собою подію **B**.
2) **Рівність**: **A=B**, якщо одночасно виконуються наступні умови: **A⊂B** і **B⊂A**.
3) **Сума**: **A∪B** виконується тоді, коли відбувається хоч одна з цих подій
4) **Множення**: **A∩B** виконується тоді, коли виконуються обидві події (і **A**, і **B**).
5) **Різниця**: **A\B** виконується тоді, коли подія **A** відбувається, а подія **B** не відбувається.
6) **Протилежність**. Протилежна подія $\overline{A} = \Omega \setminus A$ здійснюється тільки тоді, коли не відбувається подія **A**. Вірне і наступне твердження $\overline{\overline{A}} = A$.
7) **Сумісність**. Події називаються сумісними, якщо вони можуть одночасно відбуватися, і несумісними, якщо при здійсненні однієї події інша відбутися не може, тобто вони не можуть виникнути одночасно.

3. Система подій. Множина подій **$\{A_1, A_2, ..., A_n\}$** називається системою подій **A** , якщо ця множина не містить ні однієї пари рівних подій. Система подій **A = $\{A_1, A_2, ..., A_n\}$** називається системою несумісних подій, якщо всі події у системі попарно несумісні. Система несумісних подій називається повною, якщо сума всіх подій системи є достовірною подією **Ω**.

4. Класична ймовірність. Міра можливості настання родій називається його ймовірністю. Класична ймовірність події **A** визначається як відношення кількості елементарних подій **m,** що входять до складу події **A,** до кількості всіх можливих елементарних подій **n :**

$$P(A) = \frac{m}{n} \qquad \textbf{(1)}$$

Ймовірність неможливої події **Ф** дорівнює нулю: **P(Ф)=0,** ймовірність достовірної події **Ω** дорівнює одиниці: **P(Ω)=1**, а ймовірність довільної випадкової події знаходиться між 0 і 1: **0<P(A)<1**.

Ймовірність суми подій **A∪B** визначається наступною формулою:

$$\mathbf{P(A \cup B) = P(A) + P(B) - P(A \cup B)} \qquad \textbf{(2)}$$

Якщо обидві події несумісні то формула (2) спрощується й приймає вигляд:

$$P(A \cup B) = P(A) + P(B) \qquad \textbf{(3)}$$

Цю формулу можна узагальнити на випадок суми любої кількості несумісних подій. Враховуючи, що $A \subset \overline{A} = \Omega$, і приймаючи до уваги формулу (3) отримуємо:

$$P(A) + P(\overline{A}) = 1 \qquad \textbf{(4)}$$

Події називають незалежними, якщо ймовірність настання однієї події не залежить від того, відбулась чи ні друга подія При залежних подіях **A** і **B** є зміст говорити про умовну ймовірність **P(A|B)** Події **A,** при умові, що подія **B** вже відбулась. При незалежних подіях умовна ймовірність дорівнює звичайній ймовірності **P(A|B)=P(A)**

Ймовірність множення подій $A \cap B$ визначається за наступними формулами:

$$P(A \cap B) = P(A|B) \bullet P(B) \text{ або } P(A \cap B) = P(B|A) \bullet P(A) \qquad \textbf{(5)}$$

Якщо події незалежні то формула (5) спрощується і приймає вигляд:

$$P(A \cap B) = P(A) \bullet P(B) \qquad \textbf{(6)}$$

Цю формулу можна узагальнити на випадок множення любої кількості незалежних подій.

Якщо подія **A** залежить від подій (гіпотез) повної системи подій $\mathrm{B} = \{\mathbf{B_1, B_2, ..., B_n}\}$, то ймовірність події **A** вираховується за формулою повної ймовірності:

$$P(A) = \sum_{i=1}^{n} P(A|B_i) \bullet P(B_i) \qquad \textbf{(7)}$$

В цьому випадку разом з здійсненням події **A** відбувається одна і тільки одна подія із системи B.

Якщо подія **A** відбулась, то можна вирахувати умовну ймовірність того, що разом з подією **A** здійснилась гіпотеза $\mathbf{B_i}$:

$$P(B_i|A) = \frac{P(B_i) \bullet P(A|B_i)}{P(A)} \qquad \textbf{(8)}$$

де **P(A)** - повна ймовірність події **A.** Отриману формулу називають формулою Бейеса. За допомогою формули Бейеса можна після дослідження уточнити ймовірність виникнення гіпотези $\mathbf{B_i}$. Сума ймовірностей гіпотез $\mathbf{B_i}$ має бути рівна 1.

5.- Поняття комбінаторики. При вирішенні задач теорії ймовірностей часто використовують наступні поняття комбінаторики: перестановка, сполучення та розміщення, а також правило множення та правило додавання.

Нехай дана множина N з **n** об'єктів. Нерізноманітні послідовності з усіх **n** об'єктів називають **перестановками.** Загальна кількість $\mathbf{P_n}$ різних перестановок з **n** об'єктів вираховують за формулою:

$\mathbf{P_n = n!}$ **(9)**

де **n!= 1•2•... n,** при цьому рахують, що **0! = 1.**

Сполученнями називають підмножини множини N. Загальна кількість різних сполучень C_n^m з **n** об'єктів по **m** вираховують за формулою:

$$C_n^m = \frac{n!}{m!(n-m)!} \qquad \textbf{(10)}$$

Розміщеннями називають впорядковані послідовності об'єктів підмножини множини N. Загальна кількість різних розміщень A_n^m з **n** об'єктів по **m** вираховують за формулою:

$$A_n^m = \frac{n!}{(n-m)!} = n(n-1)\ldots(n-m+1) \qquad \textbf{(11)}$$

Правило множення. Якщо треба виконати одне за іншим якісь K дій, які можна виконати відповідно $\mathbf{n_1, n_2, ..., n_k}$ способами, то всі k дій можуть бути виконані $\mathbf{n_1 \bullet n_2 \bullet ... \bullet n_k}$ способами.

Правило додавання. Якщо дві взаємо виключаючи одна одну дії можуть виконуватися відповідно **m** або **n** способами, то виконати одну з цих дій можна **m+n** способами.

3. СТАТИСТИЧНЕ СПОСТЕРЕЖЕННЯ

3.1. Статистичного спостереження

Любе дослідження в біології потребує чіткої спланованості і від першого етапу – спостереження – в значній мірі залежить подальший успіх подальшої роботи. ***Статистичне спостереження —*** це спланована, науково організована реєстрація масових даних про будь-які явища та процеси в природі. Від інших методів збирання даних біометричне спостереження відрізняється характером і масовістю даних та способами їх отримання. Крім безпосередньої реєстрації (вимірювання, підрахунок, оцінювання) широко застосовується вивчення суспільної думки на підставі опитування.

Залежно від рівня реєстрації, біометричні спостереження можуть бути первинними або вторинними. ***Первинне спостереження —*** це реєстрація вихідних даних, що надходять від об'єкта, який їх продукує. Прикладом може бути поточний облік кількості зоопланктону у водоймі; вимірювання маси тіла миші тощо. ***Вторинне спостереження*** — це збирання раніше зареєстрованих та оброблених даних, наприклад матеріалів польових звітів, результатів спостережень, гербарних зборів тощо.

Біометричні дані обов'язково мають кількісну визначеність, завдяки чому підлягають нагромадженню, зведенню та узагальненню. Останнє можливе лише за умов їх системності.

Зрозуміло, що від якості даних біометричного спостереження залежать результати подальшого дослідження. Тому вони мають відповідати певним вимогам.

Перша — це ***вірогідність даних,*** тобто їх відповідність реальному стану.

Друга вимога — це ***повнота даних*** як за їх обсягом, так і по суті.

Третя вимога — це ***своєчасність даних.*** Інформація має дійти до користувача, перш ніж застаріє, інакше вона втрачає корисність.

Четверта вимога — ця ***порівнянність даних*** у часі або у просторі. Дані можуть бути не порівнянні ***за складом сукупності.*** Гострою є проблема ***порівнянності за одиницями, які взято для вимірювання.***

П'ятою вимогою є ***доступність даних.***

Біометричне спостереження здійснюється в три етапи:

1) підготовка спостереження;

2) реєстрація даних;

3) формування бази даних.

Підготовка біометричного спостереження — найвідповідальніший етап, оскільки тут постають і вирішуються основні ***методологічні питання:*** що і як вивчатиметься, на які запитання потрібно дістати відповіді. На цьому етапі вирішуються також ***організаційні питання:*** хто, де, коли проводить спостереження і що для цього необхідно. Тобто на першому етапі складається докладний ***план спостереження,*** що охоплює методологічні та організаційні питання.

На другому етапі збирають дані. Від якості збирання залежать також точність, повнота й вірогідність даних. Третій етап передбачає контроль та нагромадження даних спостереження, а також їх збереження. На цьому етапі відпрацьовується система оперативного доступу та пошуку необхідних даних.

3.2. Програмно-методологічні питання біометричного спостереження

Розробка ***програмно-методологічних питань*** плану спостереження полягає в науково-практичному обґрунтуванні та визначенні суті явища, умов його формування та прояву. Крім того, добирається система ознак, що характеризують явище, ураховується можливість їх кількісної обробки та перевірки на точність.

Комплекс програмно-методологічних питань може бути поданий у послідовності їх виникнення та розв'язування.

Мета спостереження — здобути дані, які є підставою для узагальненої характеристики стану та розвитку явища або процесу з визначенням відповідної закономірності.

Мета спостереження визначає його об'єкт. ***Об'єкт спостереження*** — це сукупність явищ, що підлягають обстеженню. Чітке визначення суті та меж об'єкта дозволяє запобігти розбіжностям у тлумаченні результатів обстеження. Для цього застосовуються цензи. ***Ценз*** — це набір кількісних обмежувальних ознак. Скажімо, слід обґрунтувати, яка сукупність об'єктів досліджуватиметься.

Об'єкт спостереження як сукупність складається з окремих елементів — одиниць сукупності. ***Одиниця сукупності*** — це первинний елемент об'єкта, що є носієм ознак, які підлягають реєстрації.

Проте не обов’язково кожна одиниця сукупності може бути врахована. Тому в ході обстеження виокремлюють одиницю спостереження. ***Одиниця спостереження*** — це первинна одиниця, від якої дістають інформацію.

Отже, одиниця сукупності та одиниця спостереження можуть збігатися (як, зокрема, для рідкісних видів).

Визначивши носії ознак і джерела інформації, складають програму спостереження. ***Програма спостереження*** — це перелік запитань, на які потрібно дістати відповіді в результаті спостереження. Зміст та кількість

запитань формують згідно з метою спостереження і реальними можливостями його проведення (грошовими, трудовими витратами та терміном реєстрації). Складають такий перелік запитань, який на підставі щонайменшої кількості залучених даних дає якнайбільше інформації. Цього можна досягти завдяки системному підходу. Запитання слід добирати так, щоб кожне мало самостійне призначення і водночас у комбінації з рештою запитань давало додаткову суміжну та побічну інформацію. Використовуючи блоки взаємозалежних узгоджених запитань, можна дослідити також внутрішні зв'язки об'єкта дослідження. Далі ці запитання деталізують, щоб сформувати набір ознак, які характеризують зазначений об'єкт обстеження та його одиниці.

Головними вимогами щодо подання ознак є простота формулювань, чітке тлумачення і однозначність. Оскільки ознаки за формою вираження можуть бути не лише кількісними, а й атрибутивними, то в разі добору шкали вимірювання перевагу слід надати не лише більш інформативним ознакам (номінальній шкалі), а й ознакам із ширшими можливостями статистичної обробки (порядковій та метричній шкалам). На підставі сформованого переліку ознак розробляють формуляр для проведення спостереження.

Формуляр **—** це обліковий документ єдиного зразка, що містить характеристику об'єкта спостереження та дані про нього. Формулярами є звіти, бланки документів, анкети тощо. Формуляри для спостереження можуть розроблятися дослідником самостійно або бути стандартними. До прикладу стандартних формулярів можна назвати бланки геоботанічних описів, карток фенологічного спостереження тощо.

Деякі стандартні формуляри містять також коротку інструкцію щодо їх заповнення. До інших інструкція додається як окремий документ, де роз'яснюється порядок реєстрації даних, розтлумачується зміст окремих запитань або відповідей. Проте чим вдаліше формулюються запитання та ознаки, тим менше роз'яснень вони потребують. Складаючи формуляри, застосовують запитання закритого, напівзакритого та відкритого типу. До ***закритого типу*** належать запитання, до яких додається повний набір відповідей, а респондентові пропонується лише вибрати потрібні. Іноді кількість можливих відповідей обмежується (не більше або не менше як дві-три відповіді). Наприклад, під час опису умов навколишнього середовища пропонуються запитання: «Охарактеризуйте умови місцезростання»:

1. Ліс.
2. Чагарники.
3. Узлісся.
4. Лука.

Запитання ***напівзакритого типу*** містять перелік готових відповідей, а також вільний рядок для заповнення.

Наприклад: «Які тварини переважають у раціоні тварини?».

1. Безхребетні.
2. Дрібні ссавці.
3. Травоїдні.
4. Інше _____ (впишіть).

Запитання ***відкритого типу*** передбачають, що дослідник самостійно формулює відповідь. До таких належить, наприклад, запитання: «Адміністративне розміщення регіону дослідження»

Зрозуміло, що в разі організації масових обстежень з залученням декількох дослідників доцільно пропонувати запитання закритого та напівзакритого типу. Вони дають змогу, по-перше, діставати адекватні відповіді, а по-друге, — формалізувати, нагромаджувати та кількісно їх обробляти. Причому перевагу мають напівзакриті запитання, що передбачають не лише чіткі відповіді, а й дають змогу респондентові вільно висловити власну думку.

Зауважимо, що формуляр має забезпечувати не лише вхідну, а й вихідну частину інформаційної бази спостереження. Тобто, визначаючи ознаки, складаючи блоки запитань, одночасно готують макети вихідних таблиць, де немає цифрової інформації. За макетами таблиць одразу можна визначити, наскільки кожне запитання (ознака) узгоджується з іншими, передбачити, як «працюватиме» те чи інше запитання (ознака), вилучити малоінформативні запитання, обґрунтувати методику подальшої статистичної обробки решти запитань.

Програма спостереження передбачає також визначення ***виду та способу реєстрації даних.*** Здебільшого вид і спосіб спостереження залежать від його мети, суті об'єкта спостереження, обсягу та ступеня точності очікуваних результатів. Важливими для обстеження є його фінансове та трудове забезпечення, а також фактор часу.

Неодмінною умовою програм спостереження є ***наступність*** їх змісту, одиниць спостереження, методик обчислення. Це потрібно для порівняння результатів спостереження в часі та просторі.

Готуючи спостереження, слід ***забезпечити точність даних реєстрації.*** Точність результатів досягається завдяки застосуванню, з одного боку, ***системи контролю,*** а з іншого— ретельно відпрацьованого механізму збирання даних і практичного досвіду в цій роботі. Такий досвід формується під час ***пробних,*** так званих ***пілотних*** обстежень — невеликих за обсягом, що мають на меті

випробувати, уточнити програму спостереження, підвищити якість опрацювання організаційних питань.

Отже, підготовка спостереження потребує вирішення не лише програмно-методологічних, а й суто практичних (організаційних) питань.

3.3. Організаційні питання біометричного спостереження

Другою складовою плану спостереження є комплекс організаційних питань. ***Організаційні питання*** стосуються місця й часу проведення обстеження, персоналу, залученого до обстеження, його матеріально-технічного забезпечення, а також системи гарантування точності результатів

Організаційні питання тісно пов'язані з програмно-методологічними й залежать від мети та умов обстеження.

Наступним питанням є обґрунтування ***місця обстеження*** — пункту, в якому перебуває одиниця спостереження і реєструються дані.

Час спостереження (об'єктивний час) — це час, до якого належать дані спостереження. Коли об'єктом спостереження є ***процес,*** то вибирають інтервал часу, протягом якого нагромаджуються дані. Якщо об'єктом спостереження є ***певний стан,*** то вибирають ***критичний момент*** — момент часу, станом на який реєструються дані.

Час спостереження вибирають у найсприятливіший або нейтральний для об'єкта спостереження період.

Контроль даних спостереження водночас є методологічним і організаційним питанням. Якщо його розглядати як засіб ***попередження помилок,*** то питання наближається до методологічного. Якщо контроль — це ***виявлення та виправлення помилок,*** то йдеться швидше про організаційний захід.

Контроль означає насамперед перевірку даних обстежень щодо їх повноти й вірогідності.

Повнота даних контролюється, як правило, візуально: перевіряють наявність даних за всіма одиницями та позиціями.

Дані на вірогідність перевіряють засобами логічного та арифметичного контролю.

Логічний контроль — це перевірка сумісності даних, яка полягає в порівнянні взаємозалежних ознак. Зрозуміло, що логічний контроль установлює лише наявність помилки, а не її розмір. Проте іноді вдається визначити наближені межі помилки, порівнюючи дані спостереження з аналогічними даними інших спостережень чи дані в динаміці.

Встановити розмір помилки та виправити її можна засобами ***арифметичного контролю,*** тобто прямим чи побічним перерахунком зареєстрованих даних.

Зауважимо, що здебільшого встановити вірогідність даних неможливо (коли помилки виходять за межі логічного та арифметичного контролю). Отже, доводиться враховувати природу помилок. Залежно від причини виникнення розрізняють такі помилки:

репрезентативності — виникають під час вибіркового спостереження через несуцільність реєстрації даних і порушення принципів випадковості відбору;

реєстрації — виникають у разі будь-якого спостереження внаслідок перекручення фактів або неправильного їх запису.

Залежно від природи виникнення ***помилки реєстрації можуть*** бути випадковими або систематичними.

Випадкові помилки виникають внаслідок збігу випадкових обставин — неуважності реєстратора або збою апаратури. Вони викривлюють дані спостереження в той чи інший бік, проте завдяки масовості розглядуваних випадків дія таких помилок урівноважується, і вони істотно не впливають на результати.

Більш небезпечними є ***систематичні помилки,*** які виникають у результаті постійних спотворень в одному напрямі. Такі помилки істотно зміщують результати спостереження в один бік (збільшення або зменшення). Наприклад відбір тільки гарних чи великих квіток.

Систематичні помилки бувають ***незловмисними*** та ***зловмисними. Незловмисні помилки*** виникають внаслідок необґрунтованості програми спостереження, некомпетентності дослідників, неосвіченості респондентів тощо.

Зловмисні помилки виникають через свідоме викривлення фактів з певною метою (очорнити або прикрасити дійсність, підтасувати данні з певною метою).

3.4. Форми, види та способи спостереження

З огляду на різноманітність сфер спостереження та багатогранність його аспектів застосовують різні організаційні форми, види і способи збирання даних.

За ступенем охоплення спостереження бувають суцільними та несуцільними.

Суцільні спостереження — це обстеження, під час яких реєструються всі без винятку одиниці сукупності.

Несуцільні спостереження — це обстеження, що мають на меті реєструвати не всі одиниці сукупності, а лише їх певну частину. До таких спостережень належать вибіркове, основного масиву, монографічне, анкетне, моніторинг.

Обліки — суцільні спостереження масових явищ, які ґрунтуються на даних огляду, опитування та документальних записів.

Прикладом може бути облік поголів'я худоби за видами, групами і категоріями господарств, а також облік земельного фонду за видами угідь, якістю ґрунту, категоріями господарств тощо.

Спеціальні обстеження — несуцільне спостереження окремих масових явищ згідно з певною тематикою. Вони можуть бути періодичними або одноразовими.

Опитування — це, як правило, несуцільне спостереження думок, мотивів, оцінок, що реєструються зі слів респондентів.

Вибіркове спостереження — це обстеження, під час якого реєструється деяка частина одиниць сукупності, відібрана у випадковому порядку.

Обстеження основного масиву — це обстеження переважної частини одиниць сукупності, що відіграють визначальну роль у характеристиці об'єкта спостереження.

Монографічне обстеження — це ретельне обстеження окремих типових одиниць сукупності з метою їх досконалого вивчення.

Анкетне спостереження — це обстеження певної частини одиниць сукупності внаслідок неповного повернення від респондентів заповнених реєстраційних формулярів (анкет).

Моніторинг — це спеціально організоване систематичне спостереження за станом певного середовища. Наприклад, моніторинг рівня радіаційного забруднення на територіях, що постраждали внаслідок аварії на ЧАЕС.

Спостереження за часом реєстрації фактів поділяються на поточне, періодичне та одноразове.

Поточне спостереження — це систематична реєстрація фактів щодо явищ, у міру їх виникнення або збирання фактів щодо безперервного процесу. Безперервними є демографічні процеси: народжуваність, смертність. До числа явищ, які реєструються в певні моменти, належать спостереження над перельотом птахів.

Періодичне спостереження проводиться через певні (як правило, рівні) проміжки часу.

Одноразове спостереження проводиться в міру виникнення потреби в дослідженні явища чи процесу.

Спостереження здійснюються трьома способами: безпосередній облік фактів, документальний облік, опитування.

Безпосередній облік— це обстеження, під час якого обліковець особисто реєструє факти підрахунком, вимірюванням, оцінюванням, оглядом.

Документальний облік — це обстеження, коли факти реєструють за даними, наведеними в документах первинного обліку.

Опитування може здійснюватись по-різному: експедиційним способом, самореєстрацією, кореспондентським та анкетним шляхом.

Експедиційний спосіб— це реєстрація фактів спеціально підготовленими обліковцями з одночасною перевіркою точності реєстрації.

Самореєстрація означає, що факти фіксують самі респонденти після попереднього інструктажу з боку реєстраторів-обліковців.

Кореспондентський спосіб— це реєстрація фактів про явища та процеси на місцях їх виникнення спеціально підготовленими особами та надсилання результатів до відповідних інстанцій.

Окремі види та способи спостереження можуть застосовуватись у комплексі, не виключаючи один одного, залежно від складності доступу до об'єкта спостереження, ступеня підготовленості персоналу до певного методу спостереження, сучасних досягнень щодо методології та організації спостережень.

4. ЗВЕДЕННЯ ТА ГРУПУВАННЯ ДАНИХ

4.1. Суть біометричного зведення

Перехід від одиничного до загального відбувається завдяки зведенню.

Суть ***зведення*** полягає в тому, що матеріали спостереження класифікують та агрегують. Елементи сукупності за певними ознаками об'єднують у групи, класи, типи, а інформацію про них агрегують як у межах груп, так і в цілому по сукупності. Основне завдання зведення — виявити типові риси та закономірності масових явищ чи процесів.

Зведення є основою подальшого аналізу інформації. За зведеними даними обчислюють узагальнюючі показники, виконують порівняльний аналіз, а також аналіз причин групових відмінностей, вивчають взаємозв'язки між ознаками.

Результати зведення подаються у формі таблиць, макети яких розроблюються разом з програмою обробки даних.

На практиці використовуються різні технологічні схеми комп'ютерної обробки первинних даних. Спільними для всіх є дві операції: кодування даних і

перенесення їх із документів на технічні носії інформації, наприклад на магнітні диски.

4.2. Класифікації та групування

Поділ сукупностей на групи, однорідні в тому чи іншому розумінні, пов'язаний з такими діями, як систематизація, типологія, класифікація, групування. Традиційно зазначений поділ виконують за такою схемою: із множини ознак, які описують явище, добирають розмежувальні, а потім сукупність поділяють на групи та підгрупи відповідно до значень цих ознак.

Головний принцип будь-якого поділу ґрунтується на двох положеннях:

1) в один клас, групу об'єднуються елементи певною мірою подібні між собою;

2) ступінь подібності між елементами, які належать до одного класу, значно вищий, ніж між елементами, що належать до різних класів.

У кожному конкретному дослідженні вирішуються три питання:

1) що взяти за основу групування;

2) скільки груп, позицій необхідно відокремити;

3) як розмежувати групи.

Основою групування може бути будь-яка атрибутивна чи кількісна ознака, що має градації. Таку ознаку називають ***групувальною.*** Залежно від складності масового явища (процесу) та мети дослідження групувальних ознак може бути одна, дві й більше.

У практиці часто вдаються до розбиття сукупностей за атрибутивними ознаками — ***класифікації*** та ***номенклатури.*** Здебільшого йдеться про багатоступеневу класифікацію з докладною номенклатурою груп і підгруп, із чітко визначеними вимогами та умовами віднесення елементів сукупності до тієї чи іншої групи.

Кожній класифікаційній позиції надається ***код*** (шифр), який замінює її назву і є постійним засобом ідентифікації під час передавання інформації по каналах зв'язку, комп'ютерної обробки тощо.

Кожна класифікація є сталою, забезпечуючи порівнянність даних у просторі та часі.

Поряд з класифікацією для висвітлення певних аспектів конкретного дослідження використовують ***групування,*** на яке покладаються такі аналітичні функції:

1) вивчення структури та структурних зрушень;

2) визначення типів явищ, виокремлення однорідних груп і підгруп;

3) виявлення взаємозв'язків між ознаками.

Згідно з цими функціями розрізняють три види групувань: структурне, типологічне, аналітичне.

Структурне групування характеризує склад однорідної сукупності за певними ознаками. Наприклад, склад населення регіону за місцем проживання.

Типологічне групування — це поділ якісно неоднорідної сукупності на класи, типи, однорідні групи. Основне завдання такого групування — ідентифікація типів. Вибір групувальної ознаки та кількісних міжгрупових меж ґрунтується на всебічному теоретичному аналізі суті явища, його характерних рис та особливостей формування в конкретних умовах часу та простору.

Скориставшись групуванням, можна також виявити наявність та напрям зв'язку між ознаками, з яких одна розглядається як результат, інша — як фактор, що впливає на результат. Висновок про наявність зв'язку можна зробити на підставі комбінаційного поділу за цими ознаками згідно з характером розміщення частот.

Якщо результативна ознака кількісна, для кожної групи за факторною ознакою можна визначити середнє значення результативної ознаки. За наявності зв'язку між ознаками групові середні результативної ознаки систематично змінюються від групи до групи (збільшуються чи зменшуються). Таке групування називається ***аналітичним.*** Очевидно, аналітичне групування докладніше й виразніше, ніж комбінаційний поділ, описує зв'язок між ознаками.

Зауважимо, що поділ групувань на три види певною мірою відносний. Адже часто групування універсальні: одночасно виділяються типи, визначається склад сукупності й виявляється взаємозв'язок між ознаками.

4.3. Принципи формування груп

Якщо групувальна ознака неперервна, постає питання про кількість груп та межі кожної з них. Кількість груп залежить від ступеня варіації групувальної ознаки та обсягу сукупності. Так, для дискретної ознаки, діапазон варіації якої обмежений (кількість дітей у сім'ї, стать тощо), груп, як правило, стільки, скільки варіант ознаки. У разі значної варіації дискретної ознаки (кількість листків на дереві, кількість бактерій), як і неперервної (маса тіла, довжина пагона), діапазон варіації розбивається на m інтервалів.

Орієнтовно оптимальна кількість груп визначається за стандартними процедурами, зокрема за формулою Стерджеса:

$m = 1 + 2{,}30259 \lg n,$

де n — обсяг сукупності; m — число інтервалів.

Якщо сукупність містить велику кількість спостережень ($n>100$) можна використовувати формулу $m = 5 \lg n$.

Для практичного використання для визначення кількості груп зручно користуватися наступною таблицею.

Кількість спостережень n	Кількість груп m
25 — 40	5 — 6
40 — 60	6 — 8
60 — 100	7 — 10
100 — 200	8 — 12
> 200	10 — 15

Інтервали являють собою каркас групувань. На практиці їх утворюють за трьома формальними принципами: ***рівності інтервалів; кратності інтервалів; рівності частот.***

У структурних і аналітичних групуваннях найчастіше застосовують принцип рівності інтервалів. Ширина кожного інтервалу залежить від діапазону варіації ознаки x та обґрунтованого числа груп (інтервалів) m:

$$h = \frac{x_{\max} - x_{\min}}{m},$$

Визначаючи межі інтервалів, ширину h доцільно округлювати, самі межі слід позначати з такою точністю, щоб поділ елементів сукупності на групи був однозначним.

Якщо діапазон варіації ознаки надто широкий і поділ значень нерівномірний, беруть нерівні інтервали, зокрема сформовані за принципом кратності, коли ширина кожного наступного інтервалу в k раз більша (менша), ніж попереднього.

Перший та останній інтервали (або один із них) відкриті, тобто мають лише одну межу (верхню чи нижню). За допомогою відкритих інтервалів усі крайні значення ознаки, що варіює, зводяться в одну групу, завдяки чому групування стає компактним.

Інтервали типологічного групування формуються не за математичними принципами, а за якісним змістом. Межа інтервалу розглядається як умовна межа переходу кількості в нову якість. Число груп залежить від кількості існуючих типів.

Принцип рівних частот використовують нечасто і переважно в аналітичних групуваннях, щоб уникнути зважування групових середніх (дисперсійний аналіз результатів експерименту).

Групування за однією ознакою називається ***простим,*** за двома і більше ознаками — ***комбінаційним.*** У комбінаційних групуваннях ознаки ієрархічно впорядковуються за змістом чи за вагомістю.

Групи, утворені за першою ознакою, поділяються на підгрупи за другою, а ті, у свою чергу, можуть поділятися на підгрупи за третьою ознакою і т. д. На кожному етапі поділу використовується лише одна ознака, тобто відбувається послідовне описування груп. Кількість підгруп дорівнює добутку числа групувальних ознак на число градацій за кожною з них. У разі трьох і більше групувальних ознак сукупність стрімко подрібнюється, групи виявляються нечисленними, а характеристики груп — ненадійними.

Альтернативою комбінаційному групуванню є багатовимірне, коли групи утворюються за певною множиною ознак одночасно. Мірою подібності елементів є різні критерії і, як наслідок, — різні методи багатовимірного групування. Найпростішим серед них є групування за інтегральним показником, наприклад за рейтинговою оцінкою. У такому разі багатовимірне групування зводиться до простого.

Іноді доводиться перегруповувати дані, передусім щоб забезпечити порівнянність структур двох сукупностей за однією і тією самою ознакою. Результат перегрупування називають ***вторинним групуванням.*** Перегрупування виконують або об'єднанням, або розбиттям інтервалів первинного групування.

Якщо межі інтервалів первинного і вторинного групувань збігаються, частоти (частки) об'єднувальних інтервалів просто підсумовуються. Коли виконується розбиття інтервалу первинного групування, частоти поділяються між новоутвореними групами пропорційно до співвідношення частин довжини початкового інтервалу. Припускається, що всередині інтервалу поділ рівномірний.

Перегрупуванням даних можна перейти від структурного групування до типологічного.

5. СТАТИСТИЧНІ ПОКАЗНИКИ

5.1. Суть і види статистичних показників

Інформація про розміри, пропорції, зміни в часі, інші закономірності створюється, передається і зберігається у вигляді статистичних показників.

Статистичний показник — це міра, тобто єдність якісного і кількісного відображення певної властивості явища чи процесу.

Якісний зміст показника визначається суттю явища і відбивається в його назві: народжуваність, урожайність тощо. Кількісна сторона подається числом та його вимірником. Оскільки біометрія вивчає явища в природі в конкретних умовах простору і часу, значення будь-якого показника визначається щодо цих атрибутів.

Показники розрізняють за способом обчислення, ознакою часу та аналітичними функціями.

За способом обчислення розглядають ***первинні*** **і** ***похідні*** показники. ***Первинні*** визначаються зведенням даних спостереження й подаються у формі абсолютних величин (кількість і вага особин популяції). ***Похідні*** показники обчислюються на базі первинних або похідних показників. Вони мають форму середніх або відносних величин (середня маса тіла, індекс приросту популяції).

За ознакою часу показники поділяються на інтервальні та моментні. ***Інтервальні*** характеризують явище за певний час (день, декаду, місяць, рік). До ***моментних*** відносять показники, що дають кількісну характеристику явищ на певний момент часу: площа виноградних і цитрусових насаджень тощо. Інтервальні та моментні показники можуть бути як первинними, так і похідними. Наприклад, площа зрошуваних земель — первинний моментний показник, а частка таких земель у загальній площі — похідний моментний показник.

Інтервальні показники залежать від довжини періоду, за який вони обчислюються. Особливістю первинних інтервальних показників є ***адитивність,*** тобто можливість підсумовування. Похідні показники здебільшого неадитивні.

Серед біометричних показників існують пари взаємообернених показників, які паралельно характеризують одне й те саме явище. Прямий показник *x* зростає з підсиленням явища, обернений *1/x,* навпаки, зменшується.

5.2. Абсолютні величини

Абсолютні величини характеризують розміри явищ. Йдеться про обсяги сукупності чи окремих її частин (кількість елементів) та відповідні їм обсяги значень ознаки.

Абсолютні величини являють собою іменовані числа, тобто кожна з них має свою ***одиницю вимірювання:*** штуки, тонни, кіловати, тощо. Натуральні вимірники відбивають притаманні явищам фізичні властивості.

5.3. Відносні величини

Абсолютні величини мають незаперечне значення, проте поглиблений аналіз фактів потребує різного роду порівнянь. Порівнюються значення показників у часі (за одним об'єктом), у просторі (між об'єктами), співвідносяться різні ознаки одного й того самого об'єкта.

Результатом порівняння є ***відносна величина,*** яка характеризує міру кількісного співвідношення різнойменних чи однойменних показників.

Кожна відносна величина являє собою дріб, чисельником якого є порівнювана величина, а знаменником — ***база порівняння.*** Відносна величина показує, у скільки разів порівнювана величина перевищує базисну або яку частку перша становить щодо другої, іноді — скільки одиниць однієї величини припадає на 100, на 1000 і т. д. одиниць іншої, базисної величини.

Різноманітність співвідношень і пропорцій реального життя для свого відображення потребує різних за змістом відносних величин. Відповідно до аналітичних функцій відносні величини можна класифікувати так:

А. Відношення однойменних показників

1) відносні величини динаміки;
2) відносні величини просторових порівнянь;
3) відносні величини порівняння зі стандартом;
4) відносні величини структури;
5) відносні величини координації.

Б. Відношення різнойменних показників

6) відносні величини інтенсивності.

Відносні величини динаміки

Динамікою називається зміна явища в часі, а ***відносна величина динаміки*** характеризує напрям та інтенсивність зміни. Відносні величини динаміки визначаються співвідношенням значень показника за два періоди чи моменти часу. При цьому базою порівняння може бути або попередній або більш віддалений у часі рівень.

Якщо значення показника зменшується, відносна величина динаміки буде меншою за одиницю.

Передумовою обчислення відносних величин динаміки є порівнянність даних за одиницею, методикою розрахунку показника, масштабом об'єкта.

Відносні величини просторових порівнянь

Найчастіше це регіональні чи міжнародні порівняння показників економічного розвитку або життєвого рівня. Вибір бази порівняння довільний.

Головне, щоб методика розрахунку показників, що порівнюються, була однаковою.

Наприклад, на початку 90-х років в Україні з 1 га ріллі отримували продукції вдвічі менше, ніж у США.

Відносні величини порівняння зі стандартом

Важливу роль у статистичному аналізі відіграє порівняння ***фактичних значень показника з певним еталоном — нормативом, стандартом, оптимальним рівнем.*** Будь-яке відхилення відносної величини від 1 чи 100% свідчить про порушення оптимальності процесу.

Для показників, які не мають визначеного еталона (захворюваність, тощо), базою порівняння може бути максимальне чи мінімальне значення або середня по сукупності в цілому.

Відносні величини структури

Статистичні сукупності структуровані, у них завжди можна виявити певні складові. ***Відносні величини структури*** характеризують склад, структуру сукупності за тією чи іншою ознакою. Вони визначаються відношенням розмірів складових частин сукупності до загального підсумку. Скільки складових, стільки відносних величин структури. Кожну з них окремо називають ***часткою,*** або питомою вагою, виражають простим чи десятковим дробом або процентом. Наприклад, певна частина сукупності становить 1/4, або 0,25, або 25% загального обсягу сукупності.

Відносні величини структури адитивні. Сума всіх часток дорівнює одиниці. Якщо частку *j*-ї складової сукупності позначити d_j, то $\Sigma d_j=1$ або, у процентах, $100\ \Sigma d_j= 100\%$.

За допомогою відносних величин структури можна оцінити структурні зрушення, тобто зміни у складі сукупності за певний період часу. Така оцінка ґрунтується на порівнянні часток *dj* за два періоди. Аналогічно можна порівняти структуру різних за обсягом сукупностей. Різницю між відповідними частками двох сукупностей називають ***процентним пунктом*** (п.п.).

Відносні величини координації

Поглиблений аналіз структури передбачає оцінювання співвідношень, пропорцій між окремими складовими одного цілого. Такий різновид порівнянь називають ***відносною величиною координації.*** Вона показує, скільки одиниць однієї частини сукупності припадає на 1, 100 і 1000 одиниць іншої, узятої за базу порівняння.

Відносні величини інтенсивності

Особливим видом відносних показників є результат порівняння різнойменних абсолютних величин: у чисельнику — обсяги певного явища (кількість подій, фактів), у знаменнику — обсяг середовища, якому це явище (подія) властиве. У кожному конкретному випадку таке співвідношення характеризує ***інтенсивність поширення явища в середовищі,*** а тому називається ***відносною величиною інтенсивності.*** На відміну від відносних величин групи А відносні величини інтенсивності іменовані одиницями вимірювання чисельника і знаменника співвідношення. Наприклад, щільність популяції — особин на 1 км^2, ефективність фотосинтезу— Дж на 1 м^3 тощо.

У порівняльному аналізі використовуються кратні співвідношення не лише абсолютних величин. Комплексна й всебічна характеристика закономірностей передбачає порівняння середніх і відносних величин.

5.4. Середні величини

Середня величина є узагальнюючою мірою ознаки, що варіює, у статистичній сукупності. Показник у формі середньої характеризує рівень ознаки в розрахунку на одиницю сукупності. Як уже зазначалося, значення ознаки *j*-го елемента поєднує в собі як спільні для всієї сукупності типові риси, так і притаманні лише цьому елементу індивідуальні особливості. Абстрагуючись від індивідуальних особливостей окремих елементів, можна виявити те загальне, типове, що властиве всій сукупності.

Саме в середній взаємно компенсуються індивідуальні відмінності елементів та узагальнюються типові риси. Типовість середньої пов'язана з однорідністю сукупності. Середня характеризуватиме типовий рівень лише за умови, що сукупність якісно однорідна. У неоднорідній сукупності, за влучним висловом П. Самуельсона, осереднюються «тигри та кицьки», що лише створює ілюзію «благодаті» і не віддзеркалює реалій.

Взаємозв'язок індивідуальних значень ознаки та середньої — це діалектична єдність загального і окремого. Замінюючи множину індивідуальних значень, середня не змінює ***визначальної властивості*** сукупності — загального обсягу явища. Зв'язок визначальної властивості з елементами сукупності описується функцією $f(x_1, x_2, \ldots x_n)$, яка виражає певну математичну дію над емпіричними значеннями ознаки (підсумовування, множення, степенювання, коренювання) і визначає вид середньої. Так, у разі підсумовування значень ознаки визначальну властивість забезпечує середня арифметична, при множенні — середня геометрична і т. д.

Отже, при обчисленні середніх у біологічних дослідженнях необхідно чітко усвідомити визначальну властивість сукупності та логіко-математичну суть —***логічну формулу*** — показника.

$$M = \sqrt[k]{\frac{\sum x_i^k}{n}}$$

Чисельник логічної формули середньої являє собою обсяг значень (визначальну властивість) ознаки, що варіює, а знаменник — обсяг сукупності. Як правило, визначальна властивість — це реальна абсолютна чи відносна величина, яка має самостійне значення в аналізі. У кожному конкретному випадку для реалізації логічної формули використовується певний вид середньої, зокрема:

а) середня арифметична (k=1);

б) середня гармонічна(k=-1);

в) середня геометрична;

г) середня квадратична (k=2) і т. д.

Залежно від характеру первинної інформації середня будь-якого виду може бути простою чи зваженою. Позначається середня символом *x* (риска над символом означає осереднення індивідуальних значень) і вимірюється в тих самих одиницях, що й ознака.

Середня арифметична

Оскільки для більшості явищ характерна адитивність обсягів, то найпоширенішою є арифметична середня, яка обчислюється діленням загального обсягу значень ознаки на обсяг сукупності. За первинними, незгрупованими даними обчислюється ***середня арифметична проста:***

$$\bar{x} = \frac{\sum_1^n x_i}{n}.$$

Моментні показники замінюються середніми як півсума значень на початок і кінець періоду. Якщо моментів більш ніж два, а інтервали часу між ними рівні, то в чисельнику до півсуми крайніх значень додають усі проміжні, а знаменником є число інтервалів, яке на одиницю менше від числа значень ознаки. Таку формулу називають ***середньою хронологічною:***

$$\bar{x} = \frac{\frac{x_1 + x_n}{2} + x_2 + x_3 + ... + x_{n-1}}{n-1}.$$

У великих за обсягом сукупностях окремі значення ознаки (варіанти) можуть повторюватись. У такому разі їх можна об'єднати в групи (j = 1, 2, ..., m), а обсяг значень ознаки визначити як суму добутків варіант x_j на відповідні їм частоти f_j, тобто **як** $\sum_{j}^{m} x_j f_j$. Такий процес множення у статистиці називають ***зважуванням,*** *а* число елементів сукупності з однаковими варіантами — ***вагами.*** Сама назва «ваги» відбиває факт різновагомості окремих варіант. Значення ознаки осереднюються за формулою ***середньої арифметичної зваженої:***

$$\bar{x} = \frac{\sum_{1}^{m} x_j f_j}{\sum_{1}^{m} f_j}.$$

Вагами можуть бути частоти або частки (відносні величини структури), іноді інші величини (абсолютні показники).

Формально між середньою арифметичною простою і середньою арифметичною зваженою немає принципових відмінностей. Адже багаторазове (f раз) підсумовування значень однієї варіанти замінюється множенням варіант x на вагу f. Проте функціонально середня зважена більш навантажена, оскільки враховує поширеність, повторюваність кожної варіанти і певною мірою відображує склад сукупності. Значення середньої зваженої залежить не лише від значень варіант, а й від структури сукупності. Чим більшу вагу мають малі значення ознаки, тим менша середня, і навпаки. Наприклад, незважаючи на той факт, що в двох регіонах мешкають люди різного віку, у тому регіоні, де більше дітей, середній вік населення буде менший. На цю властивість середніх слід зважати при використанні їх у порівняльному аналізі сукупностей, склад яких істотно різний. У таких випадках, аби елімінувати (усунути) вплив структури сукупності на середню, вдаються до пошуку стандартизованих ваг.

У структурованій сукупності при розрахунку середньої зваженої варіантами можуть бути як окремі значення ознаки, так і ***групові середні*** $\bar{x}_j$, кожна з яких має відповідну вагу у вигляді групових частот f_j:

$$\bar{x} = \frac{\sum_{1}^{m} \bar{x}_j f_j}{\sum_{1}^{m} f_j}$$

Обчислену так середню на відміну від групових називають ***загальною.***

Середня арифметична має певні властивості, які розкривають її суть.

1. Алгебраїчна сума відхилень окремих варіант ознаки від середньої дорівнює нулю:

$$\sum_{1}^{m}\left(x_j - \bar{x}\right)f_j = 0$$

тобто в середній взаємно компенсуються додатні та від'ємні відхилення окремих варіант.

2. Якщо всі варіанти збільшити (зменшити) на одну й ту саму величину *A* або в *A* раз, то й середня зміниться аналогічно.

3. Значення середньої залежить не від абсолютних значень ваг, а від пропорцій між ними. При пропорційній зміні всіх ваг середня не зміниться. Згідно з цією властивістю замість абсолютних ваг – частот f_j – можна використати відносні ваги у вигляді часток $d_j = \dfrac{f_j}{\sum_{1}^{m} f_j}$, або процентів $100d_j$:

$\bar{x} = \sum_{1}^{m} x_j d_j$, або у процентах, $\bar{x} = \dfrac{\sum_{1}^{m} x_j d_j}{100}$.

Середня гармонічна

При розрахунку середньої з обернених показників використовують середню гармонічну. ***Середню гармонічну просту*** обраховують за наступною формулою:

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

Середню гармонічну зважену обраховують за наступною формулою:

$$\bar{x} = \frac{\sum_{1}^{m} Z_j}{\sum_{1}^{m}\frac{1}{x}Z_j}$$

Середня геометрична

Якщо визначальна властивість сукупності формується як добуток індивідуальних значень ознаки, використовується середня геометрична:

$$\bar{x} = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \ldots \cdot x_n} = \sqrt[n]{\prod_{1}^{n} x_j},$$

де П — символ добутку; x_j — відносні величини динаміки, виражені кратним відношенням *j*-го значення показника до попереднього (*j*- 1)-го.

Коли часові інтервали не однакові, розрахунок виконують за формулою ***середньої геометричної зваженої:***

$$\bar{x} = \sqrt[n]{\prod_{1}^{m} x_j^{n_j}},$$

де *n;* — часовий інтервал, $\sum n_j = n, m$ — кількість інтервалів.

Середня квадратична

Головною сферою застосування середньої квадратичної є вимірювання варіації та площ.

$$\bar{x}_q = \sqrt{\frac{\sum x_i^2}{n}}, \text{ або } \bar{x}_q = \sqrt{\frac{\sum \left(x_i^2\right) n_i}{n}};$$

Середня кубічна

Середнє кубічне застосовується для характеристики ознак, що є об'ємом.

$$\bar{x}_Q = \sqrt[3]{\frac{\sum x_i^3}{n}}, \text{ або } \bar{x}_Q = \sqrt[3]{\sum \frac{\left(x_i^3\right) n_i}{n}};$$

6. РЯДИ РОЗПОДІЛУ. АНАЛІЗ ВАРІАЦІЙ ТА ФОРМИ РОЗПОДІЛУ

6.1. Закономірність розподілу

Статистична сукупність формується під впливом причин та умов, з одного боку — типових, спільних для всіх елементів сукупності, а з іншого — випадкових, індивідуальних. Ці чинники взаємозв'язані, а їх спільна взаємодія визначає як індивідуальні значення ознак, так і розподіл останніх у межах сукупності. Характерні властивості структури статистичної сукупності відбиваються в рядах розподілу.

Ряд розподілу складається з двох елементів: ***варіант*** — значень групувальної ознаки x_j та ***частот (часток)*** f_j**.** Саме у співвідношенні варіант і частот виявляється ***закономірність розподілу.***

Залежно від статистичної природи варіант ***ряди розподілу*** поділяються на ***атрибутивні*** та ***варіаційні.*** Частотними характеристиками будь-якого ряду є абсолютна чисельність *j*-ї групи — ***частота*** f_j та відносна частота *j*-ї групи — ***частка*** d_j**.** Очевидно, $\sum_{1}^{m} f_j = n$, а $\sum_{1}^{m} d_j = 1$, або 100%. Додатковою характеристикою варіаційних рядів є кумулятивна частота (частка), що являє собою результат послідовного об'єднання груп і підсумовування відповідних їм

частот (часток). ***Кумулятивна частота*** S_{f_j} (частка S_{d_j}) характеризує обсяг сукупності зі значеннями варіант, які не перевищують x_j (табл.).

Варіаційний ряд може бути дискретним або інтервальним. Якщо варіаційний ряд інтервальний з нерівними інтервалами, то його частотні характеристики непорівнянні. Тоді, аналізуючи розподіл, використовують ***щільність*** частоти (частки) на одиницю інтервалу, тобто $g_j = f_j : h_j$. або $g_j = f_j : \mathrm{d}_j$

Таблиця

Частотні характеристики рядів розподілу

Значення варіант x_j	Частоти f_j	Частки d_j	Кумулятивні	
			частоти S_{f_j}	частки S_{d_j}
x_1	f_1	d_1	f_1	d_1
x_2	f_2	d_2	$f_1 + f_2$	$d_1 + d_2$
x_3	f_3	d_3	$f_1 + f_2 + f_3$	$d_1 + d_2 + d_3$
…	…	…	…	…
x_m	f_m	d_m	$\sum f_j$	1
Разом	$\sum f_j$	1		

Отже, поглиблений аналіз закономірностей розподілу передбачає характеристику зазначених особливостей сукупності, зокрема:

а) визначення типового рівня ознаки, який є центром тяжіння;

б) вимірювання варіації ознаки, ступеня згрупованості індивідуальних значень ознаки навколо центра розподілу;

в) оцінювання особливостей варіації, ступеня її відхилення від симетрії;

г) оцінювання нерівномірності розподілу значень ознаки між окремими елементами сукупності, тобто ступінь їх концентрації.

Базою аналізу закономірностей розподілу є варіаційний ряд — дискретний або інтервальний —з рівними інтервалами.

6.2. Характеристики центра розподілу

Центром тяжіння будь-якої статистичної сукупності є типовий рівень ознаки, узагальнююча характеристика всього розмаїття її індивідуальних значень. Такою характеристикою є ***середня величина*** $\bar{x}$. За даними ряду розподілу середня обчислюється як арифметична зважена; вагами є частоти f_j або частки d_j:

$$\bar{x} = \frac{\sum_{1}^{m} x_j f_j}{\sum_{1}^{m} f_j}, \quad \bar{x} = \sum_{1}^{m} x_j d_j$$

де j — номер групи; m — число груп.

В інтервальних рядах, припускаючи рівномірний розподіл елементів сукупності в межах j-го інтервалу, як варіанту x_j використовують середину інтервалу. При цьому ширину відкритого інтервалу умовно вважають такою самою, як сусіднього закритого інтервалу.

Окрім типового рівня важливе значення має домінанта, тобто найбільш поширене значення ознаки. Таке значення називають ***модою*** (*Mo*). У дискретному ряду моду визначають безпосередньо за найбільшою частотою (часткою).

В інтервальному ряду за тим самим принципом визначається модальний інтервал, а в разі потреби конкретне модальне значення в середині інтервалу обчислюється за інтерполяційною формулою:

$$Mo = x_0 + h\frac{f_{mo} - f_{mo-1}}{(f_{mo} - f_{mo-1}) + (f_{mo} - f_{mo+1})},$$

де x_0 та h — відповідно нижня межа та ширина модального інтервалу, f_{mo}, f_{mo-1}, f_{mo+1} — частоти (частки) відповідно модального, передмодального та післямодального інтервалів.

Для моди як домінанти число відхилень (x - Mo) мінімальне. Оскільки мода не залежить від крайніх значень ознаки, то її доцільно використовувати тоді, коли ряд розподілу має невизначені межі.

Характеристикою центра розподілу вважається також ***медіана*** (*Me*) — значення ознаки, яке припадає на середину впорядкованого ряду, поділяє його навпіл — на дві рівні за обсягом частини. Визначаючи медіану, використовують кумулятивні частоти S_{f_j} .або частки S_{d_j}. У дискретному ряду медіаною буде значення ознаки, кумулятивна частота якого перевищує половину обсягу сукупності, тобто $S_{f_j} \geq 0{,}5\sum_{1}^{m} f_j$ (для кумулятивної частки $S_{d_j} \geq 0{,}5$).

В інтервальному ряду за цим принципом визначають медіанний інтервал, а значення медіани в середині інтервалу, як і значення моди, обчислюють за інтерполяційною формулою:

$$Me = x_0 + h\frac{0{,}5\sum_{1}^{m} f_j - S_{f_{me-1}}}{f_{me}},$$

де x_0 та h — відповідно нижня межа та ширина медіанного інтервалу; f_{me} — частота медіанного інтервалу; $S_{f_{me-1}}$ — кумулятивна частота передмедіанного інтервалу.

У симетричному розподілі всі три зазначені характеристики центра розподілу однакові: Мо = Ме = $\bar{x}$, у помірно асиметричному відстань медіани до середньої втричі менша за відстань середньої до моди, тобто $3(\bar{x} - Me) \approx \bar{x} - Mo$.

Медіана, як і мода, не залежить від крайніх значень ознаки; сума модулів відхилень варіант від медіани мінімальна, тобто вона має властивість лінійного мінімуму:

$$\sum_{1}^{m} \left| x_j - Me \right| f_j = \min .$$

Окрім моди і медіани, в аналізі закономірностей розподілу використовуються також квартилі та децилі. ***Квартилі*** — це варіанти, які поділяють обсяги сукупності на чотири рівні частини, ***децилі*** — на десять рівних частин. Ці характеристики визначаються на основі кумулятивних частот (часток) за аналогією з медіаною, яка є другим квартилем або п'ятим децилем.

6.3. Характеристики варіації

В одних сукупностях індивідуальні значення ознаки щільно групуються навколо центра розподілу, в інших — значно відхиляються. Чим менші відхилення, тим однорідніша сукупність, а отже, тим більш надійні й типові характеристики центра розподілу, передусім середня величина. Вимірювання ступеня коливання ознаки, її варіації — невід'ємна складова аналізу закономірностей розподілу. Міри варіації широко використовуються у біологічній науці і практичній діяльності.

На основі характеристик варіації оцінюється інтенсивність структурних зрушень, щільність взаємозв'язків екологічних явищ, точність результатів вибіркового обстеження.

Для вимірювання та оцінювання варіації використовуються абсолютні та відносні характеристики. До абсолютних належать: варіаційний розмах, середнє лінійне та середнє квадратичне відхилення, дисперсії; відносні характеристики подаються низкою коефіцієнтів варіації, локалізації, концентрації.

Варіаційний розмах R — це різниця між максимальним і мінімальним значеннями ознаки: $R = x_{\max} - x_{\min}$. Він характеризує діапазон варіації, наприклад родючості ґрунтів у регіоні тощо. Безперечною перевагою варіаційного розмаху як міри варіації є простота його обчислення й тлумачення.

Проте, коли частоти крайніх варіант надто малі, варіаційний розмах неадекватно характеризує варіацію. У таких випадках використовують квартильні або децильні розмахи. Квартильний розмах $R_Q = Q_3 - Q_1$ охоплює 50% обсягу сукупності, децильний $R_{D_2} = D_8 - D_2$ — 60% або $R_{D_1} = D_9 - D_1$ — 80%.

Інші абсолютні характеристики варіації враховують усі відхилення значень ознаки від центра розподілу, поданого середньою величиною. Оскільки алгебраїчна сума відхилень $\sum_1^m (x_j - \bar{x}) f_j = 0$, то використовуються або модулі відхилень $\sum_1^m |x_j - \bar{x}| f_j$ або квадрати відхилень $\sum (x_j - \bar{x})^2 f_j$. Узагальнюючою характеристикою варіації є ***середнє*** відхилення:

а) *лінійне*

$$\bar{l} = \frac{\sum_1^m |x_j - \bar{x}| f_j}{\sum_1^m f_j};$$

б) *квадратичне, або стандартне*

$$\sigma = \sqrt{\frac{\sum_1^m (x_j - \bar{x})^2 f_j}{\sum_1^m f_j}};$$

в) *дисперсія (середній квадрат відхилень)*

$$\sigma^2 = \frac{\sum_1^m (x_j - \bar{x})^2 f_j}{\sum_1^m f_j}.$$

На підставі первинних, незгрупованих даних наведені характеристики обчислюють за принципом незваженої середньої:

$$\bar{l} = \frac{\sum_1^n |x - \bar{x}|}{n} \text{ або } \sigma = \sqrt{\frac{\sum_1^n (x - \bar{x})^2}{n}}.$$

Середнє лінійне $\bar{l}$ та середнє квадратичне відхилення σ є безпосередніми мірами варіації. Це іменовані числа (в одиницях вимірювання ознаки), за змістом вони ідентичні, проте завдяки математичним властивостям $\sigma > \bar{l}$. Коли обсяг сукупності досить великий і розподіл ознаки, що варіює, наближається до нормального, то $\sigma = 1{,}25\bar{l}$, а $R = 6\sigma$. Значення ознаки в межах $\bar{x} \pm \sigma$ мають 68,3% обсягу сукупності, у межах $\bar{x} \pm 2\sigma$ — 95,4%, у межах $\bar{x} \pm 3\sigma$ — 99,7%. Це відоме

«правило трьох сигм». При значній асиметрії розподілу розрахунок σ не має сенсу.

Очевидний взаємозв'язок середнього квадратичного відхилення та дисперсії: $\sigma = \sqrt{\sigma^2}$. Дисперсія входить до більшості теорем теорії ймовірностей, які є фундаментом математичної статистики, і широко використовується для вимірювання зв'язку й перевірки статистичних гіпотез.

При порівнянні варіації різних ознак або однієї ознаки в різних сукупностях використовуються коефіцієнти варіації *V*. Вони визначаються відношенням абсолютних іменованих характеристик варіації (σ, $\bar{l}$, *R*) до центра розподілу, найчастіше виражаються у процентах. Значення цих коефіцієнтів залежить від того, яка саме абсолютна характеристика варіації використовується. Отже, маємо ***коефіцієнти варіації:***

а) лінійний

$$V_{\bar{l}} = \frac{\bar{l}}{\bar{x}} 100\% ;$$

б) квадратичний

$$V_{\sigma} = \frac{\sigma}{\bar{x}} 100\% ;$$

в) осциляції

$$V_R = \frac{R}{\bar{x}}.$$

Якщо центр розподілу поданий медіаною, то за відносну міру варіації беруть ***квартильний коефіцієнт варіації***

$$V_Q = \frac{Q_3 - Q_1}{2Me}.$$

Для оцінювання ступеня варіації застосовують також співвідношення децилів. Так, ***коефіцієнт децильної диференціації*** показує кратність співвідношення дев'ятого та першого децилів:

$$V_D = \frac{D_9}{D_1}$$

6.4. Характеристики форми розподілу

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

Однорідність сукупності — передумова використання інших статистичних методів (середніх величин, регресійного аналізу тощо). ***Однорідними*** вважаються такі сукупності, елементи яких мають спільні властивості і належать до одного типу, класу. При цьому однорідність означає

не повну тотожність властивостей елементів, а лише наявність у них спільного в істотному, головному.

В однорідних сукупностях розподіли одновершинні (одномодальні). Багатовершинність свідчить про неоднорідний склад сукупності, про різнотиповість окремих складових. У такому разі необхідно перегрупувати дані, виокремити однорідні групи. ***Критерієм однорідності*** сукупності вважається квадратичний коефіцієнт варіації, який завдяки властивостям σ в симетричному розподілі становить $V_\sigma = 0{,}33$.

З-поміж одновершинних розподілів є симетричні та асиметричні (скошені), гостро- та плосковершинні. У ***симетричному розподілі*** рівновіддалені від центра значення ознаки мають однакові частоти, в ***асиметричному*** — вершина розподілу зміщена. Напрям асиметрії протилежний напряму зміщення вершини. Якщо вершина зміщена ліворуч, маємо правосторонню асиметрію, і навпаки. Зазначимо, що асиметрія виникає внаслідок обмеженої варіації в одному напрямі або під впливом домінуючої причини розвитку, яка призводить до зміщення центра розподілу. Ступінь асиметрії різний — від помірного до значного.

Як уже зазначалося, у симетричному розподілі характеристики центра — середня, мода, медіана — мають однакові значення, в асиметричному між ними існують певні розбіжності. У разі правосторонньої асиметрії $\bar{x} > Me > Mo$, а в разі лівосторонньої, навпаки, $\bar{x} < Me < Mo$. Чим більша асиметрія, тим більше відхилення ($\bar{x}$ - *Mo*). Очевидно, найпростішою ***мірою асиметрії*** є відносне відхилення $A = \dfrac{\bar{x} - Mo}{\sigma}$, яке характеризує напрям і міру скошеності в середині розподілу: при правосторонній асиметрії $A > 0$, при лівосторонній — $A < 0$.

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

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

Асиметрія та ***ексцес*** — дві пов'язані з варіацією властивості форми розподілу. Комплексне їх оцінювання виконується на базі ***центральних моментів розподілу.*** Алгебраїчне центральний момент розподілу — це середня арифметична *k-го* ступеня відхилення індивідуальних значень ознаки від середньої:

$$\mu_k = \frac{\sum_1^m \left(x_j - \bar{x}\right)^k f_j}{\sum_1^m f_j}$$

Очевидно, що момент 2-го порядку є дисперсією, яка характеризує варіацію. Моменти 3-го і 4-го порядків характеризують відповідно асиметрію та ексцес. У симетричному розподілі $\mu_3 = 0$. Чим більша скошеність ряду, тим більше значення μ_3. Для того щоб характеристика скошеності не залежала від масштабу вимірювання ознаки, для порівняння ступеня асиметрії різних розподілів використовується стандартизований момент $A_s = \frac{\mu_3}{\sigma^3}$, який на відміну від коефіцієнта скошеності залежить від крайніх значень ознаки. При правосторонній асиметрії коефіцієнт $A_s > 0$, при лівосторонній $A_s < 0$. Звідси правостороння асиметрія називається додатною, а лівостороння — від'ємною. Уважається, що при $A_s < 0{,}25$ асиметрія низька, якщо A_s . не перевищує 0,5 — середня, при $A_s > 0{,}5$ — висока.

Для вимірювання ексцесу використовується стандартизований момент 4-го порядку $E_k = \frac{\mu_4}{\sigma^4}$. У симетричному, близькому до нормального розподілі $E_k = 3$. Очевидно, при гостровершинному розподілі $E_k > 3$. при плосковершинному $E_k < 3$.

Аналіз закономірностей розподілу можна поглибити, описати його певною функцією.

Не менш важливими у статистичному аналізі є характеристика нерівномірності розподілу певної ознаки між окремими складовими сукупності, а також оцінка концентрації значень ознаки в окремих її частинах.

Якщо розподіл значень ознаки в сукупності рівномірний, то частки однакові — $d_j = D_j$, відхилення часток свідчать про певну концентрацію. Верхня межа суми відхилень $\sum \left|d_j - D_j\right| = 2$, а тому ***коефіцієнт концентрації*** обчислюється як півсума модулів відхилень:

$$K = \frac{1}{2}\sum_{j=1}^{m} \left|d_j - D_j\right| .$$

Значення коефіцієнта коливаються в межах від нуля (рівномірний розподіл) до одиниці (повна концентрація). Чим більший ступінь концентрації, тим більше значення коефіцієнта *К*.

Коефіцієнти концентрації широко використовуються в регіональному аналізі для оцінювання рівномірності територіального розподілу особин, кормової бази тощо. За кожним регіоном визначається також ***коефіцієнт локалізації***

$$L_j = \frac{D_j}{d_j} 100\%$$

який характеризує співвідношення часток.

Порівняння структур на основі відхилень часток доцільне в рядах з нерівними інтервалами, а надто в атрибутивних рядах.

За аналогією з коефіцієнтом концентрації обчислюється ***коефіцієнт подібності (схожості) структур*** двох сукупностей:

$$p = 1 - \frac{1}{2}\sum_1^m \left| d_j - d_k \right| .$$

Якщо структури однакові, P= 1; якщо абсолютно протилежні, P = 0. Чим більше схожі структури, тим більше значення P.

Структура будь-якої статистичної сукупності динамічна. Змінюються склад і рівень ресурсів, вікова й статева структура популяцій тощо. Зміна часток окремих складових сукупності свідчить про структурні зрушення. ***Інтенсивність структурних зрушень*** оцінюється за допомогою середнього лінійного $\bar{l}_d$ або середнього квадратичного σ_d відхилень часток:

$$\bar{l}_d = \frac{\sum_1^m \left| d_{j1} - d_{j0} \right|}{m};$$

$$\sigma_d = \sqrt{\frac{\sum_1^m \left(d_{j1} - d_{j0} \right)^2}{m}} .$$

де d_{j0} та d_{j1} — частки відповідно базисного та поточного періоду;

m — число складових сукупності.

6.5. Види та взаємозв'язок дисперсій

Дисперсія посідає особливе місце у статистичному аналізі біологічних явищ. На відміну від інших характеристик варіації завдяки своїм математичним властивостям вона є невіддільним і важливим елементом інших статистичних методів, зокрема дисперсійного аналізу.

Для ознак метричної шкали дисперсія — це середній квадрат відхилень індивідуальних значень ознаки від середньої:

$$\sigma^2 = \frac{\sum_1^m \left(x_j - \bar{x} \right)^2 f_j}{\sum_1^m f_j} .$$

Як і будь-яка середня, дисперсія має певні математичні властивості. Сформулюємо найважливіші з них.

1. Якщо всі значення варіанта, зменшити на сталу величину *A,* то дисперсія не зміниться:

$$\sigma^2_{(x-A)} = \sigma^2_x .$$

2. Якщо всі значення варіант *x,* змінити в *A* раз, то дисперсія зміниться в A^2 раз:

$$\sigma^2_{xA} = \sigma^2_x A^2 .$$

3. Якщо частоти замінити частками, дисперсія не зміниться.

Дисперсія альтернативної ознаки обчислюється як добуток часток: $\sigma^2 = d_1 d_0$, де d_1 — частка елементів сукупності, яким властива ознака, d_0 — частка решти елементів (d_0=1- d_1).

Дисперсія альтернативної ознаки широко використовується при проектуванні вибіркових обстежень, обробці даних соціологічних опитувань, статистичному контролі якості продукції тощо. За відсутності первинних даних про розподіл сукупності припускають, що $d_1 = d_0 = 0{,}5$ і використовують максимальне значення дисперсії $\sigma^2 = 0{,}5 \cdot 0{,}5 = 0{,}25$.

Якщо сукупність розбито на групи за певною ознакою *x,* то для будь-якої іншої ознаки *y* можна обчислити дисперсію як у цілому по сукупності, так і в кожній групі. Центром розподілу сукупності в цілому є загальна середня $\bar{y} = \frac{\sum_1^n y}{n}$, центром розподілу *j*-й групі — групова середня $\bar{y}_j = \frac{\sum_1^{f_j} y}{f_j}$. Відхилення індивідуальних значень ознаки від загальної середньої *y* можна подати як дві складові: $(y - \bar{y}) = (y - \bar{y}_j) + (\bar{y}_j - \bar{y})$.

Узагальнюючими характеристиками цих відхилень є дисперсії: загальна, групова та міжгрупова.

Загальна дисперсія характеризує варіацію ознаки *y* навколо загальної середньої:

$$\sigma^2 = \frac{\sum_1^n (y - \bar{y})^2}{n} .$$

Групова дисперсія характеризує варіацію відносно групової середньої:

$$\sigma_j^2 = \frac{\sum_1^{f_j} (y - \bar{y})^2}{f_j} .$$

Оскільки в групи об'єднуються певною мірою схожі елементи сукупності, то варіація в групах, як правило, менша, ніж у цілому по сукупності. Якщо причинні комплекси, що формують варіацію в різних групах, неоднакові, то й групові дисперсії різняться між собою.

Узагальнюючою мірою внутрішньогрупової варіації є ***середня з групових дисперсій:***

$$\overline{\sigma^2} = \frac{\sum_1^m \sigma_j^2 f_j}{\sum_1^m f_j}.$$

Різними є й групові середні $\bar{y}_j$. Мірою варіації їх навколо загальної середньої є ***міжгрупова дисперсія***

$$\delta^2 = \frac{\sum_1^m (\bar{y}_j - y)^2 f_j}{\sum_1^m f_j}.$$

Отже, загальна дисперсія складається з двох частин. Перша характеризує внутрішньогрупову, друга — міжгрупову варіацію.

Взаємозв'язок дисперсій називається ***правилом розкладання (декомпозиції) варіації:***

$$\sigma^2 = \overline{\sigma^2} + \delta^2.$$

6.6. Характеристика розподілів, що найбільш часто зустрічаються в біометричній практиці

6.6.1. Нормальний розподіл (розподіл Гауса-Лапласа)

Нормальний розподіл найбільш часто застосовується в біометрії. Він дає досить добру модель для дійсних явищ, що характеризуються такими властивостями:

1) проявляється сильна тенденція даних до групування біля центру;

2) від’ємні і додатні відхилення від центру рівноімовірні;

3) частота відхилень різко падає у напрямі від центру.

Механізм, що лежить в основі нормального розподілу, пояснюється за допомогою центральної граничної теореми.

В теорії імовірності центральна гранична теорема сформульована Муавом і Лапласом ще у XVII столітті як розвиток закону великих чисел Я. Бернулі (1713). Ідея полягає в тому, що при сумуванні великої кількості незалежних величин отримуються нормальнорозподілені величини. І це відбувається незалежно, тобто інваріантне, від розподілу вихідних величин. Іншими словами, якщо на деяку змінну впливають багато факторів, ці впливи незалежні, відносно

малі і додаються, то сумарна величина, що отримується внаслідок цього має нормальний розподіл.

Наприклад, величезна кількість факторів визначає вагу людини (тисячі генів, хвороби, харчування тощо), таким чином можна чекати нормальний розподіл ваги людини в популяції.

Формально щільність нормального розподілу записується так:

$$\varphi(x;a,\sigma^2)=\frac{1}{\sqrt{2\pi}\cdot\sigma}e^{-\frac{(x-a)^2}{2\sigma^2}},$$

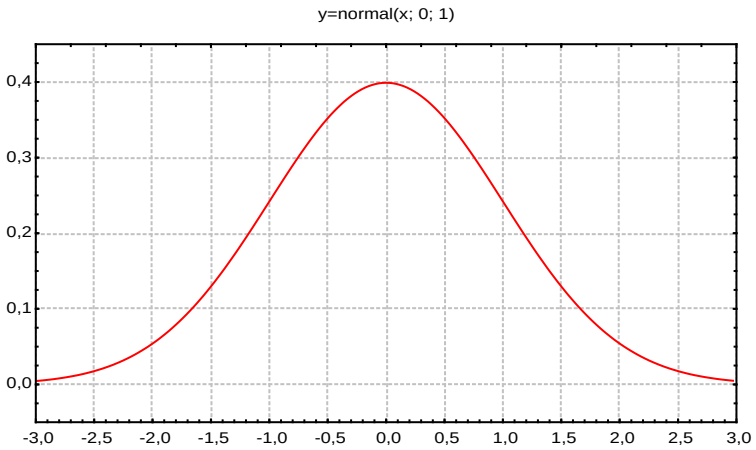
де параметри: a - середнє значення, а σ^2- (сігма) дисперсія даної випадкової величини.

Відповідна функція розподілу нормальної випадкової величини $\xi(a,\sigma^2)$ (ксі) позначається як $\Phi(x;a,\sigma^2)$ і задається співвідношенням:

$$\Phi(x;a,\sigma^2)=P\{\xi(a,\sigma^2)<x)=\frac{1}{\sqrt{2\pi}\cdot\sigma}\int_{-\infty}^{x}e^{-\frac{(t-a)^2}{2\sigma^2}}dt.$$

Нормальний розподіл з параметрами a=0 і $\sigma^2=1$ називається стандартним.

Зворотна функція стандартного нормального розподілу, що застосована до величини z, $0<z<1$, називається пробіт-перетворенням z, або просто пробітом z.



Властивості:

Середнє, мода, медіана: $E\xi=x_{\text{mod}}=x_{med}=a$;

Дисперсія: $D\xi=\sigma^2$;

Асиметрія: $\beta_1=0$;

Ексцес: $\beta_2=0$.

(асиметрія - визначає зміщення графіка від осі симетрії, ексцес - визначає гостроту піка)

При збільшенні дисперсії щільність нормального розподілу "розтікається" вздовж осі *OX* , при зменшенні дисперсії, вона навпаки стискається, концентруючись воло однієї точки - точки максимального значення, що співпадає з середнім. В граничному випадку нульової дисперсії випадкова величина вироджується і приймає єдине значення, рівне середньому.

На відстані в одне стандартне відхилення (сігма) від середнього значення концентрується 68% всієї вибірки; 2σ - 95,45%; 3σ - 99,73%.

6.6.2. Рівномірний розподіл

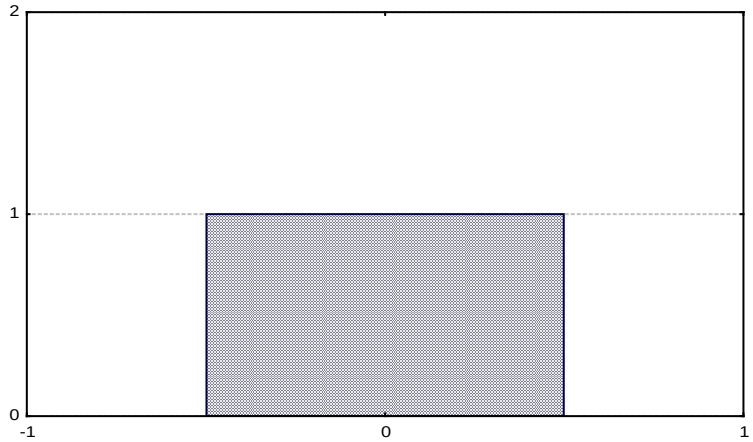
Рівномірний розподіл застосовується для опису змінних, у яких кожне значення рівноімовірне, тобто значення змінної рівномірно розподілені в деякій області.

Щільність і функція розподілу рівномірного розподілу випадкової величини, що приймає значення на відрізку [*a, b*] виражені наступними формулами:

$$f_\xi(x) = \begin{cases} \dfrac{1}{b-a}; a \le x \le b; \\ 0; x < a; x > b. \end{cases}$$

$$F_\xi = \begin{cases} 0; x \le a; \\ \dfrac{x-a}{b-a}; a < x \le b; \\ 1; x > b. \end{cases}$$

З цих формул випливає, що імовірність того, що рівномірна випадкова величина приймає значення з множини $[c,d] \subset [a,b]$, рівне $(d-c)/(b-a)$.



Властивості рівномірного розподілу:

Середнє, медіана: $E\xi = x_{med} = \dfrac{a+b}{2}$;

Дисперсія: $D\xi = \dfrac{(b-a)^2}{12}$;

Асиметрія: $\beta_1 = 0$;

Ексцес: $\beta_2 = -1{,}2$.

6.6.3. Експотенційний розподіл

Візьмемо події, що є рідкісними (наприклад частота зустрічання рідкісного виду). Якщо *T*- час між настанням рідкісних подій, що відбуваються в середньому з інтенсивністю λ (лямбда), то величина *T* має експотенційний розподіл з параметром λ.

Для цього розподілу властива відсутність післядії, або інакше марківська властивість. Ця властивість полягає в тому, що для потоку рідкісних подій час чекання наступної події завжди розподілений показниково незалежно (тобто з тим же параметром) від того, скільки часу ви її чекали.

Показниковий розподіл пов'язаний з пуасоновським розподілом: у одиничному інтервалі часу кількість подій, інтервали між якими незалежні і показниково розподілені, має розподіл Пуасона.

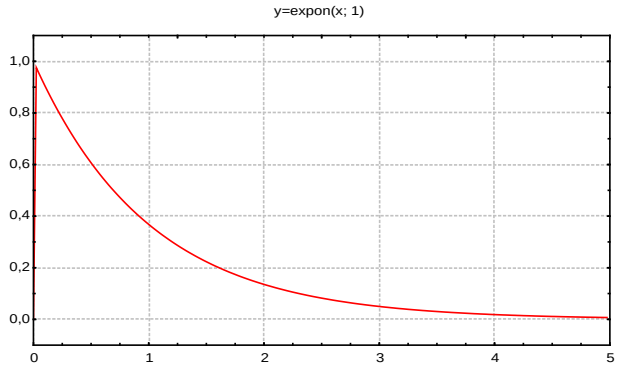
Показниковий розподіл являє собою окремий випадок розподілу Вейбула.

Якщо час не неперервний, а дискретний, то аналогом показникового розподілу є геометричний розподіл.

Щільність експотенційного розподілу визначається за формулою:

$$f_\xi(x) = \lambda_0 e^{-\lambda_0 x},\ x \geq 0.$$

Цей розподіл має тільки один параметр, що визначає його характеристики.



Властивості експотенційного розподілу:

Середнє: $E\xi = \frac{1}{\lambda_0}$;

Мода: $x_{\text{mod}} = 0$;

Медіана: $x_{med} = \frac{1}{\lambda_0} \cdot \ln 2$;

Дисперсія: $D\xi = \frac{1}{\lambda_0^2}$;

Асиметрія: $\beta_1 = 2$;

Ексцес: $\beta_2 = 6$.

6.6.4. Хі-квадрат-розподіл

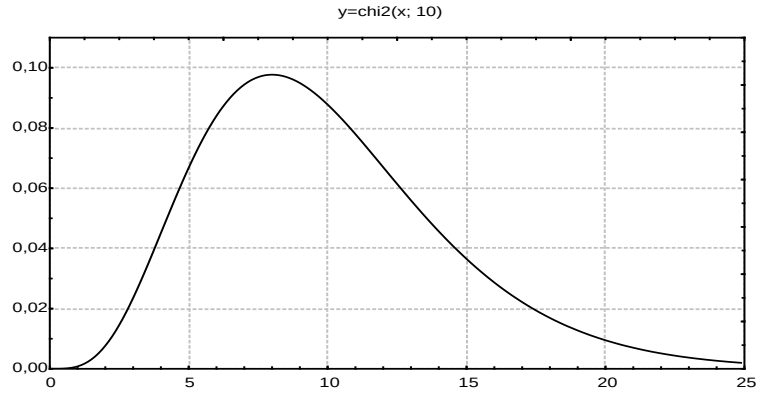
Сума квадратів m незалежних нормальних величин з середнім 0 та дисперсією 1 має хі-квадрат розподіл з m степенями свободи. Цей розподіл найбільш часто використовується при аналізі даних.

Формально щільність хі-квадрат-розподілу з m степенями свободи має вигляд:

$$f_{\chi^2(m)}(x) = \frac{1}{2^{\frac{m}{2}}\Gamma\left(\frac{m}{2}\right)} \cdot x^{\frac{m}{2}-1} e^{-\frac{x}{2}}, x > 0.$$

де Γ - гамма-функція Ейлера, $\Gamma(z) = \int_0^\infty x^{x-1} e^{-x} dx$.

При $x \le 0$ щільність функції дорівнює 0.



Властивості хі-квадрат-розподілу:

Середнє: $E\chi^2(m) = m$;

Мода: $m_{\text{mod}} = m - 2$;

Дисперсія: $D\chi^2(m) = 2m$;

Асиметрія: $\beta_1 = \frac{2^{\frac{3}{2}}}{\sqrt{m}}$;

Ексцес: $\beta_2 = \frac{12}{m}$.

6.6.5. Біноміальний розподіл

Біноміальний розподіл є найбільш важливим дискретним розподілом, що зосереджений лише в декількох точках. Цим точкам біноміальний розподіл присвоює додатні імовірності. Таким чином біноміальний розподіл відрізняється від неперервних розподілів (нормального, хі-квадрат тощо), для яких характерна нульова імовірність окремо взятим точкам, ці розподіли називаються неперервними.

Біноміальний розподіл використовується в тих випадках, коли параметр приймає дискретні значення (наприклад кількість яєць у гнізді птаха, кількість студентів на занятті тощо). Велике значення має біноміальний розподіл в теорії ігор.

При великій кількості випробовувань біноміальний розподіл переходить у нормальний.

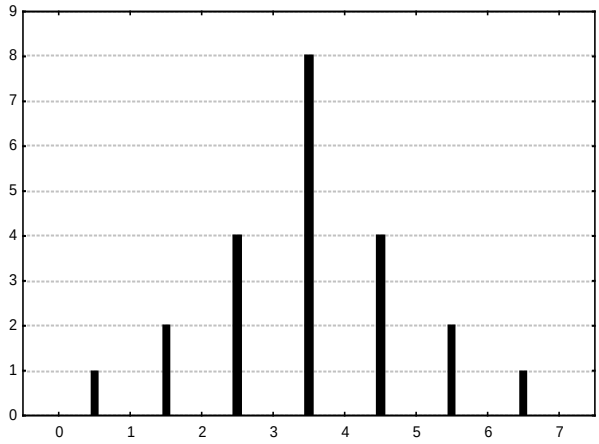
Точна формула для імовірності *m* успіхів у *n* випробовуваннях записується так:

$$f(m) = \left[\frac{n!}{m! \cdot (n-m)!} \right] \cdot p^m q^{n-m},$$

де *p*- імовірність успіху,

$q = 1 - p;\ p, q \geq 0;\ p + q = 1,$

n- кількість випробовувань, $m = 0,1...m$



Властивості біноміального розподілу:

Середнє: $E\nu_p(n) = np$;

Мода: $x_{\text{mod}} : p(n=1) - 1 \leq x_{\text{mod}} \leq p(n+1)$;

Дисперсія: $D\nu_p(n) = np(1-p)$;

Асиметрія: $\beta_1 = \frac{1-2p}{\sqrt{np(1-p)}}$;

Ексцес: $\beta_2 = \frac{1-6p(1-p)}{np(1-p)}$.

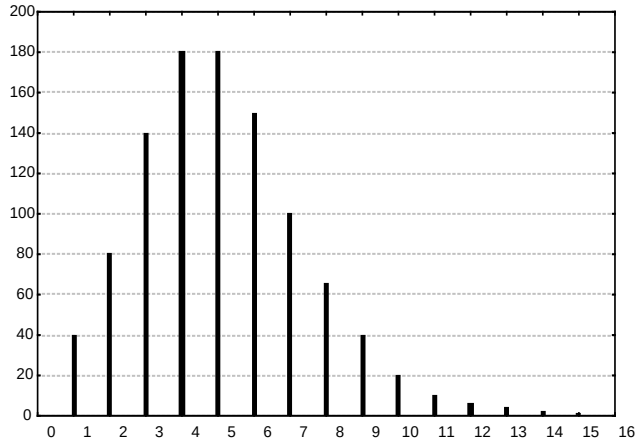
6.6.6. Розподіл Пуассона

Розподіл Пуассона іноді називають розподілом рідкісних подій. Прикладами змінних, розподілений за законом Пуассона є: кількість нещасних подій, кількість дефектів у виробничих процесах тощо.

При великій кількості подій розподіл Пуассона переходить у показниковий або експотенційний розподіл.

Розподіл Пуассона визначається за формулою:

$$f(x) = \frac{\lambda^x \cdot e^{-\lambda}}{x!}.$$



Властивості розподілу Пуассона:

Середнє: $E\nu_0 = \lambda$;

Дисперсія: $D\nu_0 = \lambda$;

Асиметрія: $\beta_1 = \frac{1}{\sqrt{\lambda}}$;

Ексцес: $\beta_2 = \frac{1}{\lambda}$.

6.6.7. Розподіл Вейбула

Розподіл Вейбула названий на честь шведського дослідника Валоді Вейбула (Walloddi Weibull), він застосовував цей розподіл для опису часу відмов різного типу у теорії надійності.

Формально щільність розподілу Вейбула записується наступною формулою:

$$f\xi(t) = \lambda_0 \alpha t^{\alpha-1} e^{-\lambda_0 t^{\alpha}}, t \geq 0.$$

Іноді щільність розподілу Вейбула записується також у вигляді:

$$f(x) = \frac{c}{b} \cdot \left(\frac{x}{b}\right)^{c-1} \cdot e^{\left(\frac{x}{b}\right)^{c}}; 0 \leq x, b > 0, c > 0,$$

де: *b*- параметр масштабу;

c- параметр форми;

e- константа Ейлера (2,718...).

α- параметр положення.

Параметр положення. Зазвичай розподіл Вейбула зосереджено на піввісі від 0 до нескінченності. Якщо замість межі 0 ввести параметр *a*, що буває необхідно на практиці, то виникає так званий трипараметричий розподіл Вейбула.

Розподіл Вейбула інтенсивно використовується в теорії надійності і страхуванні.

Експотенційний розподіл застосовується як модель, що оцінює час наробки до відмови у випадку, що імовірність відмови постійна. Якщо імовірність відмови змінюється з часом, то застосовується розподіл Вейбула.

При *c* = 1 або, в іншій параметризації, при α = 1 розподіл Вейбула переходить в експотенційний, а при α = 2 - в розподіл Релея.

Часто при проведенні аналізу надійності необхідно розглядати імовірність відмови протягом малого інтервалу часу після моменту часу *t* при умові, що до моменту *t* відмови не було.

Така функція називається функцією ризику, або функцією інтенсивності відмов, і формально визначається наступним чином:

$$h(t) = \frac{f(t)}{1 - F(t)},$$

де: *h(t)* - функція інтенсивності відмов або функція ризику в момент часу *t*;

f(t) - щільність розподілу часу відмов;

F(t) - функція розподілу часу відмов (інтеграл від щільності по інтервалу [*0,t*]).

В загальному вигляді функція інтенсивності відмов записується так:

$$\lambda(t) = \lambda_0 \alpha t^{\alpha-1},$$

де $\lambda_0 > 0; \alpha > 0$ - деякі числові параметри.

При α = 1 функція ризику дорівнює константі, що відповідає нормальній експлуатації приладу (рис.).

При α < 1 - функція ризику спадає, що відповідає припрацюванню приладу.

При α > 1 - функція ризику зростає, що відповідає старінню приладу.

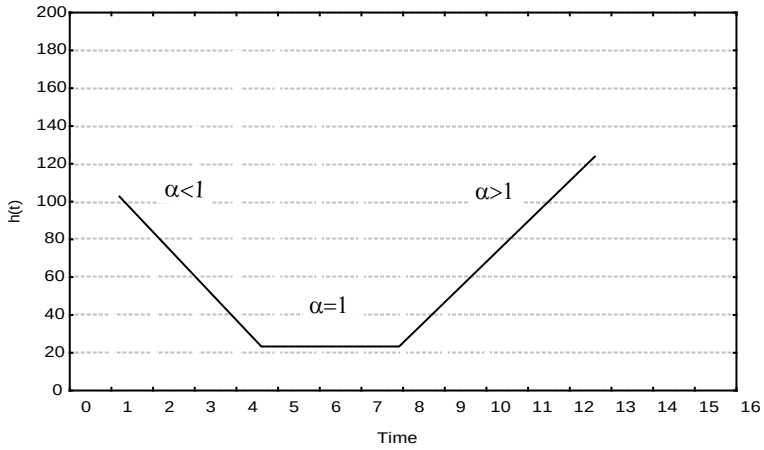
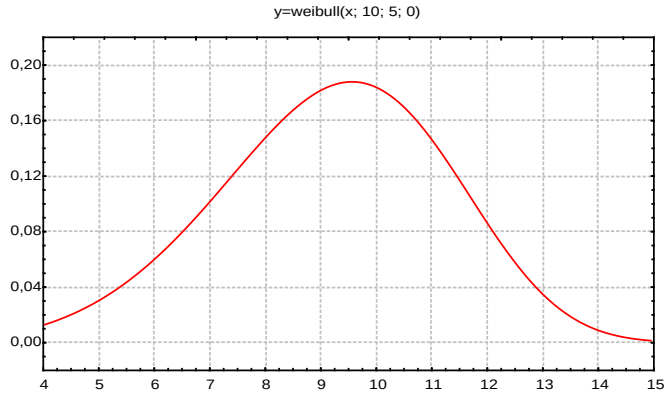
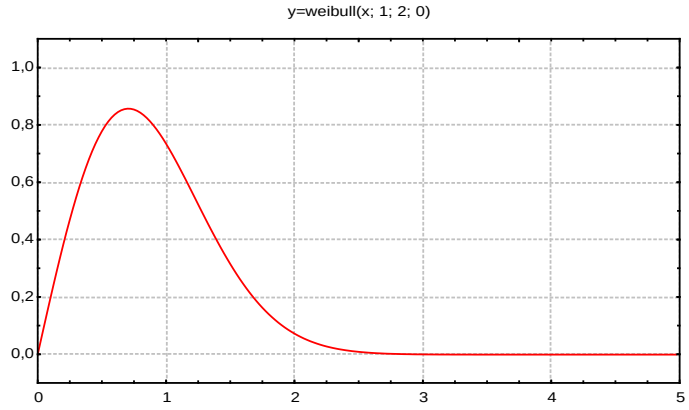
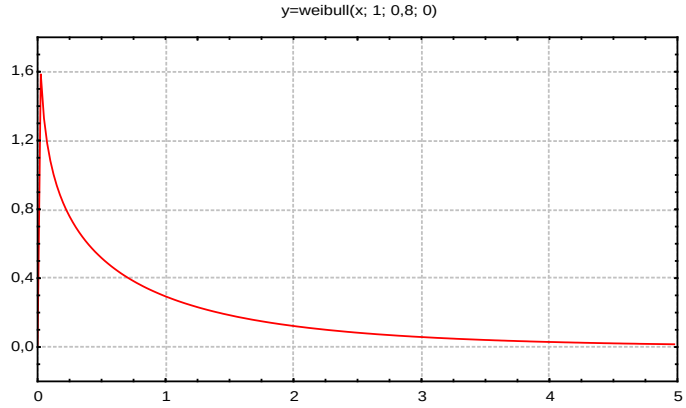


Рис. Вигляд функцій ризику при різних значеннях параметру α.







Властивості розподілу Вейбула:

Середнє: $E\xi = \lambda_0^{\frac{-1}{\alpha}} \Gamma\left(1+\frac{1}{\alpha}\right);$

Мода: $x_{\text{mod}} = \begin{cases} 0; \alpha \le 1; \\ \lambda_0^{\frac{-1}{\alpha}} \left(1-\frac{1}{\alpha}\right)^{\frac{1}{\alpha}}; \alpha > 1; \end{cases}$

Дисперсія: $D\xi = \lambda_0^{\frac{-2}{\alpha}} \cdot \left[\Gamma\left(1+\frac{2}{\alpha}\right) - \Gamma^2\left(1+\frac{1}{\alpha}\right)\right];$

Момент *k-го* порядку: $m_k = E\xi^k = \lambda_0^{\frac{-1}{\alpha}} \cdot \Gamma\left(1+\frac{k}{\alpha}\right);$

де $\Gamma(z)$ - гама-функція Ейлера, $\Gamma(z) = \int_0^{\infty} x^{x-1} e^{-x} dx$

МОДУЛЬ ІІ.
МЕТОДИ АНАЛІЗУ ДАНИХ БІОМЕТРИЧНИХ СПОСТЕРЕЖЕНЬ

7. ВИБІРКОВИЙ МЕТОД.
СТАТИСТИЧНА ПЕРЕВІРКА ГІПОТЕЗ

7.1. Суть вибіркового спостереження

Вибіркове спостереження — такий вид несуцільного спостереження, при якому обстежуються не всі елементи сукупності, що вивчається, а лише певним чином дібрана їх частина. Сукупність, з якої вибирають елементи для обстеження, називається ***генеральною,*** а сукупність, яку безпосередньо обстежують, — ***вибірковою.*** Статистичні характеристики вибіркової сукупності розглядаються як оцінки відповідних характеристик генеральної сукупності.

Практика вибіркових спостережень досить різноманітна. При обстеженні невеликої частини генеральної сукупності зменшуються помилки реєстрації, можна розширити й деталізувати програму обстеження. З іншого боку, вибіркове спостереження забезпечує економію матеріальних, трудових, фінансових ресурсів і часу.

При вивченні певного кола біологічних явищ вибіркове спостереження єдино можливе. Часом вибіркове спостереження поєднується із суцільним. Крім того, вибірковий метод використовують для прискореної обробки матеріалів суцільного спостереження та перевірки правильності даних одноразових обстежень.

Об'єктивною гарантією того, що вибірка репрезентує (представляє) всю сукупність, є додержання наукових принципів організації та проведення спостереження, насамперед неупередженого, об'єктивного підходу до вибору елементів для обстеження. Принцип випадковості вибору забезпечує всім елементам генеральної сукупності рівні можливості потрапити у вибірку.

Якщо генеральна сукупність містить N елементів, а для обстеження потрібно вибрати з них частину n, то число можливих вибірок

$$C_N^n = \frac{N!}{n!(N-n)!}.$$

Усі вони мають однакову ймовірність $\frac{1}{C_N^n}$, але кожна з них несе в собі певну похибку, що відбиває факт випадковості вибору. Оскільки вибіркова сукупність не точно відтворює склад генеральної сукупності, то й вибіркові оцінки не збігаються з відповідними характеристиками генеральної сукупності. Розбіжності між ними називають ***похибками репрезентативності:*** для середньої— це різниця між генеральною $\bar{x}_0$ вибірковою $\bar{x}$ середніми, для частки

— різниця між генеральною d_0 і вибірковою p частками, для дисперсії— відношення генеральної σ_0^2 та вибіркової σ^2 дисперсій тощо.

За причинами виникнення похибки репрезентативності поділяються на тенденційні (систематичні) та випадкові. ***Тенденційні*** похибки виникають, коли при формуванні вибіркової сукупності порушений принцип випадковості (упереджений вибір елементів, недосконала основа вибірки тощо). Ці похибки для всіх елементів сукупності однонапрямлені і призводять до зсунення результатів обстеження.

Випадкові похибки — це наслідок випадковості вибору елементів для дослідження і пов'язаних з цим розбіжностей між структурами вибіркової та генеральної сукупностей щодо ознак, які вивчаються.

При організації вибіркового обстеження важливо уникнути тенденційних похибок. Незсуненість — одна з вимог до будь-якої вибіркової оцінки. Притаманних вибірковому спостереженню випадкових похибок уникнути неможливо, проте теорія вибіркового методу дає математичну основу для обчислення таких похибок та регулювання їх розміру.

Згідно з генеральною граничною теоремою за умови достатньо великого обсягу вибірки розподіл вибіркових середніх (і часток), незалежно від розподілу генеральної сукупності, асимптотичне наближається до нормального. Більшість значень вибіркових середніх зосереджується навколо генеральної середньої, а отже, найбільшу ймовірність мають відхилення, близькі до нуля. Чим більше відхилення, тим менша його ймовірність. Для будь-якої ймовірності існує межа відхилень вибіркової середньої від генеральної. Використовуючи властивості нормального розподілу, для однієї конкретної вибірки можна визначити:

• похибки репрезентативності — середню та граничну для взятої ймовірності;

• імовірність того, що похибка вибірки не перевищить допустимого рівня;

• обсяг вибірки, який забезпечить потрібну точність результатів для взятої ймовірності.

Кінцева мета будь-якого вибіркового спостереження — поширення його характеристик на генеральну сукупність. Для середньої та частки визначаються межі можливих їх значень у генеральній сукупності з певною ймовірністю — довірчі межі. Якщо метою вибіркового обстеження є визначення обсягових показників генеральної сукупності — обсягів значень ознаки $\sum_1^N x_i$, то вибіркова середня поширюється на генеральну сукупність прямим перерахунком: $\bar{x}N = \sum_1^N x_i$

7.2. Вибіркові оцінки середньої та частки

У статистиці використовують два типи оцінок параметрів генеральної сукупності — точкові та інтервальні. ***Точкова оцінка*** — це значення параметра за даними вибірки: вибіркова середня *x* та вибіркова частка *p*. ***Інтервальною оцінкою*** називають інтервал значень параметра, розрахований за даними вибірки для певної ймовірності, тобто ***довірчий інтервал.*** Чим менший довірчий інтервал, тим точніша вибіркова оцінка.

Межі довірчого інтервалу визначаються на основі точкової оцінки та граничної похибки вибірки $\Delta = t\mu$:

для середньої

$\bar{x} - t\mu \le \bar{x}_0 \le \bar{x} + t\mu$;

для частки

$p - t\mu \le d_0 \le p + t\mu$,

де μ — стандартна (середня) похибка вибірки; t – квантіль розподілу ймовірностей (довірче число).

Стандартна похибка вибірки μ є середнім квадратичним відхиленням вибіркових оцінок від значення параметра в генеральній сукупності. Як доведено в теорії вибіркового методу, дисперсія вибіркових середніх у *n* раз менша від дисперсії ознаки в генеральній сукупності, тобто $\mu^2 = \sigma_0^2 / n$. Оскільки на практиці генеральна дисперсія ознаки σ_0^2 невідома, у розрахунках можна використати вибіркову незсунену оцінку дисперсії: для повторної вибірки $\sigma^2 \frac{n}{n-1}$, для безповторної $\sigma^2 \frac{N-n}{N-1}$. Отже, формули стандартної похибки:

для повторної вибірки

$$\mu = \sqrt{\frac{\sigma^2}{n-1}},$$

для безповторної вибірки

$$\mu = \sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)}.$$

Щодо практичного використання наведених формул слід урахувати таке:

а) дисперсія частки $\sigma^2 = p(1-p) = pq$, де p і q — частки вибіркової сукупності, яким відповідно властива і невластива ознака;

б) у великих за обсягом сукупностях (30 і більше одиниць) поправка $\frac{n}{n-1}$ не вносить істотних змін у розрахунки, а тому береться до уваги лише у вибірках з невеликою кількістю елементів;

в) коригуючий множник для безповторної вибірки $\sqrt{\frac{N-n}{N-1}} \approx \sqrt{1-\frac{n}{N}}$, тобто при малих величинах $\frac{n}{N}$ (наприклад, для 2 чи 5%-ної вибірки) наближається до 1, а тому розрахунок можна виконувати за формулою для повторної вибірки; при 10%-ній вибірці коригуючий множник становить 0,949, при 20%-ній — 0,894.

Гранична похибка вибірки $\Delta = t\mu$ — це максимально можлива похибка для взятої ймовірності *F(x)*. Довірче число *t* показує, як співвідносяться гранична та стандартна похибки. Як бачимо з рис. 6.1, з імовірністю 0,683 гранична похибка не вийде за межі стандартної $\Delta = \pm 1\mu$, з імовірністю 0,954 вона не перевищить $\pm 2\mu$, з імовірністю 0,997 — $\pm 3\mu$. На практиці найчастіше застосовують імовірність 0,954 (на рис. 7.1 незаштрихована частина площини).

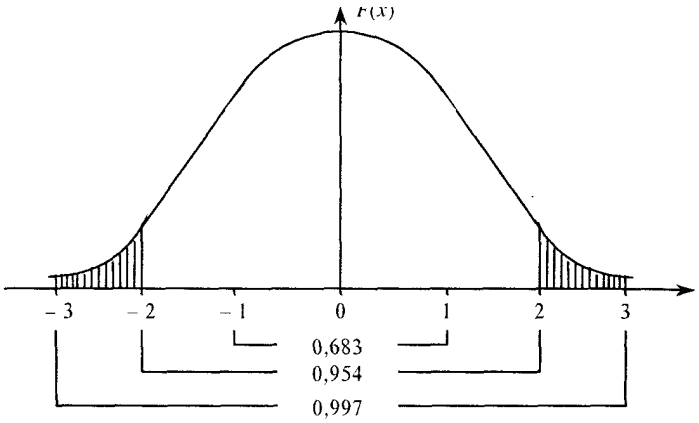


Рис. 7.1. Співвідношення ймовірностей та ширини довірчих меж

З урахуванням сказаного формули граничних похибок середньої та частки записують так:

	Повторна вибірка	Безповторна вибірка
Для середньої	$\Delta_{\bar{x}} = t\sqrt{\frac{\sigma^2}{n}}$;	$\Delta_{\bar{x}} = t\sqrt{\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)}$;
Для частки	$\Delta_p = t\sqrt{\frac{pq}{n}}$;	$\Delta_p = t\sqrt{\frac{pq}{n}\left(1-\frac{n}{N}\right)}$.

Як видно з формул, розмір граничної похибки залежить:

• від варіації ознаки σ^2;

• обсягу вибірки *n*;

• частки вибірки в генеральній сукупності $\frac{n}{N}$;

• узятого рівня ймовірності, якому відповідає квантиль *t*.

Чим більша варіація ознаки в генеральній сукупності, тим більша в середньому похибка вибірки. Залежність похибки від обсягу вибіркової сукупності обернено пропорційна. Щоб зменшити похибку вибірки вдвічі, обсяг останньої має зрости в 4 рази. При безповторному доборі похибка буде тим менша, чим більша частка обстеженої сукупності $\frac{n}{N}$. Очевидно, при суцільному спостереженні похибка репрезентативності відсутня $(\Delta = 0)$. При малих вибірках (n < 30), у розрахунках стандартних похибок використовують вибіркові оцінки дисперсій $s^2 = \sigma^2 \frac{n}{n-1}$.

Квантилі *t* визначають за розподілом ймовірностей Стьюдента. У таблиці II наведено деякі значення квантилів *t* розподілу Стьюдента для ймовірності 0,95 і ***числа ступенів свободи,*** тобто числа незалежних величин, необхідних для визначення даної характеристики, $k = n - 1$. При $n > 30$ квантилі розподілу Стьюдента і нормального розподілу збігаються.

У статистичному аналізі часто постає потреба порівняти похибки вибірки різних ознак або однієї і тієї самої ознаки в різних сукупностях.

Такі порівняння виконують за допомогою ***відносної похибки,*** яка показує, на скільки процентів вибіркова оцінка може відхилятися від параметра генеральної сукупності. Відносна стандартна похибка середньої — це коефіцієнт варіації вибіркових середніх:

$$V_\mu = \frac{\mu_{\bar{x}}}{\bar{x}} 100.$$

Її розмір можна визначити також на основі коефіцієнта варіації ознаки V_x:

для повторної вибірки

$$V_\mu = 100 \frac{V_x}{\sqrt{n}};$$

для безповторної вибірки

$$V_\mu = 100 \frac{V_x}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}.$$

Вибіркову похибку частки Δp також слід порівнювати з часткою *p*. Адже одна і та сама похибка Δp= 2% для p = 80% є малою, для p = 40% — допустимою, для p= 10% — завеликою. Відносну похибку частки обчислюють за формулою

$$V\mu = \frac{\mu_p}{p} = \frac{\sqrt{\frac{pq}{n}}}{p} = \sqrt{\frac{q}{np}}$$

Отже, відносну похибку можна використати для порівняння вибіркових оцінок різних ознак. На практиці достатнім рівнем точності вважається $V_\mu \le 5\%$. Іноді використовують граничну відносну похибку, яка враховує ймовірність статистичного висновку $V_\Delta = tV\mu$.

Дуже часто на практиці використовується ***показник точності оцінок***, який показує рівень точності вибіркових оцінок різних ознак. Показник точності оцінок розраховують за формулою:

$$Cs = \frac{Cv}{\sqrt{n}},$$

де *Cv* – коефіцієнт варіації ознаки, а *n* – об'єм вибірки.

7.3. Різновиди вибірок

Формування вибірки — не безладний процес. Ця дія виконується за певними правилами. Передусім визначається основа вибірки. У сукупностях, які складаються з «фізичних» елементів, одиниця основи може репрезентувати або окремий елемент сукупності, або певне їх угруповання. Наприклад, вивчається структура популяції рослини, що розмножується вегетативно. Загальна кількість особин *N* агрегована у *M* клон, кожен клон має *Nj* особин. Одиницею основи вибірки може бути особина або клон. Відповідно формується вибіркова сукупність: у першому випадку вибирається *n* особин із загального їх числа *N*, у другому — *m* клонів із загального їх числа *M*.

Найпростішою основою вибірки є перелік елементів генеральної сукупності, пронумерований від 1 до *N*. Простими вважаються також набори звітів, анкет, карток тощо.

На практиці досліджувані сукупності мають, як правило, не одну, а низку альтернативних основ для вибірки. Наукове обґрунтування та правильний вибір основи — перша передумова забезпечення репрезентативності результатів вибіркового спостереження.

Від основи вибірки залежить спосіб добору елементів сукупності для обстеження. Найчастіше використовують способи добору: простий випадковий, механічний, розшарований (районований), серійний.

Простий випадковий добір провадиться жеребкуванням або за допомогою таблиць випадкових чисел. Це класичний спосіб формування

вибіркової сукупності, який передбачає попередню досить складну підготовку до формування вибірки. Для жеребкування на кожну одиницю генеральної сукупності необхідно заготувати відповідну фішку; при використанні таблиць випадкових чисел усі елементи цієї сукупності мають бути пронумеровані. У великих за обсягом сукупностях така робота здебільшого недоцільна, а часом і неможлива. Тому на практиці застосовуються інші різновиди випадкових вибірок.

Механічний добір. Основа вибірки — упорядкована множина елементів сукупності. Добір елементів здійснюється через рівні інтервали. Крок інтервалу обчислюється діленням обсягу сукупності *N* на передбачений обсяг вибірки *n.* Початковий елемент вибірки визначається як випадкове число всередині першого інтервалу, другий елемент залежить від початкового числа й кроку інтервалу. Так, для частки вибірки $\frac{n}{N} = 0{,}05$ кроком інтервалу є число $\frac{n}{N} = \frac{1}{0{,}05} = 20$, тобто у вибірку має потрапити кожний двадцятий елемент. Якщо початковий елемент — випадкове число 7, то другим елементом буде 7 + 20 = 27, третім — 27 + 20 = 47 і т. д.

Механічна вибірка порівняно з простою випадковою ефективніша, її простіше здійснити. Проте за наявності циклічних коливань значень ознаки, цикл коливань яких збігається з інтервалом, можливий зсув вибіркових оцінок. Похибку механічної вибірки обчислюють за формулою похибки безповторної вибірки.

Вивчаючи безперервні в часі процеси, зокрема екологічні явища, проводять ***моментні спостереження.*** Суть їх — у періодичній фіксації стану процесу на певні моменти часу, які вибирають за схемою випадкової або механічної вибірки (через певні інтервали часу).

Щодо повноти охоплення елементів сукупності, то моментне спостереження суцільне, воно вибіркове впродовж часу, бо охоплює не весь час роботи устаткування, а лише певні моменти. У разі правильної організації моментні обстеження забезпечують досить точні результати швидко і з меншими витратами, ніж при суцільному спостереженні.

Розшарований (районований, типовий) добір — це спосіб формування вибірки з урахуванням структури генеральної сукупності. На відміну від простого випадкового та механічного добору, які проводяться в цілому по генеральній сукупності, розшарований передбачає її попередню структуризацію й незалежний добір елементів у кожній складовій. Обсягом розшарованої

вибірки є сума частинних вибірок n_j , тобто $n = \sum_j^m n_j$, де m число складових (груп, типових районів тощо).

Похибку розшарованої вибірки обчислюють, використовуючи середню з групових дисперсій $\overline{\sigma^2}$. Якщо сформовані групи об'єднують «схожі» елементи, а групові середні величини помітно різні, варіація ознаки в групах буде значно меншою, ніж по сукупності. У такому разі $\overline{\sigma^2} < \sigma^2$, а отже, похибка розшарованої вибірки порівняно з простою випадковою чи механічною буде менша:

$$\Delta = t\sqrt{\frac{\overline{\sigma^2}}{n}\left(1-\frac{n}{N}\right)}.$$

Для того щоб забезпечити більшу точність розшарованої вибірки, слід обґрунтувати ознаку розшарування сукупності, число складових частин *m,* обсяг частинних вибірок n_j і спосіб добору. Зменшення варіації ознаки при розшаруванні сукупності можливе за умови, що ознака розшарування сукупності корелює з ознакою, характеристики якої оцінюються. Ці ознаки співвідносяться як причина й наслідок.

Відповідно до правила розкладання дисперсій $\overline{\sigma^2} = \sigma^2 - \delta^2$ або $\overline{\sigma^2} = \sigma^2\left(1-\eta^2\right)$. Отже, розшарування сукупності зменшує похибку вибірки на частку $\left(1-\eta^2\right)$. Чим щільніший зв'язок між ознаками, тим помітніше зменшення похибки. При $\eta^2 = 0{,}50$ похибка вибірки зменшується вдвічі, при $\eta^2 = 0{,}66$ — утричі.

У практиці вибіркових спостережень застосовують різні способи визначення обсягу вибіркової сукупності *n* та її складових n_j. Найпростіший з них, коли всі *m* груп подані однаковою кількістю елементів:

$$n_j = \frac{n}{m}.$$

Проте застосування цього способу обмежене. Якщо чисельності груп у генеральній сукупності N_j дуже різні, може виникнути ситуація, коли $n_j > N_j$.

Найчастіше застосовують пропорційний добір, який передбачає однакове для всіх складових представництво, тобто частки $\frac{n_j}{N_j}$ однакові й обсяг частинної вибірки залежить від обсягу відповідної складової сукупності:

$$n_j = \frac{n_j}{N_j} N_j.$$

Оптимальним щодо мінімізації похибки є добір, пропорційний до середнього квадратичного відхилення:

$$n_j = \frac{N_j \sigma_j}{\sum_1^m N_j \sigma_j} n .$$

Очевидно, що обсяг вибірки залежить від рівня варіації ознаки в окремих складових генеральної сукупності. Однорідні групи подаються меншим числом елементів, неоднорідні — більшим. Відсутність даних про варіацію ускладнює практичну реалізацію такого способу вибірки.

Різновидом розшарованої вибірки є ***метод квот,*** коли обсяг частинних вибірок n_j визначається завчасно. Цей спосіб поширений при вивченні рідкісних видів, мисливської фауни тощо. Так, при вивченні копитних, що передбачає відстріл, установлюються квоти, наприклад відібрати 10 особин. В який спосіб «заповнити квоти», дослідник вирішує сам. Метод квот не гарантує незсуненості вибіркових оцінок.

Серійна вибірка. Одиниця основи вибірки — серія елементів. Серії складаються з одиниць, які пов'язані або територіально (райони, популяції), або організаційно (клони, стада тощо). Вибіркова сукупність серій формується за схемами механічної або простої випадкової вибірки. Дібрана серія розглядається як одне ціле, обстеженню підлягають усі без винятку елементи серії. При обчисленні похибки вибірки враховується міжсерійна варіація:

$$\Delta = t\sqrt{\frac{\delta^2}{m}\left(1-\frac{m}{M}\right)} .$$

де δ^2 — міжсерійна дисперсія; m та M— число серій відповідно у вибірці та генеральній сукупності.

Похибка серійної вибірки буде меншою порівняно з похибкою простої випадкової чи механічної вибірки в тому разі, якщо серії більш-менш однорідні й варіація серійних середніх незначна. Зростання міжсерійної варіації призводить до збільшення похибки вибірки.

Використання того чи іншого способу формування вибіркової сукупності залежить від мети вибіркового обстеження, можливостей його організації та проведення. Іноді поєднуються різні способи добору: механічний і серійний, розшарований і механічний, випадковий і серійний.

Таке поєднання можливе в рамках ***багатоступеневої вибірки.*** Ступенів може бути два, три й більше. Кожний із них має свою, відмінну від інших основу вибірки. Відповідно поділяються й одиниці вибірки: першого ступеня, другого і т. ін. Повнота охоплення основи й схема добору одиниць на різних ступенях різняться.

Наприклад, сукупність містить K одиниць першого ступеня, які складаються з M одиниць другого ступеня, ті, у свою чергу, об'єднують N_j

одиниць третього ступеня. Саме така триступенева вибірка застосовується при організації обстеження виду. Наприклад, одиниці першого ступеня — це популяції; одиниці другого ступеня — клони; одиниці третього ступеня — особини.

Отже, вибір елементів для безпосереднього обстеження здійснюється на останньому, третьому ступені формування вибіркової сукупності. Частка її відносно до генеральної сукупності залежить від часток вибірки на всіх ступенях. Якщо припустити, що до вибірки потрапила одна з десяти популяцій (d_1 = 0,10), у цих популяціях відібраний кожен п'ятий клон (d_2 = 0,20), а у відібраних клонах обстежуються 4% особин (d_3 = 0,04), то частка вибіркової сукупності в генеральній становить:

$d = d_1 d_2 d_3 = 0,1 \cdot 0,02 \cdot 0,04 = 0,0008$,

тобто обстеженню підлягає 0,08% особин.

Багатоступенева вибірка значно зменшує витрати на обстеження й порівняно з іншими вибірками більш ефективна.

Якщо обстежують сукупність за двома й більше ознаками, які різняться варіацією, ефективною є ***багатофазна вибірка.*** Суть її в тому, що для різних ознак формуються вибіркові сукупності різного обсягу. На відміну від багатоступеневої вибірки багатофазна використовує для всіх ознак одну й ту саму основу вибірки, проте програма обстеження різна.

Вибіркові сукупності формуються поетапно — фазами. З генеральної сукупності утворюється первинна вибірка, а з первинної — підвибірка і т. д. На кожній наступній фазі обсяг підвибірки зменшується, а програма обстеження розширюється. Вибіркові оцінки кожної фази використовуються як додаткова інформація на наступних фазах, що підвищує точність результатів вибіркового обстеження.

При організації багатофазної вибірки можливі комбінації різних способів і видів вибірки. Багатофазна вибірка поєднується з багатоступеневою, а також із суцільним спостереженням.

7.4. Визначення обсягу вибірки

У процесі проектування вибіркових спостережень визначають ***мінімально достатній обсяг вибірки,*** при якому вибіркові оцінки репрезентували б основні властивості генеральної сукупності. Занадто великий обсяг вибірки потребує зайвих витрат, а занадто малий призведе до збільшення похибки репрезентативності. Теорія вибіркового методу дає змогу науково обґрунтувати достатній обсяг вибірки.

Згідно з формулою граничної похибки вибірки $\Delta = t\sqrt{\frac{\sigma^2}{n}}$ обсяг вибірки

$$n = \frac{t^2\sigma^2}{\Delta^2}$$

тобто залежить від ступеня однорідності генеральної сукупності, імовірності, з якою гарантується результат, і необхідної точності вибіркової оцінки. Практичне використання цієї формули ускладнюється через відсутність оцінки варіації.

Коли розрахований обсяг вибіркової сукупності *n* перевищує 5% обсягу генеральної сукупності *N*, його коригують на «безповторність вибірки». Скоригований обсяг вибірки

$$n' = \frac{n}{1+\frac{n}{N}}.$$

Щодо точності вибіркового обстеження, то доцільно контролювати відносну граничну похибку V_Δ. У такому разі мірою варіації ознаки є коефіцієнт варіації V_x і тоді:

$$n = \frac{t^2V_x^2}{V_\Delta^2}.$$

Необхідний обсяг вибірки можна розрахувати також на основі відносної похибки вибірки для частки:

$$n = \frac{t^2q}{V_\Delta^2 p}.$$

Очевидно, чим більша частка *p*, тим менший обсяг вибірки забезпечить необхідну точність результатів обстеження, і навпаки: для малих значень *p* обсяг вибірки збільшується.

У таблиці наведено обсяги вибірки, які забезпечують точність результатів обстеження малопоширених явищ з відносною стандартною похибкою, меншою за 10%.

Таблиця

Достатній обсяг вибірки для вивчення малопоширених явищ

p	*q/p*	*n* при $V_\mu \le 10\%$
0.20	4,0	400
0,15	5,7	570
0,12	7,3	730
0,10	9,0	900
0,09	10,1	1010
0,08	11,5	1150

У практиці вибіркових обстежень одночасно вивчаються кілька ознак. Якщо бажаний ступінь точності визначати для кожної ознаки окремо, то результатом розрахунків стане низка значень обсягу вибірки. З метою їх узгодження використовують або максимальний обсяг *n* (і тоді решта ознак оцінюється «надто точно»). або обсяг головної ознаки.

7.5. Статистична перевірка гіпотез

Статистична гіпотеза — це певне припущення щодо властивостей генеральної сукупності, яке можна перевірити, спираючись на результати вибіркового спостереження. Суть перевірки гіпотез полягає в тому, щоб визначити, узгоджуються чи ні результати вибірки з гіпотезою, випадковими чи невипадковими є розбіжності між гіпотезою і даними вибірки.

Найчастіше гіпотеза, яку належить перевірити, формулюється як відсутність розбіжності (нульова розбіжність) між невідомим параметром генеральної сукупності *G* і заданою величиною *A,* а тому її позначають H_0. Зміст гіпотези записують після двокрапки, наприклад H_0: $G = A$.

Кожній нульовій гіпотезі протиставляють альтернативну H_a. При формулюванні H_a враховується вагомість відхилень (*G*-*A*): для додатних відхилень H_a: $G > A$, для від'ємних — H_a : $G<A$, для тих і інших — $H_a : G \neq A$.

Якщо вибіркові дані суперечать гіпотезі H_0, вона відхиляється, коли ці дані узгоджуються з гіпотезою H_0, вона не відхиляється. Спираючись на результати вибірки, статистична перевірка гіпотез неминуче пов'язана з ризиком прийняття помилкового рішення: ризик I — відхилення правильної нульової гіпотези, ризик II — невідхилення нульової гіпотези, коли насправді правильною є альтернативна. Ці ризики конкуруючі, і зменшення ймовірності α одного зумовлює збільшення ймовірності β іншого. Оскільки уникнути ризиків неможливо, а наслідки їх, як правило, різновагомі, то в кожному конкретному дослідженні прагнуть мінімізувати той ризик, який пов'язаний з більшими втратами. Імовірності ризиків наведено в таблиці.

Таблиці

Імовірність ризиків помилкових рішень при перевірці гіпотез

Правильна гіпотеза	Прийнята гіпотеза	
	H_0	H_a
H_0	1- α	α
H_a	β	1- β

Правило, за яким гіпотеза H_0 відхиляється або не відхиляється (приймається), називається ***статистичним критерієм.*** Математичною основою будь-якого критерію є статистична характеристика *Z*, значення якої визначається за даними вибірки, а закон розподілу відомий. Кожне значення характеристики *Z* має певну ймовірність *F* (*Z*). Якщо вибіркове значення *Z* малоймовірне, гіпотеза H_0 відхиляється.

Межу малоймовірності *Z* називають ***рівнем істотності*** α. Очевидно, що α — це ймовірність ризику I, а тому залежно від змісту гіпотези H_0 і наслідків її відхилення рівень істотності визначають у кожному конкретному дослідженні. Зазвичай вибирають один із рівнів α, для яких табульовані значення статистичних характеристик критеріїв. Це α = 0,10; 0,05; 0,025; 0,01.

Значення статистичної характеристики критерію $Z_{1-\alpha}$ поділяє множину вибіркових значень *Z* на дві частини: а) область допустимих значень і б) критичну область. Якщо вибіркове значення *Z* потрапляє у критичну область, гіпотеза H_0 відхиляється, якщо в область допустимих значень — не відхиляється. Саме тому значення $Z_{1-\alpha}$ називають критичним.

Залежно від того, як сформульована альтернативна гіпотеза, критична область може бути односторонньою (ліво- чи правосторонньою) або двосторонньою (рис. 7.2).

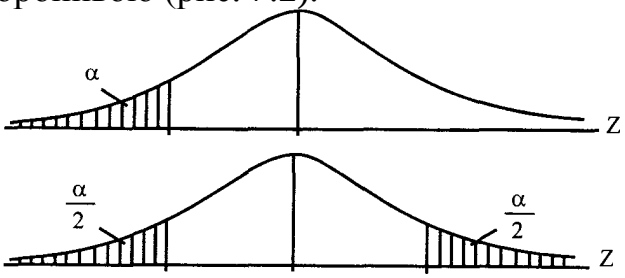


Рис. 7.2. Лівостороння та двостороння критичні області

Статистичною характеристикою гіпотези $H_0 : \bar{x}_1 = \bar{x}_2$, є нормоване відхилення середніх

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

яке підпорядковане розподілу Стьюдента з числом ступенів свободи $k = n_1 + n_2 - 2$.

У разі двосторонньої перевірки гіпотези, коли $H_a : \bar{x}_1 \neq \bar{x}_2$, використовують критичне значення для $\frac{\alpha}{2}$, наприклад при α = 0,05 це буде $t_{0,975}\ (k)$.

Отже, статистична гіпотеза перевіряється в такій послідовності:

а) формулюють нульову H_0 та альтернативну H_a гіпотези;

б) вибирають статистичну характеристику ***Z,*** за значеннями якої перевіряють правильність гіпотези H_0;

в) визначають рівень істотності а і відповідне йому критичне значення $Z_{1-\alpha}$; залежно від формулювання гіпотез H_0 і H_a критична область може бути одно- або двосторонньою;

г) за результатами вибірки розраховують фактичне (вибіркове) значення статистичної характеристики Z, яке порівнюють з критичним $Z_{1-\alpha}$, якщо $Z > Z_{1-\alpha}$, гіпотеза H_0 відхиляється, при $Z < Z_{1-\alpha}$ — не відхиляється.

Процедура перевірки гіпотез використовується при порівнянні вибіркових характеристик (середньої, частки, дисперсії) з відповідними нормативами, порівнянні характеристик двох вибіркових сукупностей, оцінюванні істотності розбіжностей двох розподілів, у дисперсійному та кореляційному аналізі.

8. МЕТОДИ АНАЛІЗУ ВЗАЄМОЗВ'ЯЗКІВ

8.1. Види взаємозв'язків

Усі явища навколишнього світу – взаємозв'язані й взаємозумовлені. У складному переплетенні всеохоплюючого взаємозв'язку будь-яке явище є наслідком дії певної множини причин і водночас — причиною інших явищ. Причини та наслідки пов'язані неперервними ланцюгами прямо або опосередковано, що схематично ілюструє рис. 8.1. Так, незалежне в межах зображеного графіку зв'язку явище x_1 є причиною явищ x_2, x_3, x_5. Із них явище x_3, у свою чергу, впливає на x_4 , а x_4 на x_5.

Поряд із причинними існують зв'язки паралельних явищ, на які впливає спільна причина. На рис. 8.1 це зв'язок між x_2 і x_3, які мають спільну причину x_1.

Визначальна мета вимірювання взаємозв'язків — виявити і дати кількісну характеристику причинних зв'язків. Суть причинного зв'язку полягає в тому, що за певних умов одне явище спричинює інше. Причина сама по собі не визначає наслідку, останній залежить також від умов, в яких діє причина. Вивчаючи закономірності зв'язку, причини та умови об'єднують в одне поняття «фактор». Відповідно ***ознаки,*** які характеризують фактори, називаються ***факторними,*** а ті, що характеризують наслідки, — ***результативними.***

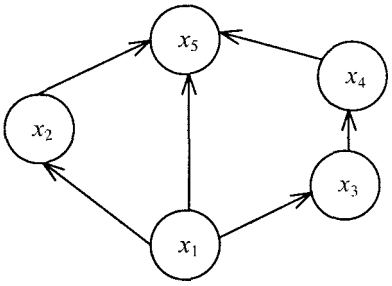


Рис. 8.1. Графік взаємозв'язків

Аналіз характеру взаємозв'язків та оцінювання сили впливу факторів на результат є передумовою розробки науково обґрунтованих рішень, прогнозування й регулювання складних явищ і процесів.

Розрізняють два типи зв'язків — ***функціональні*** та ***стохастичні.*** У разі функціонального зв'язку кожному значенню фактора *x* відповідає одне або кілька чітко визначених значень *y.* Такою, наприклад, є залежність довжини ртутного стовпчика від температури навколишнього середовища. Знаючи *x,* можна в кожному окремому випадку точно визначити результат *y.*

До функціонального типу належать зв'язки між показниками — адитивні *(a + b + c*) або мультиплікативні (*a = be, c = a/b),* а також залежність середніх величин від структури сукупності.

На відміну від функціональних, стохастичні зв'язки неоднозначні. Наприклад, залежність захворюваності населення від екологічного стану довкілля. На забруднених радіонуклідами територіях, як і на інших, стан здоров'я мешканців коливається від «тяжко хворого» до «практично здорового». Проте в середньому в таких регіонах порівняно з екологічно чистими захворюваність значно вища.

Стохастичні зв'язки виявляються як узгодженість варіації двох чи більше ознак. У ланці зв'язку «*x* → *y*» кожному значенню ознаки *x* відповідає певна множина значень ознаки *y,* які утворюють так званий ***умовний розподіл.*** Стохастичний зв'язок, відбиваючи множинність причин і наслідків, виявляється в зміні умовних розподілів, що схематично ілюструє табл. 8.1.

Якщо умовні розподіли замінюються одним параметром — середньою $\bar{y}_i$, то такий зв'язок називають ***кореляційним.*** Отже, кореляційний зв'язок є різновидом стохастичного і виявляється зміною середніх умовних розподілів.

Таблиця 8.1

Види взаємозв'язків і їх особливості

Факторна ознака x_i	Результативна ознака *y* за наявності зв'язку		
	функціонального	стохастичного	кореляційного
x_1	y_1	$y_1 y_2$	$\bar{y}_1$

x_2	y_2	$y_1 y_2 y_3$	$\bar{y}_2$
x_3	y_3	$y_2 y_3 y_4$	$\bar{y}_3$
…	…	…	…
x_m	y_m	$y_{m-1} y_m$	$\bar{y}_m$

8.2. Регресійний аналіз

Під ***регресією*** розуміють математично виражену взаємозалежність між ознаками. Важливою характеристикою кореляційного зв'язку є ***лінія регресії***— емпірична в моделі аналітичного групування і теоретична в моделі регресійного аналізу. ***Емпірична лінія регресії*** представлена груповими середніми результативної ознаки $\bar{y}_j$, кожна з яких належить до відповідного інтервалу значень групувального фактора x_j. ***Теоретична лінія регресії*** описується певною функцією *Y=f(x), якy* називають ***рівнянням регресії, а Y — теоретичним рівнем результативної ознаки.***

На відміну від емпіричної, теоретична лінія регресії неперервна. Так. уважають, що маса дорослої людини в кілограмах має бути на 100 одиниць менша за її зріст у сантиметрах. Співвідношення між масою і зростом можна записати у вигляді рівняння:

Y = -100 + *x*, де *У* — маса, *x* — зріст.

Безперечно, така форма зв'язку між масою та зростом людини надто спрощена. Насправді збільшення маси не жорстко пропорційне до збільшення зросту. Люди одного зросту мають різну масу, проте в середньому зі збільшенням зросту маса зростає. Для точнішого відображення зв'язку між цими ознаками в рівняння слід увести другий параметр, який був би коефіцієнтом пропорційності при *x,* тобто У= - 100 + *bx.*

Рівняння регресії в такому вигляді описує числове співвідношення варіації ознак *x* і *y* в середньому. Коефіцієнт пропорційності при цьому відіграє визначальну роль. Він показує, на скільки одиниць у середньому змінюється *y* зі зміною *x* на одиницю. У разі прямого зв'язку *b* — величина додатна, у разі оберненого — від'ємна.

Подаючи *y* як функцію *x,* тим самим абстрагуються від множинності причин, штучно спрощуючи механізм формування варіації *y.* Аналіз причинних комплексів здійснюється за допомогою множинної регресії.

Різні явища по-різному реагують на зміну факторів. Для того щоб відобразити характерні особливості зв'язку конкретних явищ. статистика використовує різні за функціональним видом регресійні рівняння. Якщо зі зміною фактора *x* результат *y* змінюється більш-менш рівномірно, такий зв'язок

описується лінійною функцією $Y = a + bx$. Коли йдеться про нерівномірне співвідношення варіацій взаємозв'язаних ознак (наприклад, коли прирости значень y зі зміною x прискорені чи сповільнені або напрям зв'язку змінюється), застосовують нелінійні регресії, зокрема:

степеневу	$Y = ax^b$
гіперболічну	$Y = a + \frac{b}{x}$
параболічну	$Y = a + bx + cx^2$ тощо

Вибір та обґрунтування функціонального виду регресії ґрунтується на теоретичному аналізі суті зв'язку. Нехай вивчається зв'язок між урожайністю та кількістю опадів. Надто мала і надто велика кількість опадів спричинюють зниження врожайності, максимальний її рівень можливий за умови оптимальної кількості опадів, тобто зі збільшенням факторної ознаки (опади) урожайність спершу зростає, а потім зменшується. Залежність такого роду описується параболою $Y = a + bx + cx^2$.

Зауважимо, що теоретичний аналіз суті зв'язку, хоча й дуже важливий, лише окреслює особливості форми регресії і не може точно визначити її функціонального виду. До того ж у конкретних умовах простору і часу межі варіації взаємозв'язаних ознак x і y значно вужчі за теоретично можливі. І якщо кривина регресії невелика, то в межах фактичної варіації ознак зв'язок між ними досить точно описується лінійною функцією. Цим значною мірою пояснюється широке застосування лінійних рівнянь регресії:

$Y = a + bx$.

Параметр b ***(коефіцієнт регресії)*** — величина іменована, має розмірність результативної ознаки і розглядається як ***ефект впливу*** x на y. Параметр a — вільний член рівняння регресії, це значення y при $x = 0$. Якщо межі варіації x не містять нуля, то цей параметр має лише розрахункове значення.

Параметри рівняння регресії визначаються методом найменших квадратів, основна умова якого — мінімізація суми квадратів відхилень емпіричних значень y від ***теоретичних*** Y:

$$\sum (y - Y)^2 = \min .$$

Математично доведено, що значення параметрів a та b, при яких мінімізується сума квадратів відхилень, визначаються із системи нормальних рівнянь:

$$\begin{cases} \sum y = na + b\sum x, \\ \sum xy = a\sum x + b\sum x^2. \end{cases}$$

Розв'язавши цю систему, знаходимо такі значення параметрів:

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \sum x \sum y},$$

$$a = \bar{y} - b\bar{x}.$$

Рівняння регресії відбиває закон зв'язку між *x* і *y* не для окремих елементів сукупності, а для сукупності в цілому: закон, який абстрагує вплив інших факторів, виходить з принципу «за інших однакових умов». Вплив інших окрім *x* факторів зумовлює відхилення емпіричних значень *y* від теоретичних у той чи інший бік. Відхилення (*y* - *Y*) називають ***залишками*** і позначають символом *e.* Залишки, як правило, менші за відхилення від середньої, тобто $(y - Y) \le (y - \bar{y})$.

Відповідно ***загальна дисперсія***

$$\sigma_y^2 = \frac{1}{n}\sum_1^n (y - \bar{y})^2$$

залишкова дисперсія

$$\sigma_e^2 = \frac{1}{n}\sum_1^n (y - Y)^2 .$$

У невеликих за обсягом сукупностях коефіцієнт регресії схильний до випадкових коливань. Тому слід перевірити його істотність. Коли зв'язок лінійний, істотність коефіцієнта регресії перевіряють за допомогою *t*- критерію (Стьюдента), статистична характеристика якого для гіпотези $H_0 : b = 0$ визначається відношенням коефіцієнта регресії *b* до власної стандартної похибки μ_b, тобто $t = b / \mu_b$.

Стандартна похибка коефіцієнта регресії залежить від варіації факторної ознаки σ_x^2, залишкової дисперсії σ_e^2, і числа ступенів свободи *k = n - m,* де *m* — кількість параметрів рівняння регресії:

$$\mu_b = \sqrt{\frac{\sigma_e^2}{\sigma_x^2 (n - m)}} .$$

Для коефіцієнта регресії, як і для будь-якої іншої випадкової величини, визначаються довірчі межі $b \pm t\mu_b$.

Важливою характеристикою регресійної моделі є відносний ефект впливу фактора *x* на результат *y* — ***коефіцієнт еластичності:***

$$\gamma = b\frac{\bar{x}}{\bar{y}} .$$

Він показує, на скільки процентів у середньому змінюється результат *y* зі зміною фактора *x* на 1%.

Оцінити відносний ефект впливу фактора *x* на результат *y* можна безпосередньо на основі степеневої функції $Y = ax^b$ параметр *b* якої є коефіцієнтом еластичності. Степенева функція зводиться до лінійного виду

логарифмуванням $\lg Y = \lg a + b \lg x$. До класу степеневих належать функції споживання, виробничі функції тощо.

8.3. Оцінка щільності та перевірка істотності кореляційного зв'язку

Поряд із визначенням характеру зв'язку та ефектів впливу факторів *x* на результат *y* важливе значення має оцінка щільності зв'язку, тобто оцінка узгодженості варіації взаємозв'язаних ознак. Якщо вплив факторної ознаки *x* на результативну *y* значний, це виявиться в закономірній зміні значень *y* зі зміною значень x, тобто фактор *x* своїм впливом формує варіацію *y*. За відсутності зв'язку варіація *y* не залежить від варіації x.

Для оцінювання щільності зв'язку статистика використовує низку коефіцієнтів з такими спільними властивостями:

• за відсутності будь-якого зв'язку значення коефіцієнта наближається до нуля; при функціональному зв'язку — до одиниці;

• за наявності кореляційного зв'язку коефіцієнт виражається дробом, який за абсолютною величиною тим більший, чим щільніший зв'язок.

Серед мір щільності зв'язку найпоширенішим є ***коефіцієнт кореляції*** Пірсона. Позначається цей коефіцієнт символом *r*. Оскільки сфера його використання обмежується лінійною залежністю, то і в назві фігурує слово «лінійний». Обчислення лінійного коефіцієнта кореляції *r* ґрунтується на відхиленнях значень взаємозв'язаних ознак *x* і *y* від середніх.

За наявності прямого кореляційного зв'язку будь-якому значенню $x_i > \bar{x}$ відповідає значення $y_i > \bar{y}$ а $x_k < \bar{x}$ відповідає $y_k < \bar{y}$. Узгодженість варіації *x* і *y* схематично показано на рис. 7.2 у вигляді кореляційного поля зі зміщеною системою координат.

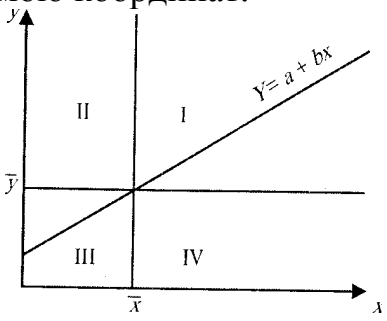


Рис. 8.2. Узгодженість варіації взаємозв'язаних ознак

Точка, координатами якої є середні $\bar{x}$ і $\bar{y}$, поділяє кореляційне поле на чотири квадранти, в яких по-різному поєднуються знаки відхилень від середніх:

Квадрант	$(x - \bar{x})$	$(y - \bar{y})$

I	+	+
II	-	+
III	-	-
IV	+	-

Для точок, розміщених у I та III квадрантах, добуток $(x-\bar{x})(y-\bar{y})$ додатний, а для точок з квадрантів II і IV — від'ємний. Чим щільніший зв'язок між ознаками *x* і *y,* тим більша алгебраїчна сума добутків відхилень $\sum_1^n (x-\bar{x})(y-\bar{y})$. Гранична сума цих добутків дорівнює $\sqrt{\sum_1^n (x-\bar{x})^2 \sum_1^n (y-\bar{y})^2}$.

Коефіцієнт кореляції визначається відношенням зазначених сум:

$$r = \frac{\sum_1^n (x-\bar{x})(y-\bar{y})}{\sqrt{\sum_1^n (x-\bar{x})^2 \sum_1^n (y-\bar{y})^2}}.$$

Очевидно, що в разі функціонального зв'язку фактична сума відхилень дорівнює граничній, а коефіцієнт кореляції $r = \pm 1$; при кореляційному зв'язку абсолютне його значення буде тим більшим, чим щільніший зв'язок.

Коефіцієнт кореляції, оцінюючи щільність зв'язку, указує також на його напрям: коли зв'язок прямий, *r*— величина додатна, а коли він зворотний — від'ємна. Знаки коефіцієнтів кореляції і регресії однакові, величини їх взаємозв'язані функціонально:

$$r = b\frac{\sigma_x}{\sigma_y}; \qquad b = r\frac{\sigma_y}{\sigma_x}.$$

Завдяки цьому один коефіцієнт можна обчислити, знаючи інший.

Вимірювання щільності нелінійного зв'язку ґрунтується на співвідношенні варіацій теоретичних та емпіричних (фактичних) значень результативної ознаки *y.* Як зазначалося в підрозд. 5.6, відхилення індивідуального значення ознаки *y* від середньої $(y-\bar{y})$ можна розкласти на дві складові. У регресійному аналізі це відхилення від лінії регресії (*y* - *Y*) та відхилення лінії регресії від середньої $(Y-\bar{y})$.

Відхилення $(Y-\bar{y})$ є наслідком дії фактора *x*, відхилення (*y*-*Y*) — наслідком дії інших факторів. Взаємозв'язок факторної та залишкової варіацій описується правилом декомпозиції варіації:

$$\sigma_y^2 = \delta_Y^2 + \sigma_e^2$$

де $\sigma_y^2 = \frac{1}{n}\sum_1^n (y-\bar{y})^2$ — загальна дисперсія ознаки *y;*

$\delta_Y^2 = \frac{1}{n}\sum_1^n (Y - \bar{y}) = \frac{1}{n}(a\sum x + b\sum xy) - \bar{y}^2$ — факторна дисперсія;

$\sigma_e^2 = \frac{1}{n}\sum_1^n (y - Y)^2$ — залишкова дисперсія.

Очевидно, значення факторної дисперсії δ_Y^2 буде тим більшим, чим сильніший вплив фактора *x* на *y*. Відношення факторної дисперсії до загальної розглядається як міра щільності кореляційного зв'язку і називається ***коефіцієнтом детермінації:***

$$R^2 = \frac{\delta_Y^2}{\sigma_y^2}$$

Корінь квадратний з коефіцієнта детермінації називають ***індексом кореляції R.*** Мірою щільності зв'язку є ***кореляційне відношення.***

$$\eta^2 = \frac{\delta^2}{\sigma^2},$$

де δ^2 — міжгрупова дисперсія, яка вимірює варіацію ознаки *y* під впливом фактора *x*. а σ^2 — загальна дисперсія.

Обчислення та інтерпретація коефіцієнта детермінації R^2 і кореляційного відношення η^2 показують: ці характеристики щільності зв'язку за змістом ідентичні, вони характеризують внесок фактора *x* у загальну варіацію результату *y*.

Перевірка істотності кореляційного зв'язку ґрунтується на порівнянні фактичних значень R^2 і η^2 з критичними, які могли б виникнути за відсутності зв'язку. Якщо фактичне значення R^2 чи η^2 перевищує критичне, то зв'язок між ознаками не випадковий. Гіпотеза, що перевіряється, формулюється як нульова:

$H_0 : R^2 = 0$ або $H_0 : \eta^2 = 0$.

Критичні значення характеристик щільності зв'язку для рівня істотності $\alpha = 0{,}05$ і відповідного числа ступенів свободи для факторної дисперсії k_1 і залишкової k_2, наведено в табл. III. Ступені свободи залежать від обсягу сукупності *n* та числа груп або параметрів функції *m,* тобто $k_1 = m-1$, $k_2 = n- m$.

Розглянута процедура перевірки істотності зв'язку є складовою дисперсійного аналізу, розробленого Р. Фішером. Характеристика критерію Фішера — дисперсійне відношення *F* — функціонально пов'язана з кореляційним відношенням $F\frac{\eta^2}{1-\eta^2}\frac{k_2}{k_1}$, а тому результати перевірки будуть ідентичні.

8.4. Рангова кореляція

Взаємозв'язок між ознаками, які можна зранжувати, передусім на основі бальних оцінок, вимірюється методами рангової кореляції. ***Рангами*** називають числа натурального ряду, які згідно зі значеннями ознаки надаються елементам сукупності і певним чином упорядковують її. Ранжування проводиться за кожною ознакою окремо: перший ранг надається найменшому значенню ознаки, останній — найбільшому або навпаки. Кількість рангів дорівнює обсягу сукупності. Очевидно, зі збільшенням обсягу сукупності ступінь «розпізнаваності» елементів зменшується. З огляду на те, що рангова кореляція не потребує додержання будь-яких математичних передумов щодо розподілу ознак, *зокрема* вимоги нормальності розподілу, рангові оцінки щільності зв'язку доцільно використовувати для сукупностей невеликого обсягу.

Ранги, надані елементам сукупності за ознаками x і y, позначають відповідно Rx_j та Ry_j. Залежно від ступеня зв'язку між ознаками певним чином співвідносяться й ранги. При прямому функціональному зв'язку $Rx_j = Ry_j$, тобто відхилення між рангами $d_j=Rx_j-Ry_j = 0$, отже, і сума квадратів відхилень $\sum_1^m d_j^2 = 0$. При зворотному функціональному зв'язку $\sum d_j^2 = \frac{1}{3}n(n^2-1)$, де n — число рангів. Якщо зв'язок між ознаками відсутній, $\sum_1^n d_j^2$ являє собою середню арифметичну цих крайніх значень:

$$\sum_1^n d_j^2 = \frac{1}{2}\left[0+\frac{1}{3}n(n^2-1)\right] = \frac{1}{6}n(n^2-1),$$

а отже.

$$\frac{6\sum_1^n d_j^2}{n(n^2-1)} = 1.$$

Спираючись на зазначену математичну тотожність, К. Спірмен запропонував формулу для коефіцієнта рангової кореляції:

$$\rho = 1 - \frac{6\sum_1^n d_j^2}{n(n^2-1)}$$

Цей коефіцієнт має такі самі властивості, як і лінійний коефіцієнт кореляції: змінюється в межах від - 1 до + 1, водночас оцінює щільність зв'язку та вказує на його напрям.

Якщо два і більше елементів сукупності мають однакові значення ознаки, їм надається середній ранг. Нехай, наприклад, друге за розміром значення ознаки мають три елементи сукупності (№ 2, 3, 4), тоді всім їм надається ранг —

(2+3+4)= 3, а щільність зв'язку можна оцінити за формулою лінійного коефіцієнта кореляції.

8.5. Оцінка узгодженості варіації атрибутивних ознак

Взаємозв'язки між атрибутивними ознаками аналізуються на підставі таблиць взаємної спряженості (співзалежності).

Оцінка щільності стохастичного зв'язку ґрунтується на відхиленнях частот (часток) умовного та безумовного розподілів, тобто на відхиленнях фактичних частот f_{ij}, від теоретичних F_{ij}, пропорційних до підсумкових:

$$Fy = \frac{f_{i0} f_{0j}}{n},$$

де f_{i0} — підсумкові частоти за ознакою x; f_{oj} — підсумкові частоти за ознакою y ; n — обсяг сукупності $\left(n = \sum_{i}^{m_x} f_{i0} = \sum_{j}^{m_y} f_{0j} \right)$.

Абсолютну величину відхилень фактичних частот f_{ij} від пропорційних F_{ij} характеризує квадратична спряженість χ^2 Пірсона:

$$\chi^2 = \sum_i \sum_j \frac{(f_{ij} - F_{ij})^2}{F_{ij}} = n\left[\sum_i \sum_j \frac{f_{ij}^2}{f_{i0} f_{0j}} \right].$$

За відсутності стохастичного зв'язку $\chi^2 = 0$. На основі розподілу ймовірностей χ^2 перевіряється істотність зв'язку. Критичні значення χ^2 для $\alpha = 0{,}05$ і числа ступенів свободи $k = (m_x - 1)(m_y - 1)$ наведено в табл. V.

Відносною мірою щільності стохастичного зв'язку слугує ***коефіцієнт взаємної спряженості*** (співзалежності). За умови, що $m_x = m_y$ використовують формулу Чупрова:

$$C = \sqrt{\frac{\chi^2}{n\sqrt{(m_x - 1)(m_y - 1)}}},$$

де m_x — число груп за ознакою x; m_y — число груп за ознакою y. Оскільки за відсутності зв'язку між ознаками $\chi^2 = 0$, то і $C = 0$. При функціональному зв'язку $C \to 1$. У разі, коли $m_x \neq m_y$, віддають перевагу коефіцієнту спряженості Крамера:

$$C = \sqrt{\frac{\chi^2}{n(m_{\min} - 1)}},$$

де m_{min} — мінімальне число груп (m_x або m_y).

Якщо обидві взаємозв'язані ознаки альтернативні, тобто кількість груп $m_x = m_y = 2$, то за відсутності зв'язку добутки діагональних частот однакові:

$f_{11}f_{22} = f_{12}f_{21}$ Саме на відхиленнях добутків частот ґрунтуються характеристики зв'язку:

$$\chi^2 = n\frac{(f_{11}f_{22} - f_{12}f_{21})^2}{f_{01}f_{02}f_{10}f_{20}},$$

$$C = \sqrt{\frac{\chi^2}{n}} = \frac{f_{11}f_{22} - f_{12}f_{21}}{\sqrt{f_{01}f_{02}f_{10}f_{20}}}.$$

У літературі зі статистики коефіцієнт C для 4-клітинкової таблиці називається ***коефіцієнтом контингенції*** або ***асоціації.*** Очевидно, що за змістом він ідентичний коефіцієнту взаємної спряженості, а з χ^2 пов'язаний функціонально: $\chi^2 = nC^2$.

За допомогою коефіцієнта контингенції оцінимо щільність зв'язку між шкідливою звичкою палити і хворобами легенів (табл.8.11).

Таблиця 8.11

Розподіл пацієнтів клініки за результатами легеневих проб

Наявність звички палити	Результати легеневих проб		Разом
	Аномальні	Нормальні	
Палить	20	5	25
Не палить	10	15	25
Разом	30	20	50

$$C = \frac{20\cdot 15 - 10\cdot 5}{\sqrt{30\cdot 20\cdot 25\cdot 25}} = \frac{250}{612{,}3} = 0{,}408.$$

Значення $\chi^2 = nC^2 = 50\cdot 0{,}408^2 = 8{,}32$ перевищує критичне $\chi^2_{0,95}(1) = 3{,}89$. Істотність зв'язку доведено з імовірністю 0,95.

Корисною мірою при аналізі 4-клітинкових таблиць взаємної спряженості є відношення перехресних добутків або ***відношення шансів***

$$W = \frac{f_{11}f_{22}}{f_{12}f_{21}}.$$

Відношення шансів характеризує міру відносного ризику. У нашому прикладі

$$W = \frac{15\cdot 20}{5\cdot 10} = 6.$$

Отже, імовірність легеневих хвороб у тих, хто палить, у 6 разів вища порівняно з тими, хто не палить.

Зауважимо, що методи аналізу таблиць взаємної спряженості можна використати і для кількісних ознак. Будь-які технічні перешкоди відсутні. Проте слід пам'ятати, що коефіцієнт спряженості оцінює лише узгодженість

фактичного розподілу з пропорційним. При переставлянні рядків чи стовпців значення коефіцієнта *C* не зміниться. Міри щільності кореляційного зв'язку — коефіцієнт детермінації R^2 і кореляційне відношення η^2 — оцінюють не лише узгодженість частот, а й порядок, послідовність, в якій поєднуються різні значення ознак. Отже, ці характеристики зв'язку більш потужні. А загалом вибір методу вимірювання зв'язку і характеристик його щільності має ґрунтуватись на попередньому теоретичному аналізі суті явищ, характеру взаємозв'язків, наявній інформації.

9. МЕТОДИ ВІЗУАЛЬНОГО АНАЛІЗУ

9.1. Двовимірний (2М) візуальний аналіз

9.1.1. Гістограми

Термін „гістограма” вперше ввів Карл Пірсон у 1895 р. Гістограми дозволяють побачити як розподілені значення змінних у різних інтервалах групування даних, тобто, як часто змінні приймають певне значення у певному інтервалі.

Розрізняють наступні типи гістограм: прості, складені, з подвійною віссю *Y* і „звислі стовпчики”.

Прості гістограми є звичайними стовбчастими графіками розподілу для однієї змінної.

Складені гістограми являють собою розподіл частот для декількох змінних на одному графіку.

Гістограми з подвійною віссю *Y* являють собою варіант складеної гістограми, де для задавання масштабу використовуються обидві вісі *Y*.

Гістограми „звислі стовпчики” є візуальним способом перевірки розподілу змінної. Для цього типу гістограм характерним є те, що верх стовпчика прив’язують не до вісі, а до графіка певної функції.

9.1.2. Діаграми розсіювання (скатер-діаграми)

Застосовуються для візуального спостереження залежності між двома змінними. Дані зображаються точками у двовимірному просторі.

Розрізняють наступні діаграми розсіювання: прості, складені, з подвійною віссю *Y*, частотні і діаграми Вороного.

Прості діаграми розсіювання відбивають залежність між двома змінними.

Складені діаграми розсіювання включають декілька залежностей від однієї змінної.

Діаграми розсіювання з подвійною віссю *Y* є різновидом складених, але для побудови графіків використовують обидві вісі *Y*.

Частотні діаграми розсіювання дозволяють зобразити частоти точок, що перекриваються.

Діаграми Вороного будуються таким чином, що простір розділяється ділянки, що максимально наближаються до точок спостереження, тобто будуються „зони впливу”.

9.1.3. Діаграми розмаху

Вперше цей тип графіків був запропонований Тьюкі у 1970 р., ці діаграми ще називаються графіками „ящики-вуса”. В цих графіках діапазони значень вибраної змінної (або змінних) будуються окремо для груп спостережень, що визначаються значеннями категоризуючої змінної (або групуючої змінної).

Центр (напр. медіана або середнє) і статистики діапазонів або варіацій (напр. квартилі, стандартна похибка, стандартні відхилення, тощо) вираховуються окремо для кожної групи змінних. При побудові діаграм розмаху викиди (атипові варіанти) можуть не враховуватися або позначатися окремо.

9.1.4. Стовбчасті діаграми

Цей ти графіків подібний до гістограм, але для однієї змінної будується тільки один стовпчик.

Розрізняють прості і складені стовпчасті діаграми.

9.1.5. Лінійні графіки

Являють собою звичайні графіки функцій.

Розрізняють наступні типи лінійних графіків: прості, складені, з подвійною віссю *Y* і трасировочні *XY*.

Трасировочні лінійні графіки *XY* являють собою аналоги діаграм розсіювання, у яких крапки з’єднуються лінією в порядку зчитування з файлу даних.

9.1.6. Послідовні або накладені графіки

При побудові цих графіків значення однієї змінної накладається (сумується) зі значенням іншої.

Розрізняють наступні типи послідовних графіків: лінійні, ступінчасті, стовбчасті, зонні, тощо.

9.1.7. Колові діаграми або циклограми

Вперше цей тип діаграм був запропонований Хаскелл у 1922.

Розрізняють колові діаграми частот та значень. також колові діаграми бувають з виділеними секторами.

Колові діаграми частот за своєю суттю ідентичні гістограмам.

Колові діаграми значень будують для різних змінних, в такому випадку вони відбивають внесок кожної змінної в сукупність.

9.2. Тривимірний (3М) візуальний аналіз даних

Тривимірний візуальний аналіз дозволяє аналізувати дані у тривимірному просторі. при цьому можна або зіставляти графіки для кількох змінних або проводити аналіз трьох і більше змінних одночасно. Тривимірні графіки слід відрізняти від псевдотривимірних, для яких третя площина використовується для візуального придання об'єму і не несе ніякої іншої інформації. Псевдотривимірні графіки є декоративним варіантом двовимірних.

9.2.1. 3М стовпчасті діаграми

9.2.2. 3М блокові діаграми

9.2.3. 3М стрічкові діаграми

9.2.4. 3М лінійні графіки

9.2.5. 3М діаграми сплеску

3М діаграми сплеску нагадують діаграми розсіювання, але мають перпендикуляр на нижню площину.

9.2.6. 3М графіки поверхонь

9.2.7. Карти ліній рівня

Карти ліній рівня являють собою проекцію графіків поверхні.

9.2.8. 3М діаграми діапазонів

3М діаграми діапазонів нагадують 2М діаграми діапазонів, але будуються у тривимірному просторі.

9.2.9. Діаграми „літаючі ящики”

Діаграми „літаючі ящики” побудовані у вигляді стовпців, що вільно розміщуються у тривимірному просторі. Відбивають певні діапазони значень.

9.2.10. Діаграми „літаючі блоки”

Аналогічні „літаючим ящикам” і нагадують товсті стрічки.

9.2.11. Діаграми „подвійні стрічки”

9.2.12. 3М діаграми розсіювання

9.2.13. 3М трасировочні графіки

9.2.14. Спектральні діаграми

9.3. Тернарні графіки

Тернарні графіки використовуються у випадках зв’язку трьох змінних, коли сума їх дорівнює постійній величині.

9.3.1. 2М тернарні діаграми розсіювання

9.3.2. 3М тернарні діаграми розсіювання

9.3.3. Тернарні діаграми поверхонь

9.3.4. Тернарні карти ліній

9.3.5. Тернарні зонні карти

9.3.6. Тернарні трасировочні графіки

9.4. Піктографіки

На статистичних піктографіках спостереження або окремі випробування представлені у вигляді символів з багатьма елементами.

9.4.1. Колові піктографіки

Зірки.

Промені.

Багатокутники.

9.4.2. Послідовні піктографіки

Стовпці.
Профілі.
Лінійні.

9.4.3. Колові піктографіки

9.4.4. „Обличчя Чернова”

10. ФАКТОРНИЙ АНАЛІЗ

10.1. Основна мета

Головною метою факторного аналізу є скорочення кількості змінних (редукція даних) і визначення структури взаємозалежностей між змінними тобто класифікація змінних.

10.2. Факторний аналіз як метод редукції даних

Допустимо, що ви проводите дослідження росту людини причому паралельно вимірюєте цей параметр в дюймах та сантиметрах, таким чином у вас є дві змінні. Якщо в подальшому дослідженні буде вивчатися вплив різних харчових добавок на ріст, чи є доцільність продовжувати використання обох попередніх змінних? Імовірно що ні, так як ріст людини є однією характеристикою людини незалежно в яких одиницях ви його вимірюєте.

Об’єднання двох змінних в один фактор.
Залежність між двома змінними можна виявити за допомогою діаграм розсіювання. Отримана шляхом підгонки лінійна регресія дає графічне представлення залежності, якщо визначити нову змінну на основі лінії регресії, то ця змінна буде включати найбільш характерні риси обох змінних. Таким чином ми скоротили кількість змінних з двох до однієї. Отримана змінна (фактор) є в дійсності лінійною комбінацією двох вихідних змінних.

Аналіз головних компонент.
Приклад у якому дві зкорельовані змінні об’єднанні в один фактор, показує головну ідею факторного аналізу за методом головних компонент. Якщо кількість змінних, що об’єднується в один фактор збільшити, то основний принцип не міняється.

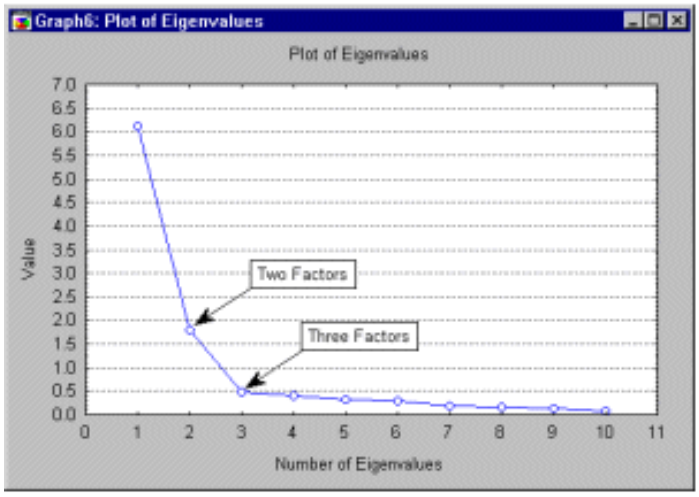
Виділення головних компонент.
В основному процедура виділення головних компонент подібна обертанню, що максимізує дисперсію (варімакс) вихідного простору змінних. Наприклад на діаграмі розсіювання можна розглядати лінію регресії, як вісь *X*, повернувши її так, що вона співпаде з прямою регресії. Цей тип обертання називається обертанням, що максимізує дисперсію, так як критерій (мета) обертання є максимізація дисперсії (мінливості) "нової" змінної (фактора) і мінімізація розсіювання навколо неї. У випадку для кількості змінних більшою за три важче це уявити це графічно. Але логіка перетворень залишається незмінною.
Декілька ортогональних факторів.
Після того, як знайдена лінія для якої дисперсія максимальна, навколо неї залишається деяке розсіювання даних, і природнім видається повторення попередньої процедури. В аналізі головних компонент так і роблять: після того, як виділений перший фактор, тобто перша лінія проведена, визначається наступна лінія максимізуючи залишкову варіацію (тобто розсіювання навколо першої прямої). і т.д. Таким чином фактори послідовно виділяються один за одним. Так як кожний наступний фактор визначається так, щоби максимізувати мінливість, що залишилася від попереднього, то фактори виявляються незалежними один від іншого, інакше кажучи незкорельованими або ортогональними.

Приклад застосування аналізу головних компонент.
Для прикладу візьмемо матрицю даних у якої дисперсії всіх змінних рівні і дорівнюють 1.0, тому загальна дисперсія дорівнює кількості змінних (10).

STATISTICA ФАКТОРНИЙ АНАЛІЗ	**Власні значення (factor.sta) Виділення: Головні компоненти**			
Значення	**Власні значення**	**% загальної дисперсії**	**Кумулята власн.значень**	**Кумулята %**
1	6.118369	61.18369	6.11837	61.1837
2	1.800682	18.00682	7.91905	79.1905
3	.472888	4.72888	8.39194	83.9194
4	.407996	4.07996	8.79993	87.9993
5	.317222	3.17222	9.11716	91.1716
6	.293300	2.93300	9.41046	94.1046
7	.195808	1.95808	9.60626	96.0626
8	.170431	1.70431	9.77670	97.7670
9	.137970	1.37970	9.91467	99.1467
10	.085334	.85334	10.00000	100.0000

У другому стовпчику (Власні значення) таблиці результатів знаходиться дисперсія нового виділеного фактора. У третьому стовпчику для кожного фактора привидиться відсоток від загальної дисперсії для кожного фактора. Дисперсії, що виділяються факторами називаються *власними значеннями фактора*.

Задача про кількість факторів.
Існує декілька способів визначення кількості факторів зокрема:
Критерій Кайзера.
Для початку можна відібрати фактори з власними значеннями більшими за 1.0. Це означає, що якщо фактор не відділяє частку дисперсії, еквівалентну хоча б дисперсії однієї змінної, то він ігнорується. Цей критерій був запропонований Кайзером (Kaiser, 1960), і є найбільш широковживаним. В попередньому прикладі згідно цього критерію слід залишить лише 2 фактора (дві головні компоненти).



Критерій кам'янистого осипища.
Критерій кам'янистого осипища є графічним методом, він був запропонований Кетелем (Cattell, 1966). Зобразимо власні значення факторів попереднього прикладу у вигляді графіка.

Кетель запропонував знайти таке місце на графіку, де зменшення власних значень зліва на право максимально уповільнюється. Допускається, що правіше цієї точки знаходиться тільки "факторіальне осипище". Що нагадує кам'янисте осипище що накопичується в нижній частині скелястого схилу. Відповідно до цього критерію можна залишити у даному прикладі 2 або 3 фактора.

Аналіз головних факторів
Перед тим як продовжити розгляд різних аспектів аналізу головних компонент, введемо поняття аналізу головних компонент. Повернемося до прикладу анкети про задоволеність життям, щоби сформувати іншу модель. Можемо уявити, що відповіді респондентів залежать від двох компонент. Спочатку вибираємо деякі підходящі загальні фактори, такі наприклад як "задоволеність своїм хобі".

Кожен пункт вимірює деяку частину цього аспекту задоволення, крім того, кожен пункт включає унікальний аспект задоволеності, не характерний для іншого пункту.

Спільності. Якщо ця модель вірна, то ви не можете чекати, що фактори будуть містити всю дисперсію в змінних, вони будуть відбивати тільки ту частину, що належить спільним факторам і розподілена по декількох змінних. На мові факторного аналізу частка дисперсії окремої змінної, що належить спільним факторам (і розподілена з іншими змінними) називається спільністю. Тому додатковою роботою для дослідника, що використовує цю модель. Є оцінка спільностей для кожної змінної, тобто частки дисперсії, що є спільною для всіх пунктів. Частка дисперсії, за яку відповідає кожний пункт тоду рівна сумарній дисперсії, що відповідає всім змінним мінус спільність. З загальної точки зору в якості оцінки спільності треба використовувати множинний коефіцієнт кореляції вибраної змінної з усіма іншими.

Головні фактори у порівнянні з головними компонентами.

Основна відмінність двох моделей факторного аналізу полягає в тому, що в аналізі головних компонент допускається, що має бути використана вся мінливість змінних, тоді як в аналізі головних факторів використовується тільки мінливість змінної, що є спільною з усіма змінними.

В більшості випадків ці два методи дають досить близькі результати. Але аналіз головних компонент частіше використовується як метод скорочення даних, в той же час як аналіз головних факторів краще застосовувати для виявлення структури даних

10.3. Факторний аналіз як метод класифікації

Повернемося до інтерпретації результатів факторного аналізу ігноруючи яким методом він проведений. Допустимо, що ви знаходитеся на тій стадії аналізу, коли в цілому знаєте скільки факторів слід виділити. Ви можете захотіти узнати значимість факторів, тобто чи можна інтерпретувати їх розумним чином. Для того що би продемонструвати яким чином це може бути зроблено, проводять дії у “зворотному порядку”, тобто починають з деякої осмисленої структури, а за тим дивляться, як це відбивається на результатах. Нижче приведена кореляційна матриця що відбиває задоволеність на роботі і дома.

STATISTICA ФАКТОРНИЙ	**Кореляції Порядкове n=100**	**видалення**	**(factor.sta) ПД**

АНАЛІЗ						
Змінна	**РАБОТА_1**	**РАБОТА_2**	**РАБОТА_3**	**ДОМ_1**	**ДОМ_2**	**ДОМ_3**
РОБОТА_1	1.00	.65	.65	.14	.15	.14
РОБОТА_2	.65	1.00	.73	.14	.18	.24
РОБОТА_3	.65	.73	1.00	.16	.24	.25
ДОМ_1	.14	.14	.16	1.00	.66	.59
ДОМ_2	.15	.18	.24	.66	1.00	.73
ДОМ_3	.14	.24	.25	.59	.73	1.00

Змінні, що відносяться до задоволеності на роботі тісніше зкорельовані між собою, а змінні, що відносяться до задоволеності удома також тісно зкорельовані між собою. Зкорельованість між обома групами незначна. Тому слід допустити існування двох відносно незалежних факторів. Один відноситься до задоволеності на роботі, а інший - удома.

Факторні навантаження.

Тепер проведемо аналіз головних компонент і розглянемо рішення з двома факторами. Для цього розглянемо кореляції між змінними і двома факторами, ці кореляції називаються факторними навантаженнями.

STATISTICA ФАКТОРНИЙ АНАЛІЗ	**Факторні навантаження(Нема обертання) Головні компоненти**	
Змінна	**Фактор 1**	**Фактор 2**
РОБОТА_1	.654384	.564143
РОБОТА_2	.715256	.541444
РОБОТА_3	.741688	.508212
ДОМ_1	.634120	-.563123
ДОМ_2	.706267	-.572658
ДОМ_3	.707446	-.525602
Загальна дисперсія	2.891313	1.791000
Частка загальної дисп.	.481885	.298500

Імовірніше, що перший фактор більш корелює з змінними, як інший. Це і слід було чекати, так як фактори виділяються послідовно і містять все менше і менше від загальної дисперсії.

Обертання факторної структури.

Ви можете зобразити факторні навантаження у вигляді діаграми розсіювання. На цій діаграмі кожна змінна представлена точкою. Можна повернути вісі в

любому напрямі без зміни відносного положення точок, але дійсні координати (тобто факторні навантаження) безперечно зміняться. Якщо повернути діаграму розсіювання для попереднього приклада на 45 град. То досягається більш чіткіше уявлення про навантаження, що визначають змінні, задоволеність на роботі і дома.

Методи обертання.

Існують різні методи обертання факторів. Метою цих методів є отримання зрозумілої (інтерпретуємої) матриці навантажень. Типовими методами обертання є стратегії варімакс, квартимакс та еквімакс.

Ідея обертання за методом варімакс була розглянута вище. Нижче приведена таблиця навантажень на повернуті фактори.

STATISTICA ФАКТОРНИЙ АНАЛІЗ	**Факторні навантаження(Варімакс нормаліз.) Виділення: Головні компоненти**	
Змінні	**Фактор 1**	**Фактор 2**
РАБОТА_1	.862443	.051643
РАБОТА_2	.890267	.110351
РАБОТА_3	.886055	.152603
ДОМ_1	.062145	.845786
ДОМ_2	.107230	.902913
ДОМ_3	.140876	.869995
Загальна дисперсія	2.356684	2.325629
Частка загальної дисп.	.392781	.387605

Після обертання у попередньому прикладі картина стає чіткішою і легко інтерпретується

12. РЯДИ ДИНАМІКИ. АНАЛІЗ ІНТЕНСИВНОСТІ ТА ТЕНДЕНЦІЙ РОЗВИТКУ

12.1. Суть і складові елементи динамічного ряду

Явища в природі безперервно змінюються. Протягом певного часу — місяць за місяцем, рік за роком — змінюються кількість населення, структура популяції, рівень продуктивності тощо. Аналіз розвитку — одне з важливих завдань статистики. Інформаційною базою його слугують динамічні (часові, хронологічні) ряди.

Динамічний ряд — це послідовність чисел, які характеризують зміну того чи іншого явища. Елементами динамічного ряду є перелік хронологічних дат (моментів) або інтервалів часу і конкретні значення відповідних статистичних показників, які називаються ***рівнями*** ряду.

При вивченні динаміки важливі не лише числові значення рівнів, а і їх послідовність. Як правило, часові інтервали між рівнями однакові (доба, декада, календарний місяць, квартал, рік). Узявши будь-який інтервал за одиницю, послідовність рівнів записуємо так: $y_1, y_2, y_3, \ldots, y_n$.

Залежно від статистичної природи показника-рівня розрізняють динамічні ряди первинні й похідні, ряди абсолютних, середніх і відносних величин. За ознакою часу динамічні ряди поділяються на інтервальні та моментні. Рівень ***моментного ряду*** фіксує стан явища на певний момент часу *t*, наприклад кількість особин на початок року, студентів — на 1 вересня і т. д. В ***інтервальному ряді*** рівень — це агрегований результат процесу й залежить від тривалості часового інтервалу: вилов риби за сезон тощо. Зауважимо, що й похідні показники, обчислені на основі інтервальних рядів, на відміну від моментних залежать від протяжності часу (середньодобова чи середньорічна первинна продукція на одиницю біомаси).

Біологічні процеси динамічні, що виявляються сталою зміною рівнів динамічного ряду. Поряд з динамічністю їм притаманна інерційність: зберігається механізм формування явищ і характер розвитку (темпи, напрям, коливання). При значній інерційності процесу й незмінності комплексу умов його розвитку правомірно очікувати в майбутньому ті властивості й характер розвитку, які були виявлені в минулому. Діалектична єдність мінливості і сталості, динамічності та інерційності формує характер динаміки, уможливлюючи статистичне прогнозування екологічних процесів.

При вивченні закономірностей розвитку статистика розв'язує низку завдань: вимірює інтенсивність динаміки, виявляє й описує тенденції, оцінює структурні зрушення, сталість і коливання рядів; виявляє фактори, які спричинюють зміни.

Передумовою аналізу будь-якого динамічного ряду є порівнянність статистичних даних, які його формують. Непорівнянність даних може зумовлюватися різними причинами:

• змінами в методології обліку та розрахунку показника, зокрема використання різних одиниць для вимірювання;

• змінами в структурі сукупності, а також територіальними змінами;

• різними критичними моментами реєстрації даних чи тривалістю періодів, до яких належать рівні.

Порівнянність даних забезпечується на етапах їх збирання та обробки. Використовують також спеціальні прийоми зведення даних до порівнянного вигляду — «статистичні ключі» зімкнення динамічних рядів.

Подолати переривчастість ряду можна двома способами. Перший — спосіб відносних рівнів, коли за базу порівняння для кожного ряду беруть кінцевий рівень. Два ряди відносних рівнів об'єднуються в один.

12.2. Характеристики інтенсивності динаміки

Швидкість та інтенсивність розвитку різних екологічних явищ значно варіюють, що позначається на структурі відповідних динамічних рядів. Для оцінювання зазначених властивостей динаміки статистика використовує низку взаємозв'язаних характеристик. Серед них: абсолютний приріст, відносний приріст, темп зростання, інші.

Розрахунок характеристик динаміки ґрунтується на порівнянні рівнів ряду. При порівнянні певної множини послідовних рівнів база порівняння може бути постійною чи змінною. За постійну базу вибирається або початковий рівень ряду, або рівень, який уважається вихідним для розвитку явища, що вивчається. Характеристики динаміки, обчислені відносно постійної бази, називаються ***базисними.*** Якщо кожний рівень ряду y_t, порівнюється з попереднім y_{t-1} , характеристики динаміки називаються ***ланцюговими.*** Схематично варіанти порівняння ілюструє рис. 12.1.

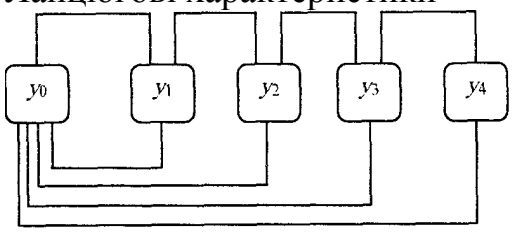


Рис. 12.1. Схеми порівняння при обчисленні ланцюгових і базисних характеристик динаміки

Абсолютний приріст Δ_t характеризує абсолютний розмір збільшення (чи зменшення) рівня ряду y_t, за певний часовий інтервал і обчислюється як різниця рівнів ряду:

базисний приріст $\Delta_t = y_t - y_0$;

ланцюговий приріст $\Delta_t = y_t - y_{t-1}$.

Знак «+», «-» свідчить про напрям динаміки.

Темп зростання k_t показує, у скільки разів рівень y, більший (менший) від рівня, узятого за базу порівняння. Він являє собою кратне відношення рівнів:

базисний темп $k_t = \frac{y_t}{y_0}$;

ланцюговий темп $k_t = \frac{y_t}{y_{t-1}}$.

При збільшенні рівня k_t>1, при зменшенні — k_t < 1. Темпи зростання виражаються як у коефіцієнтах, так і в процентах.

Ланцюгові Δ_t і k_t відображують відповідно абсолютну і відносну швидкість динаміки. Вони взаємозв'язані. Якщо подати $y_t = y_{t-1} + \Delta_t$, то

$$k_t = \frac{y_{t-1} + \Delta_t}{y_{t-1}} = 1 + \frac{\Delta_t}{y_{t-1}}.$$

Отже, при стабільній абсолютній швидкості темпи зростання зменшуватимуться. Стабільні темпи зростання можливі за умови прискорення абсолютної швидкості.

Величину $\frac{\Delta_t}{y_{t-1}}$ називають відносним прискоренням або ***темпом приросту*** і позначають символом T_t. Вона функціонально пов'язана з темпом зростання, але на відміну від останнього завжди виражається в процентах:

$$T_t = 100(k_t - 1).$$

Отже, темп приросту показує, на скільки процентів рівень у більший (менший) від бази порівняння.

Співвідношенням абсолютного приросту і темпу приросту визначається ***абсолютне значення 1% приросту.*** Нескладні алгебраїчні перетворення цього відношення показують, що воно становить соту частину рівня, узятого за базу порівняння:

$$A_t = \frac{\Delta_t}{T_t} = \frac{y_t - y_{t-1}}{100\left(\frac{y_t - y_{t-1}}{y_{t-1}}\right)} = \frac{y_{t-1}}{100}.$$

Ланцюгові й базисні характеристики динаміки взаємопов'язані:

а) сума ланцюгових абсолютних приростів дорівнює кінцевому базисному:

$$\sum_1^n \Delta_t = \sum_1^n (y_t - y_{t-1}) = y_n - y_0$$

б) добуток ланцюгових темпів зростання дорівнює кінцевому базисному:

$$k_1 k_2 \ldots k_n = \prod_1^n k_t = K_n = \frac{y_n}{y_0}$$

Щодо темпів приросту, то вони не мають таких властивостей, як абсолютні прирости чи темпи зростання. Ланцюгові й базисні темпи приросту співвідносяться через темпи зростання.

Якщо швидкість розвитку в межах періоду, що вивчається, неоднакова, порівнянням однойменних характеристик швидкості вимірюється прискорення чи уповільнення динаміки. На базі абсолютних приростів оцінюються ***абсолютне та відносне прискорення.*** Абсолютне — це різниця між абсолютними приростами: $\delta_t = \Delta_t - \Delta_{t-1}$. Прискорення характеризується додатною величиною $\delta_t > 0$, уповільнення — від'ємною $\delta_t < 0$.

Порівняння темпів зростання дає ***коефіцієнт прискорення (уповільнення)*** відносної швидкості розвитку. Для наочності та зручності їх тлумачення дільником є більший за значенням темп зростання.

У статистичному аналізі порівнюється також інтенсивність динаміки в різних рядах. Відношення темпів зростання *k':k"* називають ***коефіцієнтом випередження.*** За допомогою останнього порівнюють відносну швидкість динамічних рядів однакового змісту по різних об'єктах (регіони, країни тощо) або різного змісту по одному об'єкту.

Щодо темпів приросту, то співвідношення їх використовують лише для взаємозв'язаних показників *x* і *y*. Таке співвідношення називають ***емпіричнім коефіцієнтом еластичності*** $\gamma = T_y : T_x$; він показує, на скільки процентів змінюється *y* зі зміною *x* на 1%.

12.3. Середня абсолютна та відносна швидкість розвитку

З плином часу змінюються, варіюють рівні динамічних рядів і обчислені на їх основі абсолютні прирости та темпи зростання. Постає потреба узагальнення притаманних динамічному ряду властивостей, визначення типових характеристик розвитку. Такими характеристиками є середні величини. Зауважимо, що динамічна середня буде типовою характеристикою лише за умови однорідності ряду, коли причинний комплекс формування закономірностей розвитку більш-менш стабільний.

Середні рівні використовують насамперед для узагальнення коливних рядів. Наприклад, при аналізі динаміки сільськогосподарського виробництва оперують не річними, а більш сталими середньорічними показниками за певні періоди. Середні рівні необхідні також для забезпечення порівнянності чисельника і знаменника при побудові динамічних рядів похідних показників. Наприклад, виробництво продукції на одного працюючого. Обсяг продукції — інтервальний показник, а кількість працюючих — моментний. Щоб забезпечити

порівнянність цих показників, слід обчислити середньорічну кількість працюючих.

Метод обчислення середнього рівня динамічного ряду залежить від статистичної структури показника. В інтервальному ряді абсолютних величин, рівні якого динамічно адитивні, використовується середня арифметична проста:

$$\bar{y} = \frac{1}{n}\sum_{1}^{n} y_t \ .$$

де n — число рівнів ряду.

У моментному ряді, за припущення про рівномірну зміну показника між датами, середня розраховується як півсума значень на початок і кінець періоду:

$$\bar{y} = \frac{y_0 + y_n}{2} .$$

Якщо в моментному ряді $n > 2$ і між суміжними датами однакові інтервали, розрахунок виконується за формулою ***середньої хронологічної:***

$$\bar{y} = \frac{\frac{y_1 + y_n}{2} + \sum_{2}^{n-1} y_t}{n-1} .$$

Обґрунтування та розрахунок такої середньої наведено в підрозд. 4.4.

У моментних рядах з різними інтервалами між датами розраховується середня арифметична зважена:

$$\bar{y} = \frac{1}{\sum D_t}\sum_{1}^{m} y_t D_t \ ,$$

де D_t, — інтервал часу між датами, m — кількість інтервалів.

Середній абсолютний приріст (абсолютна швидкість динаміки) обчислюється діленням загального приросту за весь період на довжину цього періоду у відповідних одиницях часу (рік, квартал, місяць тощо):

$$\bar{\Delta} = \frac{y_n - y_0}{n} = \frac{\sum_{t=1}^{n} \Delta_t}{n} .$$

При обчисленні середнього темпу зростання враховується правило складних процентів, за якими змінюється відносна швидкість динаміки (нагромаджується приріст на приріст). Тому ***середній темп зростання*** обчислюється за формулою середньої геометричної з ланцюгових темпів зростання:

$$\bar{k} = \sqrt[n]{k_1 k_2 \ldots k_n} = \sqrt[n]{\prod_{t=1}^{n} k_t} \ ,$$

де n — кількість темпів зростання за однакові інтервали часу.

Урахувавши взаємозв'язок ланцюгових і базисних темпів зростання, формулу середньої геометричної можна записати так:

$$\bar{k} = \sqrt[n]{K_n} = \sqrt[n]{\frac{y_n}{y_0}}.$$

Розрахунок можна виконувати за допомогою логарифмів:

$$\lg \bar{k} = \frac{1}{n}\sum \lg k_t \text{ або } \lg \bar{k}_t = \frac{1}{n}(\lg y_n - \lg y_0).$$

Отже, середній темп зростання можна обчислити на основі:

• ланцюгових темпів зростання k_t;

• кінцевого (за весь період) темпу зростання K_n;

• кінцевого y_n і базисного y_0 рівнів ряду.

При інтерпретації середньої абсолютної чи відносної швидкості динаміки необхідно вказувати часовий інтервал, до якого належать середні, та часову одиницю вимірювання (рік, квартал, місяць, доба тощо).

12.4. Характеристика основної тенденції розвитку

Будь-який динамічний ряд у межах періоду з більш-менш стабільними умовами розвитку виявляє певну закономірність зміни рівнів —***загальну тенденцію.*** Одним рядам притаманна тенденція до зростання, іншим — до зниження рівнів. Зростання чи зниження рівнів динамічного ряду, у свою чергу, відбувається по-різному: рівномірно, прискорено чи уповільнено. Нерідко ряди динаміки через коливання рівнів не виявляють чітко вираженої тенденції.

Щоб виявити й схарактеризувати основну тенденцію, застосовують різні способи згладжування та аналітичного вирівнювання динамічних рядів.

Суть ***згладжування*** полягає в укрупненні інтервалів часу та заміні первинного ряду рядом середніх по інтервалах. У середніх взаємозрівноважуються коливання рівнів первинного ряду, внаслідок чого тенденція розвитку вирізняється чіткіше.

Залежно від схеми формування інтервалів розрізняють ступінчасті та ковзні (плинні) середні. Ряди цих середніх схематично зображено на рис. 12.2 для інтервалу $m = 3$. Очевидно, що ковзна середня більш гнучка і може краще відбити особливості тенденції.

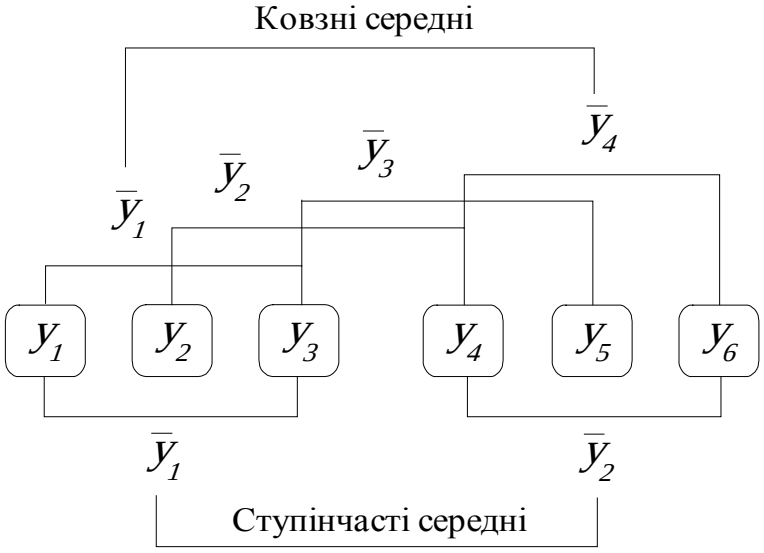


Рис. 12.2. Схеми утворення інтервалів згладжування динамічних рядів

При розрахунку ***ковзних середніх*** кожний наступний інтервал утворюється на основі попереднього заміною одного рівня. Оскільки середня $\bar{y}_j$ належить до середини інтервалу, то доцільно формувати інтервали з непарного числа рівнів первинного ряду. У разі парного числа рівнів необхідна додаткова процедура центрування (усереднення кожної пари значень $\bar{y}_j$).

Ряд ковзних середніх коротший за первинний на (*m* - 1) рівнів, що потребує уважного ставлення до вибору ширини інтервалу *m*. Якщо первинному ряду динаміки притаманна певна періодичність коливань, то інтервал згладжування має бути рівним або кратним періоду коливань.

Метод ковзних середніх застосовують також для попередньої обробки дуже коливних динамічних рядів; можливе подвійне згладжування.

При *аналітичному вирівнюванні динамічного ряду* фактичні значення y_t замінюються обчисленими на основі певної функції *Y= f(t),* яку називають ***трендовим рівнянням*** (*t* — змінна часу, *Y*— теоретичний рівень ряду).

Вибір типу функції ґрунтується на теоретичному аналізі суті явища, яке вивчається, і характері його динаміки. Зазвичай перевага надається функціям, параметри яких мають чіткий біологічний зміст і вимірюють абсолютну чи відносну швидкість розвитку. Суттєвою підмогою при виборі функцій є аналіз ланцюгових характеристик інтенсивності динаміки. Якщо ланцюгові абсолютні прирости відносно стабільні, не мають чіткої тенденції до зростання чи зменшення, вирівнювання ряду виконується на основі лінійної функції: $Y_t = a + bt$. Якщо ж відносно стабільними є ланцюгові темпи приросту, то найбільш адекватною такому характеру динаміки є експонента $Y_t = ab^t$. У

зазначених функціях t — порядковий номер періоду (дати), a — рівень ряду при $t = 0$. Параметр b характеризує швидкість динаміки: середню абсолютну в лінійній функції і середню відносну — в експоненті. Коли характеристики швидкості розвитку зростають (чи зменшуються), використовуються інші функції (парабола 2-го степеня, модифікована експонента тощо).

Параметри трендових рівнянь визначають методом найменших квадратів. Згідно з умовою мінімізації суми квадратів відхилень фактичних рівнів ряду y_t від теоретичних Y_t параметри визначаються розв'язуванням системи нормальних рівнянь. Для лінійної функції вона записується так:

$$na + b\sum t = \sum y,$$

$$a\sum t + b\sum t^2 = \sum yt.$$

Система рівнянь спрощується, якщо початок відліку часу (t = 0) перенести в середину динамічного ряду. Тоді значення ї, розміщені вище середини, будуть від'ємними, а нижче — додатними. При непарному числі членів ряду (наприклад, n = 5) змінній t надаються значення з інтервалом одиниця: -2, -1, 0, 1, 2; при парному: -2,5, -1,5, -0,5, 0,5, 1,5, 2,5. В обох випадках $\sum t = 0$, а система рівнянь набирає вигляду

$$na = \sum y,$$

$$b\sum t^2 = \sum yt$$

Отже, $a = \frac{\sum y}{n}$, $b = \frac{\sum yt}{\sum t^2}$. Значення $\sum t^2$ можна визначити за формулами:

для непарного числа членів ряду

$$\sum t^2 = \frac{n(n^2 - 1)}{12};$$

для парного числа членів ряду

$$\sum t^2 = \frac{n(n^2 - 1)}{3}.$$

Продовження виявленої тенденції за межі ряду динаміки називають ***екстраполяцією тренду.*** Це один із методів статистичного прогнозування, передумовою використання якого є незмінність причинного комплексу, що формує тенденцію. Прогнозний, очікуваний рівень $Y_{t+\nu}$ залежить від бази прогнозування та періоду упередження ν.

Метод екстраполяції дає точковий прогноз. На практиці, як правило, визначають довірчі межі прогнозного рівня $Y_{t+\nu} \pm ts_p$, де s_p — стандартна похибка прогнозу, t- квантиль розподілу Стьюдента.

12.5. Оцінка коливань та сталості динаміки

Фактичні рівні динамічних рядів під впливом різного роду чинників варіюють, відхиляючись від основної тенденції розвитку. В одних рядах коливання мають систематичний, закономірний характер, повторюються через певні інтервали часу, в інших — не мають такого характеру і тому називаються ***випадковими.*** У конкретному ряду можуть поєднуватися систематичні та випадкові коливання.

Найпростішою оцінкою систематичних коливань є ***коефіцієнти нерівномірності,*** які обчислюються відношенням максимального і мінімального рівнів динамічного ряду до середнього. Чим більша нерівномірність процесу, тим більша різниця між цими двома коефіцієнтами.

Окремим екологічним процесам притаманні внутрішньорічні, сезонні піднесення і спади. ***Сезонні коливання*** виявляються і аналізуються на основі рядів щомісячних або щоквартальних даних.

Характер сезонних коливань описується ***«сезонною хвилею»,*** *яку* утворюють індекси сезонності. У динамічних рядах, які не виявляють чіткої тенденції розвитку, ***індекси сезонності*** є відношенням фактичних місячних (квартальних) рівнів y_t, до середньомісячного (середньоквартального) за рік $\bar{y}$, %:

$$I_c = 100\frac{y_t}{\bar{y}}.$$

Оскільки сезонні коливання з року в рік не лишаються незмінними, виявити сталу сезонну хвилю можна за допомогою середніх індексів сезонності за кілька років:

$$\bar{I}_c = \frac{1}{n}\sum_1^n I_c$$

де n — число років.

Для порівняння інтенсивності сезонних коливань різних явищ чи одного й того самого явища в різні роки використовуються узагальнюючі характеристики варіації індексів сезонності:

середнє лінійне відхилення $\bar{l}_C = \frac{1}{12}\sum_1^{12}\left|I_c - 100\right|$;

або середнє квадратичне відхилення $\sigma_t = \sqrt{\frac{1}{12}\sum_1^{12}(I_c - 100)^2}$.

Якщо спостерігається тенденція розвитку, попередньо проводиться згладжування чи вирівнювання динамічного ряду, визначаються теоретичні

рівні для кожного місяця (кварталу) року, а індекс сезонності обчислюється як відношення фактичних рівнів ряду y_t до теоретичних Y_t, тобто $I_c = 100\frac{y_t}{Y_t}$.

Абсолютною мірою випадкових коливань є ***середнє квадратичне відхилення*** s_e, яке обчислюється на основі залишкової дисперсії:

$s_e = \sqrt{s_e^2}$.

Поряд з абсолютною мірою випадкових коливань використовують відносну — ***коефіцієнт варіації*** $V_e = 100\frac{s_e}{\bar{y}}$, де $\bar{y}$ — середній рівень динамічного ряду.

Різницю 100- V_e використовують для оцінки сталості динаміки.

13. ІНДЕКСИ

13.1. Суть і функції індексів

Термін *індекс (index)* є синонімом певної узагальнюючої характеристики. Кожний індекс є співвідношенням двох значень показника, який індексується: оціночного (поточного) і взятого за базу порівняння. Тобто за статистичною природою індекс — це відносна величина, яка характеризує зміну явища в часі чи просторі або ступінь відхилення значення показника від певного стандарту (нормативу, середньої). Форми вираження індексу: коефіцієнти, проценти, проміле.

Історично індекси створювались як інструмент вивчення динаміки споживчих цін. Ще на початку XVII ст. англійський купець Т. Ман доводив переваги торгівлі з Індією, порівнюючи вартість товарів, які ввозились в Англію з Індії і Туреччини (шовк-сирець, прянощі тощо), за цінами індійськими та турецькими. Індекс цін становив 0,33, тобто закупівля зазначених товарів у Індії втричі дешевша порівняно з Туреччиною. Різницю вартостей Т. Ман визначив як суму економії від заміни торгового партнера. Такого роду розрахунки й досі вважаються найбільш логічним вираженням індексів.

Поступово коло показників, що піддавалися індексному аналізу, розширювалось, а методи аналізу вдосконалювались.

Індекс, як і будь-який статистичний показник, поєднує в собі якісний та кількісний аспекти. Назва індексу відбиває біологічний зміст показника, числове його значення — інтенсивність змін або ступінь відхилення (індекс урожайності, індекс продуктивності, продиховий індекс, тощо).

Методика розрахунку (модель) індексу залежить від мети дослідження, статистичної природи показника, ступеня агрегованості інформації. Мета

дослідження, зокрема, визначає функцію, яку виконує індекс у конкретному аналізі, та характер порівнянь. Розрізняють дві функції індексів:

1) синтетичну, пов'язану з побудовою узагальнюючих характеристик динаміки чи просторових порівнянь;

2) аналітичну, спрямовану на вивчення закономірностей динаміки, взаємозв'язків між показниками, структурних зрушень.

Так, індексний факторний аналіз передбачає оцінку впливу факторів на динаміку показника, який індексується; індексні ряди є основою моніторингу динамічних явищ.

Очевидно, що синтетична та аналітична функції індексів взаємозв'язані. Часто один і той самий індекс виконує обидві функції.

За характером порівнянь (у часі, просторі, з певним стандартом) індекси поділяються на *динамічні, територіальні, міжгрупові.*

Динамічний індекс характеризує інтенсивність динаміки; при його розрахунку базою порівняння є одне з попередніх значень показника. База порівняння ідентифікується підрядковою позначкою «0», поточне значення показника — «1».

При ***просторових порівняннях*** визначається ступінь відхилення значень показника у просторі — між об'єктами, країнами, регіонами, які ідентифікуються певними літерами; вибір бази порівняння довільний. ***Міжгруповий індекс*** характеризує відхилення від певного стандарту (еталонного, максимального чи мінімального значення) або від середнього рівня по сукупності в цілому.

Визначальними ознаками інформаційної бази індексного аналізу є ступінь агрегованості та статистична природа показника. За ступенем агрегованості інформації індекси поділяються на ***індивідуальні та зведені.*** Вони позначаються відповідно символами *і* та *I*. ***Індивідуальні індекси*** характеризують співвідношення рівнів показника окремих елементів сукупності, ***зведені***— певної множини елементів. У структурованій сукупності зведений індекс може бути груповим (субіндексом) або загальним (тотальним). Наприклад, в ієрархії динамічних індексів динаміку обсягів фотосинтезу характеризують індивідуальні індекси, окремих видів (дуб, береза), Життєвих форм (дерева, кущі) — субіндекси, лісу в цілому — загальний індекс.

Показник, динаміку чи співвідношення якого характеризує індекс, називають ***індексованою величиною,*** йому надається певний символ. Символи не випадкові, вони здебільшого відповідають початковим літерам англійських назв.

Індивідуальний індекс — це відносна величина динаміки або порівняння.

Неодмінною умовою їх обчислення є порівнянність методики вимірювання чисельника та знаменника відношення, що являє собою індекс.
Щодо зведених індексів, то розрахунок кожного з них окремо передбачає вирішення низки методологічних питань.

13.2. Методологічні основи побудови зведених індексів

Зведений індекс показує, як у середньому змінився показник по сукупності елементів. Основою побудови зведених індексів є агрегування. узагальнення інформації. Нехай за даними про накопичення біомаси (j = 1, 2, 3, ..,) за два періоди (t = 0, 1) необхідно визначити зведений індекс приросту біомаси:
(p_{10}, p_{20}, p_{30}, p_{n0}) та (p_{11}, p_{21}, p_{31}, p_{n1})
Агрегування інформації для зведеного індексу приросту біомаси I_p, можна здійснити трьома способами.

1. У вигляді відношення сум приросту біомас (індекс Дюто, 1735 р.). Поділивши чисельник і знаменник на n, цей індекс можна подати як відношення середніх величин.	$I_p = \frac{\sum p_{j1}}{\sum p_{j0}} = \frac{\overline{p}_{j1}}{\overline{p}_{j0}}.$
2. Як середню арифметичну з індивідуальних індексів приросту біомаси $i_p = \frac{p_{j1}}{p_{j0}}$ (індекс Карлі, 1751 р.).	$I_p = \frac{\sum \frac{p_{j1}}{p_{j0}}}{n} = \frac{\sum i_p}{n}.$
3. Як середню геометричну з індивідуальних індексів (Джевон, 1863 р.).	$I_p = \sqrt[n]{\frac{p_{11}}{p_{10}}\frac{p_{21}}{p_{20}}\ldots\frac{p_{n1}}{p_{n0}}}.$

Кожний з цих варіантів передбачає рівновагомість приросту біомаси у сукупності. Інакше зведений індекс неадекватно характеризуватиме сукупну зміну біомаси.
При розрахунку зведеного індексу іноді буває необхідно кожному елементу надати певну «вагу», яка б ураховувала його відносну значущість. Очевидно, що ваги повинні мати однакову розмірність, тоді зважений індекс набирає вигляду:

$$I_p = \frac{\sum \frac{p_{j1}}{p_{j0}} q_{j0} p_{j0}}{\sum q_{j0} p_{j0}} = \frac{\sum p_{j1} q_{j0}}{\sum p_{j0} q_{j0}}.$$

Це дві форми одного індексу — середньозважена та агрегатна. Основою середньозваженої форми є індекс Карлі, агрегатної — індекс Дюто.

13.3. Агрегатна форма індексів

Агрегатний індекс — це співвідношення двох агрегатів, конкретних щодо змісту й часу. ***Агрегат*** є добутком спряжених величин. Одна з цих величин індексована — у чисельнику і знаменнику вона в різних періодах, інша є вагою чи сумірником індексованої величини і фіксується на одному й тому самому рівні.

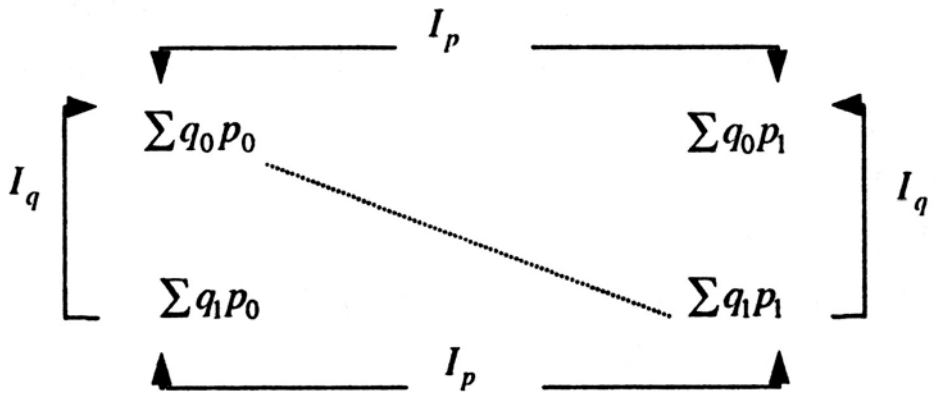


Рис. 13.1. Схема співвідношення агрегатів

На головній діагоналі матриці розміщено фактичні значення, на побічній — перехресні (умовні). По горизонталі розміщені агрегати з фіксованими вагами: у першому — на рівні базисного періоду, у другому — на рівні поточного. По вертикалі — агрегати з фіксованими сумірниками: у першому — на рівні базисного періоду, у другому — на рівні поточного. Порівняння агрегатів дає дві системи індексів — базисно-зважену (Ласпереса) та поточно-зважену (Пааше).

У базисно-зваженій системі перехресні агрегати побічної діагоналі порівнюються з базисним агрегатом головної діагоналі $\sum q_0 p_0$, у поточно-зваженій системі агрегат головної діагоналі $\sum q_1 p_1$ порівнюється з перехресними агрегатами побічної діагоналі. Схематично системи зважування показано на рис. 13.1, а формули індексів наведені в табл. 13.1.

Таблиця 13.1

Формули індексів значень і фізичного обсягу за різних систем зважування

Базисно-зважена система (Ласпереса)	Поточно-зважена система (Пааше)
$I_p = \frac{\sum p_1 q_0}{\sum p_0 q_0}$	$I_p = \frac{\sum p_1 q_1}{\sum p_0 q_1}$
$I_q = \frac{\sum q_1 p_0}{\sum q_0 p_0}$	$I_q = \frac{\sum q_1 p_1}{\sum q_0 p_1}$

Обидві системи індексів рівноправні. Реальний зміст мають не лише чисельник і знаменник індексу, а й різниця між ними. Вибір форми індексу залежить від мети дослідження та наявної інформації. Так, у зарубіжній статистиці індекс цін розраховується за Ласпересом, оскільки ґрунтується на даних про обсяги, отримані з переписів, вибіркових обстежень домогосподарств, інших джерел за минулий період. У вітчизняній статистиці при розрахунках індексів цін перевага надавалася формулі Пааше, оскільки визначальним показником була вартість поточного періоду. Індекс фізичного обсягу товарної маси, навпаки, обчислюється за формулою Ласпереса з фіксованими сумірниками на рівні базисного періоду. У такому разі динаміка цінового фактора не впливає на величину індексу. Зауважимо, що при незначній кореляції між цінами та товарною масою індекси, розраховані за Ласпересом і Пааше, практично однакові.

За наявності структурних зрушень використовують індекси із середніми вагами або усереднення різнозважених індексів за допомогою середньої геометричної:

$$I_p = \sqrt{\frac{\sum p_1 q_1}{\sum p_0 q_1} \frac{\sum p_1 q_0}{\sum p_0 q_0}}$$

Спираючись на формально-математичні критерії, яким відповідає усереднений індекс, І. Фішер назвав його «ідеальним», проте через відсутність конкретного біологічного змісту цей індекс не набув широкого практичного застосування.

13.4. Середньозважені індекси

Другою формою зведеного індексу є середньозважений з індивідуальних індексів. Використовують два види середніх — арифметичну та гармонічну.

Вибір виду середньої ґрунтується на загальних засадах: середньозважений індекс має бути тотожним відповідному індексу агрегатної форми.

У такому разі зведені індекси за Ласпересом обчислюються як середня арифметична з вагами $q_0 p_0$ а індекси за Пааше — як середня гармонічна з вагами $q_1 p_1$:

Зведені індекси за Ласпересом	Зведені індекси за Пааше
$I_p = \frac{\sum i_p p_0 q_0}{\sum p_0 q_0}$ $I_p = \frac{\sum p_1 q_1}{\sum \frac{1}{i_p} q_1 p_1}$	$I_q = \frac{\sum i_q q_0 p_0}{\sum q_0 p_0}$ $I_q = \frac{\sum q_1 p_1}{\sum \frac{1}{i_q} q_1 p_1}$

При побудові середньозважених індексів вартісні ваги можна замінити відносними величинами структури *d*, сума яких $\sum d = 1.$ У цьому разі середньозважені індекси набирають вигляду:

$$I_q = \sum i_p d_0; \quad I_p = \frac{1}{\sum \frac{1}{i_p} d_1}.$$

Ці формули підтверджують залежність значення зведеного індексу від динаміки окремих складових і пропорцій у сукупності агрегованих елементів.

Середньозважені індекси мають перевагу перед агрегатними, адже за їхньою допомогою можна вишикувати ієрархію індексів від індивідуальних на через групові (субіндекси) до загального по всій сукупності елементів. Проте їм властиві й недоліки. Якщо динаміка окремих складових сукупності протилежна, то зведений індекс не в змозі адекватно відобразити закономірність динаміки. Окрім того, середньозважений індекс визначається лише стосовно порівнянного кола елементів. Якщо ж окремі елементи сукупності відсутні в базисному чи поточному періоді, то розрахунок індивідуальних індексів неможливий. У цьому разі перевага надається індексу агрегатної форми.

Отже, за кожним індексом стоїть певне біологічне явище, що зумовлює методику його розрахунку та змістовність.

13.5. Індекси середніх величин

Поряд зі зведеними, агрегатними індексами в статистичній практиці широко використовують індекси середніх величин (індекси середньої заробітної плати, середньої врожайності тощо). Як відомо, рівень середньої залежить від значень ознаки x_j і структури сукупності:

$$x = \frac{\sum x_j f_j}{\sum f_j} = \sum x_j d_j ,$$

де f_j — частота; d_j— частка j-ї складової сукупності.

Очевидно, що й динаміка середньої визначається цими факторами: а) зміною значень ознаки x_j і б) структурними зрушеннями. Вплив кожного з них на динаміку середньої оцінюється за допомогою системи індексів середніх величин: змінного й фіксованого складу, а також структурних зрушень. У наведених формулах індексів ідентифікація складових сукупності відсутня.

Індексом змінного складу $I_{\bar{x}}$ називають індекс ***середньої величини,*** він відбиває не лише зміни значень ознаки x, а й зміни в структурі сукупності:

$$I_{\bar{x}} = \frac{\bar{x}_1}{\bar{x}_0} = \frac{\sum x_1 f_1}{\sum f_1} : \frac{\sum x_0 f_0}{\sum f_0} = \sum x_1 d_1 : \sum x_0 d_0 .$$

В *індексі фіксованого складу* I_x ваги постійні, тобто усувається вплив на динаміку середньої структурних зрушень. Величина I_x показує, як ***у середньому*** змінилися значення ознаки при незмінній, фіксованій структурі:

$$I_x = \frac{\sum x_1 f_1}{\sum f_1} : \frac{\sum x_0 f_1}{\sum f_1} = \sum x_1 d_1 : \sum x_0 d_1 .$$

Індекс структурних зрушень I_d, навпаки, показує, як змінилася середня за рахунок структурних зрушень; значення ознаки x фіксуються на постійному рівні:

$$I_d = \frac{\sum x_0 f_1}{\sum f_1} : \frac{\sum x_0 f_0}{\sum f_1} = \sum x_0 d_1 : \sum x_0 d_0 .$$

У кожній конкретній індексній системі I_d оцінює вплив на динаміку середньої того структурного фактора, який є основою поділу сукупності на складові.

Формули індексів фіксованого складу і структурних зрушень різнозважені: в I_x ваги фіксуються на рівні поточного періоду, в I_d — значення ознаки x — на рівні

базисного періоду. Саме такий варіант зважування забезпечує пов'язування цих індексів у систему:

$$I_{\bar{x}} = I_x I_d .$$

Методологічною особливістю побудови системи індексів середніх величин є порівнянність складових сукупності в часі. Проте більшість реальних сукупностей за своїм складом динамічні: одні частини сукупності зникають, інші (нові) — з'являються. Так, оновлюється видовий склад асоціації, з'являються нові види, зникають старі тощо.

Щоб оцінити вплив на динаміку середньої такого роду змін, в індексну систему вводять три індекси структурних зрушень: I_d^0 — для оцінювання впливу змін у структурі порівнянного кола складових сукупності; I_d^n —для оцінювання впливу новоутворених складових, I_d^b — для оцінювання впливу вибулих складових. Індексна система має вигляд:

$$I_{\bar{x}} = I_x^0 I_d^0 I_d^n I_d^b .$$

Індекс фіксованого складу I_x^0 обчислюється для порівнянного кола складових. Вагами всіх індексів є відносні величини структури—частки *d*. Отже,

$$I_{\bar{x}} = \frac{\sum x_1 d_1}{\sum x_0 d_0}; \qquad I_x^0 = \frac{\sum x_1 d_1^0}{\sum x_0 d_1^0};$$

$$I_d^0 = \frac{\sum x_0 d_1^0}{\sum x_0 d_0^0}; \qquad I_d^n = \frac{\sum x_1 d_1}{\sum x_1 d_1^0}; \qquad I_d^b = \frac{\sum x_0 d_0^0}{\sum x_0 d_0}.$$

13.6. Територіальні індекси

Територіальний індекс використовують як інструмент порівняння біологічних показників у просторі: за окремими країнами, територіями, регіонами, об'єктами. Особливістю цих індексів є рівноправність порівнюваних об'єктів *A* і *B*. Жоден з них не може претендувати на роль бази порівняння, а отже рівноправними слід вважати індекси як з базою порівняння *A,* так і з базою порівняння *B:*

$$I_{\frac{A}{B}} = \frac{\sum x_A f}{\sum x_B f}; \qquad I_{\frac{B}{A}} = \frac{\sum x_B f}{\sum x_A f}.$$

Де x — індексована величина, f— вага (сумірник) індексованої величини.
При фіксованих значеннях ваг (сумірників) індекси I_A і I_B обернено пропорційні.

14. СТАТИСТИКА ПРИРОДНИХ РЕСУРСІВ

14.1. Поняття про статистику навколишнього середовища

Як складовий елемент національного багатства природні ресурси потребують докладного статистичного аналізу. Розвиток науки і техніки зумовлює зростаючий вплив людини на природу.
Ураховуючи те, що такий вплив має переважно негативний і невиправданий характер, це спонукає дбайливіше ставитися до навколишнього середовища. Тому, поряд із завданнями обліку, розрахунку структури та вивчення використання в народногосподарських інтересах природних ресурсів метою статистичного дослідження постає також вивчення характеру техногенного впливу людини на довкілля, облік кількісних та якісних змін у ньому. Масштаби антропогенного (тобто зумовленого господарською діяльністю людини) впливу на природу величезні. Щороку під час переорювання полів, провадження гірничовидобувних робіт та різного будівництва переміщуються, у середньому, 4 тис. км^3 ґрунтів, з надр землі видобувається 100 млрд т руди, паливних копалин та будівельних матеріалів, на поля вноситься близько 300 млн т отрутохімікатів, на господарські потреби забирається 12% річкового стоку.
Комплекс питань, пов'язаних зі станом природи, постає на різних рівнях і часто має міжнародний, планетарний характер. Цю проблему нині вивчають багато важливих авторитетних міжнародних організацій — ООН та її підрозділи (ЮНЕСКО, ФАО), Міжнародна рада наукових союзів (МРНС), Всесвітня організація охорони здоров'я та інші.
Взаємостосунки людини і природи мають будуватися на засадах раціонального природокористування. Масштаби освоєння людською діяльністю природних ресурсів не повинні перевищувати швидкості їх оновлення, а засвоєння природних комплексів має враховувати стійкість їх внутрішньої структури. Відходи, що виникають у виробництві, мають нейтралізуватися або ліквідуватися так, аби не забруднювати середовище, а також необхідно використовувати корисні компоненти, що є у відходах. Отже, раціоналізація природокористування передбачає розгляд природних процесів і виробничої діяльності як єдиної біоекономічної системи.
Істотне погіршення якості навколишнього середовища, яке значною мірою зумовлене зростанням масштабів антропогенного впливу, зумовило потребу формування **статистики навколишнього середовища,** дані якої мають забезпечити об'єктивне уявлення про обсяги природних ресурсів, якість навколишнього середовища, інтенсивність та напрямки техногенного навантаження на природу.

Статистика навколишнього середовища є міжгалузевою за своїм характером, її інформація використовується для розробки та оцінки виконання соціально-економічних програм, вирішення питань економічної політики. Вона описує стан та тенденції зміни навколишнього середовища, до якого належать повітря, вода, ґрунт/земля, біота, що містяться в цих компонентах, а також населені пункти.

У 1972 році на засіданні Генеральної Асамблеї ООН уперше були визначені головні напрямки та принципи діяльності різних держав і закладені основи для розвитку міжнародного співробітництва в галузі статистики навколишнього середовища. Крім того, була заснована спеціальна програма ООН з навколишнього середовища (ЮНЕП), одне з довгострокових завдань якої полягало в розробці загальної системи показників статистики навколишнього середовища, яка узагальнила б досвід різних країн і забезпечила б порівняння національних даних.

14.2. Побудова системи показників статистики навколишнього середовища

Побудова системи показників статистики навколишнього середовища ґрунтується на використанні моделі «навантаження— стан—реакція», згідно з якою етап навколишнього середовища та окремих його компонентів визначається впливом соціально-економічної діяльності людини і реакцією суспільства, яка спрямована на запобігання та зменшення наслідків цієї діяльності. Зазначена система показників складається з таких підсистем:

1) показники навантаження;

2) показники стану навколишнього середовища:

3) показники реакції суспільства;

4) показники запасів та фондів окремих компонентів навколишнього середовища.

Показники навантаження поділяються на показники антропогенного та природного навантаження. Щоб оцінити антропогенне навантаження на довкілля, застосовують показники:

• видобутку (збору врожаю) окремих природних ресурсів;

• що характеризують кількість викидів і скидів забруднюючих речовин та відходів у атмосферне повітря, водні ресурси та в землю;

• що характеризують кількість використовуваних біохімічних речовин (природних та хімічних добрив, пестицидів) за розміром площі та інтенсивністю їх використання.

Природне навантаження на компоненти навколишнього середовища вивчається за допомогою показників, що відбивають масштаби та інтенсивність природних

явищ і стихійного лиха, такого як засуха, повінь, землетрус, які негативно впливають на навколишнє середовище.

Показники стану навколишнього середовища використовують для характеристики наслідків антропогенного та природного впливу на довкілля:

• показники, що відбивають кількісні зміни в природних ресурсах, а саме біологічних, відновлюваних та невідновлюваних. До них належать показники зміни площі посівів сільськогосподарських культур, популяції домашніх тварин, окремих популяцій рибних запасів, видів флори та фауни тощо. Для розрахунку таких показників застосовується підхід, що базується на оцінці зміни запасів;

• показники якості навколишнього середовища визначають фактичні якісні властивості повітря, води та землі, які виражаються показниками концентрації окремих видів забруднюючих речовин. Для такої оцінки використовується підхід, що полягає в порівнянні рівнів фактичної концентрації забруднюючих речовин зі значеннями їх нормативних показників — граничне допустимих концентрацій (ГДК), перевищення яких призводить до різних порушень стану здоров'я людей, негативно впливає на умови існування тварин і рослин;

• показники стану здоров'я населення, такі, наприклад, як захворюваність. Особлива увага приділяється вивченню стану здоров'я дітей та немовлят, які найбільше потерпають через погіршення стану навколишнього середовища.

Показники реакції суспільства на наслідки впливу соціально-економічної діяльності та природних явищ на навколишнє середовище характеризують реакцію, що має на меті змінити спрямованість несприятливих тенденцій завдяки досягненню рівноваги у співвідношенні діяльності суспільства, підтримання здоров'я екологічних систем та сталості у використанні природних ресурсів. Ці показники поділяються на такі групи:

• показники, що безпосередньо описують заходи з охорони навколишнього середовища;

• показники витрат на реалізацію заходів з охорони навколишнього середовища.

До заходів, спрямованих на охорону навколишнього середовища, належать:

1) заходи з відновлення деградованого довкілля та окремих його компонентів;

2) проведення досліджень із проблем забруднення навколишнього середовища та спостереження за забрудненням;

3) заходи із захисту та збереження природи;

4) заходи, пов'язані зі створенням державних структур для контролю за забрудненням.

Показники запасів і фондів окремих компонентів навколишнього середовища та кадастру екосистем дають оцінку запасів природних ресурсів та фондів населених пунктів і їх зміни внаслідок соціально-економічної діяльності

суспільства. Ці показники будуть розглянуті під час вивчення окремих видів природних ресурсів.

Статистика навколишнього середовища не є завершеною. З появою різноманітних проблем, що пов'язані зі станом, розвитком окремих компонентів навколишнього середовища, а також постійним зростанням розмірів соціально-економічної діяльності суспільства, статистика навколишнього середовища постійно вдосконалюється.

14.3. Показники окремих видів природних ресурсів

Оскільки окремі види природних ресурсів значно різняться між собою як за натуральною формою, так і рівнем та напрямками залучення до народногосподарського обороту, статистичні показники, необхідні для їх оцінки та характеристики, є вузькоспеціалізованими. Єдиної загальноприйнятої класифікації природних ресурсів у статистиці не існує. Нині вони вивчаються в розрізі укрупнених груп:

1) земельні фонди;
2) багатства надр;
3) водні ресурси;
4) лісові ресурси;
5) гідроенергоресурси.

З огляду на це при вивченні системи показників статистики природних ресурсів розглядаються окремі підрозділи цієї системи.

Показники статистики земельного фонду.

Існування й розвиток людського суспільства незалежно від його соціально-економічного устрою нерозривно пов'язані з землею — цим найважливішим компонентом зовнішнього середовища, роль якого в житті людей багатогранна. Зокрема, земля є головним засобом виробництва в сільському господарстві, а також виконує функцію територіального базису і в інших галузях народного господарства. Більш того, земля є загальною основою існування людини, оскільки земля — це не тільки матерія, тобто земні території, а й вода, надра, рослинність і т. ін. Уся земельна площа країни або певного регіону, включаючи внутрішні води, становить *земельний фонд.* Сукупність даних про правовий, природний та господарський стан землі має назву *земельного кадастру.* Структура кадастру передбачає той необхідний набір статистичних показників, кількісна інформація про які наповнює земельний кадастр конкретним змістом. Нині сфера кадастрового обліку обмежена землею сільськогосподарського призначення. Тому статистика земельних ресурсів є одним з розділів сільськогосподарської статистики. Докладну інформацію про земельні ресурси

містить Державна земельна книга району (або міста). У ній зазначені землекористувачі (підприємства, організації або окремі особи), структура земель за якістю, за аграрно-виробничими групами з урахуванням фізико-хімічних особливостей, що дозволяє обчислити питому вагу найважливішої частини земельних ресурсів — оранки — з урахуванням того, яка частина з них у цей час використовується в сільському господарстві, а яка — ні.

Крім характеристики складу і структури земельного фонду показники земельних ресурсів також виражають динаміку включення в господарський оборот нових земель, трансформацію освоєних земель та їх рекультивацію. Особливо велике значення для контролю за використанням земельних ресурсів мають показники їх трансформації, причому найбільший інтерес становлять показники відводу продуктивних земель з різною несільськогосподарською метою.

Зміна якості земель також може бути оцінена в динаміці. При цьому вирізняють зміни, які відбуваються під впливом натуральних процесів та результатів господарської діяльності людини. Це пов'язано з тим, що виробнича сила землі як засобу виробництва, її цінність з промислового, рекреаційного та інших поглядів є категоріями, не сталими в часі. Якість земельних ресурсів може по-ліпшитися завдяки цілеспрямованому впливу на них людської діяльності, але може й погіршитись через значне забруднення землі. Таке погіршення якості земельних ресурсів має кількісно відбиватися у зниженні ***показника економічної оцінки,*** що буде мірою втрат, які нанесені народному господарству забрудненням цього найважливішого компонента зовнішнього середовища.

Цінність такого показника полягає в тому, що, по-перше, він дозволяє цілеспрямовано планувати витрати на землеохоронні заходи і спрямовувати їх передовсім на охорону тих земель, втрати від забруднення яких особливо великі; по-друге, дає можливість раціональніше розв'язувати питання розміщення промислових підприємств, уникаючи їх концентрації в районах з високим показником втрат від забруднення землі.

Показники статистики лісових ресурсів.

Значення лісів велике і багатогранне. Як найважливіший планетарний акумулятор живої речовини ліси підтримують вуглеводневий та кисневий баланс землі, впливають на біологічний кругообіг ряду хімічних елементів. Ліси значно впливають на кліматичні умови різних географічних зон, на циркуляцію тепла в атмосфері, на запас вологи в ґрунті, води в ріках та озерах. Лісові насадження значною мірою не дають змоги розширюватись водній та вітровій ерозії.

Ліси мають велике санітарно-гігієнічне значення: крони дерев не тільки затримують тверді пильові частинки, а й детоксують різні газоподібні інгредієнти. Зі зростанням урбанізації підвищується культурно-естетичне значення лісів: вони все більшою мірою стають місцем відпочинку та туризму, а позитивний вплив лісу на здоров'я та зростання працездатності людини загальновідомий. У лісі багато ягід, лікарських рослин, горіхів, він є місцем проживання цінних промислових тварин та птахів.

Статистика лісового господарства, котра має інформацію про наявність, кількісний та якісний стан лісових ресурсів і їх охорону, розглядає ліси передусім як економічний ресурс. На підставі показників лісового господарства розробляють плани розвитку лісового господарства та лісової промисловості, організують лісогосподарське виробництво і лісоексплуатацію. Багато в чому виконання цих завдань визначається породним та віковим складом лісонасаджень. Саме тому облік лісів проводять залежно від їх народногосподарського призначення, місцезнаходження та виконуваних природоохоронних функцій. Розрізняють три групи лісів.

До першої групи належать ліси, що мають таке значення:

• водоохоронне (заборонені зони вздовж водних об'єктів, а також заборонені смуги лісів, що зберігають нерестилища цінних промислових риб):

• захисне (лісові смуги вздовж шляхів, протиерозійні ліси):

• санітарно-гігієнічне та оздоровче (зелені зони навколо міст, курортів та джерел водопостачання), а також заповідники, національні та природні парки, приальпійські ліси.

До другої групи належать ліси районів з розвиненою мережею транспорту, великою кількістю населення, тобто ті, що мають обмежене експлуатаційне значення.

Третя група об'єднує ліси багатолісових районів. Для лісів різних груп установлено різний режим користування. Так, у лісах першої групи вирубування насаджень обмежене, а ліси третьої групи здебільшого мають задовольняти потреби народного господарства щодо деревини без втрат захисних якостей лісів.

Найдокладніша інформація про кількісний та якісний стан лісових масивів міститься в матеріалах періодичного обліку лісового фонду. Під *лісовим фондом* розуміють землі, що вкриті або не вкриті лісами, але призначені для потреб лісового господарства. тобто площі, на яких мають бути вирощені ліси: вигорілі ділянки, площі загиблих лісів, галявини, розріджені частини лісових насаджень тощо. Такий облік провадиться, як правило, один раз на 5 років. При цьому весь лісовий фонд поділяється за категоріями земель на лісову та нелісову площі; остання включає в себе невикористані в лісовому господарстві

площі усередині лісового фонду (шляхи, канави, болота, яруги і т. ін.) з виокремленням груп та. категорій лісів. Запаси лісонасаджень класифікують за віковими групами та переважними породами (хвойні, м'яколистяні, твердолистяні), класами та категоріями стиглості (молоді, приспілі, спілі та перестійні ліси).

На базі цих показників можуть бути розроблені важливі узагальнюючі характеристики території та лісу:

• лісистість території: відношення площі лісу до площі регіону;

• лісові насадження на душу населення, $м^3$;

• вкрита лісом площа зеленої зони на одну особу, га:

• загальний середній приріст лісу — усього, у тому числі за групами порід, тис. $м^3$

• середній приріст на 1 га вкритої лісом площі, $м^3$/га. Отже. статистика лісових ресурсів дозволяє характеризувати наявність, склад, стан, рух. відтворення і використання лісового фонду.

На відміну від інших видів природних ресурсів корисні копалини не відновлюються, а потреба щодо них безперервно збільшується. Це потребує дбайливого ставлення до багатств надр.

Показники статистики корисних копалин.

Головне завдання обліку корисних копалин — отримання повних та достовірних даних про стан мінерально-сировинної бази підприємства, галузі та країни на початок кожного року, рівень розвіданості та підготовленості родовищ до промислового засвоєння, забезпеченість гірничодобувних підприємств розвіданими запасами. Запаси корисних копалин обліковуються згідно з класифікацією, за якою залежно від народногосподарського значення вони поділяються на дві групи, що підлягають роздільному обліку — балансові та позабалансові.

Балансові — це запаси, використання яких економічно доцільне і які відповідають кондиціям, що встановлюються для підрахунку запасів у надрах. До **позабалансових** належать запаси, використання яких при досягнутому технічному рівні економічно недоцільне через їх малу кількість, малу потужність покладів, низький вміст цінних компонентів, але які згодом можуть стати об'єктом промислового засвоєння.

Облік наявності і руху запасів відбивається статистикою в *балансі запасів,* що складається для кожного виду корисних копалин. У них ураховується наявність запасів на початок року з розподілом за рівнем розвіданості. придатності для промислового використання. Рух запасів характеризують показники видобутку та втрат корисних копалин за рік, зміни запасів за рік у зв'язку з

розвідувальними роботами, зміною кондицій, переоцінюванням та з інших причин. У балансах ураховується забезпеченість діючих, а також тих, що будуються або проектуються, гірничодобувних підприємств розвіданими запасами мінеральної сировини.

Показники статистики водних ресурсів.

Водні ресурси — це запаси поверхневих та підземних вод, а також інших водних об'єктів. Державний фонд водних ресурсів країни включає ріки, озера, водосховища, ставки, канали, підземні ріки.

Статистика водних ресурсів має забезпечити урядові органи, органи управління та планування необхідною інформацією, що відбиває водозабезпеченість окремих районів країни, водоспоживання (забір води), використання води в розподілі за галузями, призначенням, видами водних об'єктів, характеризувати обсяг скинутих стоків за водокористувачами і рівень забрудненості стоків, а також заходи щодо захисту водних ресурсів від забрудненості й ефективність цих заходів. Інакше кажучи, статистика водних ресурсів за допомогою системи показників має характеризувати наявність і використання водних ресурсів, а також зміну їх стану, що виникає під впливом господарської діяльності людини і заходів щодо охорони водних ресурсів.

Показники, що характеризують водозабезпеченість конкретної території і якість води, обчислюються на підставі даних спеціальних обстежень. Головні серед них такі:

• запас води, тис. $м^3$, усього та за видами поверхневих стоків рік;

• запас води в розрахунку на одну особу, на 1 $км^2$ території. Критерій чистоти поверхневої води встановлюється за сумою якісних показників, що використовуються для оцінки придатності її для різних водокористувачів.

Показники інших розділів статистики водних ресурсів обчислюються за даними статистичної звітності. Для характеристики водоспоживання встановлюють загальний забір води та її використання за напрямами використання, а також втрати. Обсяг стоків розраховується за якістю очищення, а також кількістю забруднювачів (у тоннах), за окремими інгредієнтами в стоках. Провадиться також облік кількості, потужності та ефективності роботи очисних споруд. Статистичні дані розробляються як у територіальному, так і галузевому розрізах. При цьому широко використовується метод угруповань.

Узагальнюючими показниками в статистиці природних ресурсів можуть бути тільки вартісні показники. Даючи економічну оцінку природних ресурсів, застосовують два методи: затратний (оцінювання в ресурсі результатів праці) та рентний (оцінювання можливих доходів від ресурсу), а також різні їх комбінації.

Основи трудового (затратного) методу закладені були С. Т. Струміліним, хоча він і обмежився лише встановленням розміру витрат на первинне освоєння ресурсів без урахування витрат на підтримання їх відтворення в процесі експлуатації. Тобто обчислювалась сукупність затрат праці на освоєння природного ресурсу. При використанні цього методу зберігається єдність підходу до оцінювання всіх елементів національного багатства як споживчих вартостей, що беруть участь у відтворенні. Але водночас вони не дістають вартісного вираження, необхідного для обміну на ринку. Найважливішими джерелами інформації про прямі та непрямі затрати праці є міжгалузеві баланси.

Сутність рентного методу полягає в установленні доходів від природного ресурсу на базі розробки диференційованої ренти. Цей метод почав застосовуватися ще в 30-ті роки для оцінювання сільськогосподарських угідь та інших природних ресурсів у балансі народного господарства колишнього СРСР за 1923—1924 рр. Розрахунки виконувались розкладанням приросту національного доходу з виділенням доходів від цих природних ресурсів. На цих самих принципах виконано більшість повоєнних переоцінок, які дозволили оцінити роль природного фактора в реальних фінансових операціях. Тут на допомогу прийшла інформація кадастрів ресурсів, про які йшлося раніше.

Але й перший, і другий методи оцінювання природних ресурсів потребують у нашій країні подальшого розвитку та докладної розробки.

ЗМІСТ

7.2. Вибіркові оцінки середньої та частки
7.3. Різновиди вибірок
7.4. Визначення обсягу вибірки
7.5. Статистична перевірка гіпотез

8. МЕТОДИ АНАЛІЗУ ВЗАЄМОЗВ'ЯЗКІВ

8.1. Види взаємозв'язків
8.2. Регресійний аналіз
8.3. Оцінка щільності та перевірка істотності кореляційного зв'язку
8.4. Рангова кореляція
8.5. Оцінка узгодженості варіації атрибутивних ознак

9. МЕТОДИ ВІЗУАЛЬНОГО АНАЛІЗУ
9.1. Двовимірний (2М) візуальний аналіз
9.2. Тривимірний (3М) візуальний аналіз даних
9.3. Тернарні графіки
9.4. Піктографіки

10. ФАКТОРНИЙ АНАЛІЗ
10.1. Основна мета
10.2. Факторний аналіз як метод редукції даних
10.3. Факторний аналіз як метод класифікації

12. РЯДИ ДИНАМІКИ.

АНАЛІЗ ІНТЕНСИВНОСТІ ТА ТЕНДЕНЦІЙ РОЗВИТКУ
12.1. Суть і складові елементи динамічного ряду
12.2. Характеристики інтенсивності динаміки
12.3. Середня абсолютна та відносна швидкість розвитку
12.4. Характеристика основної тенденції розвитку
12.5. Оцінка коливань та сталості динаміки

13. ІНДЕКСИ
13.1. Суть і функції індексів
13.2. Методологічні основи побудови зведених індексів
13.3. Агрегатна форма індексів
13.4. Середньозважені індекси
13.5. Індекси середніх величин
13.6. Територіальні індекси

14. СТАТИСТИКА ПРИРОДНИХ РЕСУРСІВ
14.1. Поняття про статистику навколишнього середовища
14.2. Побудова системи показників статистики навколишнього середовища
14.3. Показники окремих видів природних ресурсів